Oracle Maximum
Availability Architecture

EXADATA HEALTH AND RESOURCE USAGE MONITORING

ORACLE WHITE PAPER | NOVEMBER 2014

Table of Contents

Introduction

This document will demonstrate an end to end approach to health and resource utilization monitoring for Oracle Exadata Environments. In an addition to technical details a troubleshooting methodology will be explored that allows administrators to quickly identify and correct issues in an expeditious manner.

Methodology

The Purpose of this document is to provide a methodology to troubleshoot issues on Oracle Exadata Systems. The document will take a “rule out” approach in that components of the system will be verified as performing correctly to eliminate their role in the incident. There will be five areas of concentration in the overall system diagnosis

1. Steps to take before problems occur that can assist in troubleshooting
2. Changes made to the system
3. Quick analysis
4. Baseline comparison
5. Advanced diagnostics

For clarity the flow of the steps are displayed in the flow chart in Figure 1.1.

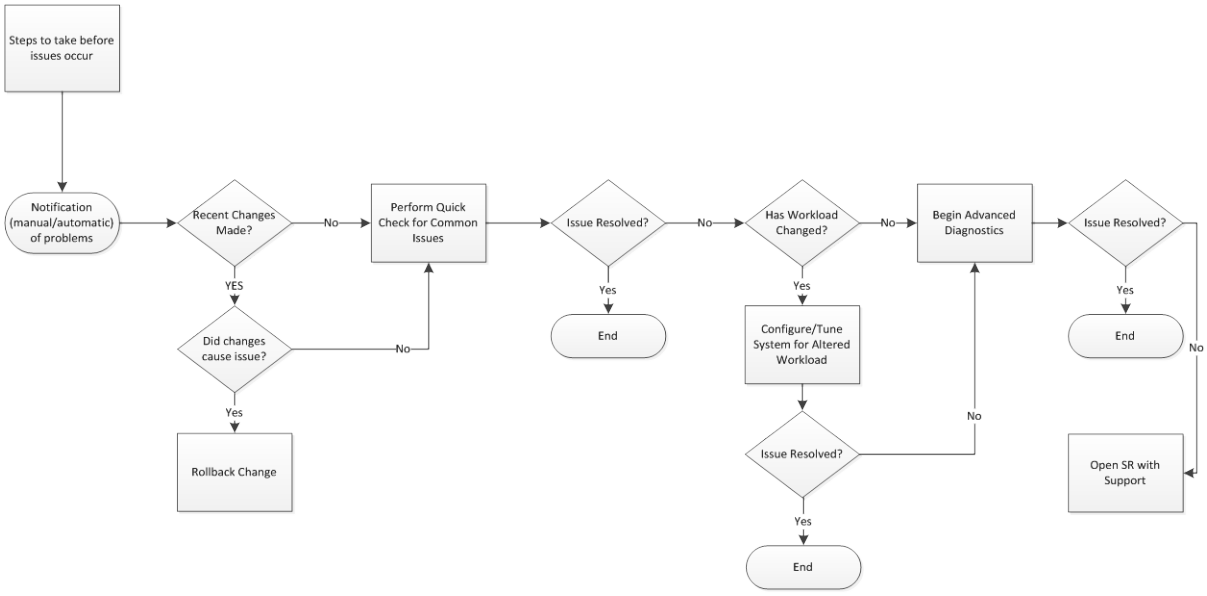


Figure 1.1

Steps to follow before problems occur

Taking time to prepare for the eventuality of issues occurring can drastically reduce the time it takes to diagnose the problems. The following sections will list the steps necessary to ensure that important information is collected.

Exachk

The Exachk and Enterprise Manager HealthCheck Plug-in output assists customers to more easily review and cross reference current collected data against supported version levels as well as recommended Oracle Exadata best practices. The Oracle Exadata Database Machine Exachk and HealthCheck tools are run manually or scheduled to help collect data regarding key software, hardware, firmware, and configurations.

Exachk can be executed as often as desired and should be run regularly (example: monthly) as part of any business' planned maintenance program for an Oracle Exadata Database Machine. ***Customers should check for the latest Exachk and HealthCheck versions available*** since this tool is updated on a regular basis in order to include additional checks against current best practices or known critical issues. See Oracle Support Document 1070954.1: "Oracle Exadata Database Machine Exachk or HealthCheck" and Exadata best practices in Oracle Support Document 757552.1: "Oracle Exadata Best Practices"

If Exachk is setup in Daemon mode then a job can be scheduled in Enterprise Manager to ensure its run on a schedule basis. See Appendix 1.1 for detail configuration instructions. For detailed information on setting up Exachk in Daemon mode see the Exachk User Guide in Oracle Support Document 1070954.1: "Oracle Exadata Database Machine Exachk or HealthCheck"

Having Exachk run via EM not only feeds information into the Enterprise Manager Health Check Plugin but it makes the output of the command viable from the console:

Job

```
Collections and audit checks log file is
/home/oracle/exachk/output/exachk_slcc12adm03_LOADTEST_071414_104129/log/exachk.log


Checking for prompts in /home/oracle/.bash_profile on slcc12adm03 for oracle user...



Checking for prompts in /home/oracle/.bash_profile on slcc12adm04 for oracle user...


Starting to run exachk in background on slcc12adm04

============================================================
                  Node name - slcc12adm03
============================================================

Collecting - ASM Diskgroup Attributes
Collecting - ASM initialization parameters
Collecting - Active sessions load balance for LOADTEST database
Collecting - Archived Destination Status for LOADTEST database
Collecting - Cluster Interconnect Config for LOADTEST database
Collecting - Database Archive Destinations for LOADTEST database
Collecting - Database Files for LOADTEST database
Collecting - Database Instance Settings for LOADTEST database
Collecting - Database Parameters for LOADTEST database
Collecting - Database Parameters for LOADTEST database
Collecting - Database Properties for LOADTEST database
Collecting - Database Registry for LOADTEST database
Collecting - Database Sequences for LOADTEST database
Collecting - Database Undocumented Parameters for LOADTEST database
Collecting - Database Undocumented Parameters for LOADTEST database
Collecting - Database Workload Services for LOADTEST database
Collecting - Dataguard Status for LOADTEST database
Collecting - Log Sequence Numbers for LOADTEST database
Collecting - Process for shipping Redo to standby for LOADTEST database
Collecting - RDBMS Feature Usage for LOADTEST database
Collecting - Redo Log information for LOADTEST database
Collecting - Standby redo log creation status before switchover for LOADTEST database
Collecting - CPU Information
Collecting - Clusterware and RDBMS software version
Collecting - Compute node PCI bus slot speed for infiniband HCAs
Collecting - Kernel parameters
Collecting - Maximum number of semaphore sets on system
Collecting - Maximum number of semaphores on system
Collecting - Maximum number of semaphores per semaphore set
Collecting - Patches for Grid Infrastructure
Collecting - Patches for RDBMS Home
Collecting - RDBMS patch inventory
Collecting - number of semaphore operations per semop system call
Preparing to run root privileged commands on DATABASE SERVER[0m slcc12adm03.
```

AWR Baselines

AWR baselines allow a specified range of snapshots to be retained and used for performance comparisons. These baselines are not deleted like normal AWR snapshots when the retention period is reached.

There are two types of baselines that should be created for reference on an Exadata machine. Each database instance on each computer node should have a moving baseline as well as static baselines that capture the different workloads that may occur in the environment. For example, if a weekly batch load occurs on Monday nights from 7 to 10 pm, it would be very useful to have a baseline of the instance during that time frame.

Static Baselines

An AWR static baseline is made by selecting two snapshots. AWR data that exists between these two snapshots is saved for evaluation. For example, a baseline could be created for a batch processing job, normal OLTP processing, or during peak workloads. AWR baselines are retained until manually deleted. To create a static baseline, execute the following as a user with DBA privileges, replacing the snapshot id's as appropriate:

```
exec dbms_workload_repository.create_baseline (start_snap_id => <starting snapshot id>,
end_snap_id => <ending snapshot id>, baseline_name => 'Normal Baseline');
```

Baselines can be created in Enterprise Manager from the AWR Baseline Page under the database/instance target home, as depicted in Figure 1.2. For complete screen by screen steps see Appendix 1.2

A single static baseline is created using two user specified time periods or alternatively two AWR snapshot ids. This is chosen via radio buttons on the Create Baseline page. See Figure 1.3 for an example.

It is also possible to create a repeating static baseline. The repeating baseline is on a user defined schedule, for example a daily or weekly snapshot of significant activity as depicted in Figure 1.4.

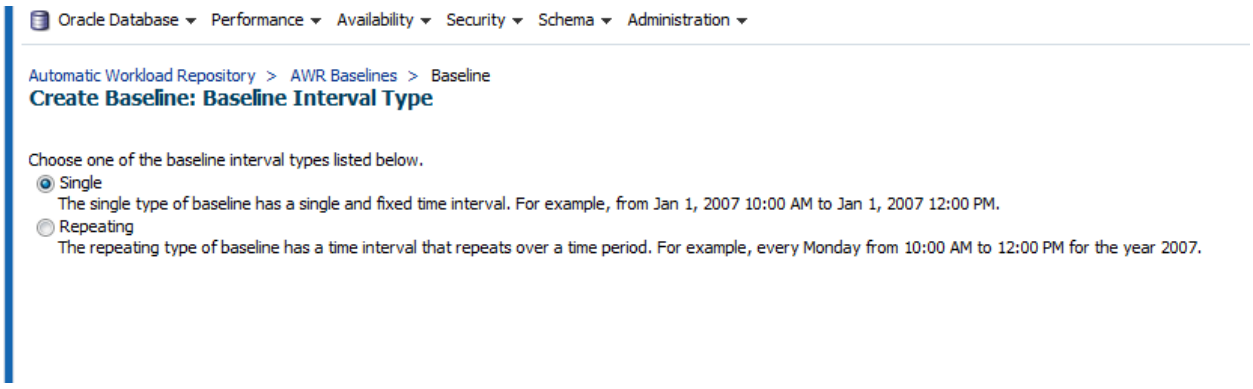


Figure 1.2

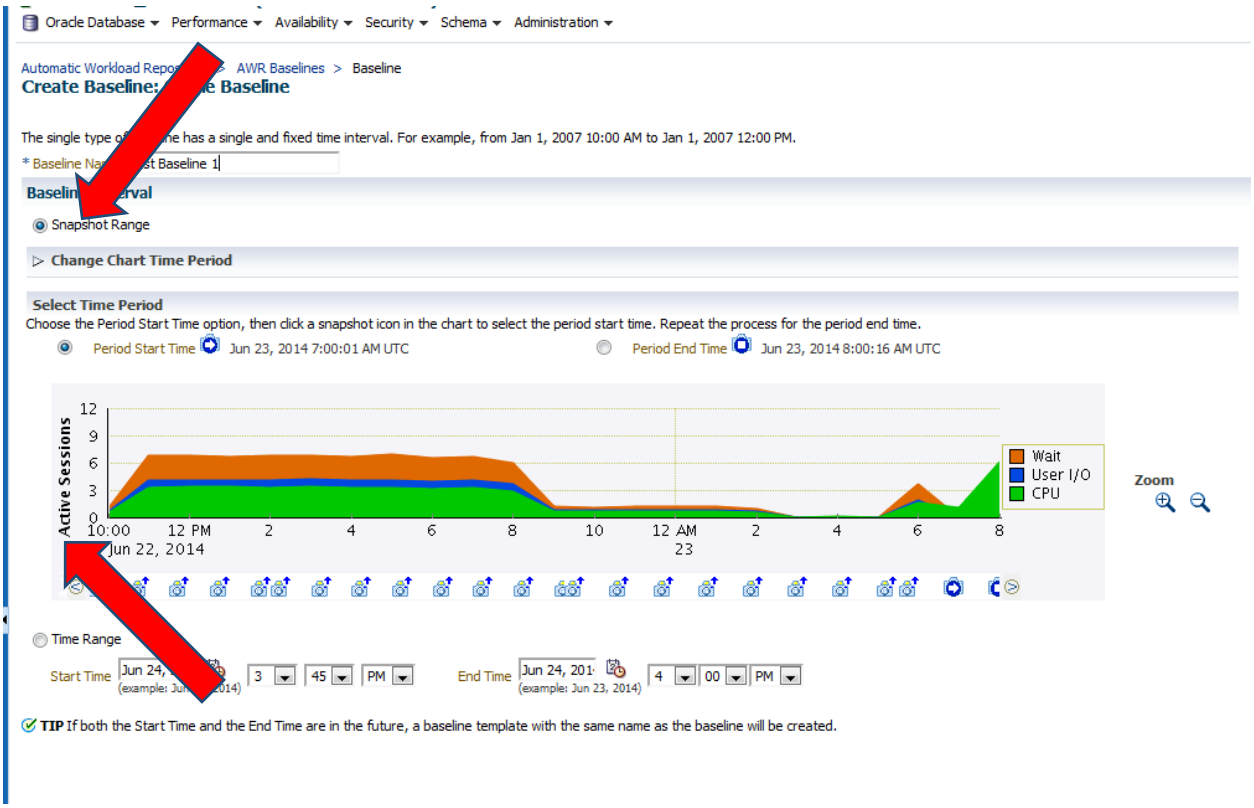


Figure 1.3

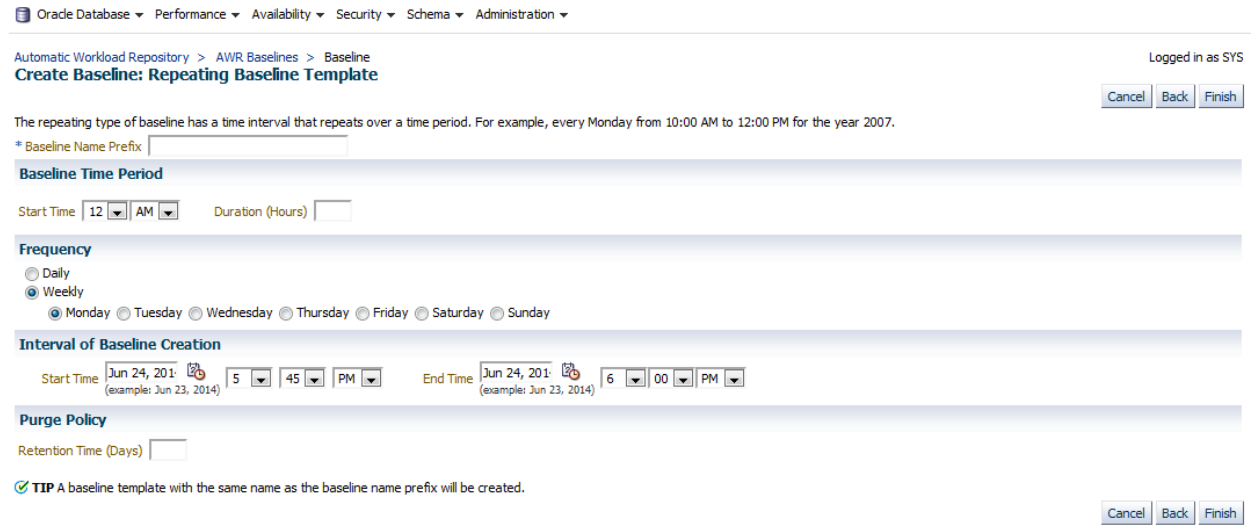


Figure 1.4

Moving Baselines

A moving baseline collects the same information as a static baseline. Note that by default the window is the AWR retention period (8 days). However, since the AWR snapshots will age out, the data will change. It is a good idea to make the AWR retention period as long as possible to ensure all necessary information is available. The maximum size of the moving window is the size of the AWR retention period. Therefore, it is recommended to change the AWR retention period to a minimum of thirty days, as in the following example:

```
exec dbms_workload_repository.modify_baseline_window_size(30);
```

Note that this is just a starting point. Consideration should be given in determining the appropriate interval based on the application workload cycle.

The moving window can be modified in Enterprise Manager as seen in Figure 1.5. For complete screen by screen steps see Appendix 1.3

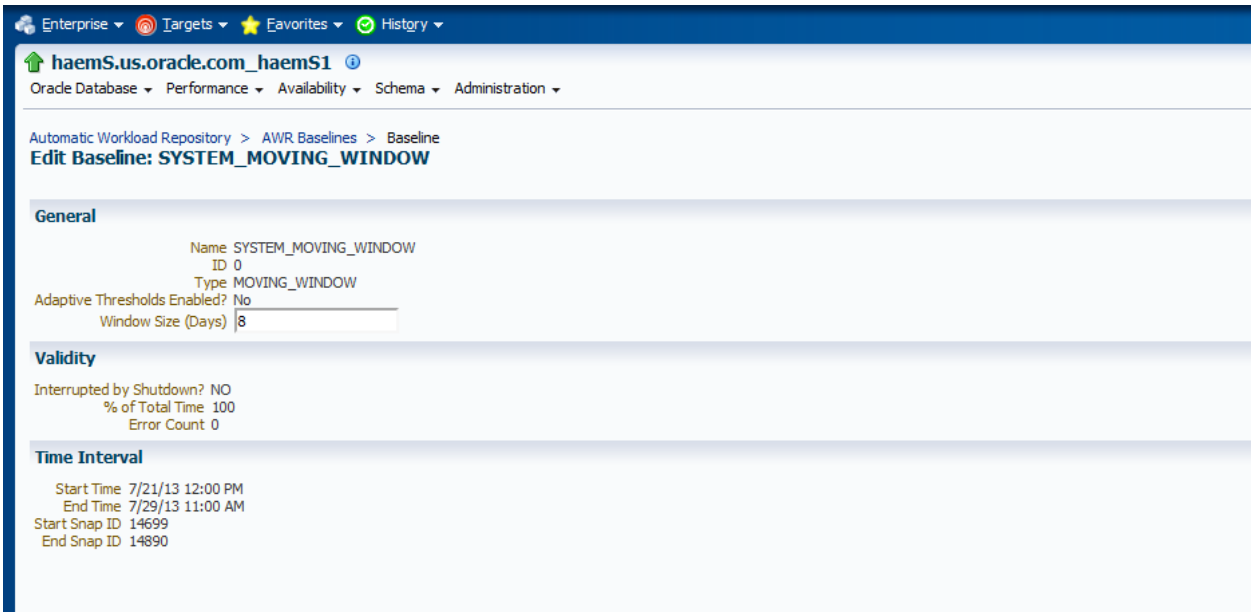


Figure 1.5

Configuration File Copies

Having time-stamped incremental copies of critical files will make it much easier to determine if changes have occurred in the environment. In essence, these can provide an audit trail of changes on various parts of the Exadata environment. A good practice is to create a script that regularly makes copies to the local file system. Although many times these files may also be backed up to tape, the time to retrieve various iterations from tape may prove too costly. A minimum list of files for which copies should be maintained includes:

- All instance/database init and spfiles
- Database password files
- ASM init files
- ASM password files

Enterprise Manager can be used to accomplish this task as well. By creating scheduled jobs for each applicable target, the above files can be copied on a regular interval. For example, follow the steps in Appendix 1.4 to create a job to create a weekly copy of the database spfile.

Tunable Parameters Snapshots

The UNIX kernel tunable parameters have a major impact on the behavior of a server. It is a good idea to have a backup of the tunable parameters to quickly compare if any changes have occurred. This can be done manually or setup via the process described above to periodically take a snapshot in Enterprise Manager. This information should be collected on each Compute node. The following command will list the parameters:

```
# sysctl -a
```

The following command will save these parameters to a time stamped file. Replace the date at the end of the file with an appropriate timestamp:

```
# sysctl -a > /root/kernel_parameters/sysctl_073013
```

To execute this on all nodes simultaneously use the dcli command, where dbs_group is a file containing all the Compute nodes, one per line:

```
# dcli -l root -g ./dbs_group "sysctl -a > /root/sysctl_`date +%m%d%y`"
```

The dcli utility allows the user to run the same command on multiple nodes. More detailed information on the command is available at:

http://www.oracle.com/technetwork/articles/oem/exadata-commands-part3-402445.html

The -g option specifies a text file that lists all the nodes on which you wish to execute the command. In this instance, the command should be run on each Compute node and Cell server.

How to configure DB parameters

Database parameters can be changed from the Database target home page in Enterprise Manager. Go to the Target homepage and select Administration-> Initialization Parameters. Parameters can then be edited as desired.

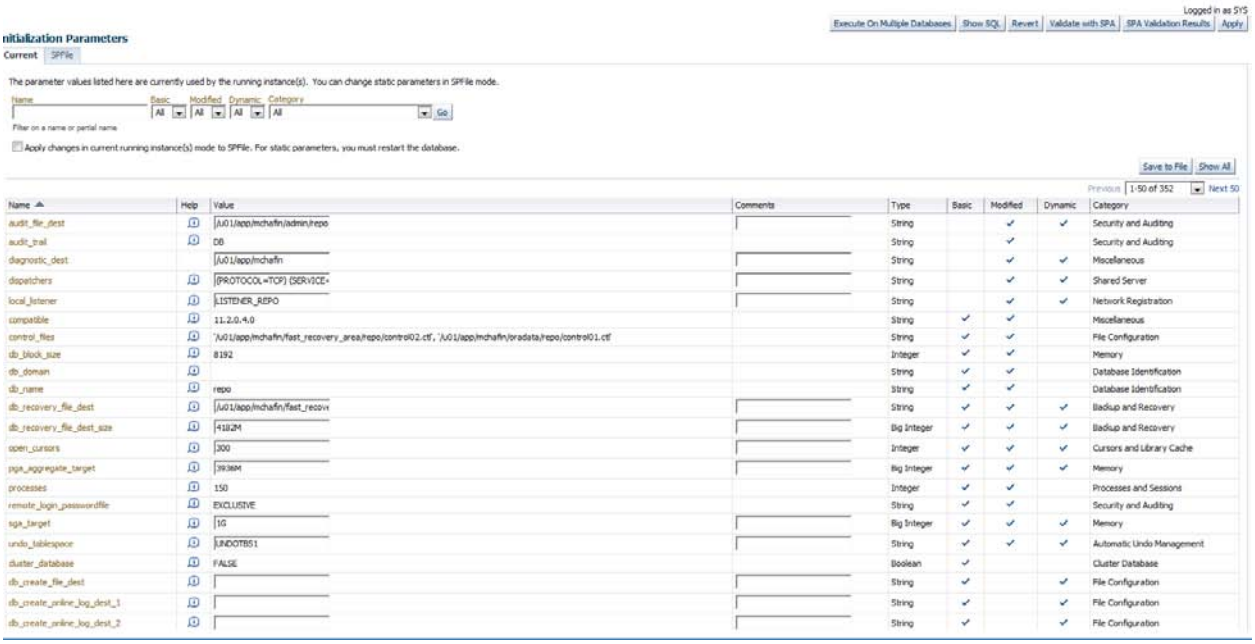


Figure 1.6

- DW Capacity Planning
    - Exadata Best Practices in: MOS note 1274318.1 and MOS note 1347995.1.

Parameter	X3-2	X3-8
parallel_max_servers	240	1280
parallel_min_servers	96	512

parallel_degree_policy	Manual	Manual
parallel_degree_limit	16	24
parallel_servers_target	128	512
sga_target	16G	128G
pga_aggregate_target	16G	256G

- OLTP Capacity Planning

- Exadata Best Practices in: in MOS note 1274318.1 and MOS note 1347995.1.

Parameter	X3-2	X3-8
parallel_max_servers	240	1280
parallel_min_servers	0	0
sga_target	24G	128G
Pga_aggregate_target	16G	64G

I/O Resource Manager

Exadata systems allow greater database consolidation without the complexities of slicing, distributing, and managing workloads across disks and diskgroups, while also providing better bandwidth and space utilization. With all databases sharing the same disks, I/O Resource Management (IORM) is the mechanism used to maintain a predictable level of performance amongst your consolidated databases.

With IORM you can guarantee a percentage of I/O operations for each database on your system and also limit them as necessary. Allocations are distributed as a percentage of total available I/O operations per storage cell and all databases processes are regulated by IORM, so when evaluating how best IORM will work for your system you need to consider not only the behavior of your applications foreground user processes but also the background database processes such as the database writers.

For additional detail about IORM see the Oracle Support Document 13390769.1: “Master Note for Oracle Database Resource Manager” and Oracle MAA Best Practices for IORM: http://www.oracle.com/webfolder/technetwork/Exadata/MAA-BestP/IORM/IORM.pdf.

The following are some of the tools you can use to monitor I/O usage and Resource Management:

metric_iorm.pl

This script, obtained through the master note listed above, can be used to see the immediate effects of throttling a database using IORM. It's a great tool to use when evaluating your database I/O needs and the results of your IORM plans. Here's an example showing mostly small I/O utilization and some time waiting in the queue. If cell I/O is being saturated or databases are being throttled through IORM, qtimes will increase.

```
Database: DBM
Utilization:     Small=22%    Large=5%          ← database IO utilization
Flash Cache:     IOPS=2068
Disk Throughput: MBPS=9
Small I/O's:     IOPS=2154    Avg qtime=0.0ms
Large I/O's:     IOPS=0.1     Avg qtime=1.2ms   ← time spent waiting in the queue
```

See Appendix 1.5 for detailed AWR screenshots demonstrating AWR effectiveness.

Enterprise Manager Exadata Resource Management

Exadata storage cell I/O resources can be managed and monitored using the Manage I/O Resources page of Enterprise Manager. The page is available from the storage cell main page shown below.

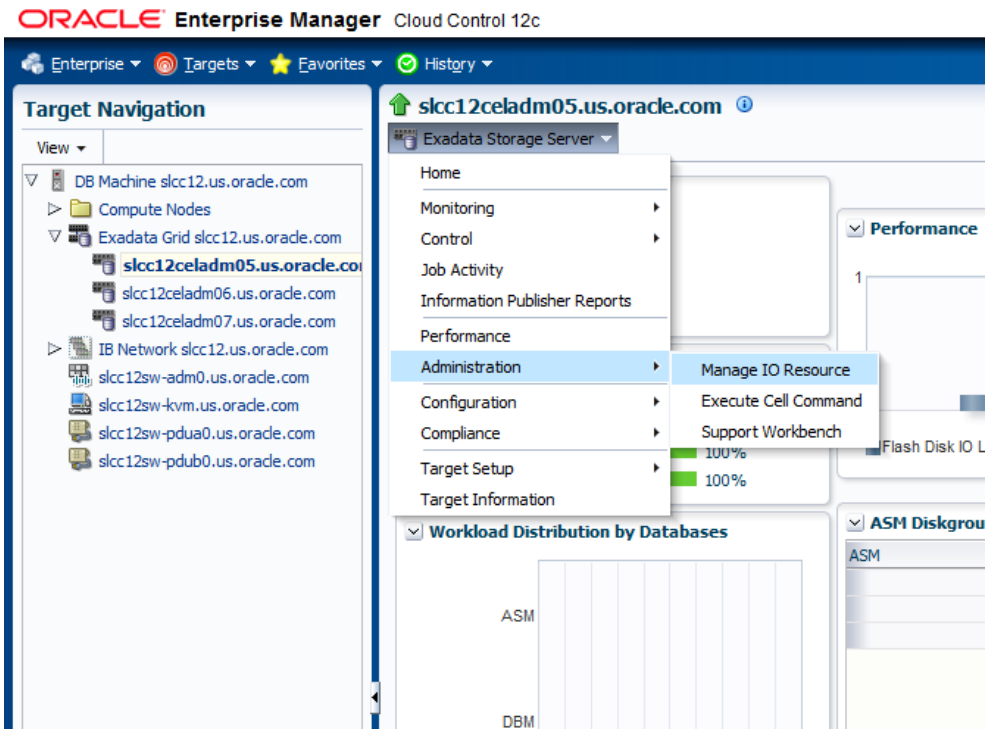


Figure 1.7

The Manage I/O Resource page allows you to configure resources as shown here.

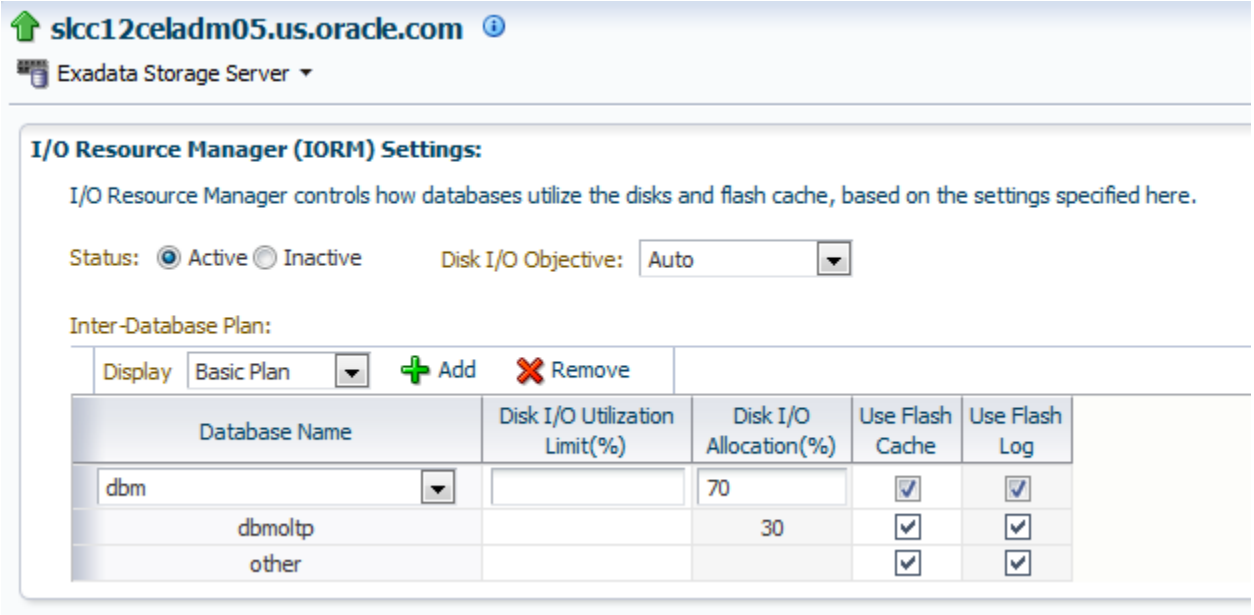


Figure 1.8

It also allows you to monitor the I/O usage of your cells as well as any I/O throttling. These graphs are great for determining baselines and long term trend analysis. In the example below you can see databases being throttled due to I/O Resource Management.

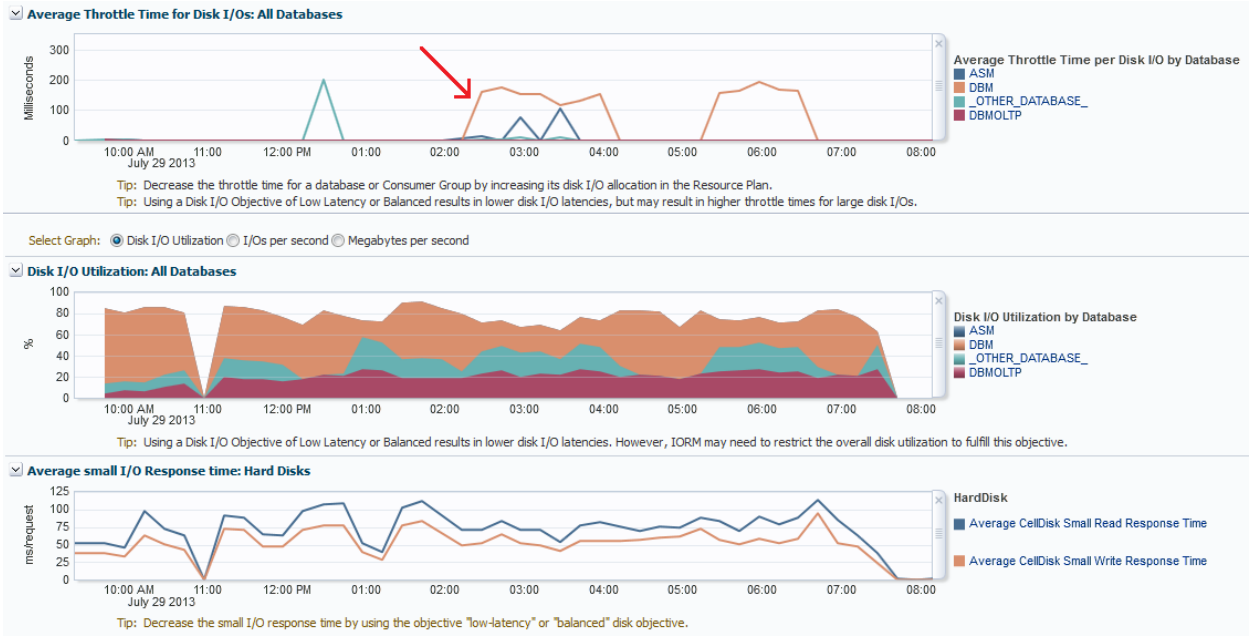


Figure 1.9

Setting limits

The effect of an IORM Plan can be very subtle particularly when hard limits are not used, because the storage cell can freely redistribute unused I/O to other databases as needed. These subtle latencies may typically be brief and transient events not registering on some charts. When limits are used, the storage cell will not allocate any addition free I/O resources to the database so the effects will be much more noticeable. Limits are great for keeping low priority databases contained and isolated so as not to interfere with other more critical databases.

Database Metrics

Database metrics can be used for trend analysis and also alerting when defined thresholds are reached.

Here's a graph showing small I/O read latency which will increase when throttling occurs:

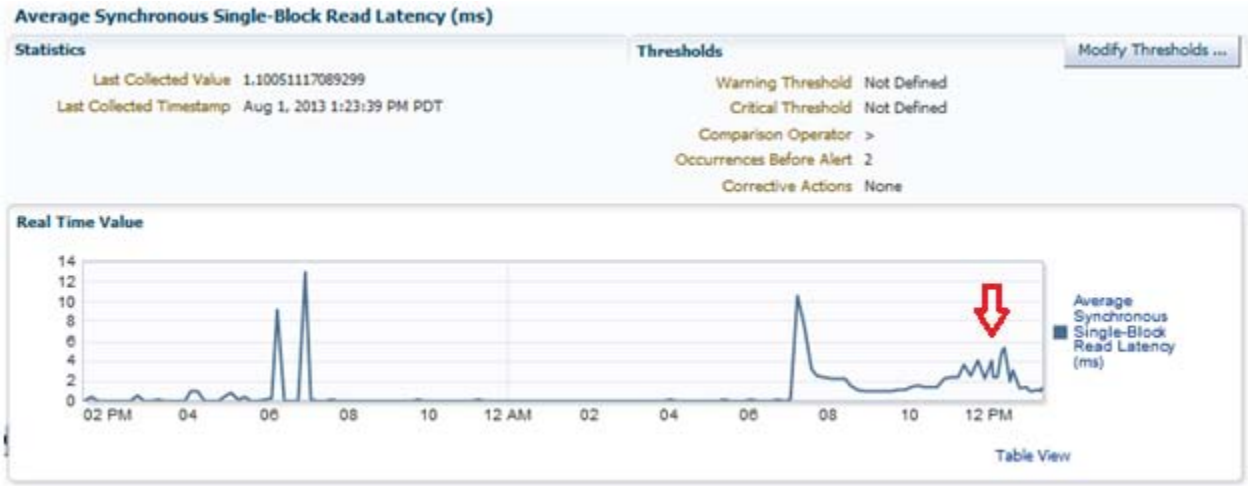


Figure 1.10

Here are I/O requests for the same time period. Notice how they drop when throttling occurs:

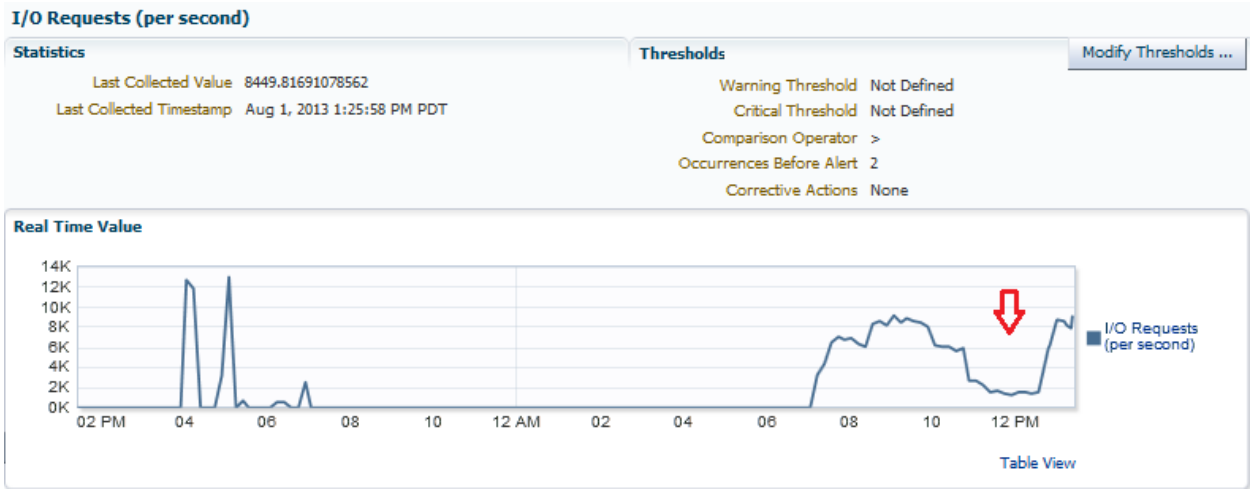


Figure 1.11

Refer to the following resources:

- **Managing I/O Resources (Exadata doc)**
- **Use IORM metric scripts (MOS)**
- **IORM recommended patches (MOS)**
-

Configure HugePages

- HugePages reduces the page table size and process startup time.
- If PageTables in /proc/meminfo is > 2% of physical memory size, set operating system parameter HugePages equal to the sum of all shared memory segments. (LINUX ONLY)
- Allocate enough HugePages to fit all of your SGA.
- See MOS Note 361323.1 for how to set HugePages.
- See MOS Note 401749.1 for how to determine amount of shared memory in use.

Configure consolidated environments

- Watch out when there are multiple databases.
- Previous configuration recommendations represent the sum of all database settings on the same node.
- Refer to Exadata Consolidation Best Practice Paper for more information on:
  - How to configure DB parameters.
  - How to configure HugePages.
  - And more …

Initial quick checks to perform

Once any changes to the environment have been ruled out as consequential to the problem at hand, a series of quick checks can be run to help isolate the nature of the issue. The checks listed in the table below are quick checks that should take no longer than a few minutes to complete. More information on performing each of the checks follows the table.

	Compute nodes	Cell servers	Database	ASM	ClusterWare	InfiniBand Switch
Component Up	X	X	X	X	X	
Alert Log			X	X	X	
OS System Log	X	X				
CPU	X	X				
Memory Utilization	X	X				
ILOM Errors	X	X				X
All Networks Up	X	X				X
Disk Status		X				
CheckHWnFWProfile	X	X				
DB Services/Listener Up					X	
Exachk	X					

Component Up

Perhaps the simplest check to perform is determining if the Exadata Machine Components are up. Verifying that all Hardware and Software components are up and available provides a solid platform from which to start further investigation.

In Enterprise Manager the Database Machine home page an overall graphical representation of the RAC environment provides an easy way to identify down components.

The Exadata Database Machine homepage is displayed with state information about each of the components. Screen by screen steps are listed in Appendix 3.1.

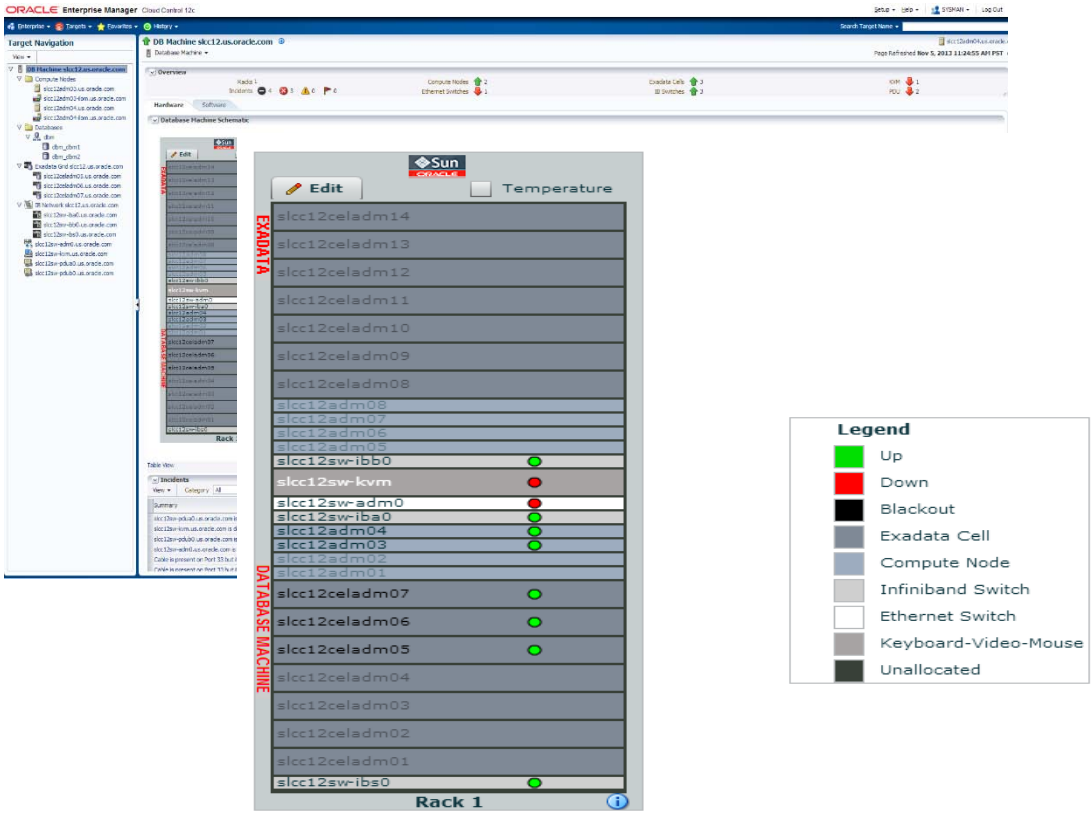


Figure 3.1

Logging into Compute nodes and Cell servers can also be used to verify components are up and running. It's also a good idea to confirm that the components are at the appropriate run level. The run level command will verify this. All Compute nodes and Cell servers should be at run level three, as is depicted in the sample output following the command.

```
# runlevel
 N 3
```

Running crsctl stat res -t will give you a quick view of all CRS resources:

```
# ./crsctl stat res -t
--------------------------------------------------------------------------------
NAME           TARGET  STATE        SERVER                   STATE_DETAILS
--------------------------------------------------------------------------------
Local Resources
--------------------------------------------------------------------------------
ora.DATA.dg
```

```
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
ora.DBFS_DG.dg
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
ora.LISTENER.lsnr
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
ora.RECO.dg
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
ora.asm
               ONLINE  ONLINE       slcb01db07               Started
               ONLINE  ONLINE       slcb01db08               Started
ora.gsd
               OFFLINE OFFLINE      slcb01db07
               OFFLINE OFFLINE      slcb01db08
ora.net1.network
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
ora.ons
               ONLINE  ONLINE       slcb01db07
               ONLINE  ONLINE       slcb01db08
```

Database and ASM status can easily be checked with the srvctl command from a Compute node.

To confirm the database is running and is in an open state, issue the following command, replacing <DATABASE NAME> as appropriate. Note the sample notional output below the command.

```
srvctl status database -d <DATABASE NAME> -v
Instance <INSTANCE NAME> is running on node <SERVER NAME>. Instance status: Open.
Instance <INSTANCE NAME> is running on node <SERVER NAME>. Instance status: Open.
```

To check ASM issue the following command. The output will be similar to the notional output depicted below the command.

```
srvctl status asm -v
ASM is running on <NODE1>,<NODE2>
Detailed state on node <NODE1>: Started
Detailed state on node <NODE2>: Started
```

The check that all Cell services are online:

```
# dcli –l root -g cell_group cellcli -e list cell
slcc12celadm05: slcc12celadm05   online
slcc12celadm06: slcc12celadm06   online
slcc12celadm07: slcc12celadm07   online
```

Incident Manager

Incident Manager provides administrators the ability to identify, monitor, and resolve issues quickly and efficiently. Incident Manager uses the following three-level hierarchy to group and categorize issues.

Event

A significant occurrence of interest on a target that has been detected by Enterprise Manager.

Incident

A set of significant events or combination of related events that pertain to the same issue.

Problem

The underlying root cause of incidents. Currently, this represents critical errors in Oracle software that are the underlying root cause of diagnostic incidents.

Incidents created for Exadata components can be viewed on the Database Machine home page.

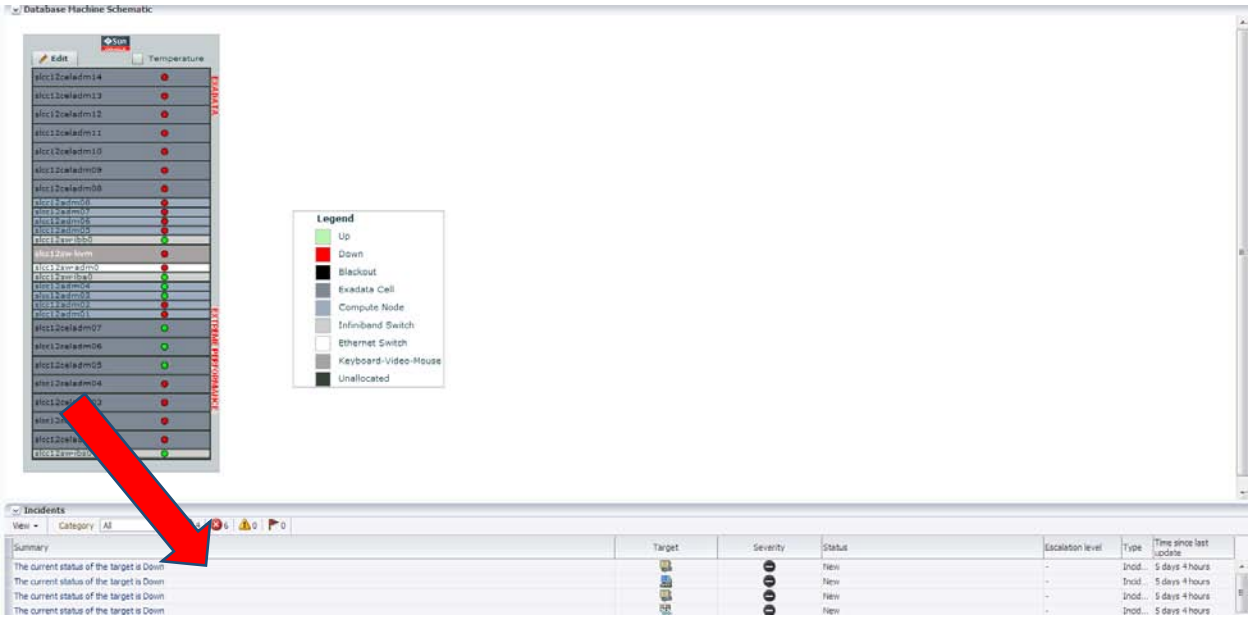


Figure 3.2

Clicking on individual incidents will bring up the incident manager screen.

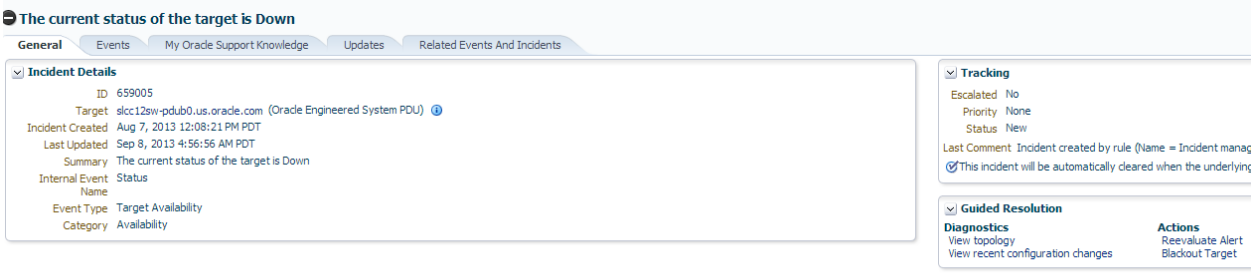


Figure 3.3

Alert Log

The alert log is a chronological ordered file of messages and errors written by Oracle components such as database, ASM, and clusterware. Alert logs are located in the following directories:

- Database
  - <DIAGNOSTIC DIRECTORY>/diag/rdbms/<database name>/<INSTANCE NAME>/trace/alert_<INSTANCE NAME>.log
- ASM
  - <DIAGNOSTIC DIRECTORY>/diag/asm/+asm/<ASM INSTANCE>/trace/alert_<ASM INSTANCE>.log
- Clusterware
  - <CLUSTWARE HOME>/log/<HOSTNAME>/alert<HOSTNAME>.log

If there is an issue at the software level, most of the time it will present itself in one of these files. In addition, some hardware level events may be identified as well such as disk or network issues.

For more detailed information on database operations see the following document:

http://docs.oracle.com/cd/E11882_01/server.112/e10897/toc.htm

For database and ASM Enterprise Manager targets, the Alert log metrics can be viewed on the instance target page. See Appendix 3.2 for screen by screen navigation

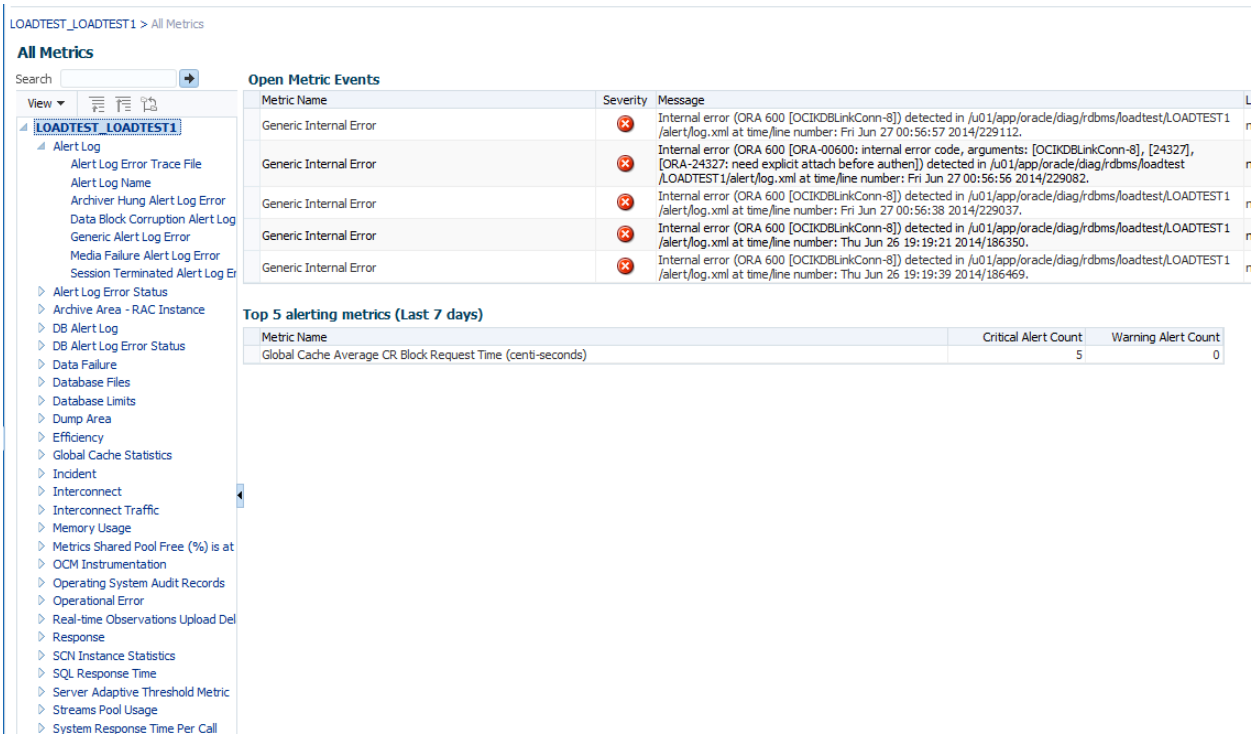


Figure 3.4

If a critical issue is detected, an Enterprise Manager Incident will be created. This can be viewed directly on the main target page.

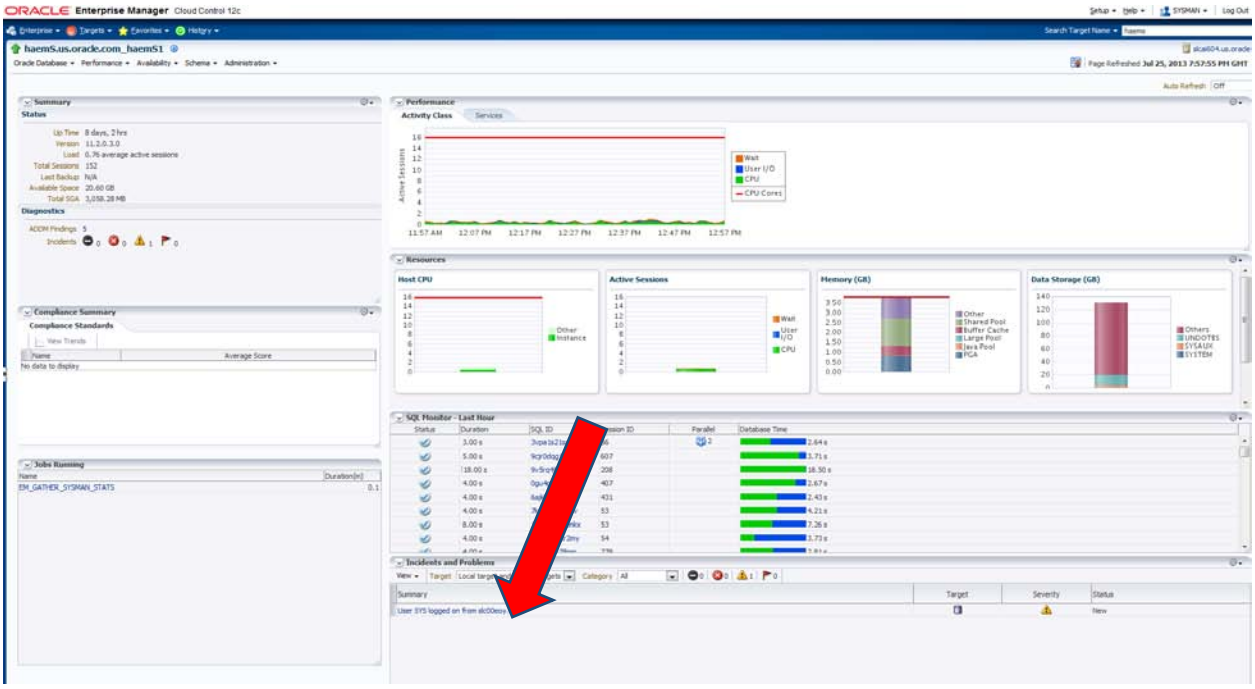


Figure 3.5

OS System Log

The system log is the place where most OS related messages and errors are published. This is an excellent starting point to investigate hardware or operating system issues. The standard location for this file on Linux is:

/var/log/messages

The log file location is configurable. To view log file configuration, see the following file:

/etc/syslog.conf

CPU Utilization

Compute node CPU utilization can be measured through many different tools including top, AWR, iostat, vmstat, etc. They all report the same number and % CPU utilization, typically averaged over a set period of time. Choose whichever tool is most convenient, but when interpreting the data ensure you allow for Intel CPU Hyper-Threading.

The Intel CPUs used in all Exadata models run with two threads per CPU core. This helps to boost overall performance, but the second thread is not as powerful as the first. The operating system assumes that all threads are equal and thus overstates the CPU capacity available to the operating system. We need to allow for this. Here is an approximate rule of thumb that can be used to estimate actual CPU utilization, but note that this can vary with different workloads:

- For CPU utilization less than 50%, multiply by 1.7.
- For CPU utilization over 50%, assume 85% plus (util-50%)* 0.3.

Here is a table that summarizes the effect:

Measured Utilization	Actual Utilization
10%	17%
20%	34%
30%	51%
40%	68%
50%	85%

60%	88%
70%	91%
80%	94%
90%	97%
100%	100%

Actual utilization gives an approximate measure of how much spare CPU is available for growth.

Plan to keep actual utilization below 85% for response time sensitive workloads. If utilization is over 85%, use IORM or instance caging to prioritize workloads appropriately and prevent CPU starvation of essential processes.

Note that it is common and appropriate to have CPU or I/O utilization reach 100% for large parallel workloads that seek maximum batch or reporting performance. Data Warehouses are an important example of this. In mixed-use systems, batch and reporting performance needs to be traded off against interactive response times. If interactive response time is important, then the degree of parallelism used for batch and reporting will need to be restricted so that batch and reporting don't consume excessive resources on the system. In a mixed workload environment, the best practice should be to enable IROM with Objective=balance. This favors small I/Os over large ones but not to the degree that Data Warehouse workloads never complete.

Enterprise Manager provides a rich set of tools for evaluating resource utilization and capacity. The below information is only a small sample of the capabilities provided. More information on the Exadata Plug-in for Enterprise Manager is available here: Managing Oracle Exadata with Oracle Enterprise Manager 12c

CPU utilization is available in Enterprise Manager for Compute node targets. From the DB Machine target home page select the desired target. This will bring up the target home page that gives high level CPU information:

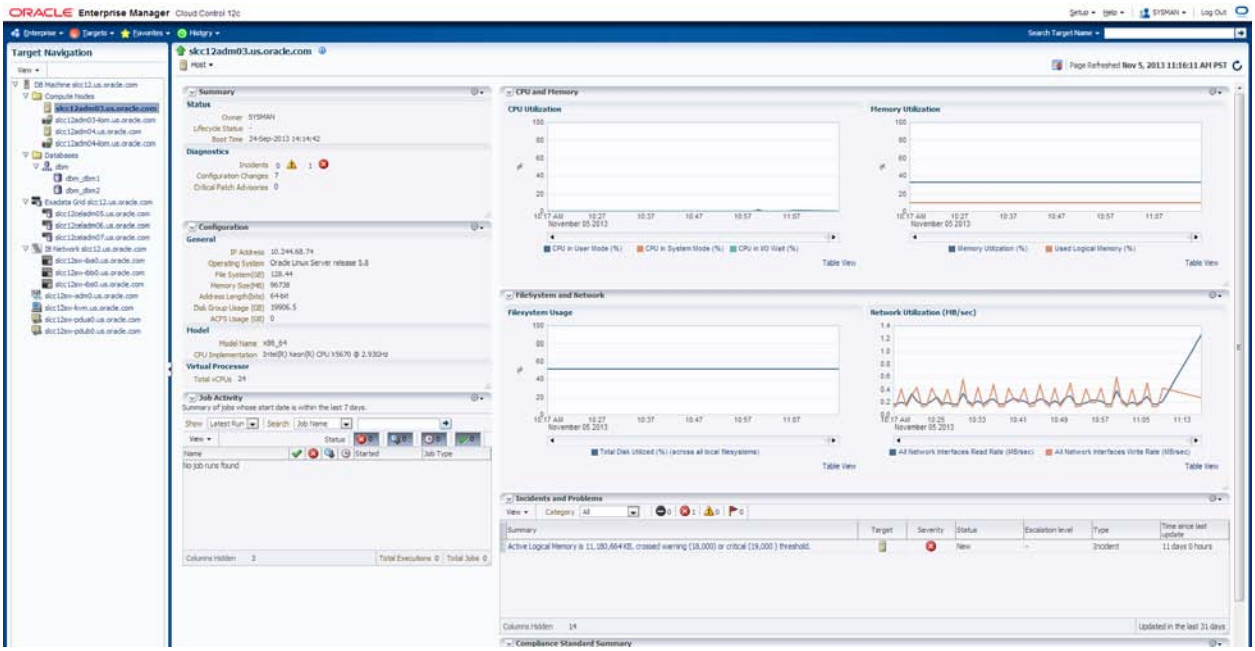


Figure 3.6

A more detailed view of CPU performance can then be obtained for Compute node targets by selecting Host→ Monitoring→ CPU Details:

Figure 3.7

Examine TOP kswapd

Examine kswapd and system CPU usage from OSW TOP data.

On a Compute node, go to /opt/oracle.Exawatcher/osw/archive/oswtop.

Check if kswapd consumes a full core and if system CPU usage is high.

Example of swapping from a real customer case

```
top - 14:29:08 up 7 days, 17:27,  0 users,  load average: 33.53, 26.46, 21.14
Tasks: 573 total,  35 running, 538 sleeping,   0 stopped,   0 zombie
Cpu(s):  6.3%us, 93.3%sy,  0.0%ni,  0.3%id,  0.1%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  74027752k total, 73698984k used,   328768k free,   116300k buffers
Swap: 16771852k total,  4087308k used, 12684544k free, 16334716k cached
  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
 1049 root      20  -5     0    0    0 R 100.1  0.0   6:33.55 [kswapd0]
 4192 root      20  -5     0    0    0 R 98.7  0.0  47:48.07 [krdsd]
 1953 oracle    25   0 17.2g 1.2g 219m R 86.2  1.6   3:16.17 ora_p061_cbaprdX4
 1929 oracle    25   0 17.2g 1.2g 220m R 84.6  1.6   3:15.81 ora_p049_cbaprdX4
```

Swap Activity can be viewed easily with Enterprise Manager by viewing the “Switch/Swap Activity” metric group. The metric group is disabled by default. However, it should be enabled and appropriate thresholds should be set to proactively identify excessive swap activity.

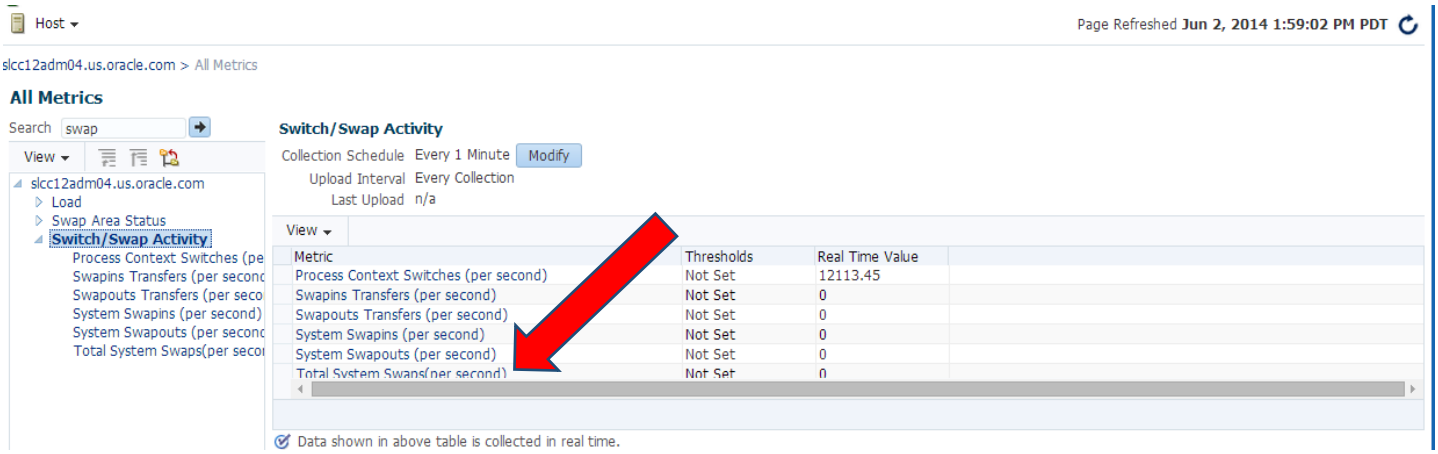


Figure 3.8

Memory Utilization

Memory utilization can be monitored using the /proc/meminfo virtual file. Add up the “MemFree:” and “Cached:” metrics to get an indication of the total available free memory. Linux will free memory from cache when necessary and so this can be regarded as part of free memory. The Exadata databases do not use the Linux page cache for database I/Os and so we need a relatively small Linux page cache. Here is an example:

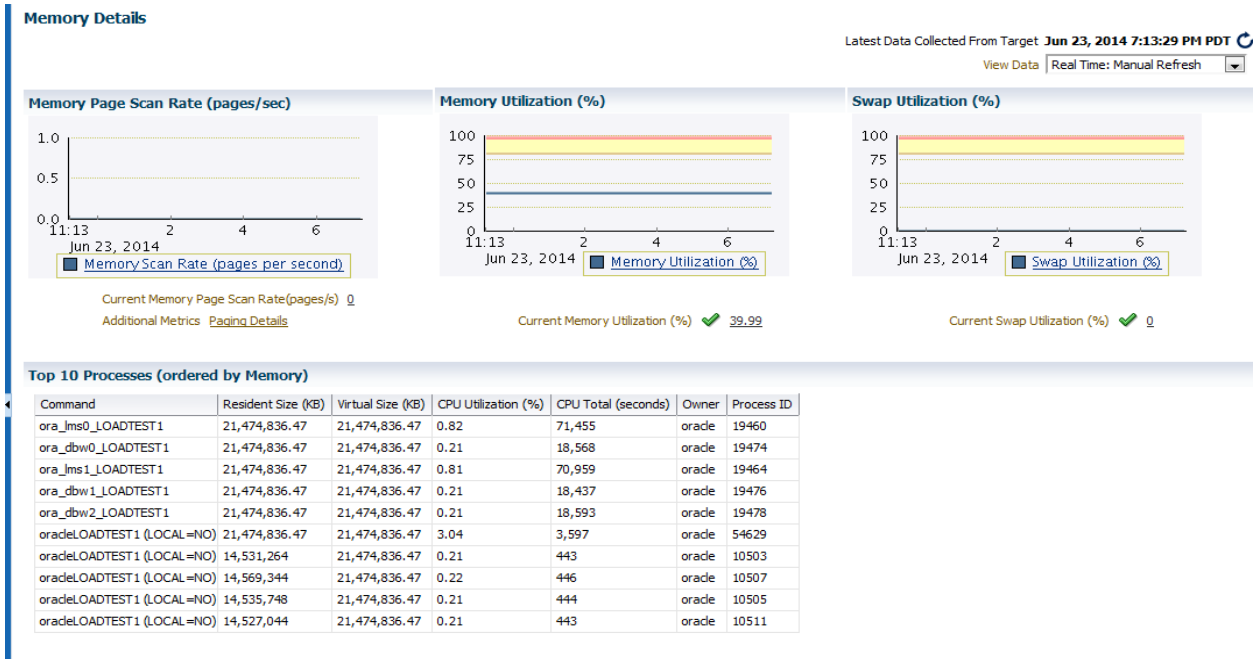
```
cat /proc/meminfo | egrep '^MemTotal:|^MemFree:|^Cached:'

MemTotal:       1058421596 kB
```

MemFree: 488324496 kB

Cached: 972268 kB

Metric	Size (kB)	% of Total
MemTotal:	1,058,421,596	100.0%
MemFree:	488,324,496	46.1%
Cached:	972,268	0.1%
Free Memory (derived)	489,296,764	46.2%

Memory utilization is accessed in Enterprise Manager on the same screen as CPU utilization shown in Figure 3.8. On Compute nodes additional memory utilization information can be accessed from the Compute node home page via Host → Monitoring → Memory Details:



Command	Resident Size (KB)	Virtual Size (KB)	CPU Utilization (%)	CPU Total (seconds)	Owner	Process ID
ora_lms0_LOADTEST1	21,474,836.47	21,474,836.47	0.82	71,455	oracle	19460
ora_dbw0_LOADTEST1	21,474,836.47	21,474,836.47	0.21	18,568	oracle	19474
ora_lms1_LOADTEST1	21,474,836.47	21,474,836.47	0.81	70,959	oracle	19464
ora_dbw1_LOADTEST1	21,474,836.47	21,474,836.47	0.21	18,437	oracle	19476
ora_dbw2_LOADTEST1	21,474,836.47	21,474,836.47	0.21	18,593	oracle	19478
oracleLOADTEST1 (LOCAL=NO)	21,474,836.47	21,474,836.47	3.04	3,597	oracle	54629
oracleLOADTEST1 (LOCAL=NO)	14,531,264	21,474,836.47	0.21	443	oracle	10503
oracleLOADTEST1 (LOCAL=NO)	14,569,344	21,474,836.47	0.22	446	oracle	10507
oracleLOADTEST1 (LOCAL=NO)	14,535,748	21,474,836.47	0.21	444	oracle	10505
oracleLOADTEST1 (LOCAL=NO)	14,527,044	21,474,836.47	0.21	443	oracle	10511

Figure 3.9

Compute and Cell nodes should also be checked to ensure huge pages are configured. The following command will display that information:

```
# grep ^Huge /proc/meminfo

HugePages_Total:   22960
HugePages_Free:     2056
```

```
HugePages_Rsvd:     2016
HugePages_Surp:        0
Hugepagesize:       2048 kB
```

Examine VMSTAT

Review memory activities from OSW VMSTAT data.

- On a Compute node, go to /opt/oracle.Exawatcher/osw/archive/oswvmstat.
- Zero swapping is needed to achieve stable and good system performance.

Example of swapping: On a healthy system the swpd column would contain only 0's.

```
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu------
r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
29  2 3837960 345492 219940 17162256   28 50166    28 50302 25320 27636 71 11 10  8  0
36  2 4083500 330308 179616 17082456  204 54676   204 54726 18572 28731 62  7 13 18  0
29  1 4084292 337328 117000 16605404  102   42   161   180 32837 28761 79  6 14  1  0
16  0 4084548 331532 116952 16554832  163   51   166   188 11681 17849 47 36 18  0  0
34  0 4085052 331460 116696 16491880    6  102     6   206 13546 13483 37 52 12  0  0
30  0 4086416 332476 116336 16398348  108  290   108   492 2200 4224  8 90  1  0  0
31  2 4087308 329096 116296 16337248    1  178     1   241 1664 3576  6 93  0  0  0
35  1 4087812 329276 116040 16269804   25  102    25   224 1587 2876  6 93  1  0  0
38  1 4088556 329152 115968 16195300  204  154   242   285 1309 3790  7 92  0  0  0
43  1 4089292 342020 114728 16110568  924  128  1187   230 1614 3739  8 92  0  0  0
29  0 4090956 328376 113476 15915564  451  332   525   573 3933 6770  4 94  2  0  0
37  0 4092040 328608 111428 15804568  112  154   117   426 1251 2881  1 98  0  0  0
34  0 4093000 331672 101392 15697876  435  154   479   255 1651 3203  1 98  0  0  0
```

ILOM Events

ILOM (Integrated Lights Out Manager) is a dedicated service processor that is used to manage and monitor servers. Each Cell server, Compute node, and InfiniBand switch will have a dedicated ILOM. There are several places to view errors and messages with ILOM. The first is with the web management console. From within the web console select "Open Problems."

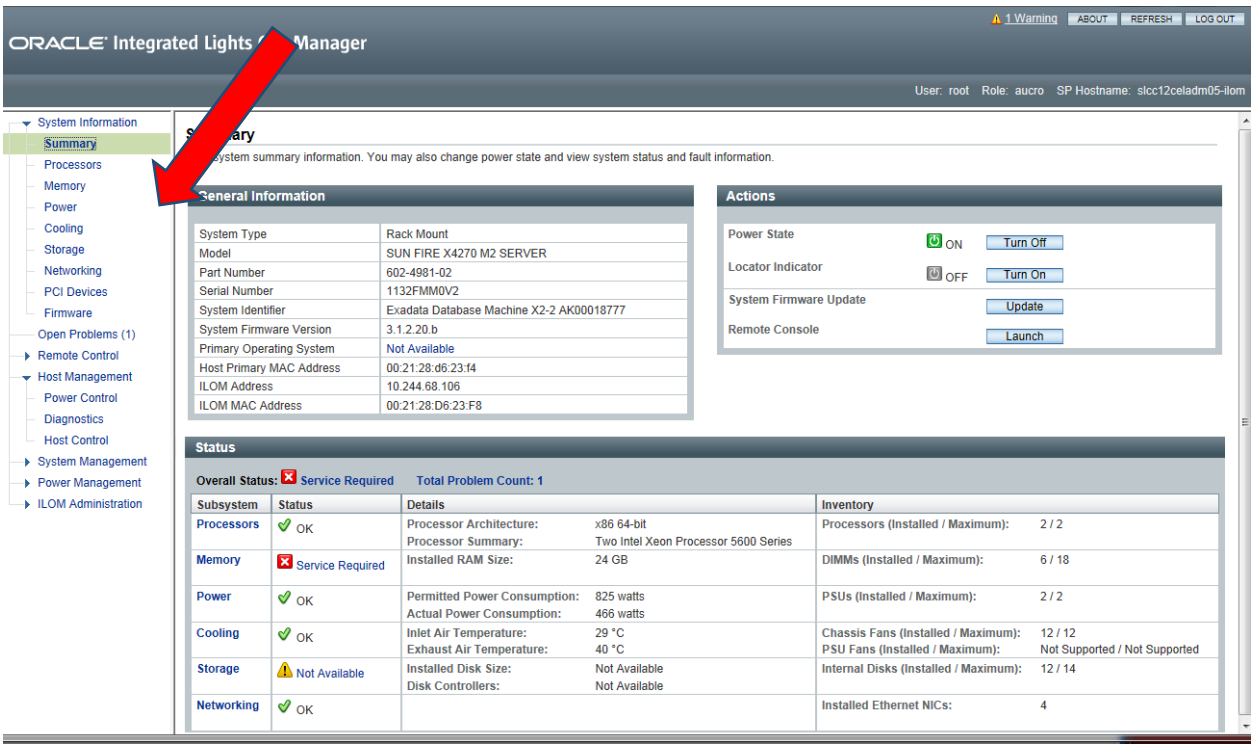


Figure 3.10

This will Display any faulted components such as CPUs, DIMMs, Fans, etc., as seen in Figure 3.11

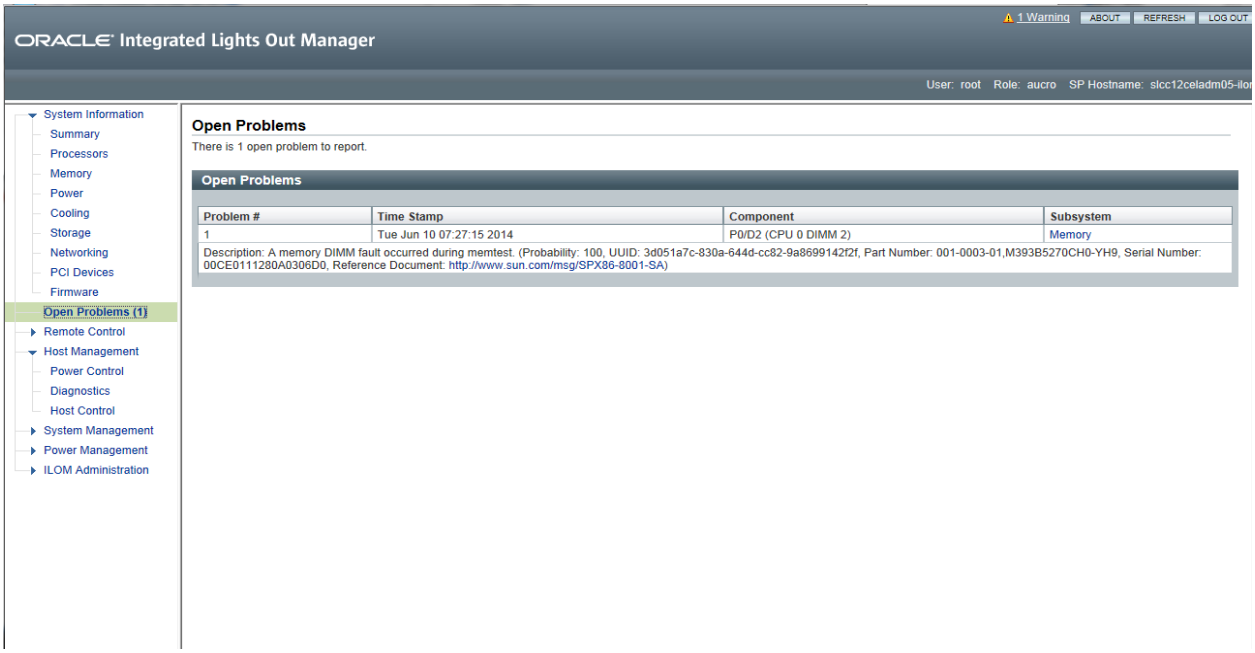


Figure 3.11

You can also view the System Event Log to get additional information that is not classified as a fault (i.e. fans over/under speed, temperature sensor warnings, etc.). Select Logs under the ILOM Administration drop down. The filter drop down box also allows the user to filter by type.

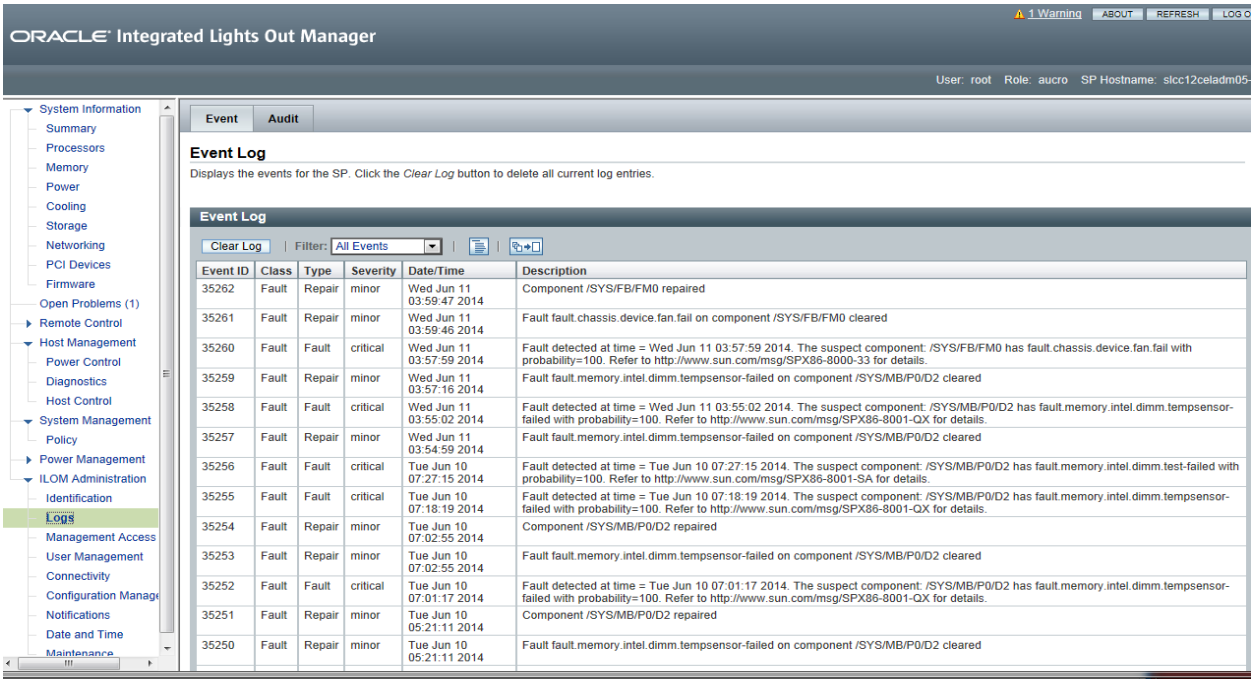


Figure 3.12

It is also possible to view the system event log from the ILOM host using the ipmitool. For example, to list the last ten events in the log file, issue the following command. Sample output follows the command:

```
# ipmitool sel list 10
  1 | 10/01/2008 | 11:32:43 | Power Supply #0x04 | Failure detected | Asserted
  2 | 07/13/2009 | 15:02:31 | Power Supply #0x05 | Failure detected | Asserted
  3 | 11/11/2009 | 13:18:20 | Fan #0x09 | Transition to Degraded
  4 | 11/11/2009 | 13:18:20 | Fan #0x09 | Transition to Running
  5 | 11/11/2009 | 13:22:36 | Fan #0x09 | Transition to Running
  6 | 11/11/2009 | 13:22:37 | Fan #0x09 | Transition to Degraded
  7 | 11/11/2009 | 13:22:37 | Fan #0x09 | Transition to Running
  8 | 11/11/2009 | 13:22:38 | Fan #0x09 | Transition to Running
  9 | 11/11/2009 | 13:22:39 | Fan #0x09 | Transition to Degraded
  a | 11/11/2009 | 13:22:39 | Fan #0x09 | Transition to Running
```

Network Status

Performing a complete diagnostic of network performance on an Exadata System is out of the scope of this white paper. However there are some simple checks that can be performed to ensure all networks are up.

On the Compute nodes, check the status of the VIPS from the cluster level using the following command, replacing node1 with the name of the Compute node. Representative output follows:

```
$ srvctl status vip -n node1
  VIP slcb0107-vip is enabled
  VIP slcb0107-vip is running on node: node1
```

From the operating system, the following command can be run to confirm all networks are up. The command returns no output if all defined networks are not in a down state:

```
dcli -l root -g ./all_group "ifconfig -a | grep DOWN"
```

Oracle Enterprise Manager is also a good source to check network status. From the host target home page select Host→Monitoring→All Metrics. See Appendix 3.3 for complete screen to screen navigation

The page will display all network interfaces active on the server and their statistics.

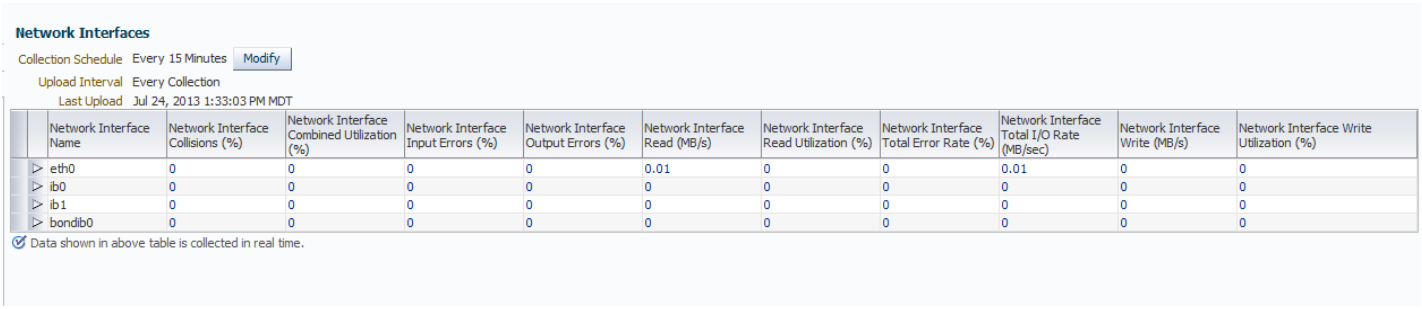


Figure 3.12

Disk Status

The MegaCli command can be used to provide a quick glimpse as to whether all disks are online. Below is a sample command that shows the status of all disks. The group file specified is a text file listing all Compute nodes and Storage nodes.

```
# dcli -g all_group -l root /opt/MegaRAID/MegaCli/MegaCli64 AdpAllInfo -aALL | grep
"Device Present" -A 8

slcb01db07: Device Present
slcb01db07: ================
slcb01db07: Virtual Drives    : 1
slcb01db07: Degraded        : 0
slcb01db07: Offline         : 0
slcb01db07: Physical Devices  : 5
slcb01db07: Disks           : 4
slcb01db07: Critical Disks  : 0
slcb01db07: Failed Disks    : 0
--
slcb01db08: Device Present
slcb01db08: ================
slcb01db08: Virtual Drives    : 1
slcb01db08: Degraded        : 0
slcb01db08: Offline         : 0
slcb01db08: Physical Devices  : 5
slcb01db08: Disks           : 4
slcb01db08: Critical Disks  : 0
slcb01db08: Failed Disks    : 0
--
slcb01cel12: Device Present
slcb01cel12: ================
slcb01cel12: Virtual Drives    : 12
slcb01cel12: Degraded        : 0
slcb01cel12: Offline         : 0
slcb01cel12: Physical Devices  : 14
slcb01cel12: Disks           : 12
slcb01cel12: Critical Disks  : 0
slcb01cel12: Failed Disks    : 0
--
slcb01cel13: Device Present
slcb01cel13: ================
slcb01cel13: Virtual Drives    : 12
slcb01cel13: Degraded        : 0
slcb01cel13: Offline         : 0
slcb01cel13: Physical Devices  : 14
slcb01cel13: Disks           : 12
slcb01cel13: Critical Disks  : 0
slcb01cel13: Failed Disks    : 0
```

If any of the drives show as degraded or offline further action should be taken to rectify the offending disk. In the event ASR has not already created a SR for the issue one should be created manually.

CheckHWnFWProfile

CheckHWnFWProfile is a program that validates whether hardware and firmware on the Compute nodes and Storage Nodes are all supported configurations. This only takes a few seconds to run and can help identify issues such as unsupported disks as demonstrated below. Note that Exachk will also execute this command to check for issues.

```
# dcli -l root -g ./all_group "/opt/oracle.SupportTools/CheckHWnFWProfile"


slcb01db07: [SUCCESS] The hardware and firmware profile matches one of the supported
profiles

slcb01db08: [SUCCESS] The hardware and firmware profile matches one of the supported
profiles

slcb01cel12: [INFO] All drives are not identical

slcb01cel12: [ERROR] Incompatible mix of disk models. All models must be from the
list: No supported models found.

slcb01cel13: [INFO] All drives are not identical

slcb01cel13: [ERROR] Incompatible mix of disk models. All models must be from the
list: No supported models found.
```

Services

If any services are defined they should be checked as well. This can be done with the lsnrctl command. Checking the scan listener will verify that all appropriate instances have handlers up and running. In the below example you can see the service haem_dbfs.us.oracle.com has a service handler on two instances in the cluster.

```
$ lsnrctl status LISTENER_SCAN2

LSNRCTL for Linux: Version 11.2.0.3.0 - Production on 24-JUL-2013 13:37:51

Copyright (c) 1991, 2011, Oracle.  All rights reserved.
Connecting to (DESCRIPTION=(ADDRESS=(PROTOCOL=IPC)(KEY=LISTENER_SCAN2)))
STATUS of the LISTENER
------------------------
Alias                     LISTENER_SCAN2
Version                   TNSLSNR for Linux: Version 11.2.0.3.0 - Production
Start Date                18-JUN-2013 19:54:18
Uptime                    35 days 17 hr. 43 min. 34 sec
Trace Level               off
Security                  ON: Local OS Authentication
SNMP                      OFF
Listener Parameter File   /u01/app/11.2.0/grid/network/admin/listener.ora
Listener Log File         /u01/app/11.2.0/grid/log/diag/tnslsnr/slcai604/listener_scan2/alert/log.xml
Listening Endpoints Summary...
  (DESCRIPTION=(ADDRESS=(PROTOCOL=ipc)(KEY=LISTENER_SCAN2)))
  (DESCRIPTION=(ADDRESS=(PROTOCOL=tcp)(HOST=0.0.0.0)(PORT=1525)))
Services Summary...
Service "haem_dbfs.us.oracle.com" has 2 instance(s).
  Instance "haemS1", status READY, has 1 handler(s) for this service...
  Instance "haemS2", status READY, has 1 handler(s) for this service...
```

Database Free Buffer Waits

A very important metric to monitor is the “free buffer wait” wait event time. Free buffer waits indicate that a database process was not able to find a free buffer into which to perform a read operation. This occurs when the DBWR process can’t write blocks to storage fast enough. “Free buffer waits” are an indication that the write rate of the I/O system is maxed out or is close to being maxed out. If this statistic appears in the top 5 wait events, then proactive action should be taken to reduce the write rate or increase the I/O capacity of storage.

Exachk

The Exadata Healthchecks Plug-in can be used within Enterprise Manager to display the latest Exachk output. Information on how to configure the Healthchecks Plug-in can be found here:

http://docs.oracle.com/cd/E11857_01/install.111/e23441/pichk.htm

Once configured, the Healthchecks target becomes a quick reference point to identify any areas not conforming to MAA best practices.

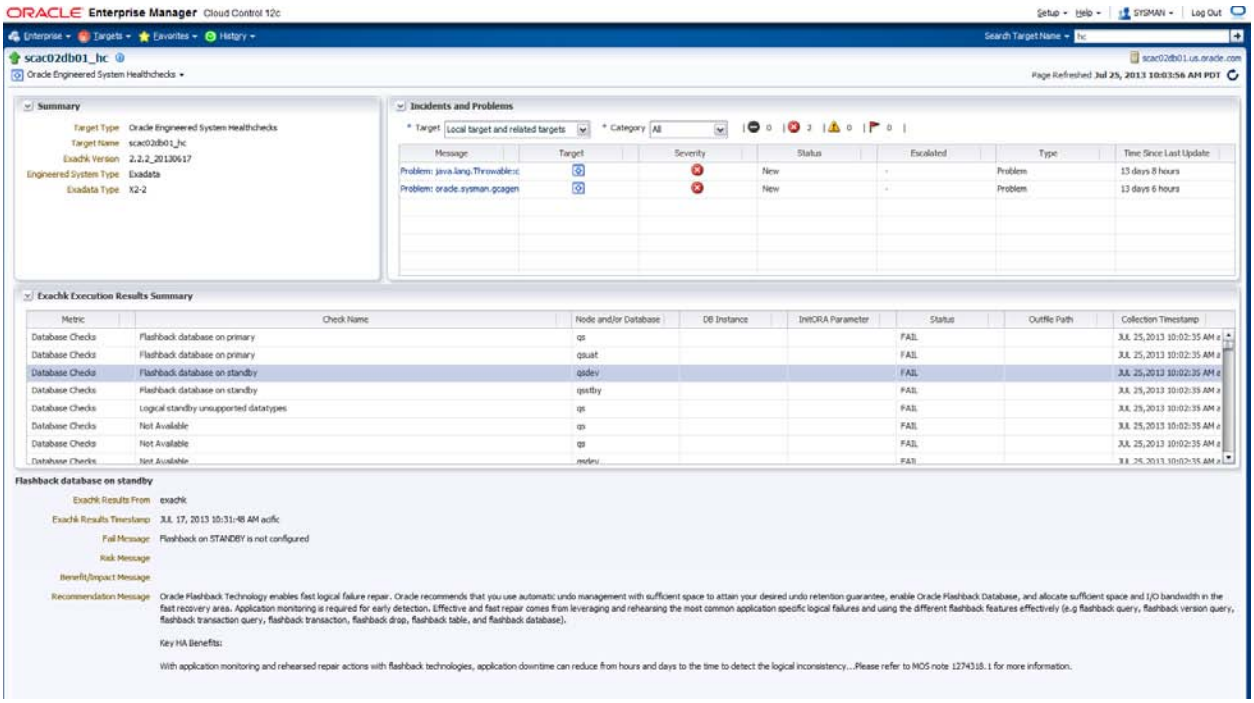


Figure 3.14

Starting with Exachk 2.2.1, the capability exists to compare two Exachk outputs. This can be useful in identifying changes.

See Section I for more information.

Have Changes Occurred in the Environment?

Changes to the environment can often have unintended side effects. Identifying recent changes to a system is an important first step to help pinpoint a common source of issues. If proper change management processes are in place then identifying changes should be quick and easy. Otherwise it may be necessary to begin investigating possible sources of change. These could include:

- Recent Oracle patching (Operating System, Database, Cell server, Clusterware, etc.)
- Newly deployed applications
- Code changes to existing applications
- Other changes in usage (i.e. new users added)
- Oracle configuration changes
- Operating system configuration changes
- Migration to a new platform
- Expansion of the environment
- Addition of other InfiniBand devices to the fabric
- Changes in resource management plans

Depending on separation of duties, checking all sources of change could be as easy as talking to one person or to many teams in large, siloed organizations.

If changes are identified to have occurred on the system, steps should be taken to ensure the changes are not related to the identified problem. If the changes are determined to be causing negative consequences, then analysis should be performed to identify the best course of action. This could include rolling back the change, increasing capacity, modifying code, etc.

Use baseline data to troubleshoot issues

Compare Configuration Files

Changes in configuration files can easily cause issues in a system. A Simple diff command can reveal recent changes made to files. Following the Suggestions from Section “Steps to follow before problems occur” will ensure backups have been made of critical files before problems arise, enabling comparison.

Spfile and password file information is binary which prevents the diff command from comparing them. However, by using the strings command the ASCII data can be exported in order to perform a comparison:

```
$ strings spfileemrep.ora > spfileemrep.ora.txt
$ strings spfileemrep.ora_072513_0100 > spfileemrep.ora_072513_0100.txt
$ diff spfileemrep.ora.txt spfileemrep.ora_072513_0100.txt
31c31
< *.processes=300
---
> *.processes=600
35c35
```

```
< *.sga_target=2147483648
---
> *.sga_target=1147483648
```

Differences in spfile/init files will also be available when running compare period reports in Enterprise Manager. However, the data is only as granular as the collected AWR snapshots.

Checking changes to the kernel tunable parameters

It is a good idea to compare the kernel settings from a known good copy. The below command will perform a comparison between an earlier baseline copy of the kernel parameters and the current configuration. The dbs_group file is simply a text file listing all the Compute nodes. Replace <baseline kernel configuration file> with the appropriate file.

```
# dcli -l root -g ./dbs_group "sysctl -a > /tmp/sysctl.current;diff /root/<baseline
kernel configuration file> /tmp/sysctl.current"
```

It is normal for some parameters to change dynamically. So the above output should be carefully analyzed to determine if the delta from the diff output is relevant to the issues being experienced.

AWR Data

AWR data provides critical information necessary to troubleshooting database issues. AWR detailed analysis will specifically be discussed later in the paper; however, deviations in performance and workload can be identified quickly using the AWR baseline data described in “Steps to follow before problems occur.” Changes in key metrics such as number of users, number of transactions, redo rate, physical reads per transaction, physical writes per transaction, etc. can help quickly identify changes that have occurred in the system

To compare a current AWR report to a baseline in Enterprise Manager, choose which periods to compare. For the First Period select a static or moving baseline which provides the closest performance benchmark. For example, compare similar time periods or workload data. For the second period select snapshot that encapsulates the performance issue. Next push the “Generate Report” button. For screen by screen navigation see Appendix 4.1.

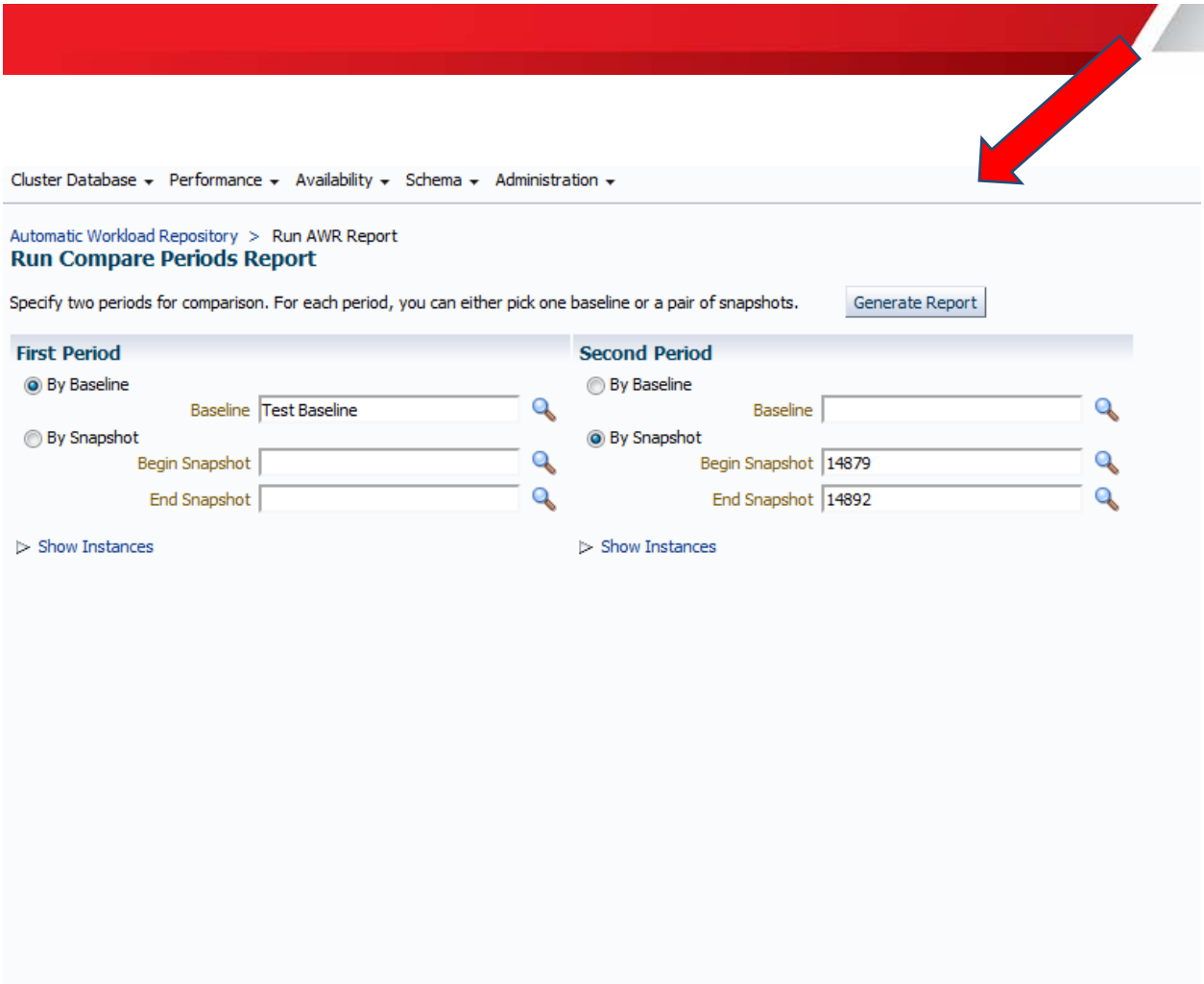


Figure 4.1

A report will be generated which will demonstrate differences in the workload and behavior of the database.

Report Summary

Host Configuration Comparison

- CPU and Memory values are from the end snapshot; averaged across all instances
- Other values are averages for all instances

	1st	2nd	Diff	%Diff
Number of CPUs:	32	32	0	0.0
Number of CPU Cores:	16	16	0	0.0
Number of CPU Sockets:	2	2	0	0.0
Physical Memory:	258064.9M	258064.9M	0M	0.0
Load at Start Snapshot:	1.81	1.97	.16	8.8
Load at End Snapshot:	1.9	1.86	-.04	-2.1
%User Time:	1.01	1	-.02	-1.0
%System Time:	.57	.56	-.01	-1.8
%Idle Time:	97.87	97.91	.04	0.0
%IO Wait Time:	.04	.04	0	0.0

Cache Sizes

- Cache Sizes are from the end snapshot; averaged across all instances

	1st (M)	2nd (M)	Diff (M)	%Diff
Memory Target				
SGA Target	2,048.0	2,048.0	0.0	*0.0*
Buffer Cache	552.0	552.0	0.0	*0.0*
Shared Pool	1,384.0	1,384.0	0.0	*0.0*
Large Pool	16.0	16.0	0.0	*0.0*
Java Pool	24.0	24.0	0.0	*0.0*
Streams Pool				
PGA Target	1,024.0	1,024.0	0.0	*0.0*
Log Buffer	10.0	10.0	0.0	*0.0*

Load Profile

	1st per sec	2nd per sec	%Diff	1st per txn	2nd per txn	%Diff
DB time:	1.1	1.1	*-3.6*	0.1	0.1	*0.0*
CPU time:	0.6	0.6	*0.0*	0.0	0.0	*0.0*
Redo size:	107,420.0	112,108.8	*4.4*	5,988.0	6,118.4	*2.2*
Logical reads:	9,380.8	9,493.7	*1.2*	522.9	518.1	*-0.9*
Block changes:	673.4	713.9	*6.0*	37.5	39.0	*3.8*
Physical reads:	139.3	119.4	*-14.3*	7.8	6.5	*-16.0*
Physical writes:	45.1	38.2	*-15.4*	2.5	2.1	*-17.1*
User calls:	88.2	98.1	*11.3*	4.9	5.4	*8.9*
Parses:	117.5	119.0	*1.2*	6.6	6.5	*-0.9*
Hard parses:	5.2	4.8	*-6.2*	0.3	0.3	*-10.3*
W/A MB processed:	3,651,674.2	4,621,325.5	*26.6*	203,558.7	252,211.0	*26.6*
Logons:	2.4	2.7	*12.7*	0.1	0.1	*15.4*
Executes:	758.5	770.7	*1.6*	42.3	42.1	*-0.5*
Transactions:	17.9	18.3	*2.1*			

	1st	2nd	Diff
% Blocks changed per Read:	7.2	7.5	*0.3*
Recursive Call %:	95.8	95.4	*-0.4*
Rollback per transaction %:	66.7	66.0	*-0.7*
Rows per Sort:	12.2	14.2	*2.0*

Figure 4.2

Advanced Diagnostics

Hardware Rule out

In this section we will go through the steps to rule out I/O performance or saturation by understanding the total number of IOPS your system should be able to achieve

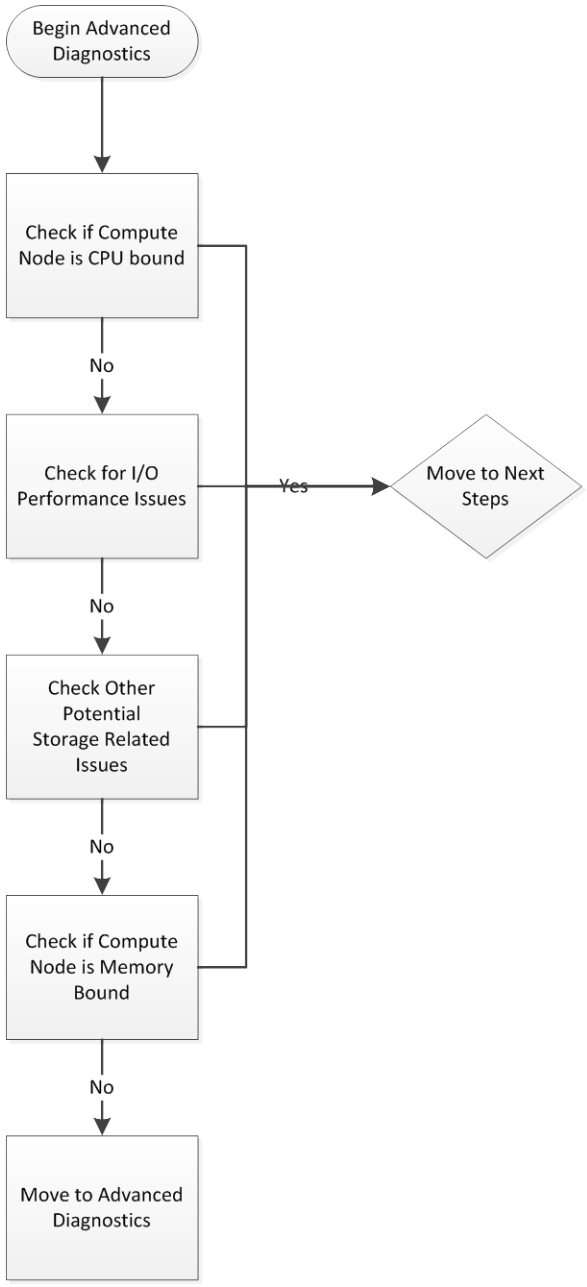


Figure 5.1

The flowchart in figure 5.1 describes the process flow followed.

The advanced diagnostics section will proceed through several areas of checks to help identify problem areas. First the Compute nodes will be checked to rule out over utilization of the CPUs. If the system is CPU saturated at the compute node it becomes hard to determine an I/O bottleneck, therefore this will be checked and eliminated first. Next how to analyze I/O load (both HDD and Flashcache) on the cell server to determine if saturation is occurring. This will be done utilizing several tools including Enterprise Manager, Exawatcher and Cell metrics. Following this identifying I/O consumption by database will be covered. Finally some miscellaneous other storage checks will be discussed followed by next steps to solving problems once they've been identified.

Check if Compute node is CPU bound

Review load average from OSW TOP data.

- On a Compute node, go to /opt/oracle.Exawatcher/osw/archive/oswtop.

Remember # of CPU cores = max load.

Refer to the Oracle Exadata Database Machine Data Sheets for CPU core count.

Evaluate load average per core = # of runnable processes per core

- Question: Is load average of 80 high?
- Answer: It depends.
    - X2-2, load/core = 80/12 ~= 6.67 runnable processes per core => yikes!
    - X2-8, load/core = 80/64 ~= 1.25 runnable processes per core => ok!

The 3 load-average values are the 1-minute, 5-minute, and 15-minute averages.

```
top - 08:29:25 up 2 days,  7:34,  0 users,  load average: 283.02, 436.25, 422.83
Cpu(s):  1.4%us, 97.1%sy,  0.0%ni,  0.4%id,  1.0%wa,  0.0%hi,  0.1%si, 0.0%st
Mem:  98848968k total, 98407464k used,   441504k free,      916k buffers
 PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
1401 root      12  -5     0    0    0 R 81.0  0.0  23:50.36 [kswapd0]
```

```
top - 08:29:36 up 2 days,  7:34,  0 users,  load average: 259.12, 426.07, 419.66
Cpu(s):  2.5%us, 78.5%sy,  0.0%ni,  4.8%id, 14.1%wa,  0.0%hi,  0.1%si, 0.0%st
Mem:  98848968k total, 98401084k used,   447884k free,      792k buffers
 PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
1401 root      10  -5     0    0    0 R 123.1  0.0  23:57.74 [kswapd0]
```

Compute load/core = 283 / 12 ~= 23 runnable processes per core

Note that Compute nodes that are CPU bound will incorrectly show high I/O wait times because the process that issues an I/O will not be immediately rescheduled when the I/O completes. Therefore CPU scheduling time will be measured as part of I/O wait times. Thus, I/O response times measured at the database level are not accurate when the CPU is maxed out. Thus it is important to have ruled out CPU contention as documented above.

I/O Performance

The characteristics of hard disks make it difficult to predict their behavior as workloads change. In particular:

- Disk controllers have a memory cache which can be used to improve I/O performance, especially for writes. The cache can become overwhelmed as write rates increase, resulting in a sudden increase in write latency. Disk utilization can be deceptive because the true disk utilization is hidden from the operating system by disk controller write caching.
- Random disk operations can take more time than serial operations because the disk head must be repositioned between each write.
- Large I/O operations can take longer to process due to the time needed for data transfer.
- As utilization increases more time is spent waiting for the disk to become free and this can significantly increase I/O latencies.
- Disks don't have a hard I/O limit. The more I/Os that are queued, the higher the I/O throughput from the disk. This is because disk controllers can perform more intelligent scheduling of I/Os when more I/Os are concurrently queued. Therefore a performance tradeoff must be made between I/O throughput and response times. Queuing more I/Os will increase system throughput at the expense of response time.

On X2, X3, and X4 systems, high performance disks can execute about 300 small I/Os per second (IOPS) without a large increase in response time (peak performance is actually above 400 IOPS), or 50,000 IOPS on a full rack. A large I/O is roughly 3 times as expensive as a small IO. You can determine approximate disk utilization by counting the total small I/Os and adding the total large I/Os multiplied by 3. Then compare this count to the 300 IOPS threshold to determine utilization. For 4TB high capacity disks, the IOPS are around 120 and the multiplier for large I/Os is about 2.

4TB high capacity disks have IOPS around 190 or 32,000 IOPS for a full rack and should also use a 2x multiplier for large I/Os. For additional Exadata capacity details please reference the Oracle Exadata Database Machine Data Sheets.

High disk latencies are not necessarily a problem – it depends on how the application is impacted. For a Data Warehouse it may be perfectly fine for the disks to be running at maximum throughput and latency when processing queries.

Fortunately, Exadata flash cache and flash log reduce disk I/Os and insulate the application from many of the effects of increased disk latencies.

- Smart flash log will allow a commit to complete quickly even if the disk redo write latency is sometimes high.
- Smart flash cache will reduce database latencies for disk reads and writes by servicing both directly from flash. Performance of the flash cache will also be affected by its size and usage as data is loaded and aged out of cache. If batch or reporting jobs occasionally make disk latencies high, flash cache insulates interactive users and keeps response times good.

If disk utilization is increasing or is expected to increase, it is important to carefully monitor the effect on the performance. Watch for:

- Increased wait event times for "cell single block physical read" and "log file parallel write."
- Increased times for "cell multiblock physical read," "log file sync," and "db file parallel write" can also be important but the wait times for these events often vary significantly even in normal workloads so they are not as useful for monitoring trends.
- Increased response times for OLTP applications or increased batch runtimes.

Check if cells are I/O bound

Check if the cell servers are I/O bound compute total HDD and FLASH throughput (MBPS) and IOPS. These rates can be obtained from Exawatcher, Enterprise Manager or directly from the cell as explained below

Refer to Oracle Exadata Database Machine Data Sheets or check Appendix 5.1 for peak numbers:

- Watch out for high latency if I/Os ever approach peak numbers.
  - High latency does NOT mean slow disks.
    - Each I/O takes long primarily due to time waiting in disk queue.
  - I/O latency can be >100ms (note disks are not slow!).
  - I/O latency depends on disk queue length so can be varied based on different workloads.
- Be aware that max MBPS and max IOPS cannot be reached simultaneously.
  - How to evaluate mixed workload?
    - Examine disk utilization - is it close to 100%?
    - Run calibrate if needed (requires Cells being quiesced).

Enterprise Manager

Enterprise Manager provides a quick and easy way to identify if Cells are I/O bound. Both on the “Aggregate Cell server performance page” as well as the “Individual Cell server performance page”. The checkbox, “Show Maximum Cell Disk Limit”, creates a maximum I/O limit line on the IOPS and MBPS graphs. This limit is calculated based on the Exadata hardware version and number of servers deployed.

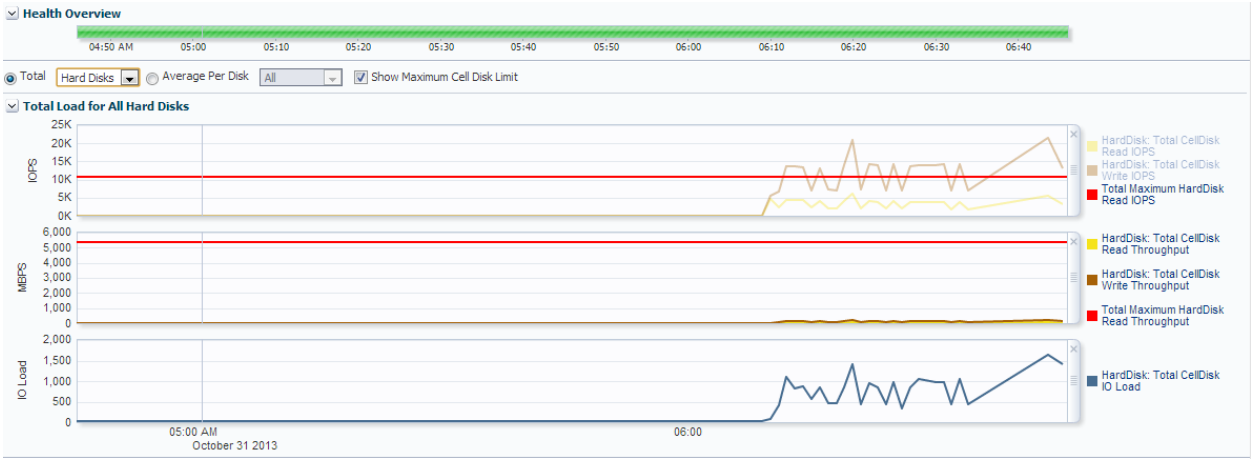


Figure 5.1

The data can also be isolated to either flash disks or hard disks by the drop box at the top of the screen as seen in Figure 5.2

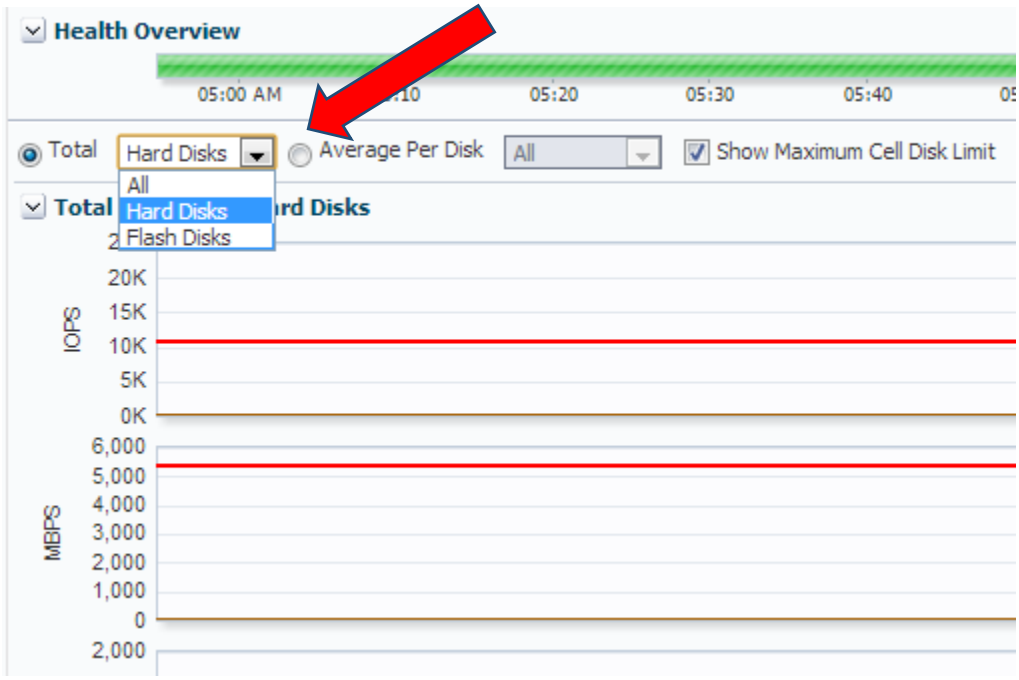


Figure 5.2

Exawatcher

Where to find Exawatcher IOSTAT

On a Cell server, go to /opt/oracle.Exawatcher/osw/archive/oswiostat

- Snapshots are taken every 5 seconds for all devices (both HDD and flash).
- Snapshots are stored in a file for each hour.
- Snapshots are retained for 7 days.

How to interpret Exawatcher IOSTAT

Column Name	What does it mean?
Device	Device name (map device back to CellDisk)
r/s	# of read I/Os per second
w/s	# of write I/Os per second
rsec/s	# of sectors (half KB) read per second
wsec/s	# of sectors (half KB) written per second
avgrq-sz	Average I/O request size in sector (half KB)
avgqu-sz	Average disk queue size
Await	Average I/O latency in milliseconds (including service time + time in queue)
Svctm	Average I/O service time in milliseconds
%util	Device bandwidth utilization(0-100%)

How to analyze Exawatcher IOSTAT

HDD Snapshot from 2TB High Capacity Cell

```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
          10.72    0.00    0.34    5.50    0.00   83.44
Device:         rrqm/s   wrqm/s   r/s   w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sda (system)     53.20    20.20 61.00 12.20 119305.60   275.20  1633.62    50.51  704.65  12.94  94.74
```

```
sdb (system)     63.00    21.60 67.40   9.60 134588.80   278.40  1751.52   185.82 2424.44  12.99 100.02
sdc              77.80     0.00 80.00   2.60 154174.40    11.60  1866.66    15.03  158.55   9.59  79.24
sdd              77.40     0.00 78.60   0.40 156766.40    12.80  1984.55    13.10  163.34   9.46  74.70
sde              53.00     0.00 64.80   0.20 129342.40     1.60  1989.91    28.14  506.29   9.95  64.68
sdf              67.60     0.00 70.00   0.40 138744.00    32.00  1971.25     9.12  130.45  10.51  73.96
sdg              88.40     0.00 83.00   0.80 159870.40    57.60  1908.45    15.11  148.26   9.76  81.76
sdh              86.80     0.00 83.00   4.60 165675.20  1640.00  1909.99    14.57  139.70   9.28  81.26
sdi              93.60     0.00 94.20   0.00 188516.80     0.00  2001.24    18.27  183.00   9.34  87.96
sdj              83.40     0.00 89.80   3.20 178910.40  1235.20  1937.05    12.79  140.77   8.79  81.76
sdk              78.20     0.00 77.80   0.60 155710.40    19.20  1986.35    11.05  138.07   9.84  77.16
sdl              75.20     0.00 79.60   1.60 155230.40    64.00  1912.49    16.34  172.51   9.67  78.50
```

What characteristics can be summarized from the snapshot?

- Workload consists primarily of reads.
- Average request size ~= 1MB reads.
- First 2 devices (i.e., system disks) are close to 100% utilization rate.
- Devices deliver close to peak throughput as listed in the Oracle Exadata Database Machine Data Sheets (85MB/sec).
- Possible Data Warehouse workload with large reads that saturate the disk subsystem.

Using Cell Disk and Flash Cache metrics

In addition to using Exawatcher, the following Cell Disk and Flash Cache metrics offer another insight into cell I/O characteristics:

- Cell Disk throughput metrics
- Cell Disk IOPS metrics
- Flash Cache throughput metrics
- Flash Cache IOPS metrics

HDD Cell Disk I/O throughput metrics are described in Monitoring Cell Disk Metrics.

What metrics show HDD Cell Disk I/O throughput

Metric	What does it mean?
CD_IO_BY_R_LG_SEC	# of MBs read per second using large I/Os
CD_IO_BY_W_LG_SEC	# of MBs written per second using large I/Os
CD_IO_BY_R_SM_SEC	# of MBs read per second using small I/Os

CD_IO_BY_W_SM_SEC	# of MBs written per second using small I/Os

Large I/Os > 128KB

Small I/Os <= 128KB

How to analyze HDD Cell Disk I/O throughput metrics

Add up all 4 metrics and compare to the datasheet numbers. If I/O cumulative number is close or exceeds datasheet numbers the environment is I/O bound

If workloads are HDD I/O throughput bound large requests tend to dominate:

```
CELLCLI> list metrichistory CD_IO_BY_R_LG_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00"  and
metricObjectName like 'CD_.*';
```

What metrics show HDD Cell Disk IOPS

Metric	What does it mean?
CD_IO_RQ_R_LG_SEC	# of large read requests per second
CD_IO_RQ_R_SM_SEC	# of small read requests per second
CD_IO_RQ_W_LG_SEC	# of large write requests per second
CD_IO_RQ_W_SM_SEC	# of small write requests per second

How to analyze Cell Disk metrics

Add up all 4 metrics and compare to Oracle Exadata Database Machine Data Sheets numbers

If workloads are HDD IOPS bound small requests tend to dominate:

```
CELLCLI> list metrichistory CD_IO_RQ_R_SM_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00"  and
metricObjectName like 'CD_.*';
```

What metrics show Flash Cache I/O throughput and IOPS

Flash Disk I/O throughput and IOPS metrics are described in Monitoring Cell Disk Metrics.

Metric	What does it mean?
CD_IO_BY_R_SM_SEC	The rate which is the number of MB read in small blocks per second from a cell disk.
CD_IO_RQ_R_SM_SEC	The rate which is the number of requests to read small blocks per second from a cell disk.

Flash Cache I/Os are always small because Flash Cache has 32K cache lines.

How to analyze Flash Cache metrics

Use the Cell Disk metrics and compare to Oracle Exadata Database Machine Data Sheets numbers:

```
CELLCLI> list metrichistory CD_IO_BY_R_SM_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00"  and
metricObjectName like 'FD_.*';

CELLCLI> list metrichistory CD_IO_RQ_R_SM_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00"  and
metricObjectName like 'FD_.*';
```

For other Flash Cache metrics, refer to Monitoring Flash Cache Metrics.

All of the above Flash and HDD I/O information can be viewed easily in Enterprise Manager as well. Simply navigate to the Exadata Grid target performance page. Select "Total" in the Show drop down list. Select the "Show Small and Large Requests" checkbox. Use the Slider or the Select Time Range button to customize the graph's timeline.

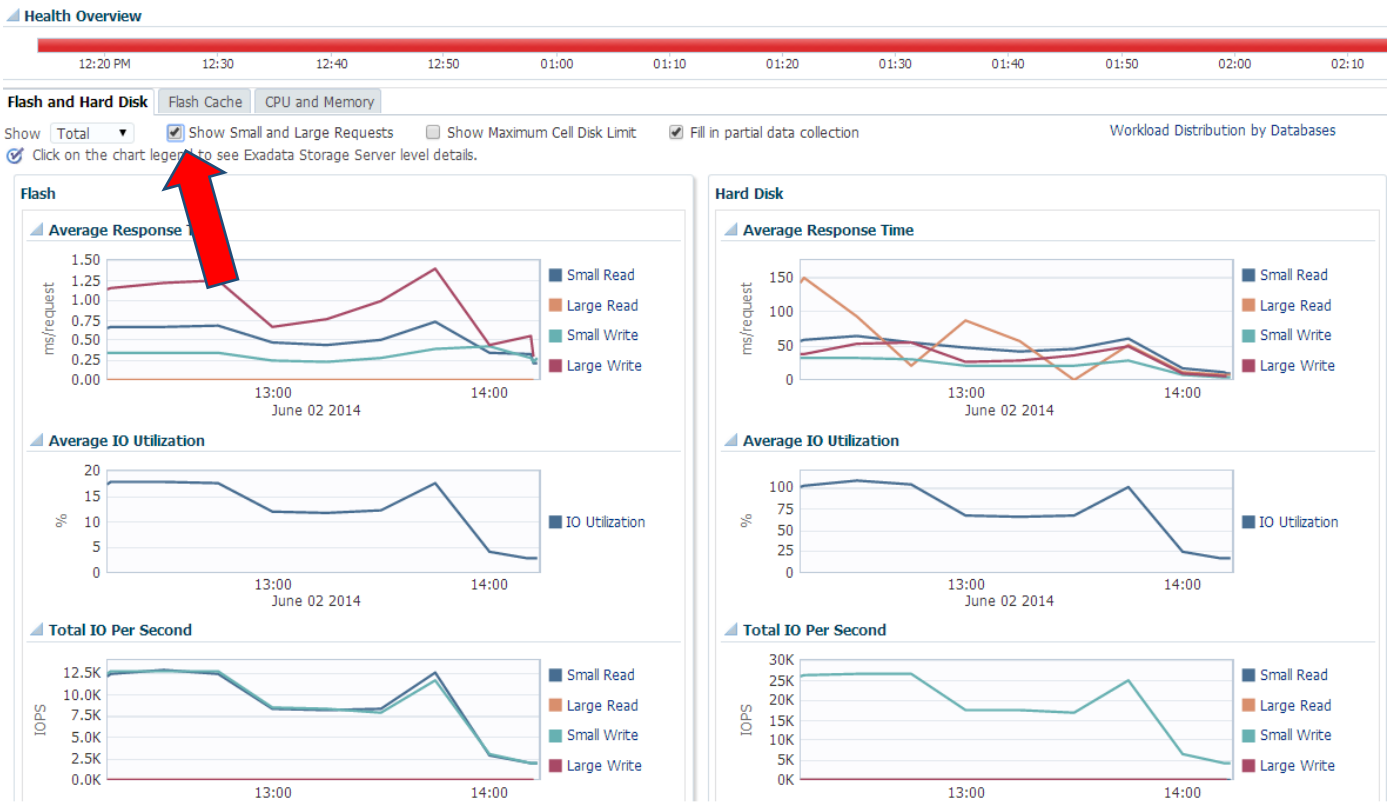


Figure 5.3

I/O Bound in a Mixed Workload

When workloads are mixed (DW & OLTP), Cells can still be I/O bound when neither MBPS nor IOPS is at peak numbers from the previous exercises.

What to do then?

Examine OSW IOSTAT util% for device bandwidth saturation.

For HDD, add up DB_IO_UTIL_LG & DB_IO_UTIL_SM for all databases to see if the total utilization approaches 100%.

Run calibrate to rule out slow disks when in doubt. (Cells must be quiesced, so this may need to be done in a rolling fashion to avoid downtime).

How to analyze Cell DB HDD I/O metrics

If Cells are found to be I/O bound through the previous exercises and there are multiple databases sharing the same Cell server then identify the top I/O consumers/databases using Cell DB I/O metrics

Enterprise Manager

I/O per database can be viewed in Enterprise Manager by selecting the Workload Distribution by Database link on the Grid Target performance page.

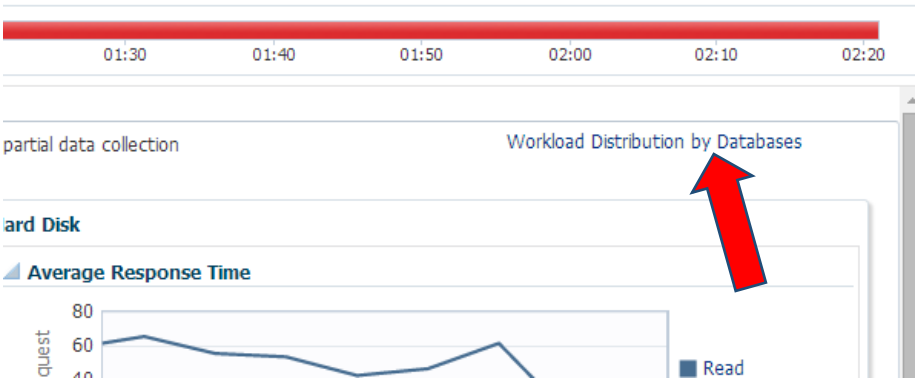


Figure 5.4

Figure 5.5

What Cell metrics show DB HDD I/O breakdown

DB HDD I/O metrics are described in Monitoring IORM with Database Metrics.

These metrics are available even if IORM is not used.

Metric	What does it mean?
DB_IO_UTIL_LG	The percentage of HDD disk resources utilized by large requests from this database.
DB_IO_UTIL_SM	The percentage of HDD disk resources utilized by small requests from this database.

Add up both metrics for each database and identify the top I/O consuming databases:

```
CELLCLI> list metrichistory DB_IO_UTIL_LG where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00“ and
metricObjectName = ‘DW_db’;

CELLCLI> list metrichistory DB_IO_UTIL_SM where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00“ and
metricObjectName = ‘DW_db’;
```

Be sure to analyze _OTHER_DATABASE_, ASM, clusterware, and other miscellaneous I/Os as well.

Sum both metrics for all databases including _OTHER_DATABASE_ and see if the cell is approaching full HDD utilization.

What metrics show DB Flash Cache I/O breakdown

DB Flash Cache I/O metrics are described in Monitoring IORM with Database Metrics.

These metrics are available even if IORM is not used.

Metric	What does it mean?
DB_FC_IO_BY_SEC	This metric shows the number of megabytes of I/O per second for this database to flash cache.
DB_FC_IO_RQ_SEC	This metric shows the number of I/O requests issued by a database to flash cache per second.

Example – Cell DB HDD I/O breakdown

Identify DBs that consume most HDD I/O resources using cell DB metrics

```
CELLCLI> list metrichistory DB_IO_UTIL_LG where collectionTime < "2011-10-25T05:13:00-
05:00" and collectionTime > "2011-10-25T05:11:00-05:00”

         DB_IO_UTIL_LG   DSS                    75 %    2011-10-25T05:11:30-05:00

         DB_IO_UTIL_LG   DW                     10 %    2011-10-25T05:11:30-05:00
```

```
DB_IO_UTIL_LG   MTXDB                5  %    2011-10-25T05:11:30-05:00
DB_IO_UTIL_LG   _OTHER_DATABASE_     0 %     2011-10-25T05:11:30-05:00
DB_IO_UTIL_LG   DSS                 43 %    2011-10-25T05:12:30-05:00
DB_IO_UTIL_LG   DW                   8  %    2011-10-25T05:12:30-05:00
DB_IO_UTIL_LG   MTXDB                6  %    2011-10-25T05:12:30-05:00
DB_IO_UTIL_LG   _OTHER_DATABASE_     0 %     2011-10-25T05:12:30-05:00
```

How to analyze DB Flash Cache I/O metrics?

Sort the DB Flash Cache I/O metrics by database and identify the top I/O consuming databases

Throughput (DW):

```
CELLCLI> list metrichistory DB_FC_IO_BY_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00” and
metricObjectName = ‘DW_db’;
```

IOPS (OLTP):

```
CELLCLI> list metrichistory DB_FC_IO_RQ_SEC where collectionTime < "2011-10-
25T04:00:00-05:00" and collectionTime > "2011-10-25T03:59:00-05:00"  and
metricObjectName = ‘OLTP_db’;
```

Or more simply you can analyze the aggregated “Flash Cache” tab on the performance page of the Exadata Grid target in Enterprise Manager.

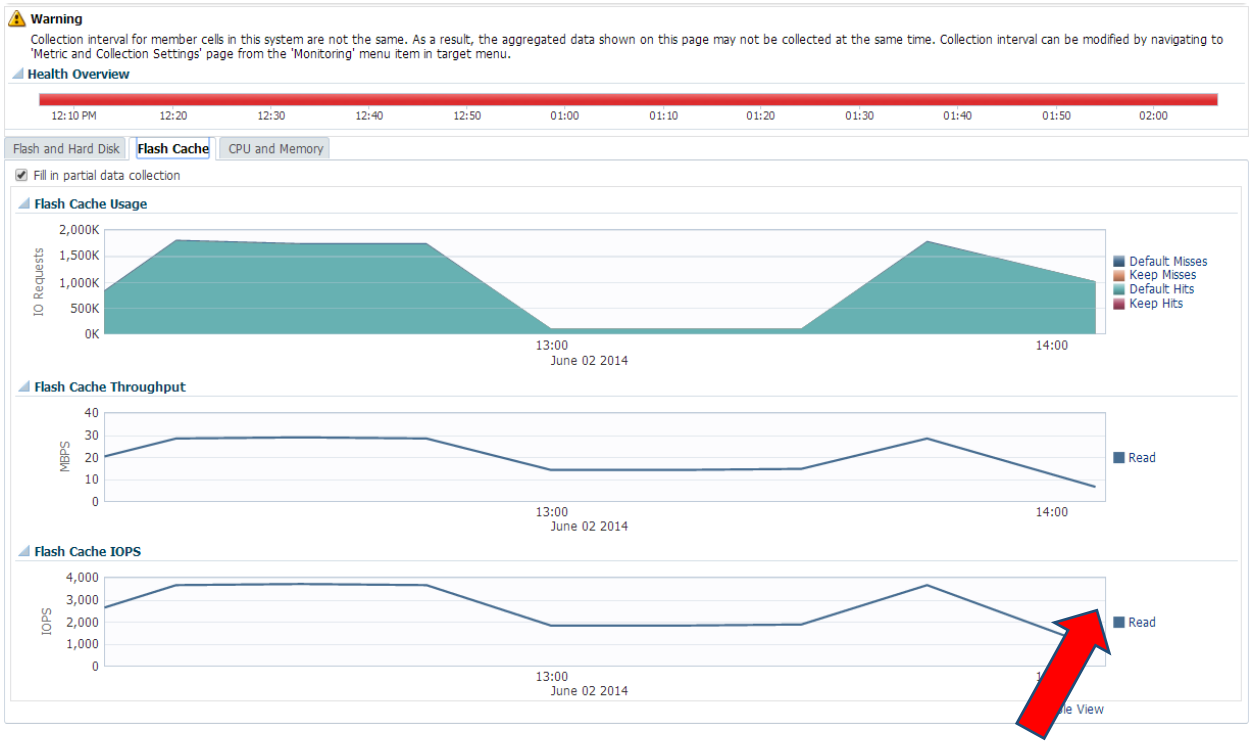


Figure 5.6

The data points can also be viewed individually by select the “Table View” link.

Performance

View | Show last known values

Time	Flash Cache Usage				Flash Cache Throughput	Flash Cache IOPS
	Default Misses	Keep Misses	Default Hits	Keep Hits	Read	Read
Jun 2, 2014 12:15:00 PM	1,828,686.97	1,806,864.50	1,806,864.50	0	28.76	3,676.56
Jun 2, 2014 12:30:00 PM	1,750,753.23	1,731,602.00	1,731,602.00	0	29.32	3,748.03
Jun 2, 2014 12:45:00 PM	1,764,650.12	1,745,687.75	1,745,687.75	0	28.87	3,690.56
Jun 2, 2014 1:00:00 PM	111,808.50	110,953.00	110,953.00	0	14.48	1,849.22
Jun 2, 2014 1:15:00 PM	111,443.50	110,588.50	110,588.50	0	14.44	1,843.14
Jun 2, 2014 1:30:00 PM	114,519.50	113,465.50	113,465.50	0	14.81	1,891.09
Jun 2, 2014 1:45:00 PM	1,807,464.46	1,791,757.12	1,791,757.12	0	28.95	3,701.13
Jun 2, 2014 2:05:10 PM	1,027,171.16	1,018,264.25	1,018,264.25	0	6.90	879.36
Jun 2, 2014 2:06:21 PM	1,027,124.16	1,018,217.25	1,018,217.25	0	6.89	879.15
Jun 2, 2014 2:07:33 PM	1,027,135.85	1,018,228.94	1,018,228.94	0	6.90	879.31

Figure 5.7

When cells are identified as I/O bound in a mixed work load environment consider the following:

Focus on the top I/O consuming databases.
Typical problem scenarios:

- Data Warehouse workloads:
  - Disks are busy but flash is idle: Can objects be marked KEEP?

- OLTP workloads:
  - Reads are fine but disks are write IOPS bound: Is checkpoint too aggressive or is buffer cache under sized?
  - Should Write Back flash cache be enable?

General strategies:

- Data Warehouse:
  - Tune TOP SQL statements.
  - Use compression if extra CPU is available: Trade CPU for disk.
- Both DW and OLTP:
  - Use IORM to prioritize I/O resources between databases and consumer groups.

Next Steps

If Cells nodes are I/O bound

Possible remedies for high I/O utilization are:

- Ensure data layout best practices are followed (see the section on SAME below)
- Make sure smart flash cache and smart flash log are configured.
- "Keep" important or high hit rate segments in buffer cache or flash cache.
- Monitor for "free buffer waits" and if these are high enable write back flash cache assuming the system is running Exadata version 11.2.3.2.1 and GI 11.2.0.3 BP9 (or higher).
- Implement compression to reduce the size of data and therefore reduce the number of I/Os that are necessary to run the application.

SAME (Stripe And Mirror Everywhere):

- When any disk or Cell is maxed out, performance will be throttled by that disk/Cell even with workload parallelization.
- Use as many Cells as possible rather than splitting a rack into smaller clusters.
- Be aware of potential performance disparity between system and data disks:
- System disks not only have user data but also have Cell's own file systems.
- System disks may run slower than data disks.
- This is more pronounced on High Capacity 3TB drives due to lower IOPS capacity when compared with High Performance 1.2TB drives.
- If all cells and disks are performing correctly with load evenly distributed, IOPS are saturated. If SLAs are not being met, add resources or begin SQL tuning.

Tune high I/O rate SQL:

- There may be a better query plan that performs fewer I/Os.
- If disks are maxed out then pay particular attention to SQL with high "unoptimized" reads.
- Optimized reads are serviced from the flash cache and don't contribute to high disk utilization.
- Check the SQL tuning guide for more information

If Compute node is CPU or memory bound

If Compute node is running out of CPU:

- Check database parameter settings against Best Practice MOS Note 757552.1
- Reduce load via tuning or adding capacity.

If Compute node is running out of memory (swapping)

- Check database parameter settings against Best Practice MOS Note 757552.1.
- Configure HugePages. (On Linux, if HugePages are not configured appropriately, this can lead to over utilization of memory. It is a "must" best practice on Exadata). See Deploying Oracle Maximum Availability Architecture with Exadata.
- Reduce load via tuning or adding capacity.

If Cell server is not I/O bound and Compute nodes are not CPU/memory bound

Are cells CPU bound?

- Exadata would automatically push load back to the Compute nodes when Cell CPUs are maxed out.

Are cells memory bound?

- Exadata Cells automatically manage their own memory to guard against swapping.

Go back to Oracle Performance Tuning Guide.

- AWR, ADDM, ASH, etc. (Covered in Section Database Diagnostic Tools).

Cell server Target Page

Enterprise Manager offers aggregate disk performance information at the Cell server level for hard disks as well as flash cache. This can provide a valuable resource to help identify potential bottlenecks as well as establish baseline I/O statistics. To access the I/O page select the Cell server target from the DB Machine Target home then select Performance from the Exadata Cell server Page

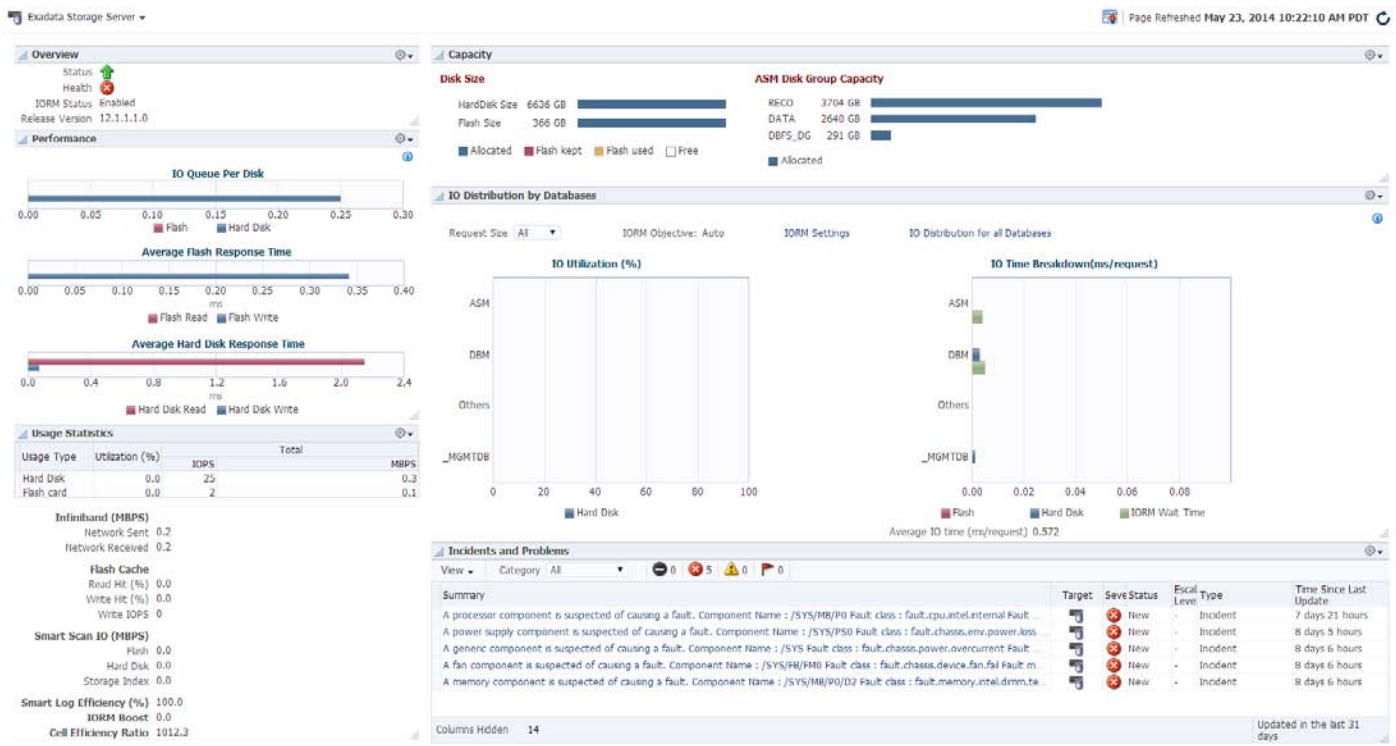


Figure 5.8

Exadata Cell Performance View

Figure 5.9 is a sample output from an Exadata Cell's performance view. Here you can observe current and historical information on memory and flash cache usage as well as CPU and workload distribution by database.

Figure 5.9

File System Information

It is also possible to view detailed information on the mounted file system on the Compute node. Screen by screen navigation can be found in Appendix 5.2. The resulting screen allows access to storage information on the file systems, ASM, local disks and databases.

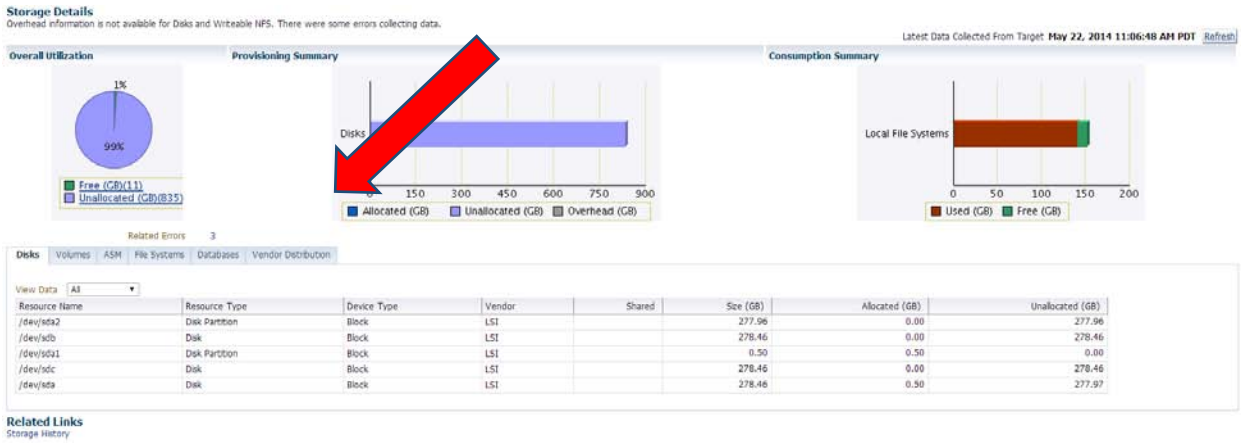


Figure 5.10

If more detailed specific ASM information is required the ASM target page itself can be viewed. To access this information from the Database Machine home page, select the Compute node on which to view the ASM target from the left hand navigation menu. See Appendix 5.3 for screen by screen navigation.

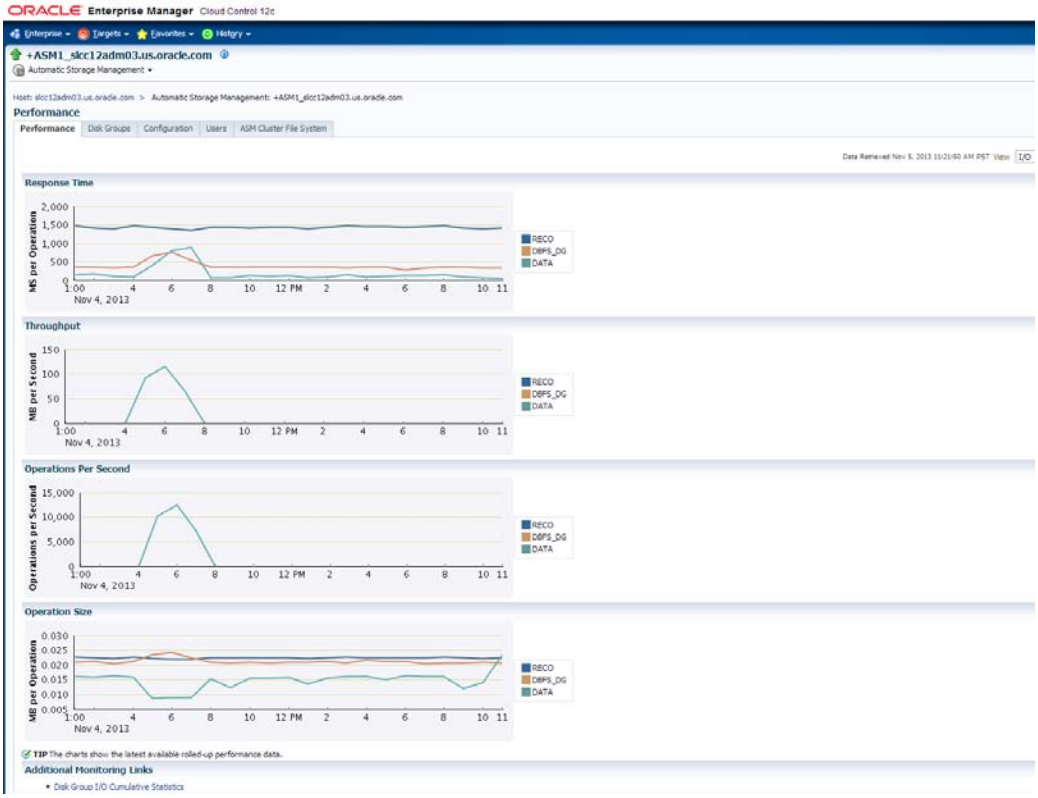


Figure 5.11

Common Performance use cases

Problem Scenario #1 – HDD are busy but flash is idle

Common DW problem scenario:
HDD disks are busy but flash is idle due to large reads issued by smart scans bypassing flash cache.

Solution:
Use flash for KEEP objects so large reads can be offloaded to flash.
Execute the following steps:

1. Run I/O intensity report @?/rdbms/admin/spawrio
2. Ensure the total size of KEEP objects does not overwhelm flash cache size.
    a. Be aware that allowed KEEP size is restricted to 80% of flash cache size.
    b. Target small tables with lots of reads for KEEP.
3. Mark each candidate table as KEEP.

4. Repeat workload and verify read I/O offload to flash.

Step 1 – Sample I/O intensity report

```
ASH Activity - Estimate of I/O wait times
-> # Samples: # of samples in ASH - approximation of DB time
-> % Activity: approximation of % Time based on ASH samples
-> ID values are based on aggregation type:
   by Wait Class: Wait Class name
   by Segments  : Owner.Object(SubObject)
                  * wildcard is used if object or subobject name is too long
   by File      : FileID-FileName

                                                        Obj.
Space        IO
Aggregation   Id                                      Type       Tablespace        # Samples
GB     Intensity  % Activity
------------- ----------------------------------------- ---------- ---------- ---------------- ---
------- ---------- ---------- by Segment       DBM01_ATS.TECS_PHC(P2011)                  TABLE
PART DBM01_D02              87,083       67.8    1,284.9        3.2
             DBM01_ATS.TECS_PHC(P2011)                  TABLE PART DBM01_16K_D0
87,083       67.8    1,284.9        3.2
             DBM01_ATS.ENTITY_ADDR                      TABLE      DBM01_D01
34,103       83.6      408.1        1.3
             DBM01_ATS.SHIPMENT_SCORE(P2011)            TABLE PART DBM01_D07
25,543       85.2      299.8        0.9
             DBM01_ATS.TECS_PHC(P2010)                  TABLE PART DBM01_16K_D0
15,006       98.3      152.6        0.6
             DBM01_ATS.TECS_PHC(P2010)                  TABLE PART DBM01_D08
15,006       98.3      152.6        0.6
             DBM01_ATS.RULE_FINDING(P2011)              TABLE PART DBM01_D07
13,079      137.6       95.0        0.5
             DBM01_ATS.XBKSHIPMENTSC*(P2011)            INDEX PART DBM01_I02
12,904      153.5       84.1        0.5

IO Intensity - by Tablespace
-> I/O Intensity calculated as IO Reqs per sec/GB allocated
-> tablespaces with >= 0.25 % of Captured IOs displayed
-> %IOPs - Running Total of % of Captured IOPs
   %Cap  - IOs as a percentage of Captured  IOPs
   %Tot  - IOs as a percentage of Total sysstat IOPs
           Sysstat IOs per Sec:    7,532.4
-> ordered by Total IOPs desc, IO Intensity desc

             IO Req     Rd Req     Wr Req     Space        IO        Read
Write               Total      Total                    Total     Read MB    Write MB
Tablespace  per Sec    per Sec    per Sec       GB     Intensity  Intensity  Intensity
%IOPs       IO MB      IO Blks %Reads %Writes    MB/s     per Sec    per Sec    %Cap   %Tot
---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ------ --
------------ ------- ------ ------- ---------- ---------- ---------- ------ ------
DBM01_D07         370.8      276.7       94.1   13,818.1        0.0        0.0        0.0
11.4   12,942,243.8  828.3M   83.4    16.6       13.8       11.5        2.3   11.4    4.9
DBM01_D01         353.1      213.2      139.9  110,880.0        0.0        0.0        0.0
22.2   27,989,013.2 1791.3M   86.5    13.5       28.6       24.4        4.1   10.8    4.7
DBM01_D06         217.2      197.0       20.2   26,444.0        0.0        0.0        0.0
28.8   15,848,043.0 1014.3M   97.1     2.9       17.2       16.7        0.5    6.7    2.9
DBM01_16K_D0      172.9      167.3        5.6   30,693.6        0.0        0.0        0.0   34.1
140,063,687.1 8964.1M   99.9     0.1      139.9      139.8        0.1    5.3    2.3
```

```
DBM01_D02          120.0      114.2        5.8    7,215.8        0.0        0.0        0.0
54.9    11,069,533.8  708.5M   96.9      3.1       13.4       13.1        0.3    3.7    1.6
```

Step 2 – How to evaluate total KEEP size

```
Id                                      Type        GB     Intensity
--------------------------------------- ---------- ------ ------------

DBM01_ATS.TECS_PHC(P2011)                 TABLE PART 67.8    1,284.9
DBM01_ATS.TECS_PHC(P2011)                 TABLE PART 67.8    1,284.9
DBM01_ATS.ENTITY_ADDR                     TABLE      83.6      408.1
```

Total KEEP size = 67.8 + 67.8 + 83.6 = 219.2 GB

Default Flash Cache size per cell = 1.6 TB

	Full Rack(X3)	Half Rack(X3)	Quarter Rack(X3)
Flash Cache Size	~22.4TB	~11.2TB	~5.6TB

Analyze the flash cache utilization rate prior to KEEP.

Ensure that newly marked KEEP objects do not trump other critical workloads effectively utilizing flash cache.

Step 3 – How to mark objects as KEEP

Run the following SQL statements:

```
ALTER TABLE TECS_PHC MODIFY PARTITION P2011 STORAGE (CELL_FLASH_CACHE KEEP);

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

Step 4 – How to verify flash usage

Repeat the same workloads.
Examine OSW IOSTAT and Cell metrics to confirm:

- Reduction in disk I/O usage (less large reads issued to HDD).
- Increase in flash I/O usage (more small reads issued to Flash Cache).

Ensure that complete workloads across all databases run faster (not slower). Watch out for potential flash cache thrashing if total KEEP size becomes too large:

- Newer KEEP cache lines will evict older KEEP cache lines – the default cache lines in the 20% of flash cache remain intact.
- Retrieve current KEEP usage using:

```
list metriccurrent attributes name, metricvalue where name like
'FC_BYKEEP_USED'
```

- If current KEEP size is close to 80% of total flash cache size, scale back on KEEP.

Problem Scenario #2 – HDD are write IOPS bound

Newer versions of the Cell Server software support Write Back writes which increases Performance. Cell server versions < 11.2.3.2.x only support Write Through mode so writes are not cached. In these earlier versions Exadata Flash Cache ensures ample read IOPS; however write performance does not improve.

What if HDD disks are write IOPS bound?

- Check if MTTR/Checkpoint is too aggressive.
  - If so, relax MTTR target.
- Check if buffer cache is undersized.
  - If so, increase buffer cache size.

Step 1 – How to examine database writes using Buffer Pool Statistics

Examine Buffer Pool Statistics

- Go to AWR->Buffer Pool Statistics
- Physical Writes include checkpoint writes, aging writes, etc

Buffer Pool Statistics

- Standard block size Pools D: default, K: keep, R: recycle
- Default Pools for other block sizes: 2k, 4k, 8k, 16k, 32k

P	Number of Buffers	Pool Hit%	Buffer Gets	Physical Reads	Physical Writes	Free Buff Wait	Writ Comp Wait	Buffer Busy Waits
D	4,748,295	91	202,981,617	18,237,002	10,511,448	0	0	4,173

Back to Buffer Pool Statistics
Back to Top

Checkpoint Activity

- Total Physical Writes: 10,513,291

MTTR Writes	Log Size Writes	Log Ckpt Writes	Other Settings Writes	Autotune Ckpt Writes	Thread Ckpt Writes
1,585,932	0	0	0	897,687	0

Figure 5.13

Step 2 – How to examine database writes using MTTR advisory

Go to AWR->MTTR Advisory
Compare ‘Size for Est (s)’ and ‘Est Total Writes’
Increase fast_start_mttr_target setting to reduce writes.

- Be aware of longer recovery time tradeoff.

For additional information, refer to Recovery Tuning in Performance Tuning Guide

MTTR Advisory

- Only rows at end snap are displayed
- Estimated Writes and I/Os are in thousands

Size for Est (s)	Dirty Limit	Est Cache Writes	Est Cache Write Fctr	Est Total Writes	Est Total Write Fctr	Est Total IOs	Est Total IO Fctr
90	150,365	98,675	2.83	98,681	2.83	153,985	1.71
450	751,829	48,173	1.38	48,178	1.38	103,482	1.15
900	1,503,658	34,873	1.00	34,879	1.00	90,182	1.00
1,211	2,023,256	29,433	0.84	29,439	0.84	84,742	0.94
1,485	2,480,007	27,763	0.80	27,769	0.80	83,072	0.92

Figure 5.14

Step 3 – How to evaluate buffer cache size

If checkpoint writes do not contribute to the total physical writes significantly, check if buffer cache may be undersized which may lead to excessive aging writes.

What are the signs for undersized buffer cache?

- If AWR Buffer Pool Advisory (based on v$db_cache_advisory) shows significant savings in reads with size increase, it will most likely reduce aging writes as well, but there is no guarantee.
- Check for long latencies in “db file parallel write.”

Increase buffer pool size if needed.

For more information, refer to Configuring and Using the Buffer Cache in Performance Tuning Guide.

Tune TOP SQLs for DW

When DW workloads are IO bound and do not fall into the common problem scenario #1 – Can the total I/Os for top SQLs be reduced?

Identify top SQLs from AWR.

- Optimized I/Os = I/Os served from flash cache + saved by storage index.
- Unoptimized I/Os = HDD I/Os (flash griddisk I/Os if configured).

Generate SQL monitor report for each top SQL.

Go back to SQL tuning guide.

- Missing index?
- Check for undersized PGA where temp segments are used instead of work areas.

When users complain that Exadata performance is slow:

- Check if Cells are I/O bound.

- Check if Compute nodes are CPU or memory bound.
- If neither is true, go back to database performance tuning and planning.

Database Diagnostic Tools

ADDM

ADDM is an excellent tool for analyzing system performance metrics over a period of time and identifying areas of focus for further tuning. In some cases, things may be running optimally and so no further tuning is needed. ADDM advice will often focus on top SQL statements and wait events. ADDM also checks for many common issues such as excessive parsing of SQL, excessive logons, and free buffer waits that may adversely affect performance.

Enterprise Manager also provides an easy to use graphical interface for ADDM. This is available by going to the database target page and selecting performance home from the performance drop down list. Then select the “Real Time ADDM Now” button on the right hand side of the screen.

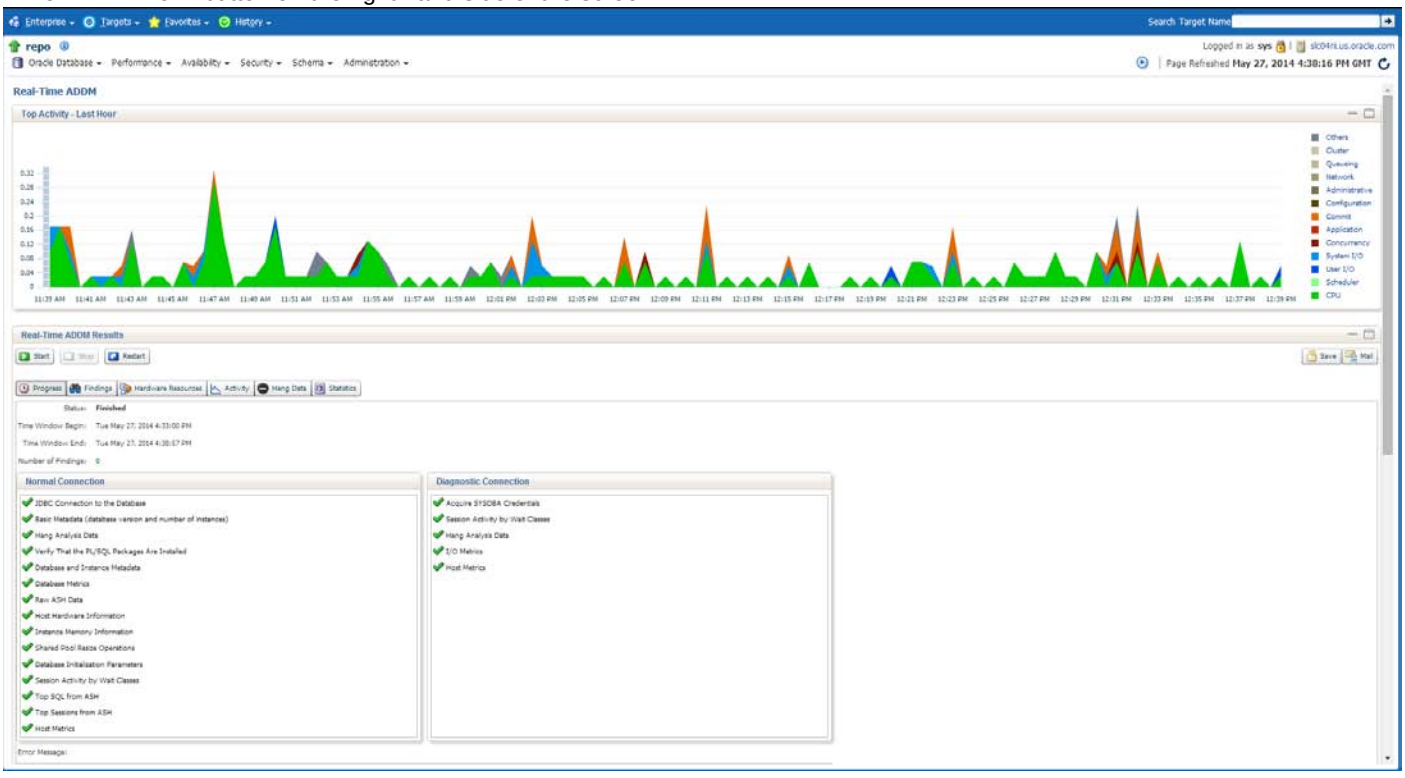


Figure 6.1

Sample output from Enterprise Manager ADDM run:

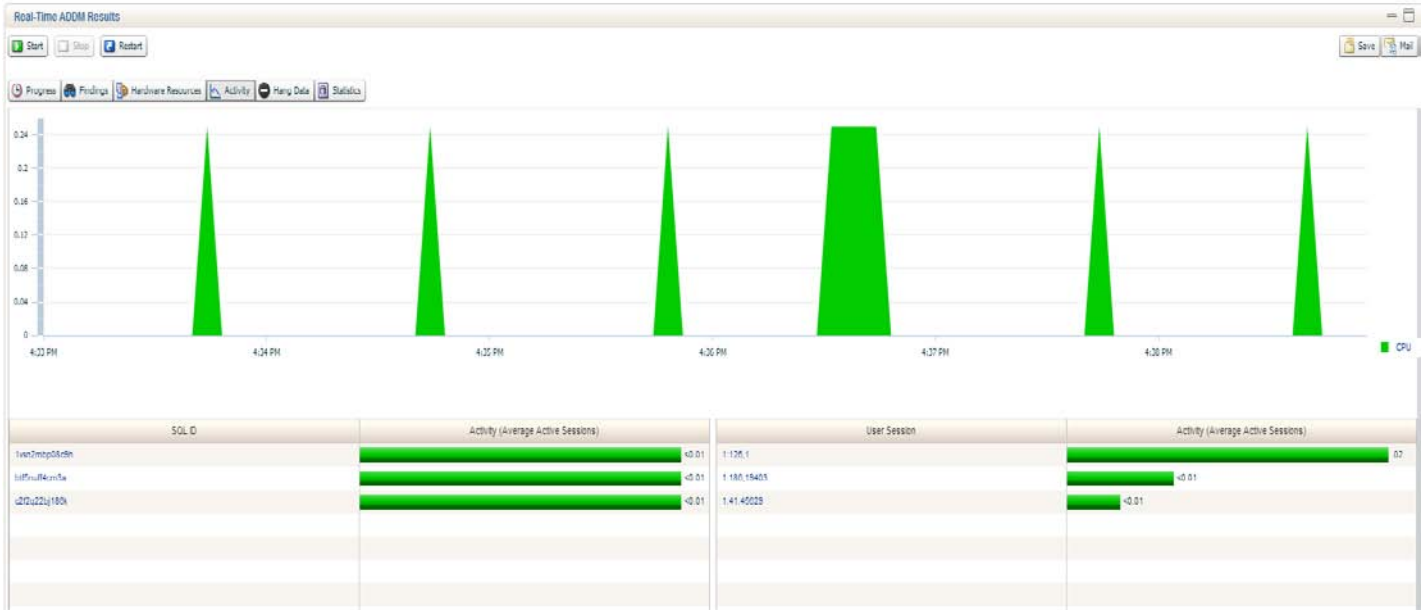


Figure 6.2

Top SQL

The top SQL relative to elapsed time, CPU utilization, buffer cache gets, and many other metrics is available in an AWR report.

The top SQL relative to elapsed time has the most direct relationship to application performance. This may be a long query running only a few times, or a short query running many times.

The top SQL with respect to CPU time and buffer cache gets is useful in finding queries that may have bad query plans.

The top SQL with respect to I/O is useful for identifying queries that are issuing excessive I/Os.

The SQL Tuning Advisor can be used to evaluate alternate query plans for a specific query and provide estimates of the performance of each plan. It may also suggest other changes that could improve its performance, such as adding indexes.

Database Top Activity Page

The Enterprise Manager Database Target Top Activity page is a graphical representation of the database Active Session History data. This page can be used to quickly find changes in the baseline signature of your database. In this example we can see a change in Configuration waits, which is the result of this database instance being throttled for IO. Notice the Database Writers become the top sessions as more time is spent flushing the buffer cache to disk.

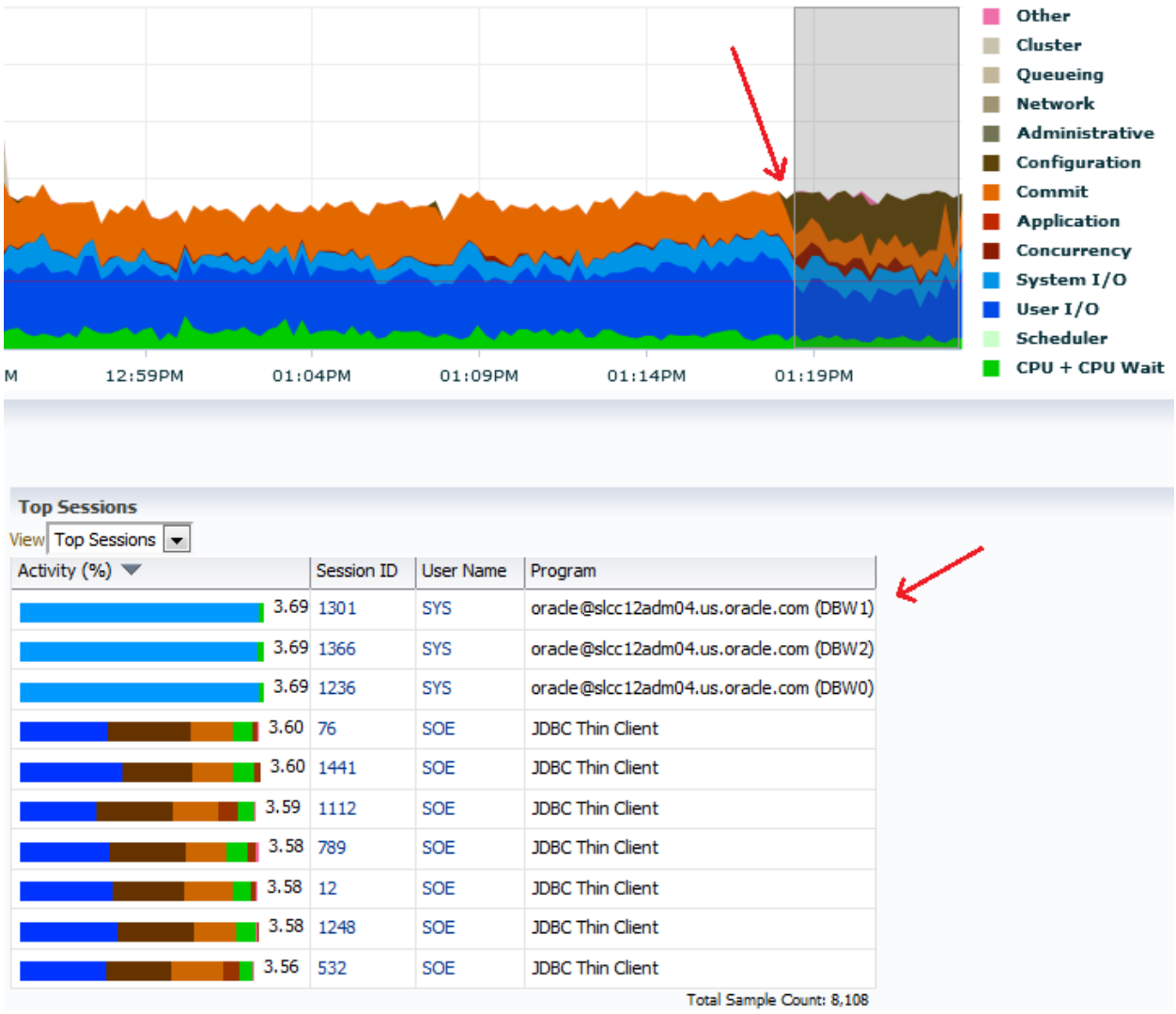


Figure 6.3

AWR Reports

AWR Reports can be obtained from Enterprise Manager. Simply go to the database target home page and Select AWR→ AWR Report.

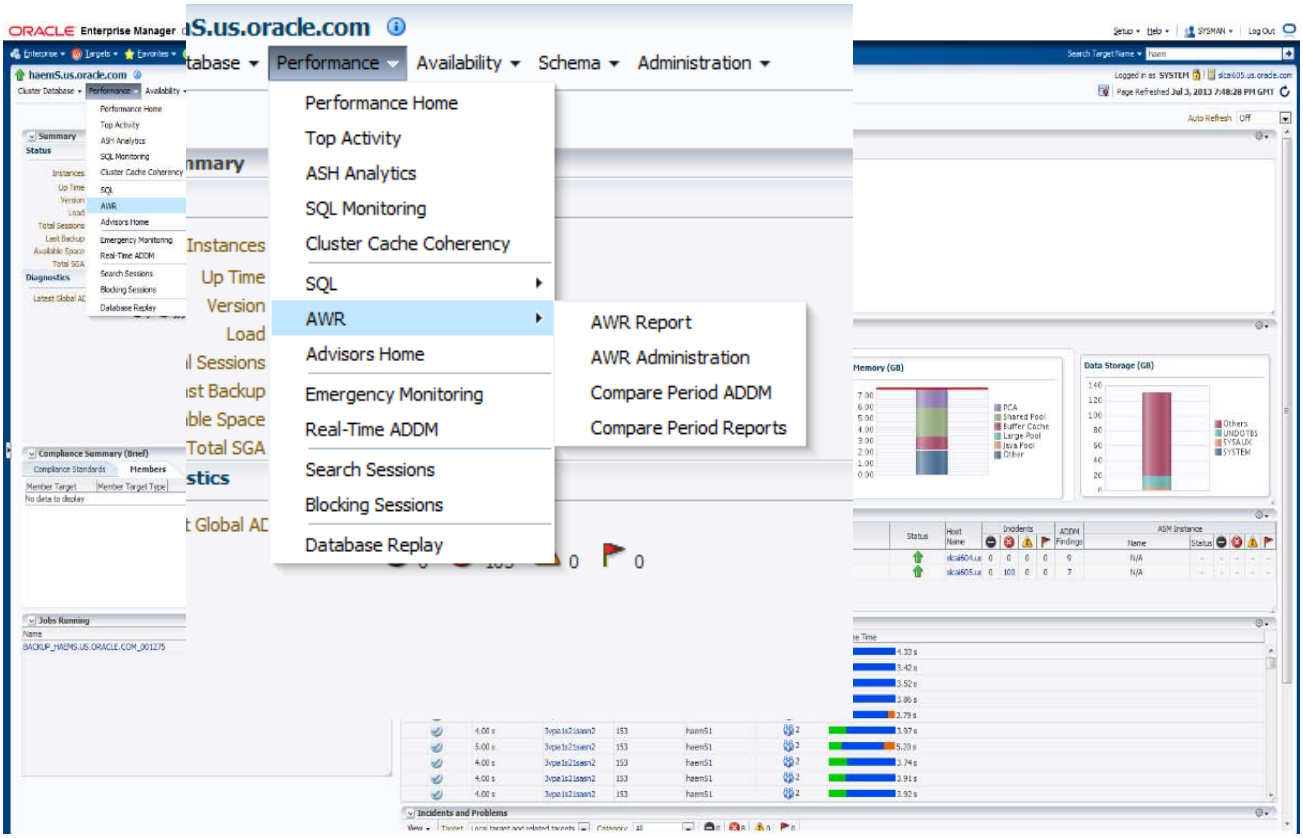


Figure 6.4

You will then have the option of selecting the time range of the report. The option is also available to run the AWR report for individual instances in the case of a RAC database.

repo

Oracle Database ▾ Performance ▾ Availability ▾ Security ▾ Schema ▾ Administration ▾

WORKLOAD REPOSITORY report for

DB Name	DB Id	Instance	Inst num	Startup Time	Release	RAC
REPO	2970173645	repo	1	02-May-14 16:05	11.2.0.4.0	NO

Host Name	Platform	CPUs	Cores	Sockets	Memory (GB)
slc04rii	Linux x86 64-bit	2	2	2	12.11

	Snap Id	Snap Time	Sessions	Cursors/Session
Begin Snap:	409	19-May-14 17:01:00	70	4.9
End Snap:	600	27-May-14 16:00:15	77	5.5
Elapsed:		11,459.26 (mins)		
DB Time:		262.44 (mins)		

Report Summary

Load Profile

	Per Second	Per Transaction	Per Exec	Per Call
DB Time(s):	0.0	0.0	0.00	0.00
DB CPU(s):	0.0	0.0	0.00	0.00
Redo size (bytes):	13,854.5	9,062.8		
Logical read (blocks):	436.7	285.7		
Block changes:	88.6	58.0		
Physical read (blocks):	22.2	14.6		
Physical write (blocks):	3.1	2.0		
Read IO requests:	1.8	1.2		
Write IO requests:	1.2	0.8		
Read IO (MB):	0.2	0.1		
Write IO (MB):	0.0	0.0		
User calls:	9.2	6.0		
Parses (SQL):	26.3	17.2		
Hard parses (SQL):	0.3	0.2		
SQL Work Area (MB):	0.2	0.2		
Logons:	0.4	0.3		
Executes (SQL):	58.2	38.1		
Rollbacks:	0.0	0.0		
Transactions:	1.5			

Instance Efficiency Percentages (Target 100%)

Buffer Nowait %:	99.99 Redo NoWait %:	100.00
Buffer Hit %:	96.23 In-memory Sort %:	100.00
Library Hit %:	99.25 Soft Parse %:	98.84
Execute to Parse %:	54.76 Latch Hit %:	99.95
Parse CPU to Parse Elapsd %:	53.90 % Non-Parse CPU:	92.72

Top 10 Foreground Events by Total Wait Time

Event	Waits	Total Wait Time (sec)	Wait Avg(ms)	% DB time	Wait Class
DB CPU		9759.1		62.0	

Figure 6.5

Top Wait Events

The top wait events are also listed in an AWR report. An event that is consuming a significant percentage of DB time will have a large effect on the performance of the database workload, so an assessment should be made as to whether the event is consistent with application baseline performance, or needs attention. Here is an example where we see particularly high “library cache: mutex X” waits:

Top 5 Timed Foreground Events

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
DB CPU		10,442		81.89	
library cache: mutex X	4,560,452	2,149	0	16.85	Concurrency
buffer busy waits	2,556,621	96	0	0.75	Concurrency
SQL*Net more data from client	3,759,147	89	0	0.70	Network
log file sync	6,624	30	5	0.23	Commit

Figure 6.6

In this case, it was not normal to have this event and a change to the application was required, but the assessment and potential solutions will vary depending on the event and the workload. The important thing is to develop a picture of what to expect with different workloads and investigate any anomalies and variations when they occur.

The total wait time for an event is the most important metric to look at. Tuning efforts that focus on reducing waits for events that consume very small fractions of the total wait time on the system will generally provide little or no benefit.

New features in Exadata Plug-in 12.1.0.6

The Oracle Exadata Plug-in provides a consolidated view of the Exadata Database Machine within Oracle Enterprise Manager. Plug-in Release 12.1.0.6.0 includes a variety of bug fixes and enhancements that allow for an even more powerful interface for Exadata. Enhancements include:

- Fine-grained performance summary for flash and hard disk with side-by-side comparison.

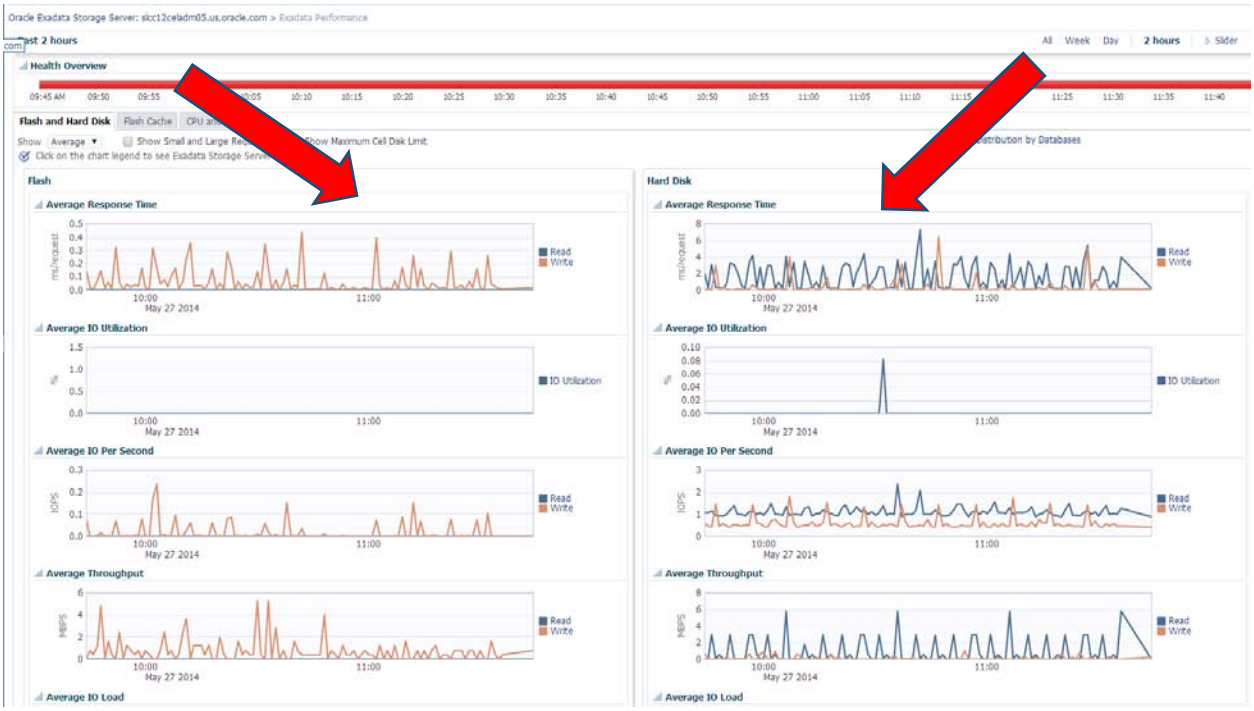
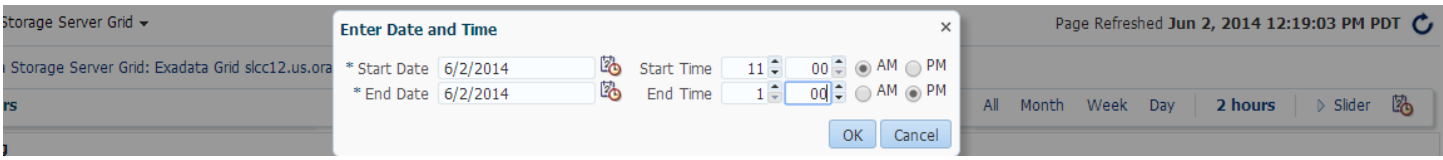


Figure 7.1

- The Ability to fine tune date ranges for graphs via hard date ranges



Or via slider

Figure 7.2

➢ New usage statistics to highlight flash cache and Smart Scan efficiency.

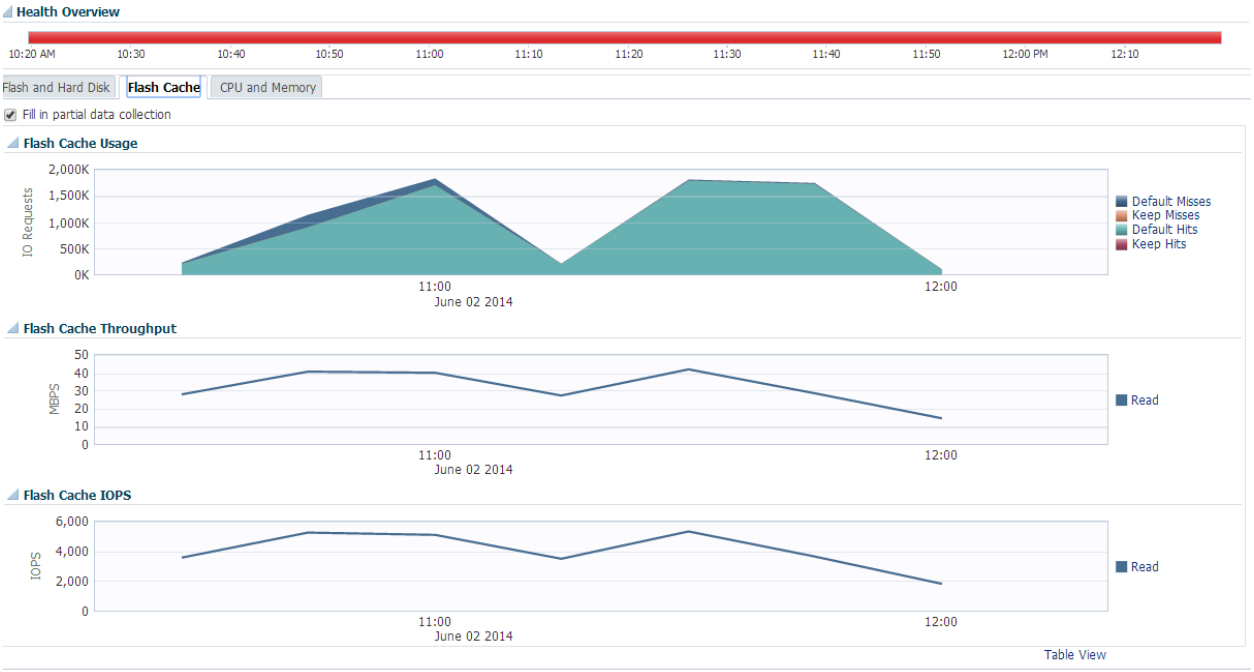


Figure 7.3

➢ Performance utilization for flash and hard disk to identify workload reaching hardware limits.

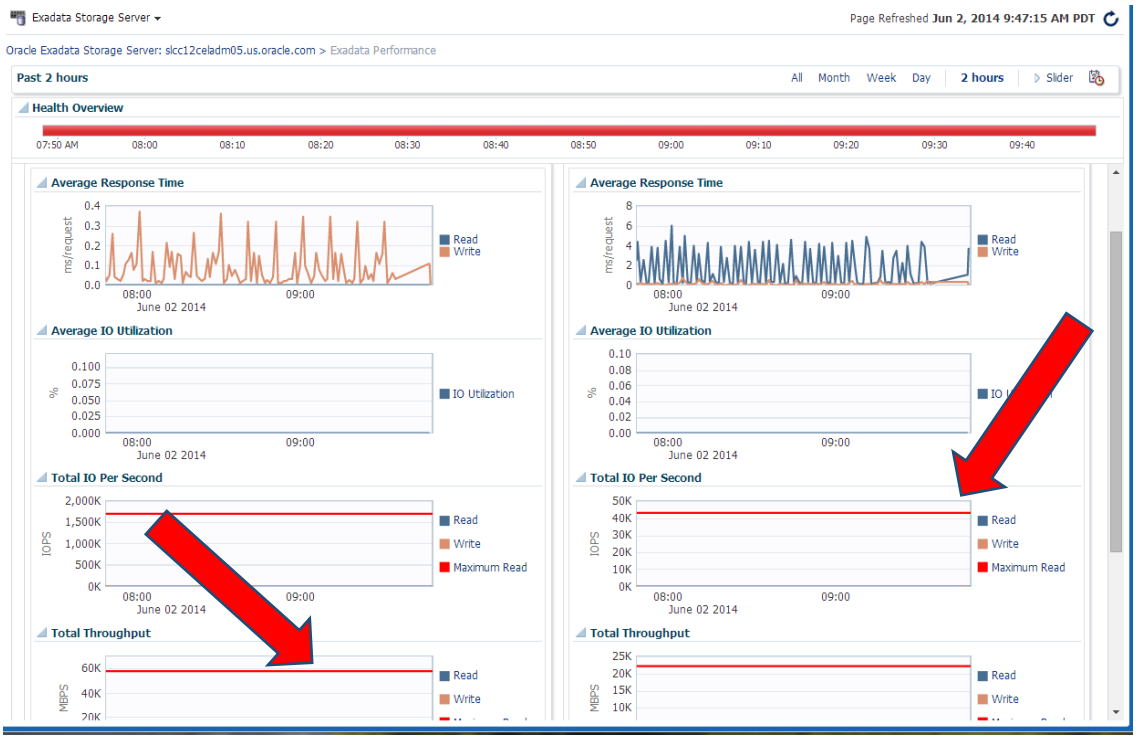


Figure 7.4

- IORM wait per database metric.

Figure 7.5

- Performance comparison between multiple Exadata Cell servers.

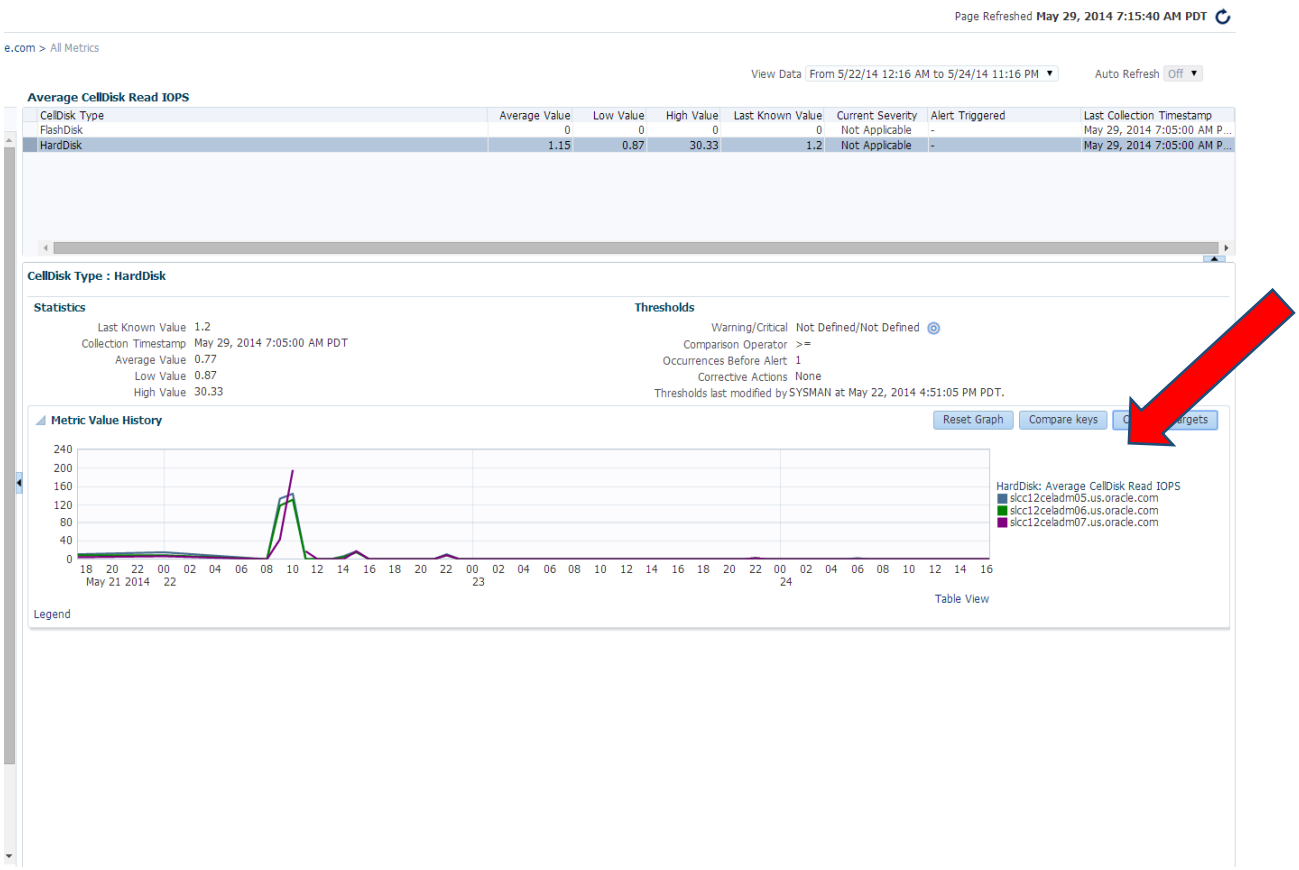

Page Refreshed May 29, 2014 7:15:40 AM PDT
e.com > All Metrics
View Data From 5/22/14 12:16 AM to 5/24/14 11:16 PM
Auto Refresh Off
Average CellDisk Read IOPS
CellDisk Type
Average Value
Low Value
High Value
Last Known Value
Current Severity
Alert Triggered
Last Collection Timestamp
FlashDisk
HardDisk
CellDisk Type : HardDisk
Statistics
Last Known Value 1.2
Collection Timestamp May 29, 2014 7:05:00 AM PDT
Average Value 0.77
Low Value 0.87
High Value 30.33
Thresholds
Warning/Critical Not Defined/Not Defined
Comparison Operator >=
Occurrences Before Alert 1
Corrective Actions None
Thresholds last modified by SYSMAN at May 22, 2014 4:51:05 PM PDT.
Metric Value History
Reset Graph
Compare keys
HardDisk: Average CellDisk Read IOPS
slcc12celadm05.us.oracle.com
slcc12celadm06.us.oracle.com
slcc12celadm07.us.oracle.com
May 21 2014
Table View
Legend

Figure 7.6

Summary

In summary, when monitoring the overall health and performance of a system it is critical to maintain a pro-active and top down approach.

Enabling pro-active health monitoring and alerting of the system, database, and application for any software or hardware errors utilizing the Oracle provided tools such as Enterprise Manager, Exadata Cell Alerts, ASR, and Exachk can assist in automatically and immediately identifying any critical errors. This enables administrators to then be able to provide fast repair actions so that no impact or outages are seen by the end users on the production system.

In addition, ensure that application service levels and response times are defined based on representative workload tests. Thresholds can then be appropriately set against these baseline data points, and therefore quickly identified in cases where resource utilization or user response times start to increase or change unexpectedly.

Key resource indicators to monitor are CPU, memory, disk space, and disk I/O performance when thresholds are exceeded for a sustained period. Momentary usage spikes do not usually significantly affect overall performance. As a rule of thumb it is better to look at fifteen minute rolling averages than instantaneous performance measurements. Rolling averages which exceed thresholds for more than 15 minutes should be reviewed. The recommended thresholds in an Exadata environment are intended to allow for sufficient headroom (approximately 20%) on the system in case of momentary workload peaks. When a key resource exhibits a sustained rise above set and expected thresholds, the system should be tuned by looking at top events, SQL, and I/O across all databases running concurrently.

References

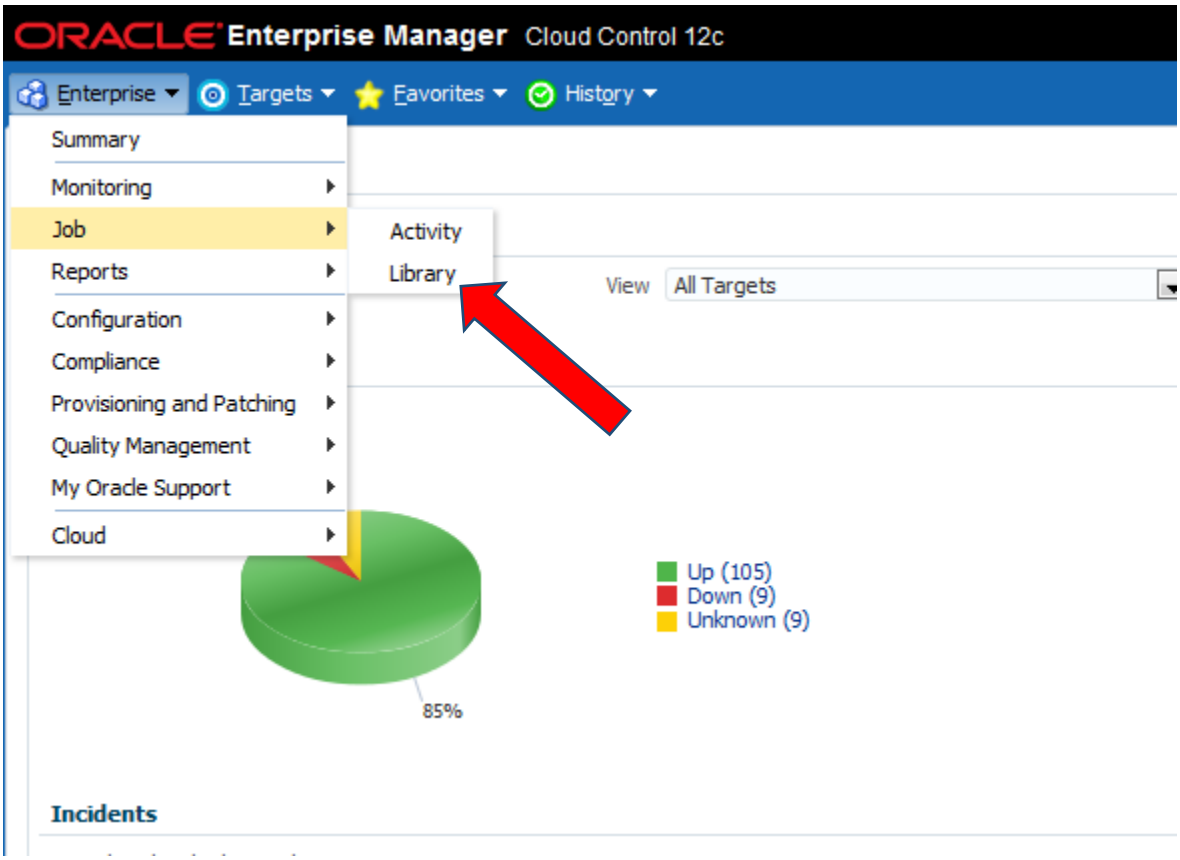
MAA Best Practices

- Deploying Oracle Maximum Availability Architecture with Exadata
- Exadata Consolidation Best Practice Paper
- http://www.oracle.com/webfolder/technetwork/Exadata/MAA-BestP/IORM/IORM.pdf

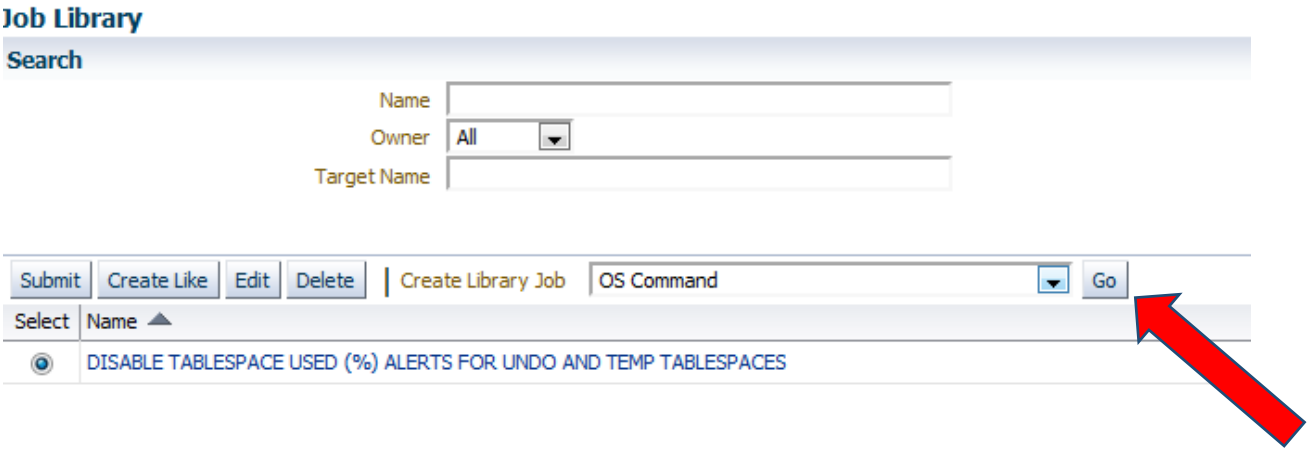
Appendices

Appendix 1.1

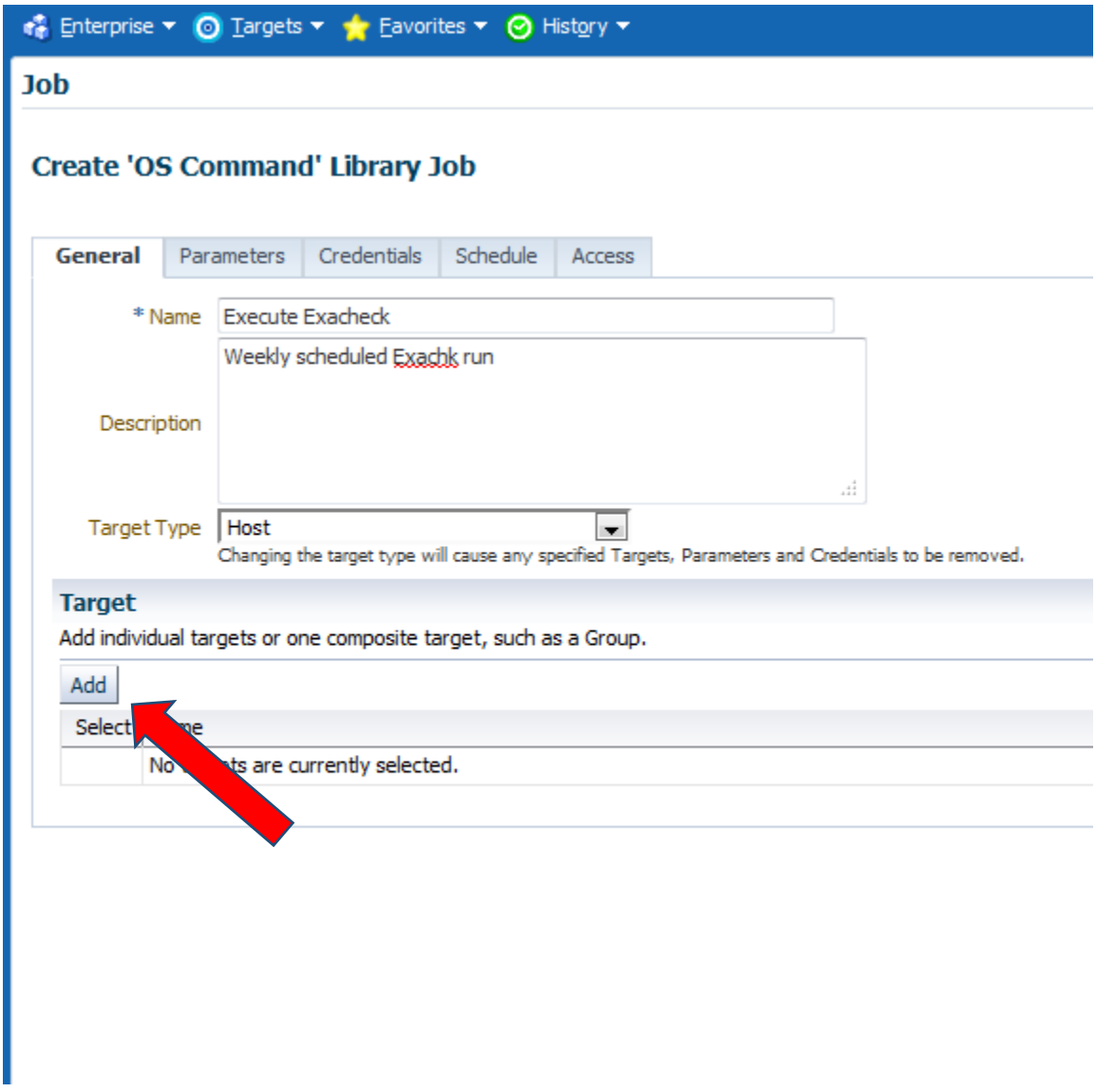
To add an Enterprise Manager job to execute Exachk on a periodic basis. Execute the following steps. First go to Enterprise → Job → Library



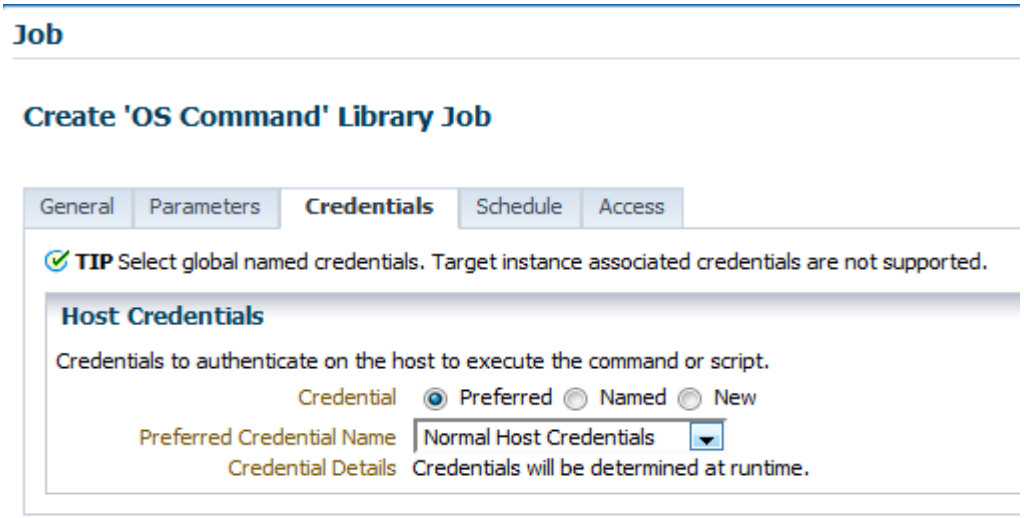
Now create a new OS Command job



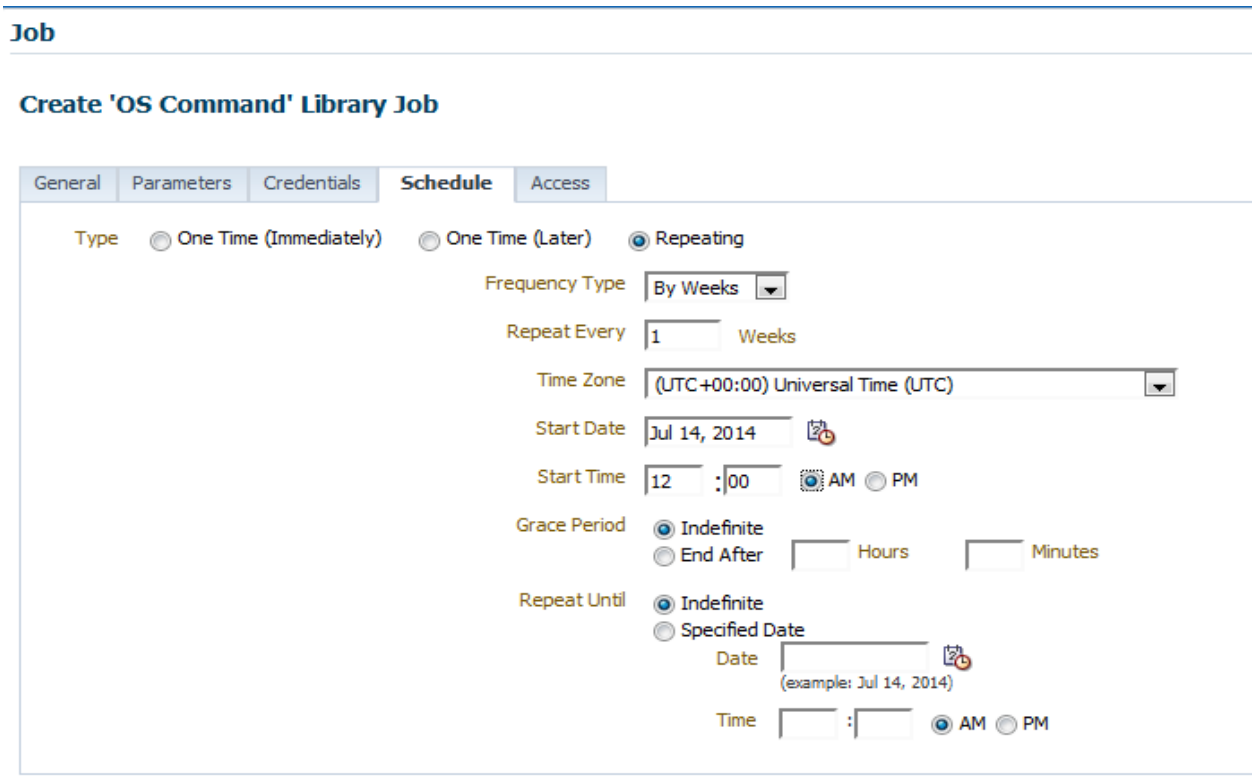
On the General tab of the “Create OS Command Library Job” screen enter a name for the job and ensure target type host is specified. Also specify which targets the job should run on. Targets can also be added after the job has been saved to the library.



On the credentials Tab specify which host credentials should be used



On the Schedule tab choose the Repeating radio button and choose the frequency that best fits your environment. At minimum Exacheck should be run monthly. In the example below a weekly job is specified.

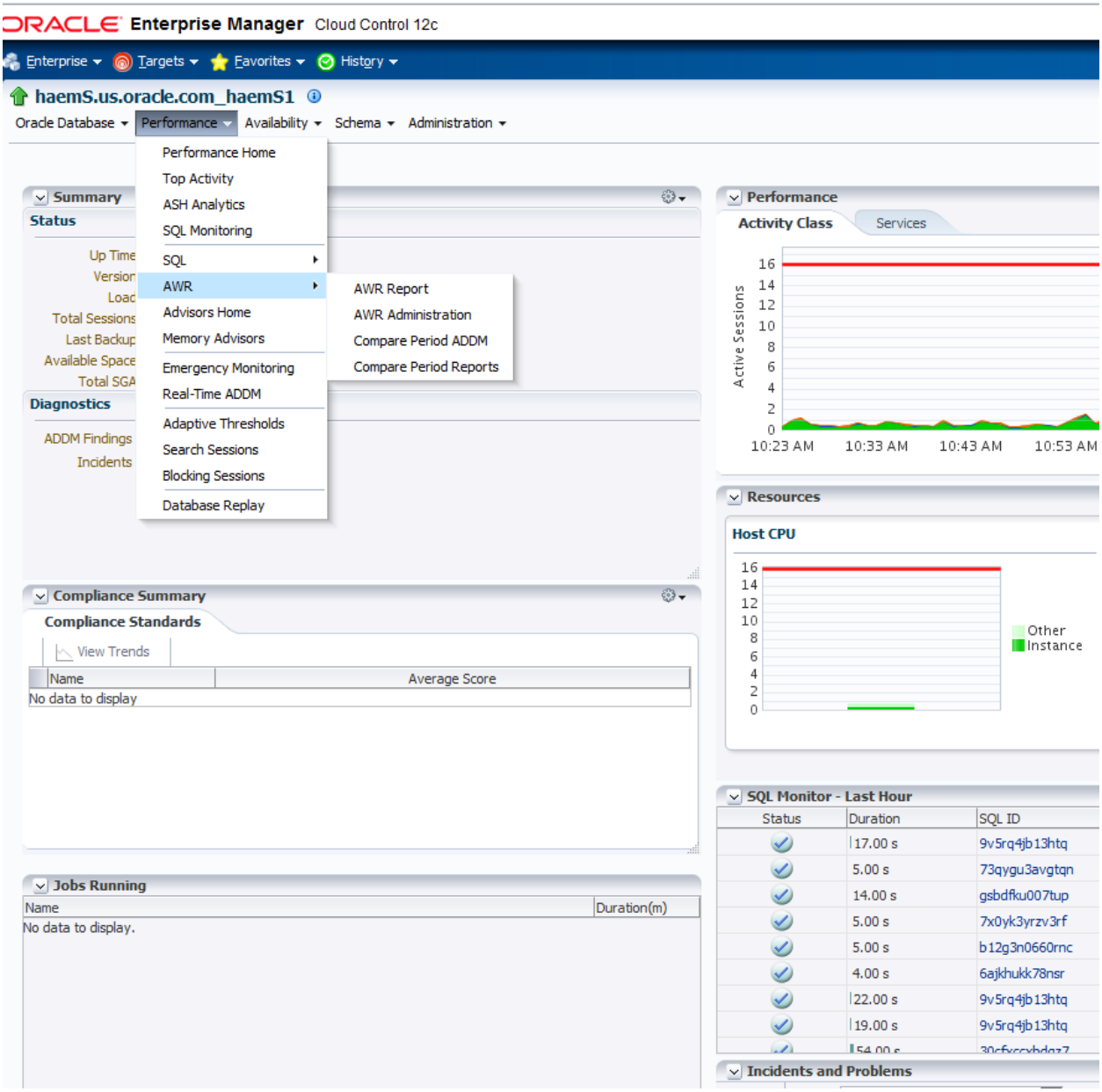


Now choose the save to Library

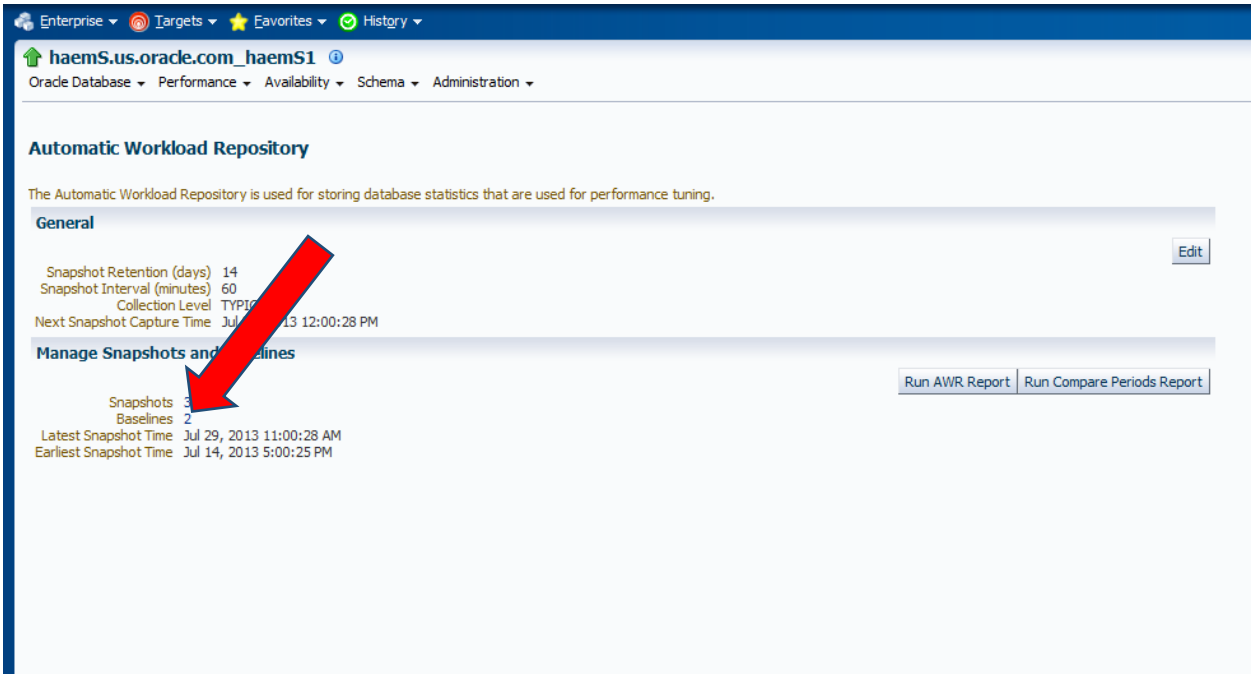

Cancel
Save to Library
Cancel
Save to Library

Appendix 1.2

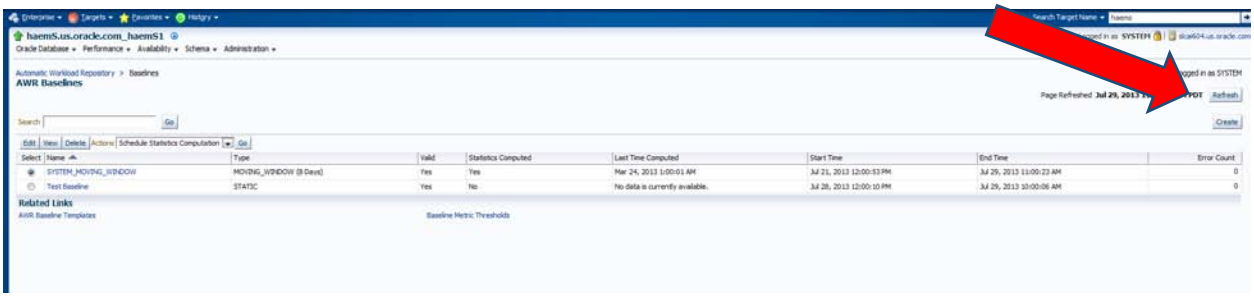
To create a static baseline in Enterprise Manager, go to the database instance target page and select Performance→ AWR→ AWR Administration.



On the Automatic Workload Repository page, select the number beside the Baselines text. This number indicates the number of configured baselines.



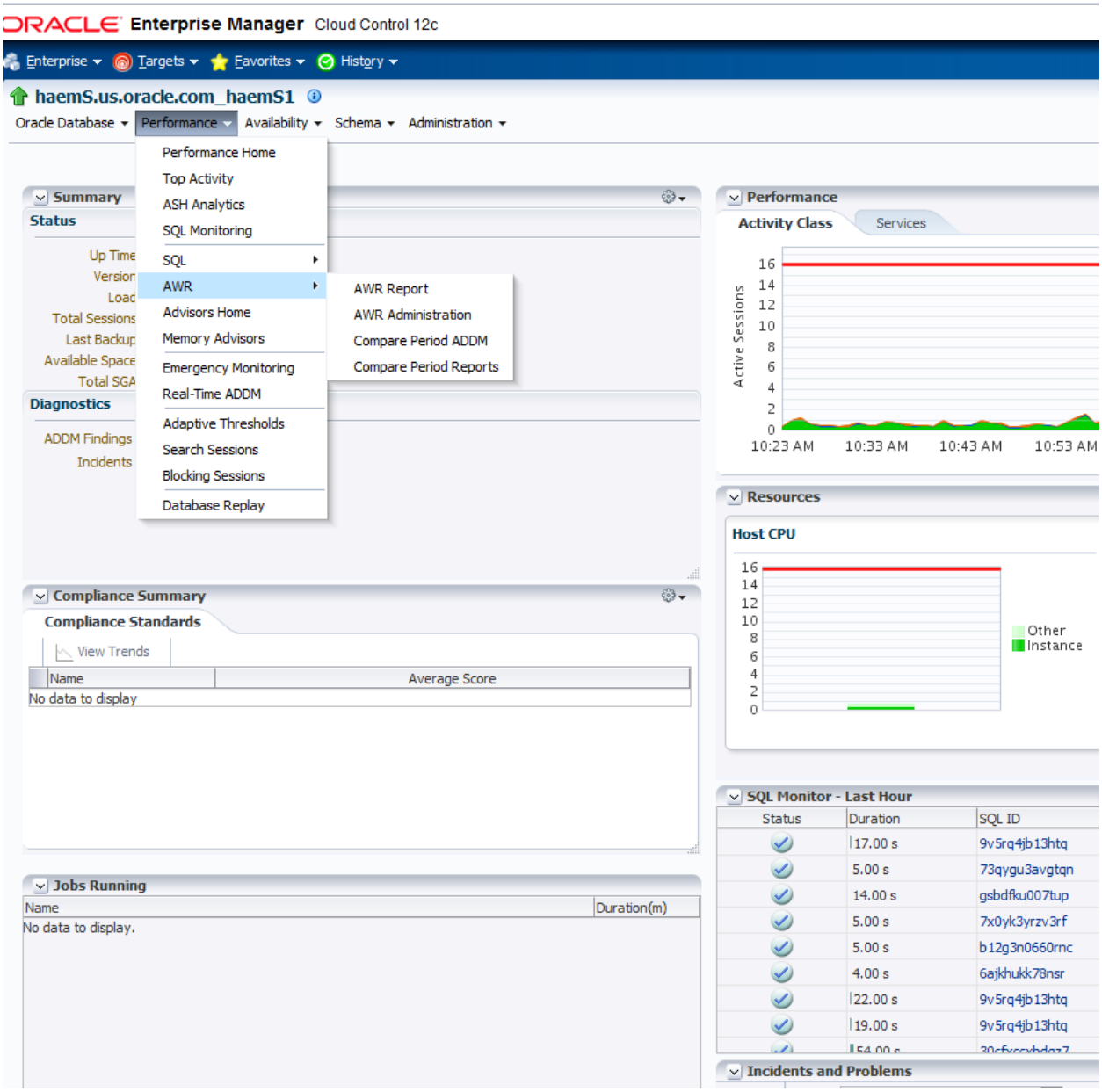
On the AWR Baselines page, select the Create button.



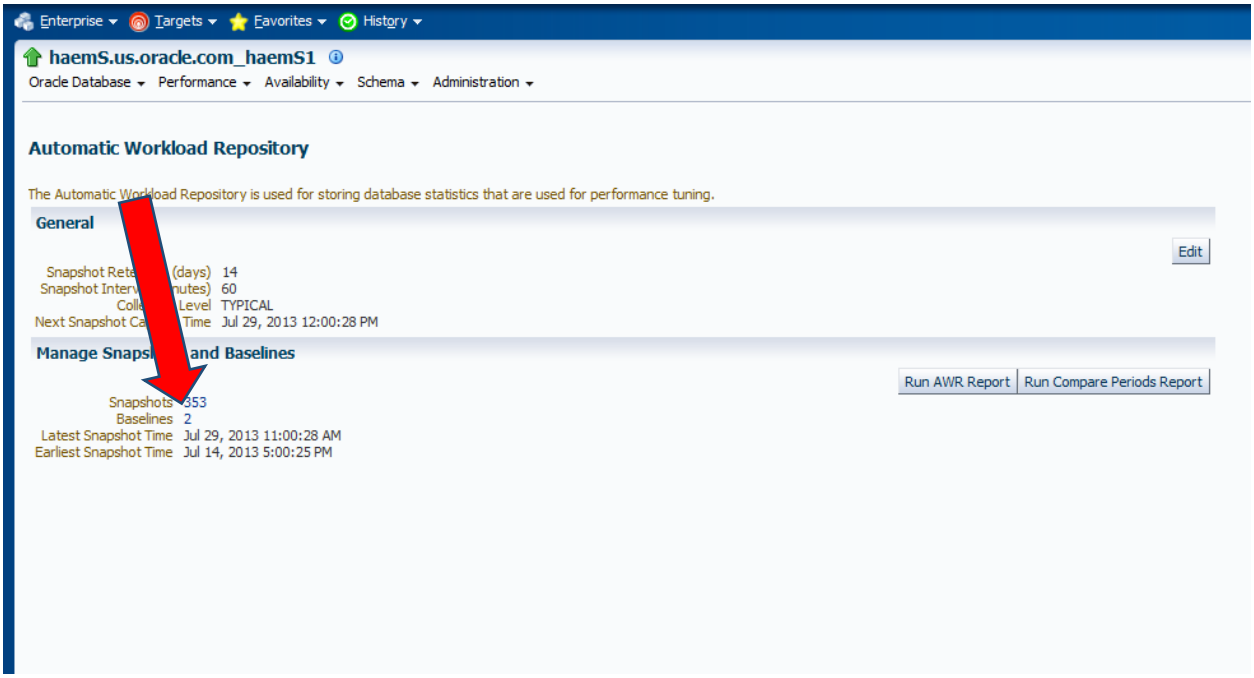
From the next screen you can create either a single static baseline or a repeating baseline. The repeating baseline is created based on a user defined schedule, for example every Monday from 9 am to 5 pm.

Appendix 1.3

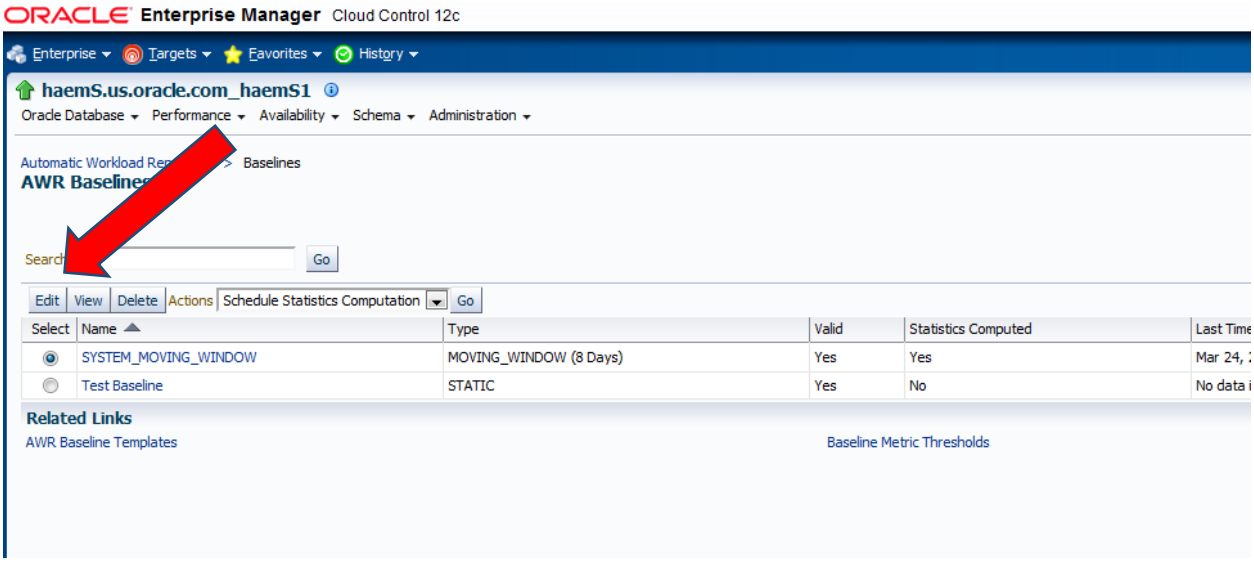
To change the moving window in Enterprise Manager, go to the AWR Baselines page baseline and select Performance→ AWR→ AWR Administration.



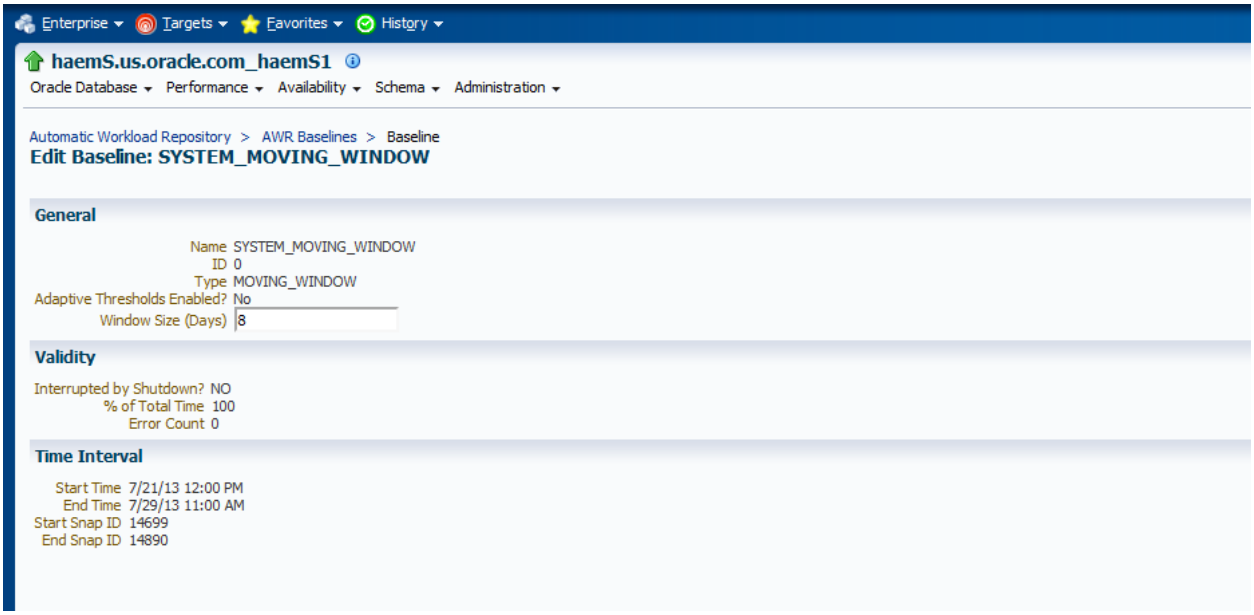
On the Automatic Workload Repository page, select the number besides the Baselines text.



On the AWR Baselines page, select the radio button next to the “SYSTEM_MOVING_WINDOW” baseline and click the Edit button.



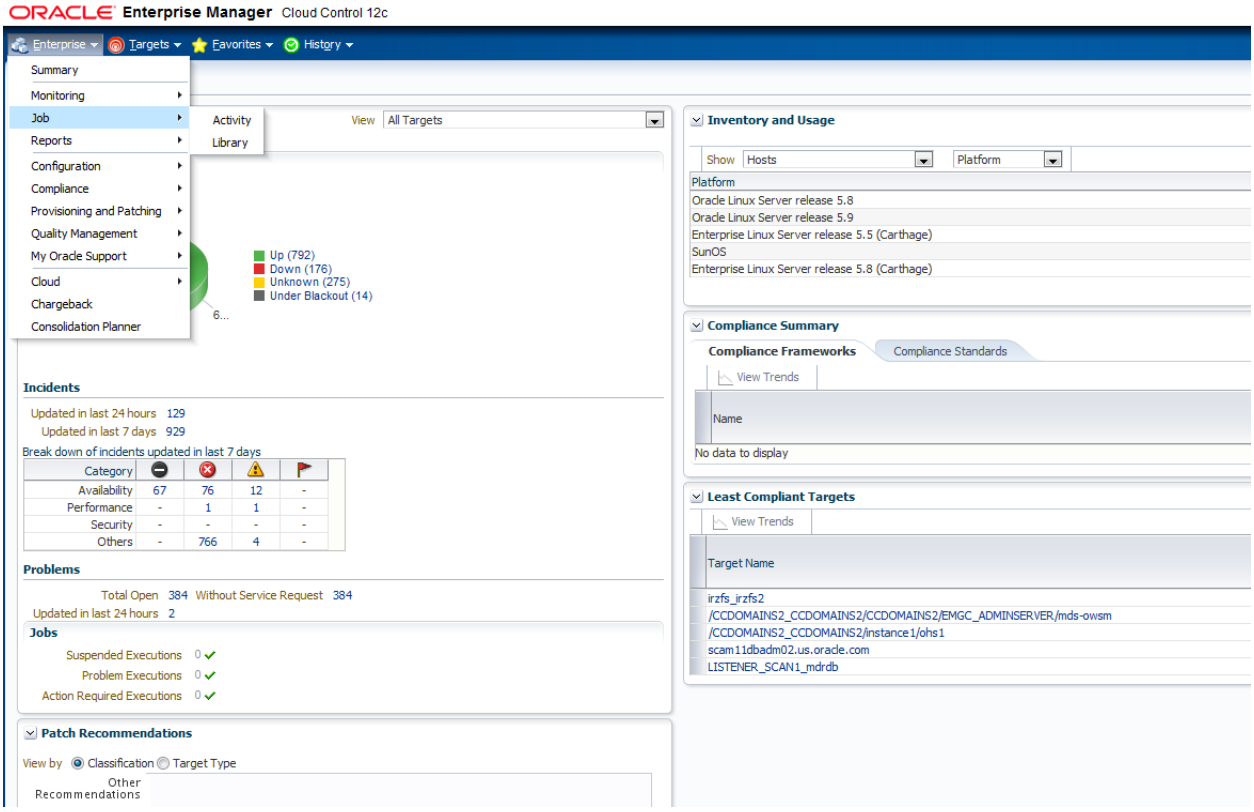
On the Edit Baseline: SYSTEM_MOVING_WINDOW page, change the Window Size (Days) value and


Enterprise
Targets
Favorites
History
haemS.us.oracle.com_haemS1
Oracle Database
Performance
Availability
Schema
Administration
Automatic Workload Repository > AWR Baselines > Baseline
Edit Baseline: SYSTEM_MOVING_WINDOW
General
Name SYSTEM_MOVING_WINDOW
ID 0
Type MOVING_WINDOW
Adaptive Thresholds Enabled? No
Window Size (Days) 8
Validity
Interrupted by Shutdown? NO
% of Total Time 100
Error Count 0
Time Interval
Start Time 7/21/13 12:00 PM
End Time 7/29/13 11:00 AM
Start Snap ID 14699
End Snap ID 14890

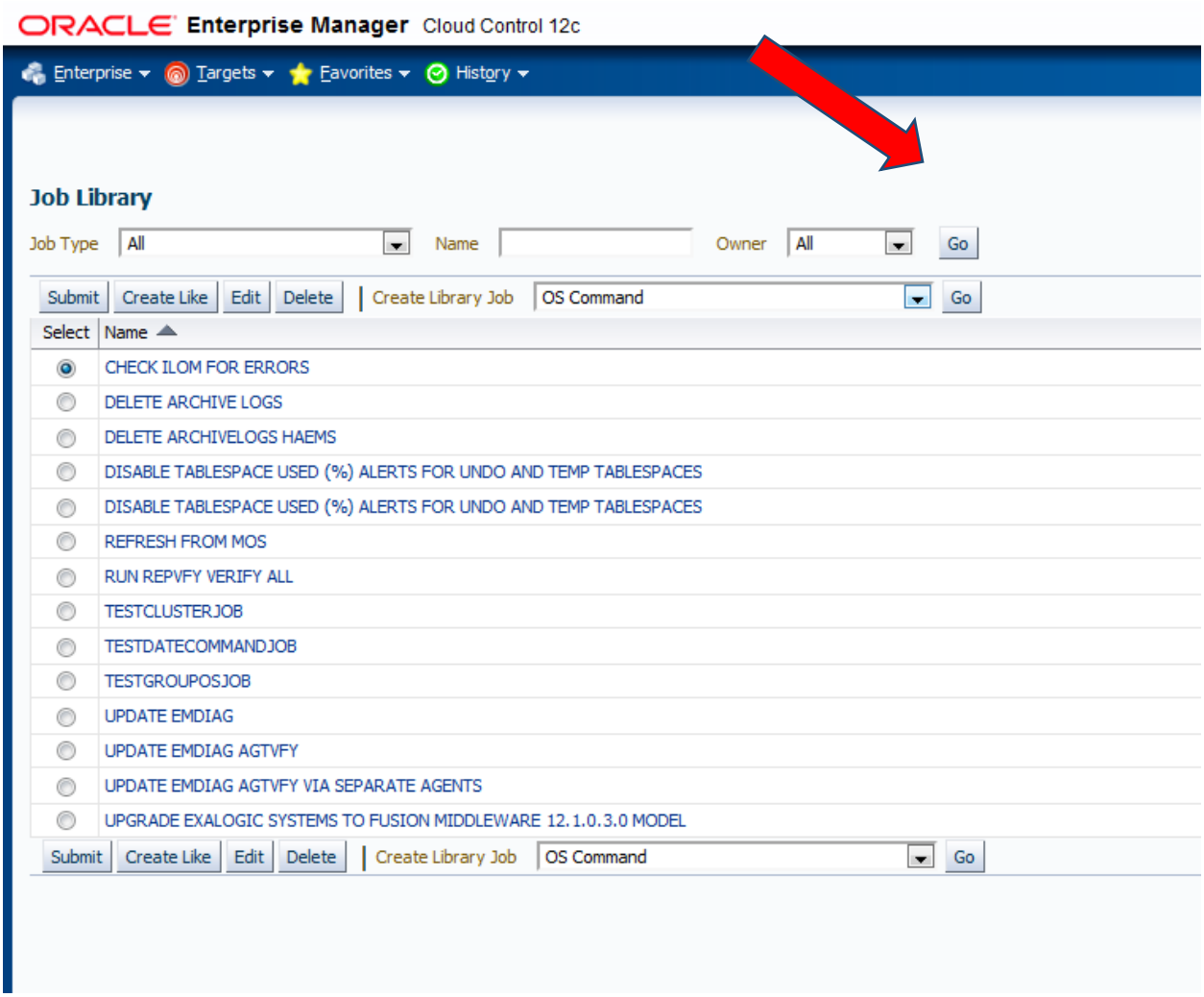
Appendix 1.4

Create EM job for spfile backup

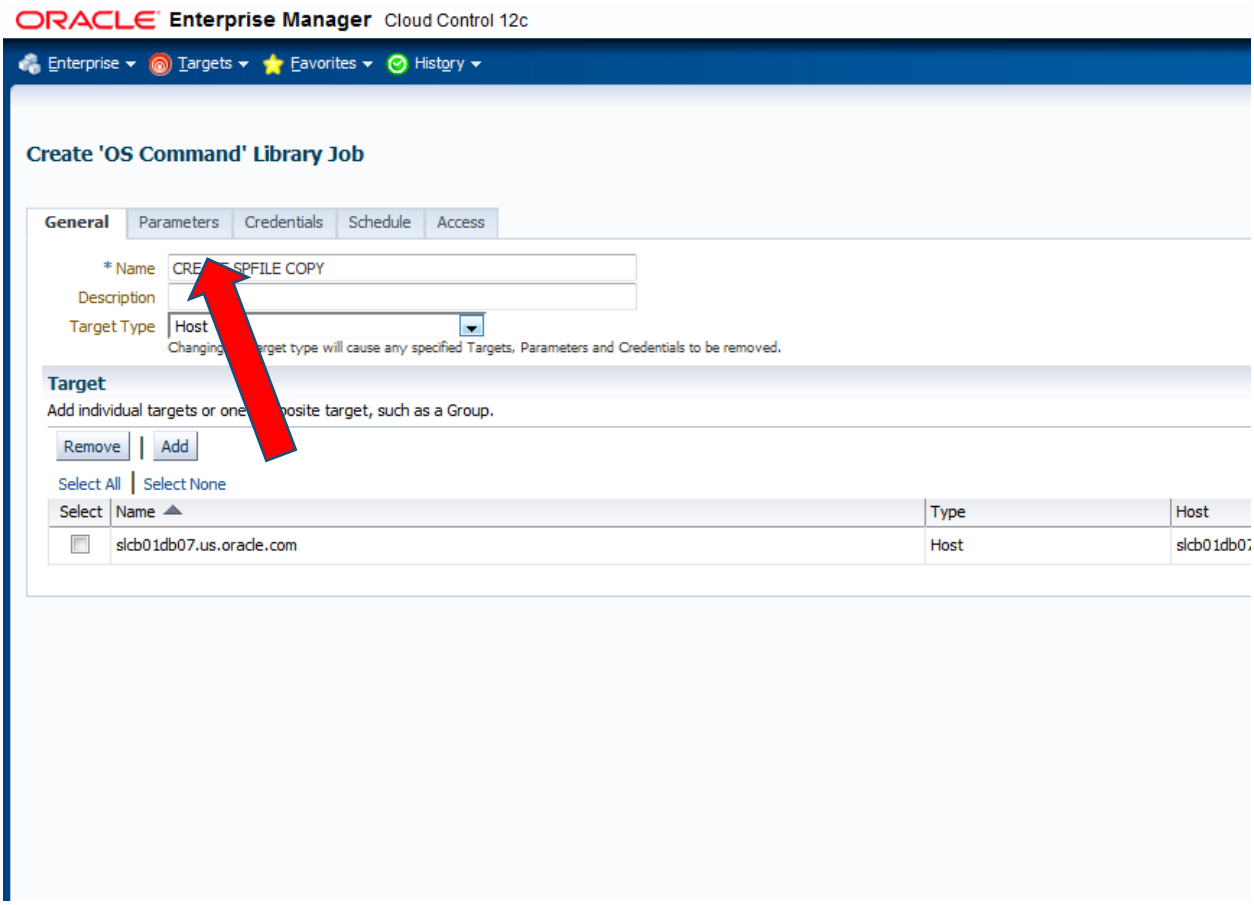
First, go to the Job Library page in Enterprise Manager by selecting Enterprise→ Job→ Library.



On the Job Library page, select “OS Command” from the Create Library Job menu and click Go to create a new job.

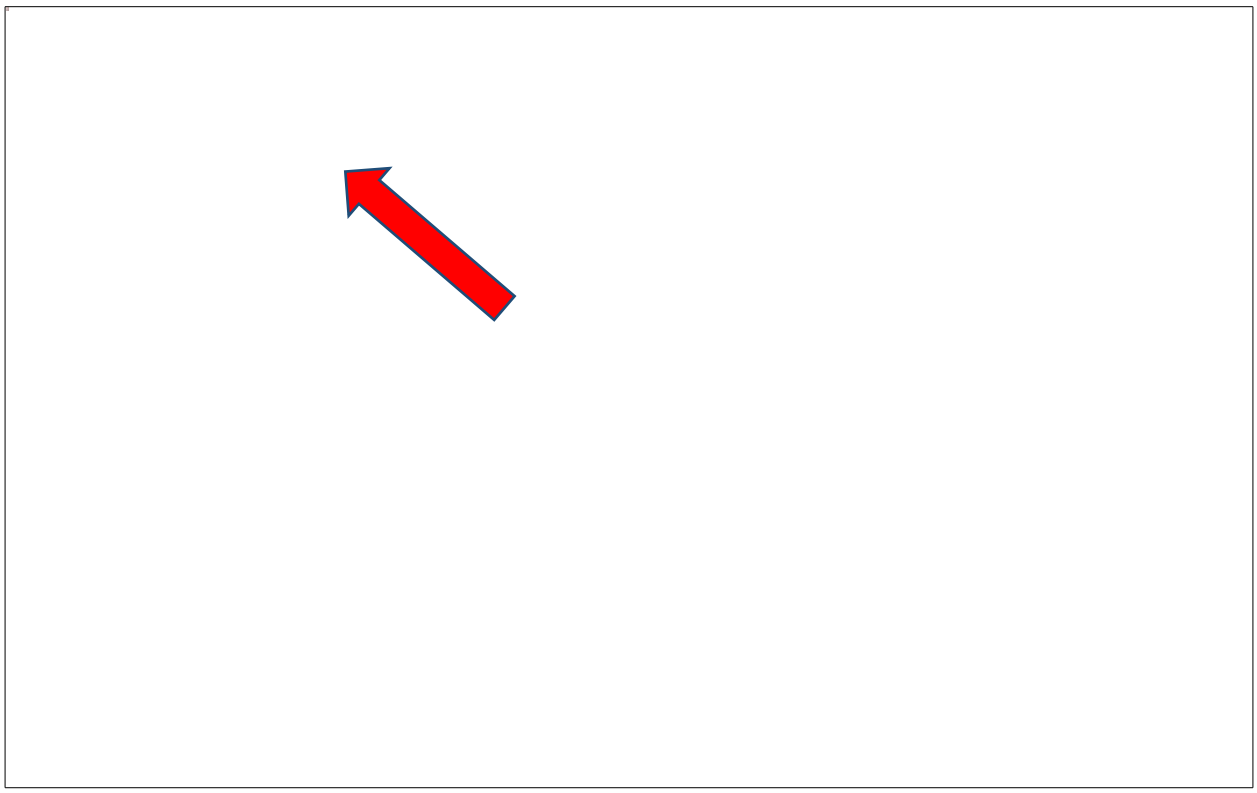

ORACLE Enterprise Manager Cloud Control 12c
Enterprise
Targets
Favorites
History
Job Library
Job Type All
Name
Owner All
Go
Submit
Create Like
Edit
Delete
Create Library Job
OS Command
Go
Select
Name
CHECK ILOM FOR ERRORS
DELETE ARCHIVE LOGS
DELETE ARCHIVELOGS HAEMS
DISABLE TABLESPACE USED (%) ALERTS FOR UNDO AND TEMP TABLESPACES
DISABLE TABLESPACE USED (%) ALERTS FOR UNDO AND TEMP TABLESPACES
REFRESH FROM MOS
RUN REPVFY VERIFY ALL
TESTCLUSTERJOB
TESTDATECOMMANDJOB
TESTGROUPOSJOB
UPDATE EMDIAG
UPDATE EMDIAG AGTVFY
UPDATE EMDIAG AGTVFY VIA SEPARATE AGENTS
UPGRADE EXALOGIC SYSTEMS TO FUSION MIDDLEWARE 12.1.0.3.0 MODEL
Submit
Create Like
Edit
Delete
Create Library Job
OS Command
Go

On the Create ‘OS Command’ Library Job page, provide the Job Name, select “Host” as Target Type, and add the target.

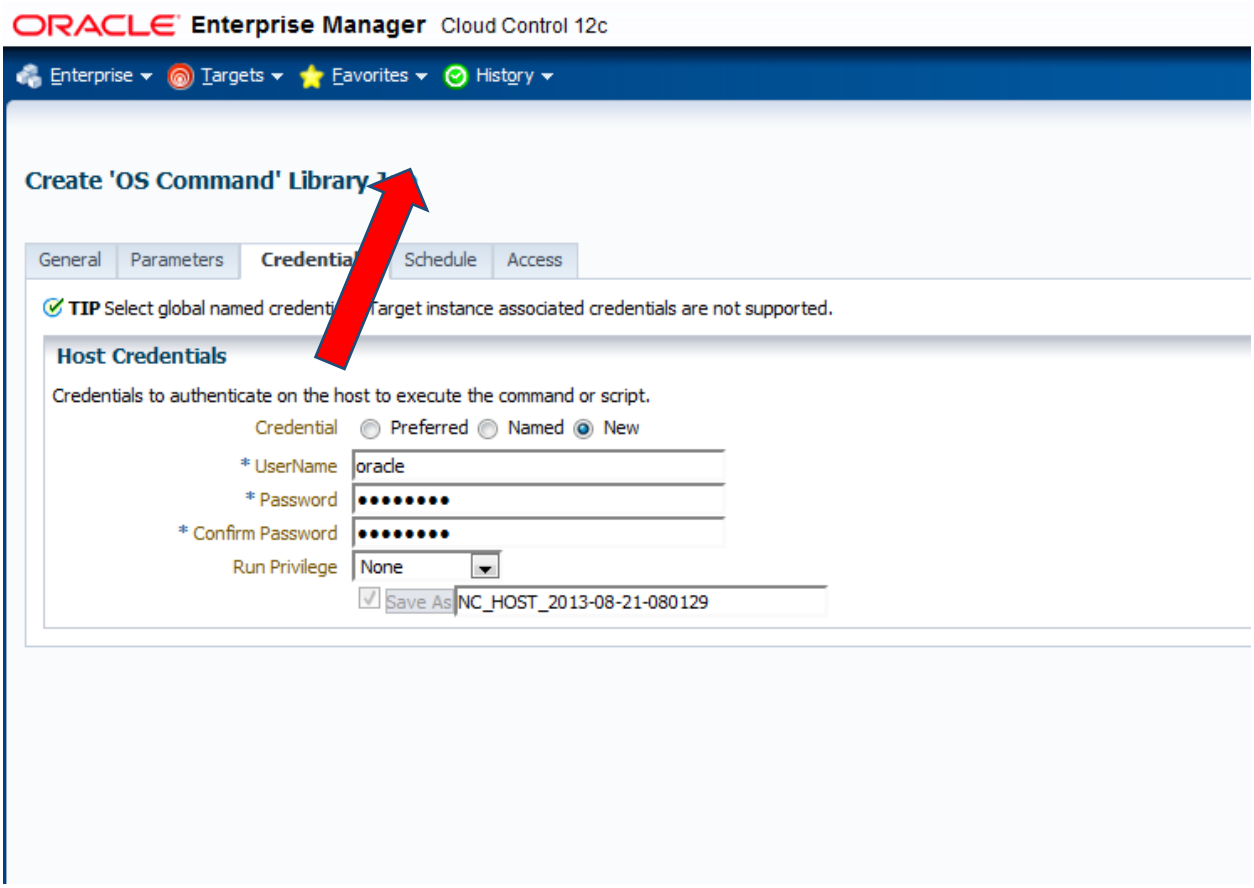


Click the Parameters tab. Select Command Type of “Script,” place the script text in the OS Script box, and select “/bin/ksh” as the Interpreter. The script used in the example is very simple:

```
export ORACLE_HOME=/u01/app/11.2.0.2/grid
export ORACLE_SID=+ASM1
cp +DATA/dbm/spfiledbm.ora /tmp/spfiledbm.ora
 export TSTAMP=`date +%%m%%d%%y`
 asmcmd cp +DATA/dbm/spfiledbm.ora /u01/app/oracle/spfiledbm.ora_$TSTAMP
```

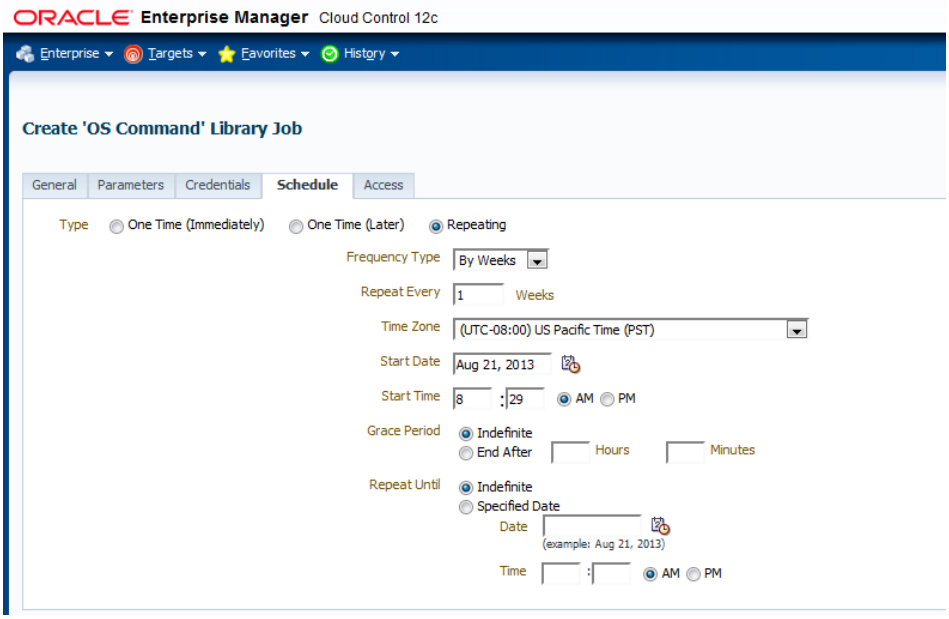


Click on the Credentials tab and Select the credentials you wish to use to run the job.



Select the Schedule tab and specify how often the job should run. At a minimum, you should try to collect changes monthly. However, in more dynamic environments a greater frequency may be necessary.

Select Save to Library and the job is saved to the job library and scheduled to run.



Appendix 1.5

Here we have baseline average I/O latencies:

IOStat by Function summary

- 'Data' columns suffixed with M,G,T,P are in multiples of 1024 other columns suffixed with K,M,G,T,P are in multiples of 1000
- ordered by (Data Read + Write) desc

Function Name	Reads: Data	Reqs per sec	Data per sec	Writes: Data	Reqs per sec	Data per sec	Waits: Count	Avg Tm(ms)
Buffer Cache Reads	23.2G	5055.94	39.5183	0M	0.00	0M	3039.7K	0.70
DBWR	1M	0.11	.001665	21.3G	3259.05	36.4068	68	22.50
LGWR	0M	0.00	0M	4.4G	1381.56	7.57219	438.2K	0.67
Others	32M	3.48	.053301	11M	1.06	.018322	2392	3.72
Direct Writes	0M	0.00	0M	0M	0.00	0M	0	
TOTAL:	23.2G	5059.53	39.5733	25.8G	4641.68	43.9973	3480.4K	0.70

Average latency increases as the database is throttled:

Function Name	Reads: Data	Reqs per sec	Data per sec	Writes: Data	Reqs per sec	Data per sec	Waits: Count	Avg Tm(ms)
Buffer Cache Reads	8.7G	1729.49	14.7892	0M	0.00	0M	1038.9K	1.84
DBWR	5M	1.25	.008327	6.6G	949.37	11.2634	748	155.12
Others	2.7G	8.71	4.55001	2.6G	5.43	4.49672	2825	35.06
LGWR	0M	0.00	0M	1.5G	609.20	2.64307	246.4K	1.24
Streams AQ	0M	0.00	0M	0M	0.00	0M	2	4.00
TOTAL:	11.3G	1739.44	19.3475	10.8G	1564.00	18.4032	1288.9K	1.88

The Top 5 wait events will typically change as well when a database is throttled. Here's the baseline:

Top 5 Timed Foreground Events

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
log file sync	902,294	4,677	5	36.18	Commit
cell single block physical read	2,498,568	3,677	1	28.44	User I/O
DB CPU		3,546		27.43	
gc current block 2-way	1,794,254	256	0	1.98	Cluster
library cache: mutex X	238,070	233	1	1.80	Concurrency

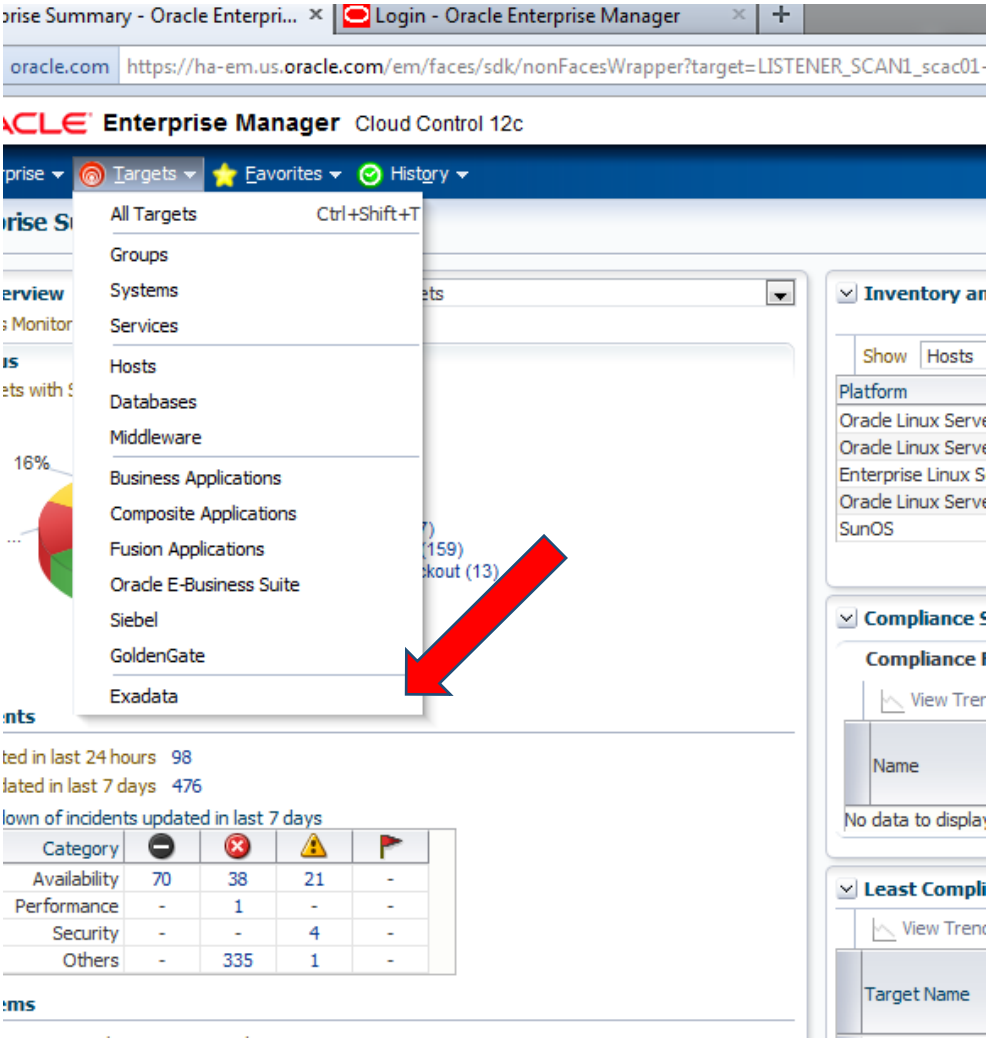
There is an increase in buffer waits for the application as the database writers spend more time writing dirty buffers to disk.

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
free buffer waits	294,844	3,329	11	24.20	Configuration
cell single block physical read	1,037,919	2,458	2	17.87	User I/O
log file sync	331,531	2,045	6	14.87	Commit
write complete waits	213	1,716	8056	12.47	Configuration
DB CPU		1,416		10.29	

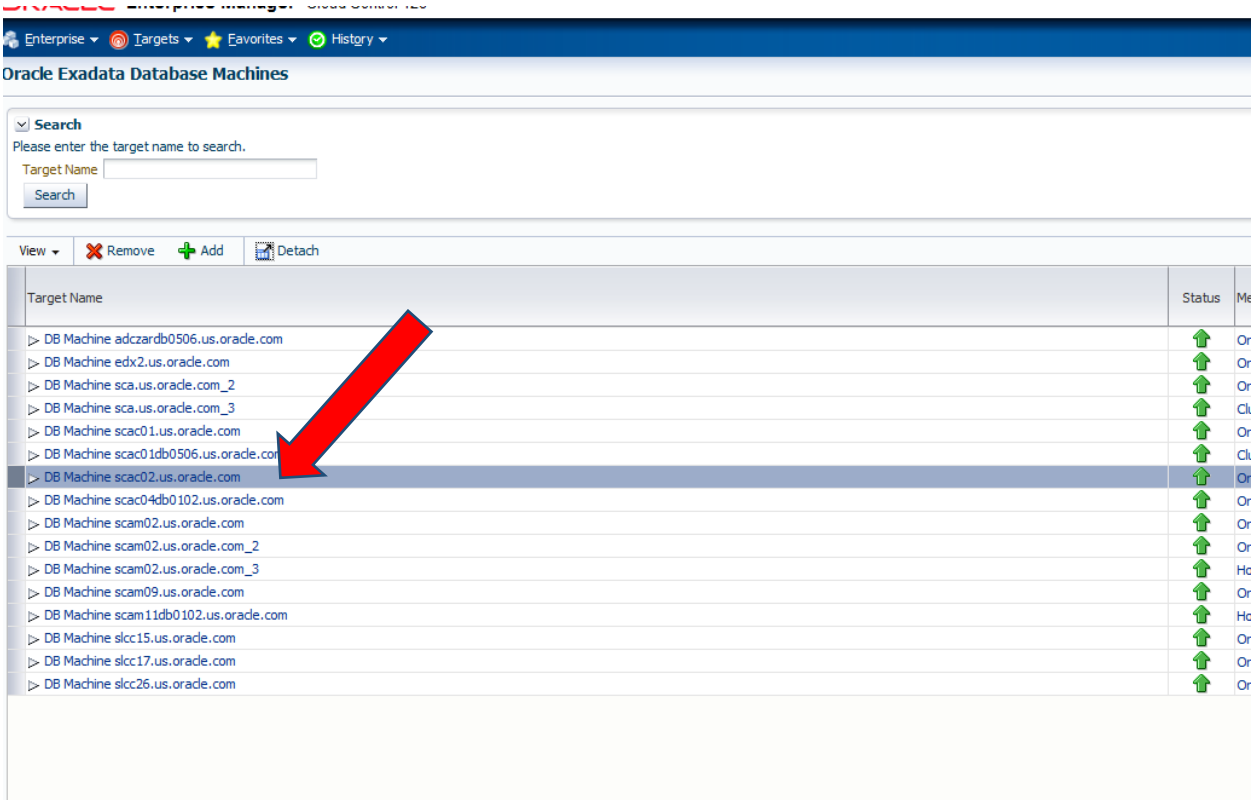
Appendix 3.1

Check target status

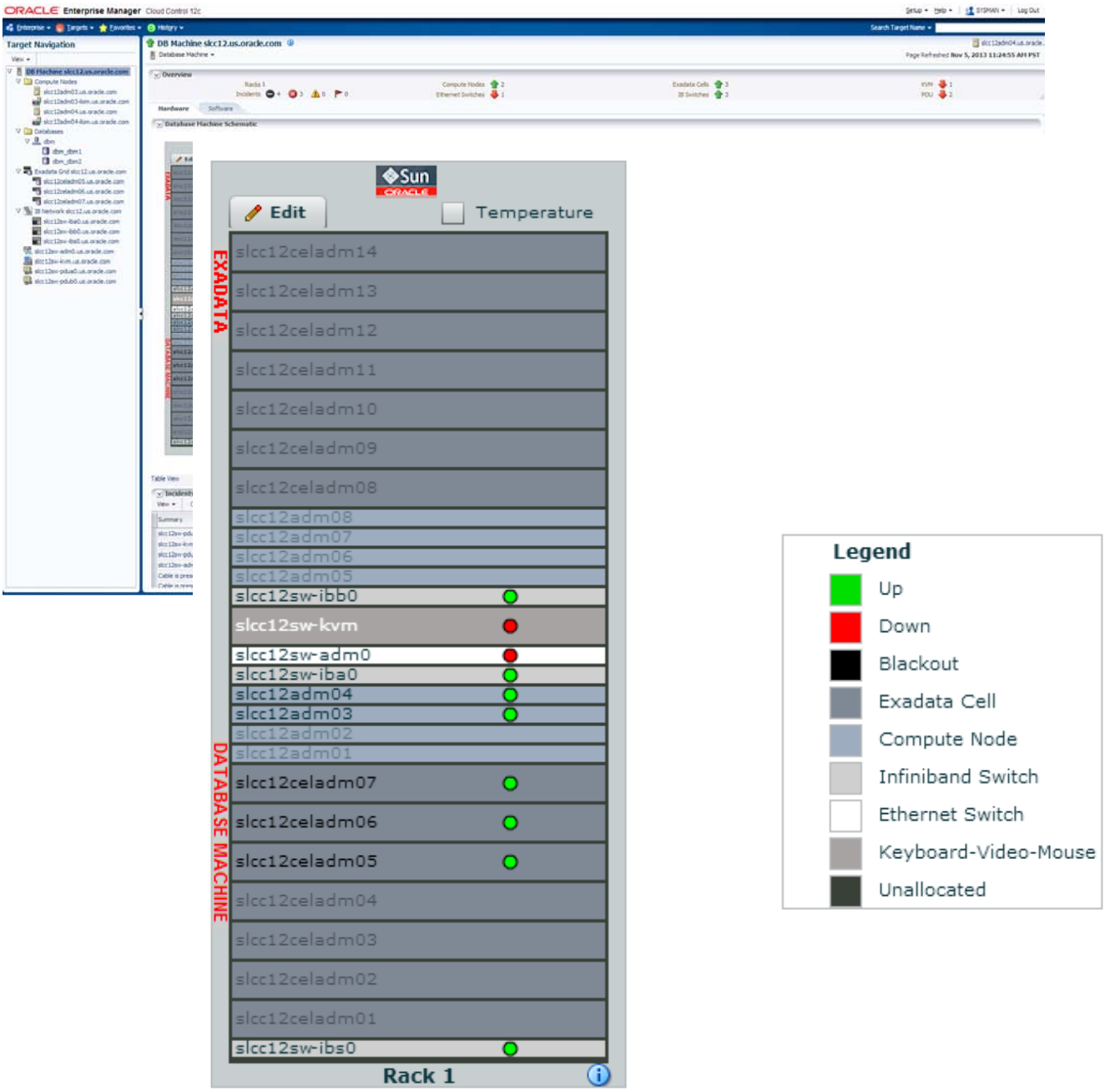
Navigate to the Exadata Target Page by selecting Targets→ Exadata.



Select the desired Exadata Database Machine.

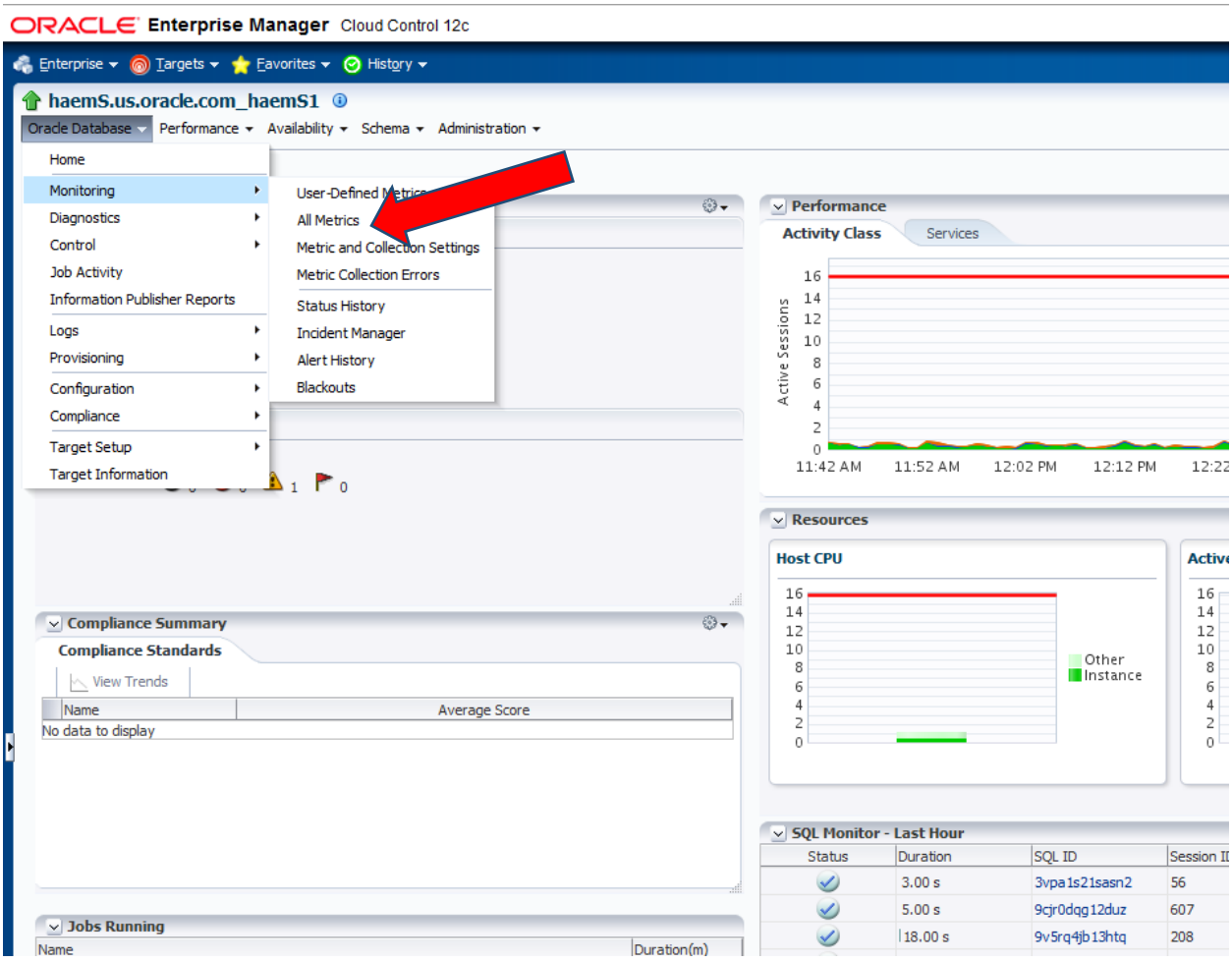


The Exadata Database Machine homepage is displayed with state information about each of the components.

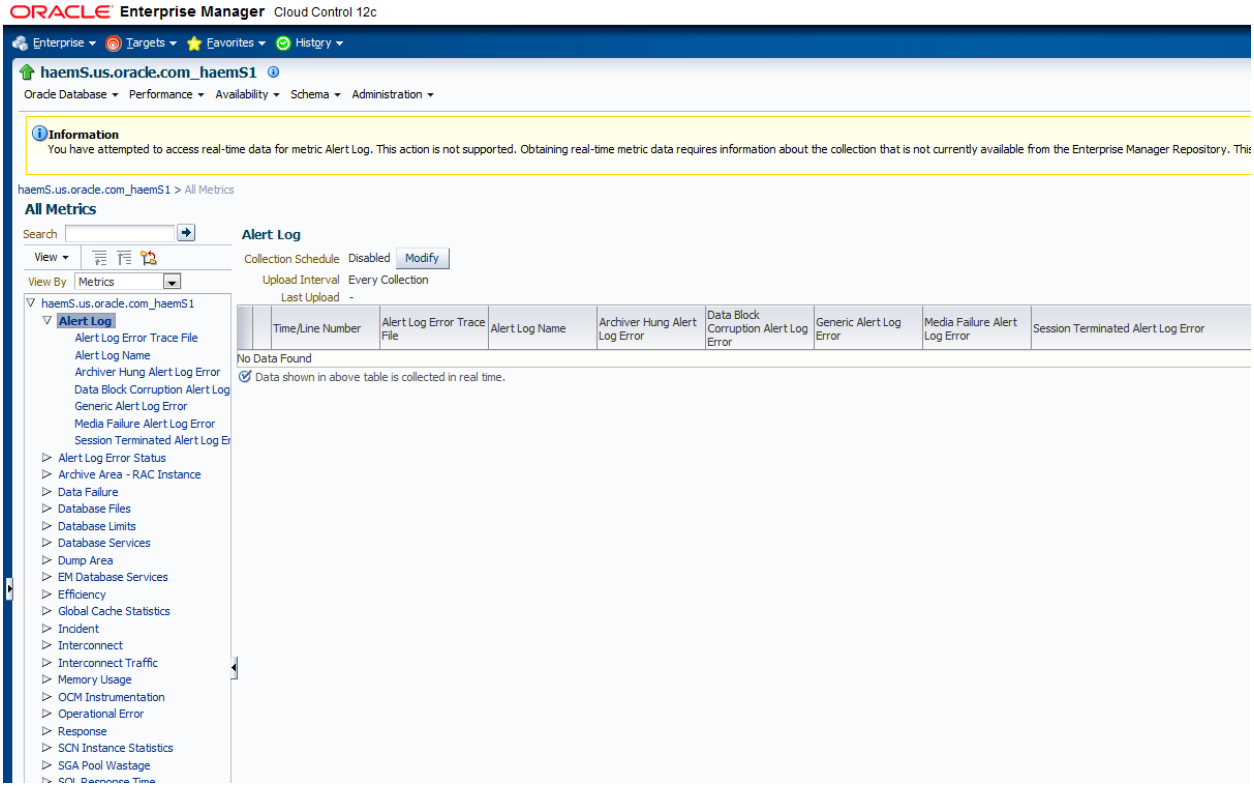


Appendix 3.2

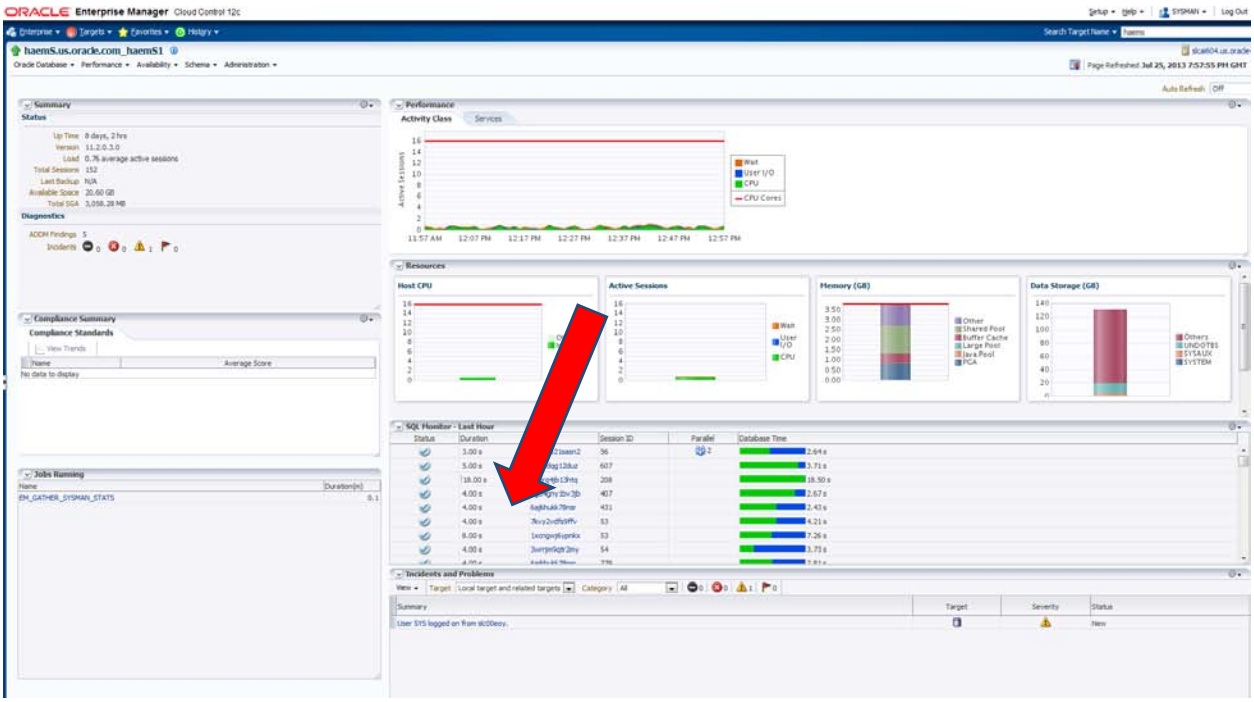
Go to the target page and then select Monitoring→ All Metrics from the target's Oracle Database, Automatic Storage Management, or Cluster menu.



From the All Metrics screen, there are several alert log metrics that can be viewed.



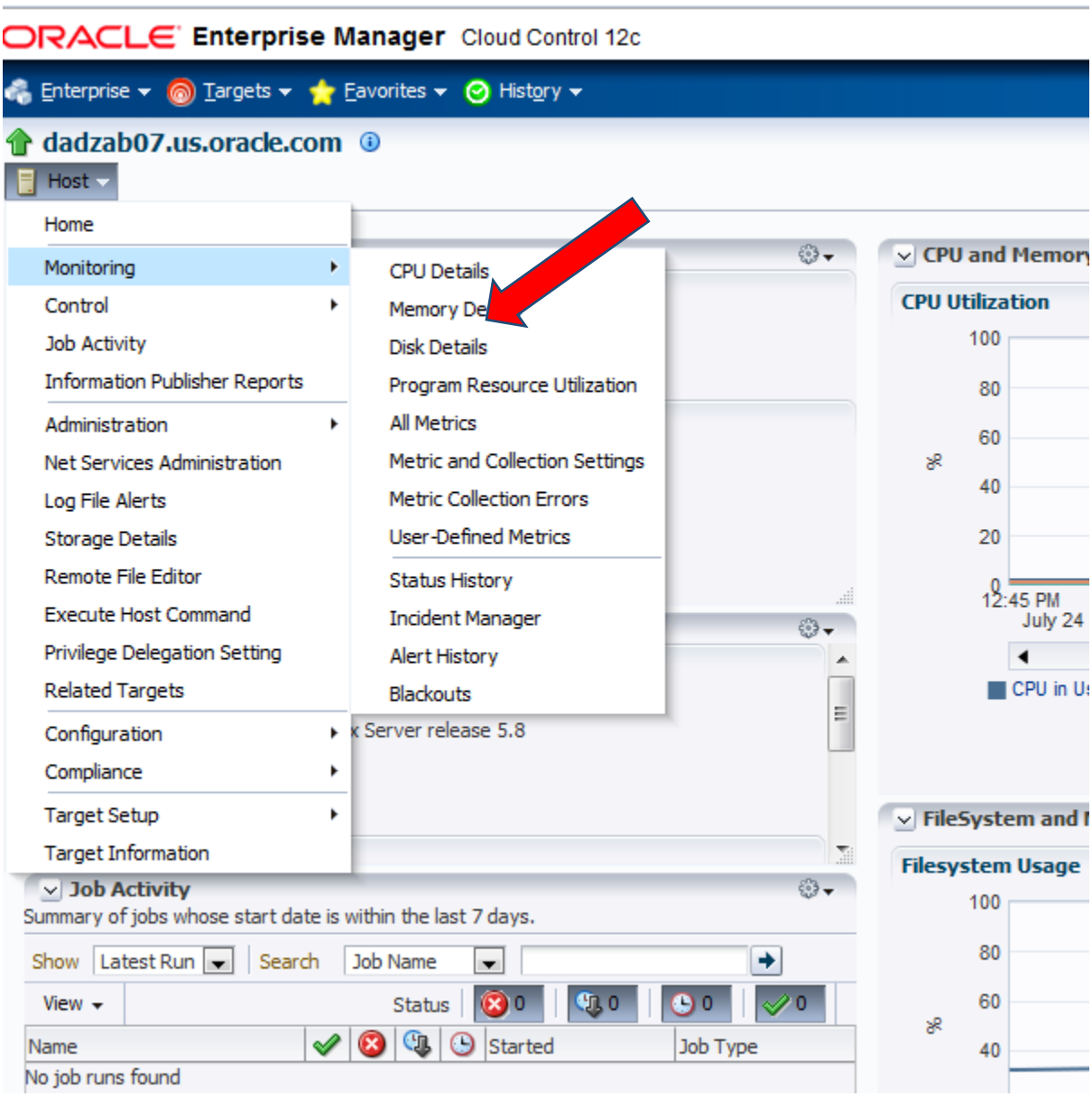
If a critical issue is detected, an Enterprise Manager Incident will be created. This can be viewed directly on the main target page.



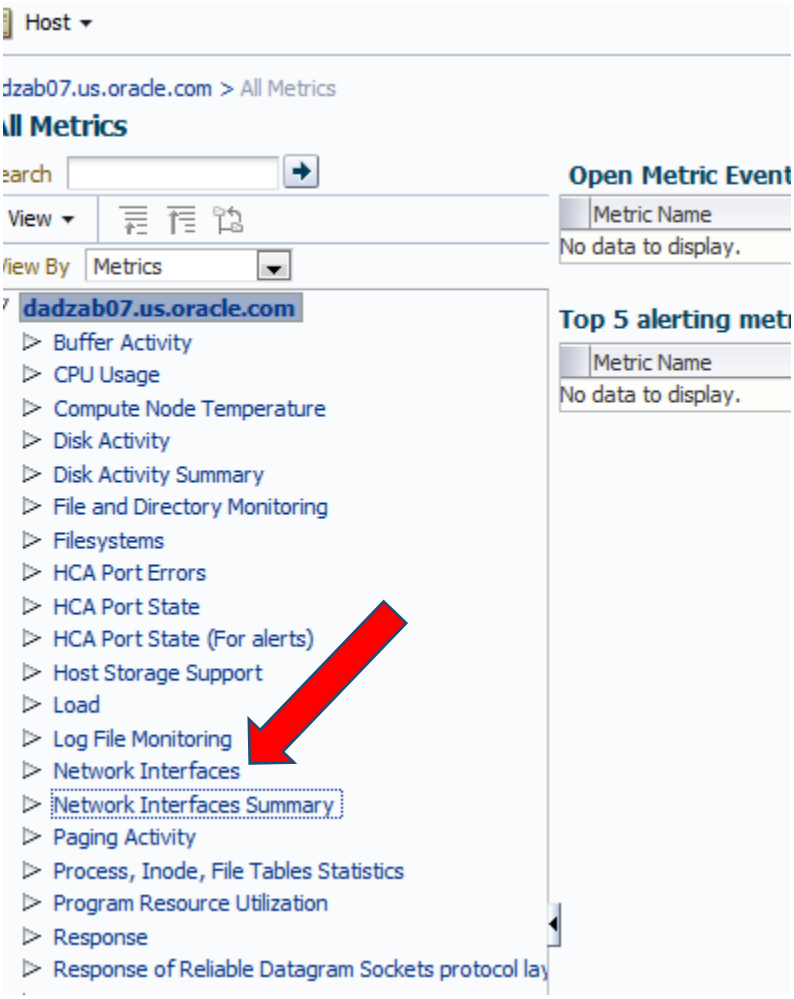
Appendix 3.3

Check network status

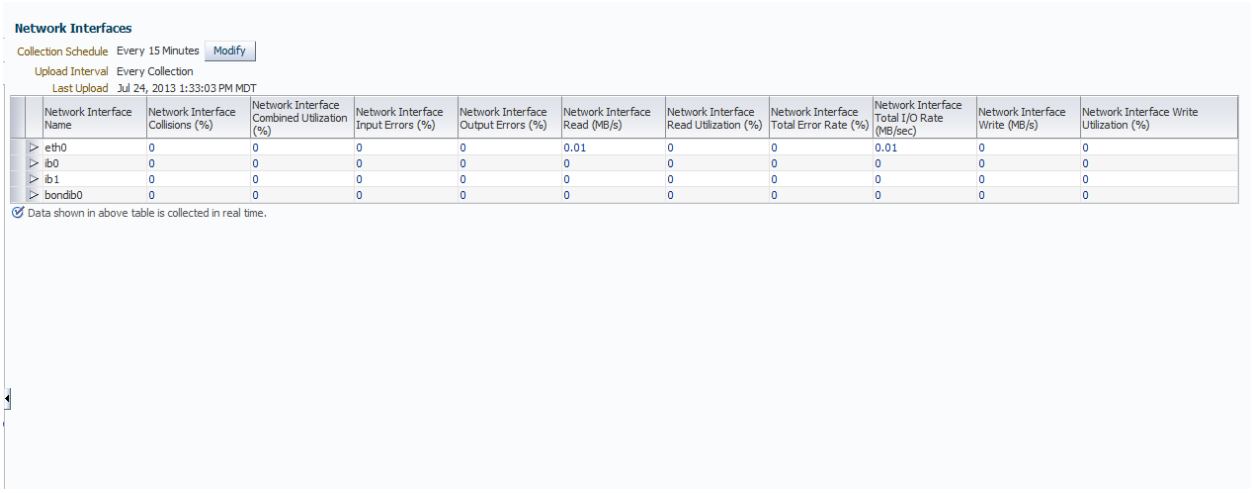
From the host target home page select Host→ Monitoring→ All Metrics.


ORACLE Enterprise Manager Cloud Control 12c
Enterprise
Targets
Favorites
History
dadzab07.us.oracle.com
Host
Home
Monitoring
Control
Job Activity
Information Publisher Reports
Administration
Net Services Administration
Log File Alerts
Storage Details
Remote File Editor
Execute Host Command
Privilege Delegation Setting
Related Targets
Configuration
Compliance
Target Setup
Target Information
CPU Details
Memory De
Disk Details
Program Resource Utilization
All Metrics
Metric and Collection Settings
Metric Collection Errors
User-Defined Metrics
Status History
Incident Manager
Alert History
Blackouts
Server release 5.8
CPU and Memory
CPU Utilization
100
80
60
40
20
0
12:45 PM
July 24
CPU in Us
FileSystem and
Filesystem Usage
100
80
60
40
Job Activity
Summary of jobs whose start date is within the last 7 days.
Show
Latest Run
Search
Job Name
View
Status
0
0
0
0
Name
Started
Job Type
No job runs found

Select Network Interfaces.



The page will display all network interfaces and their statistics



Network Interfaces

Collection Schedule Every 15 Minutes Modify
Upload Interval Every Collection
Last Upload Jul 24, 2013 1:33:03 PM MDT

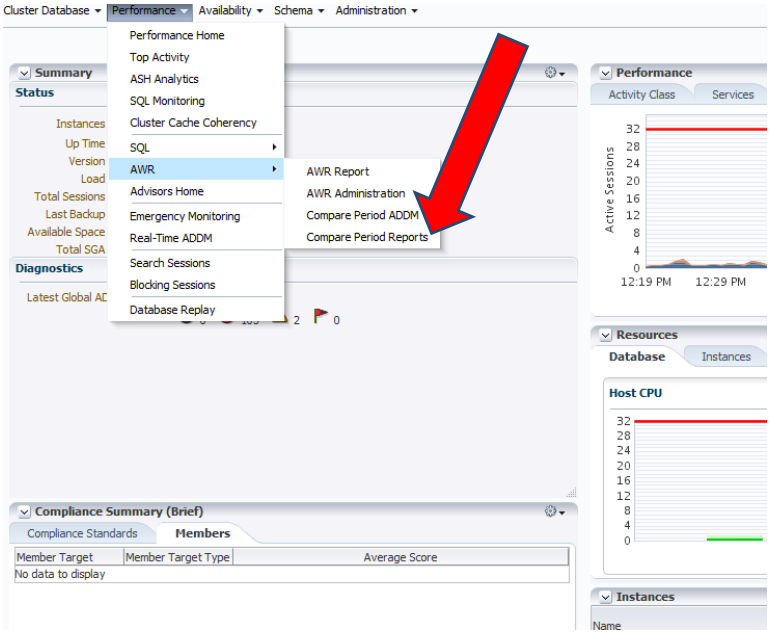
Network Interface Name	Network Interface Collisions (%)	Network Interface Combined Utilization (%)	Network Interface Input Errors (%)	Network Interface Output Errors (%)	Network Interface Read (MB/s)	Network Interface Read Utilization (%)	Network Interface Total Error Rate (%)	Network Interface Total I/O Rate (MB/sec)	Network Interface Write (MB/s)	Network Interface Write Utilization (%)
eth0	0	0	0	0	0.01	0	0	0.01	0	0
ib0	0	0	0	0	0	0	0	0	0	0
ib1	0	0	0	0	0	0	0	0	0	0
bondib0	0	0	0	0	0	0	0	0	0	0

Data shown in above table is collected in real time.

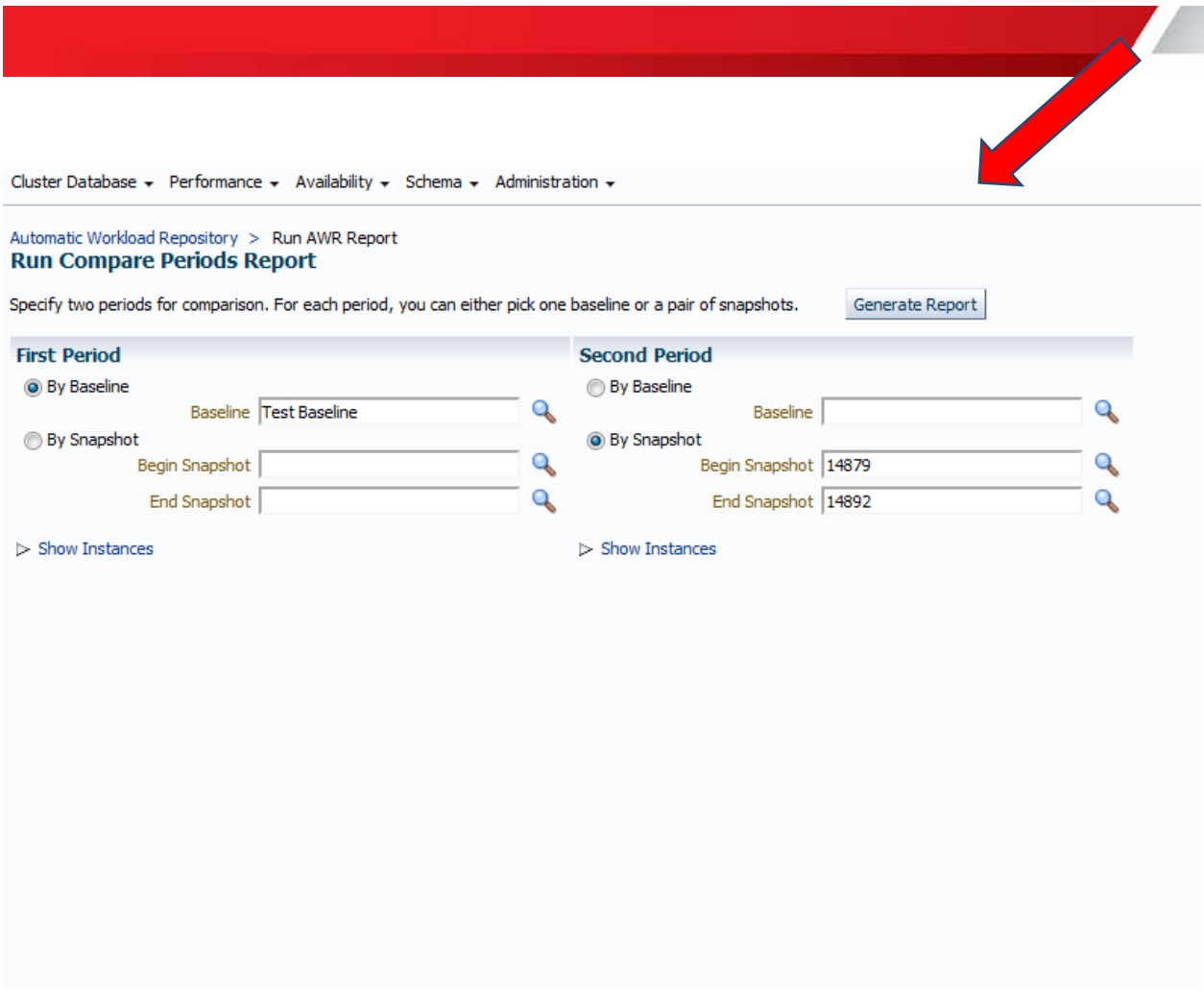
Appendix 4.1

Generate AWR comparison reports

To compare a current AWR report to a baseline in Enterprise Manager. From the database target home page select Performance→ AWR→ Compare Period Reports



Now choose which periods to compare. For the First Period select a static or moving baselines which provides the closest performance benchmark. For example comparing similar time periods or workload data. For the second period select snapshot that encapsulates the performance issue. Next push the “Generate Report” button


Cluster Database
Performance
Availability
Schema
Administration
Automatic Workload Repository > Run AWR Report
Run Compare Periods Report
Specify two periods for comparison. For each period, you can either pick one baseline or a pair of snapshots.
Generate Report
First Period
By Baseline
Baseline
Test Baseline
By Snapshot
Begin Snapshot
End Snapshot
Show Instances
Second Period
By Baseline
Baseline
By Snapshot
Begin Snapshot
14879
End Snapshot
14892
Show Instances

A report will be generated which will demonstrate differences in the workload and behavior of the database.

Report Summary

Host Configuration Comparison

- CPU and Memory values are from the end snapshot; averaged across all instances
- Other values are averages for all instances

	1st	2nd	Diff	%Diff
Number of CPUs:	32	32	0	0.0
Number of CPU Cores:	16	16	0	0.0
Number of CPU Sockets:	2	2	0	0.0
Physical Memory:	258064.9M	258064.9M	0M	0.0
Load at Start Snapshot:	1.81	1.97	.16	8.8
Load at End Snapshot:	1.9	1.86	-.04	-2.1
%User Time:	1.01	1	-.02	-1.0
%System Time:	.57	.56	-.01	-1.8
%Idle Time:	97.87	97.91	.04	0.0
%IO Wait Time:	.04	.04	0	0.0

Cache Sizes

- Cache Sizes are from the end snapshot; averaged across all instances

	1st (M)	2nd (M)	Diff (M)	%Diff
Memory Target				
SGA Target	2,048.0	2,048.0	0.0	*0.0*
Buffer Cache	552.0	552.0	0.0	*0.0*
Shared Pool	1,384.0	1,384.0	0.0	*0.0*
Large Pool	16.0	16.0	0.0	*0.0*
Java Pool	24.0	24.0	0.0	*0.0*
Streams Pool				
PGA Target	1,024.0	1,024.0	0.0	*0.0*
Log Buffer	10.0	10.0	0.0	*0.0*

Load Profile

	1st per sec	2nd per sec	%Diff	1st per txn	2nd per txn	%Diff
DB time:	1.1	1.1	*-3.6*	0.1	0.1	*0.0*
CPU time:	0.6	0.6	*0.0*	0.0	0.0	*0.0*
Redo size:	107,420.0	112,108.8	*4.4*	5,988.0	6,118.4	*2.2*
Logical reads:	9,380.8	9,493.7	*1.2*	522.9	518.1	*-0.9*
Block changes:	673.4	713.9	*6.0*	37.5	39.0	*3.8*
Physical reads:	139.3	119.4	*-14.3*	7.8	6.5	*-16.0*
Physical writes:	45.1	38.2	*-15.4*	2.5	2.1	*-17.1*
User calls:	88.2	98.1	*11.3*	4.9	5.4	*8.9*
Parses:	117.5	119.0	*1.2*	6.6	6.5	*-0.9*
Hard parses:	5.2	4.8	*-6.2*	0.3	0.3	*-10.3*
W/A MB processed:	3,651,674.2	4,621,325.5	*26.6*	203,558.7	252,211.0	*26.6*
Logons:	2.4	2.7	*12.7*	0.1	0.1	*15.4*
Executes:	758.5	770.7	*1.6*	42.3	42.1	*-0.5*
Transactions:	17.9	18.3	*2.1*			

	1st	2nd	Diff
% Blocks changed per Read:	7.2	7.5	*0.3*
Recursive Call %:	95.8	95.4	*-0.4*
Rollback per transaction %:	66.7	66.0	*-0.7*
Rows per Sort:	12.2	14.2	*2.0*

Appendix 5.1

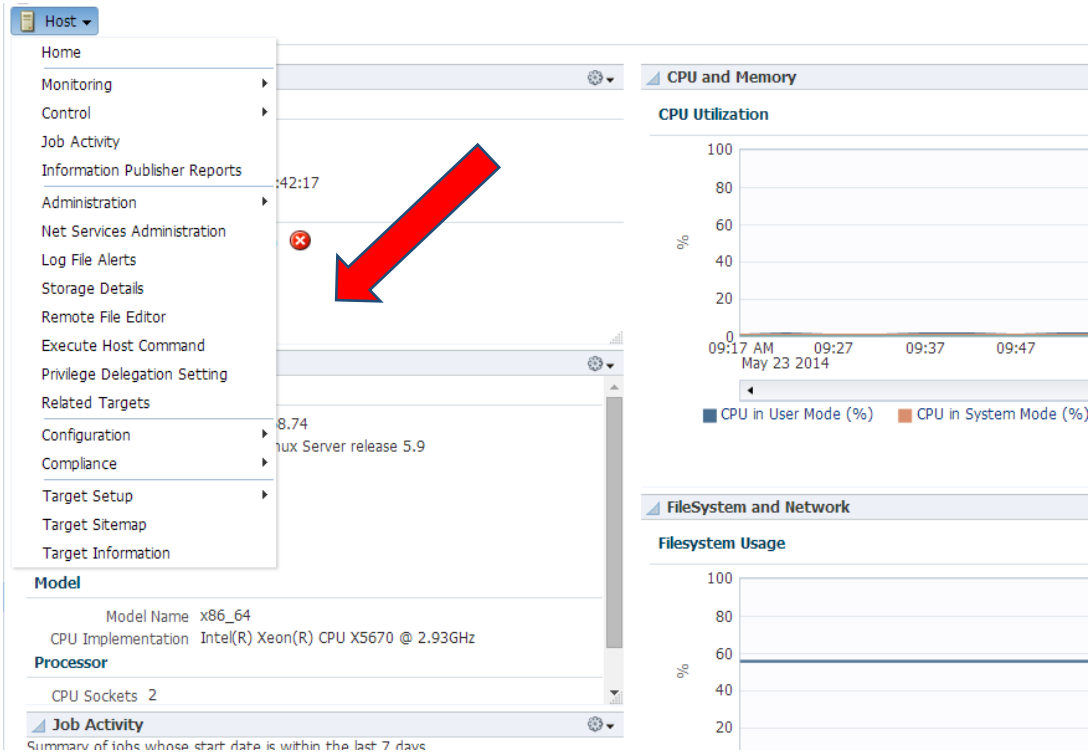
Exadata I/O Performance Maximum rates

Systems	Max throughput per device	Max IOPS per device	Max throughput per cell	Max IOPS per cell
High Performance HDD (600G/1.2TB)	152 MB/sec	297 I/Os/sec	1.8 GB/sec	3570 I/Os /sec
High Capacity HDD (3TB) (4TB)	108 MB/sec 121 MB/sec	166 I/Os/sec 190 I/Os/sec	1.3 GB/sec 1.45 GB/sec	1992 I/Os /sec 2280 I/Os /sec
Flash	228 MB/sec (read only)	7812 I/Os/sec (read only)	3.657 GB/sec (read only)	125,000 I/Os/ sec (read only)

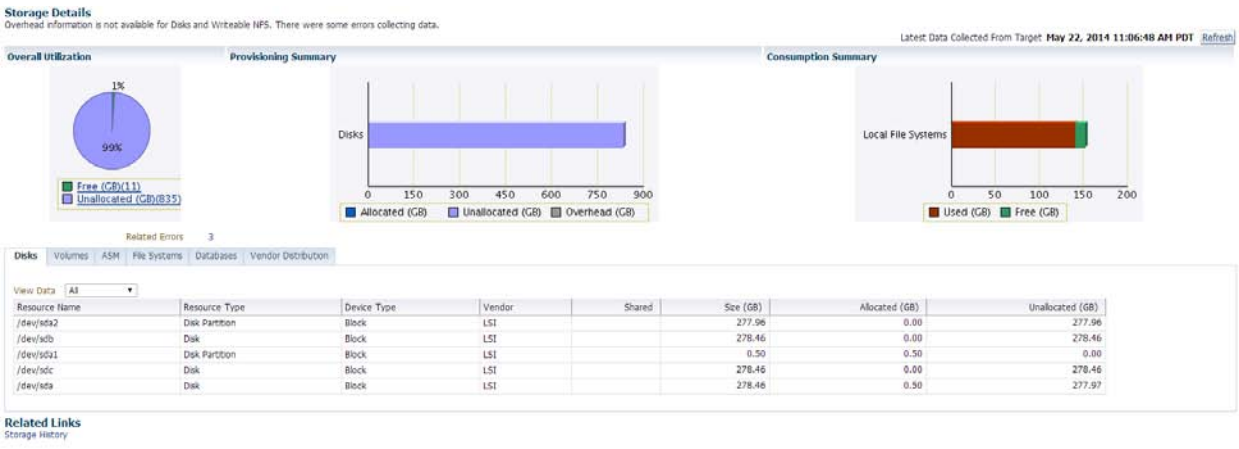
For additional information refer to: Oracle Exadata Database Machine Data Sheets

Appendix 5.2

File System Information

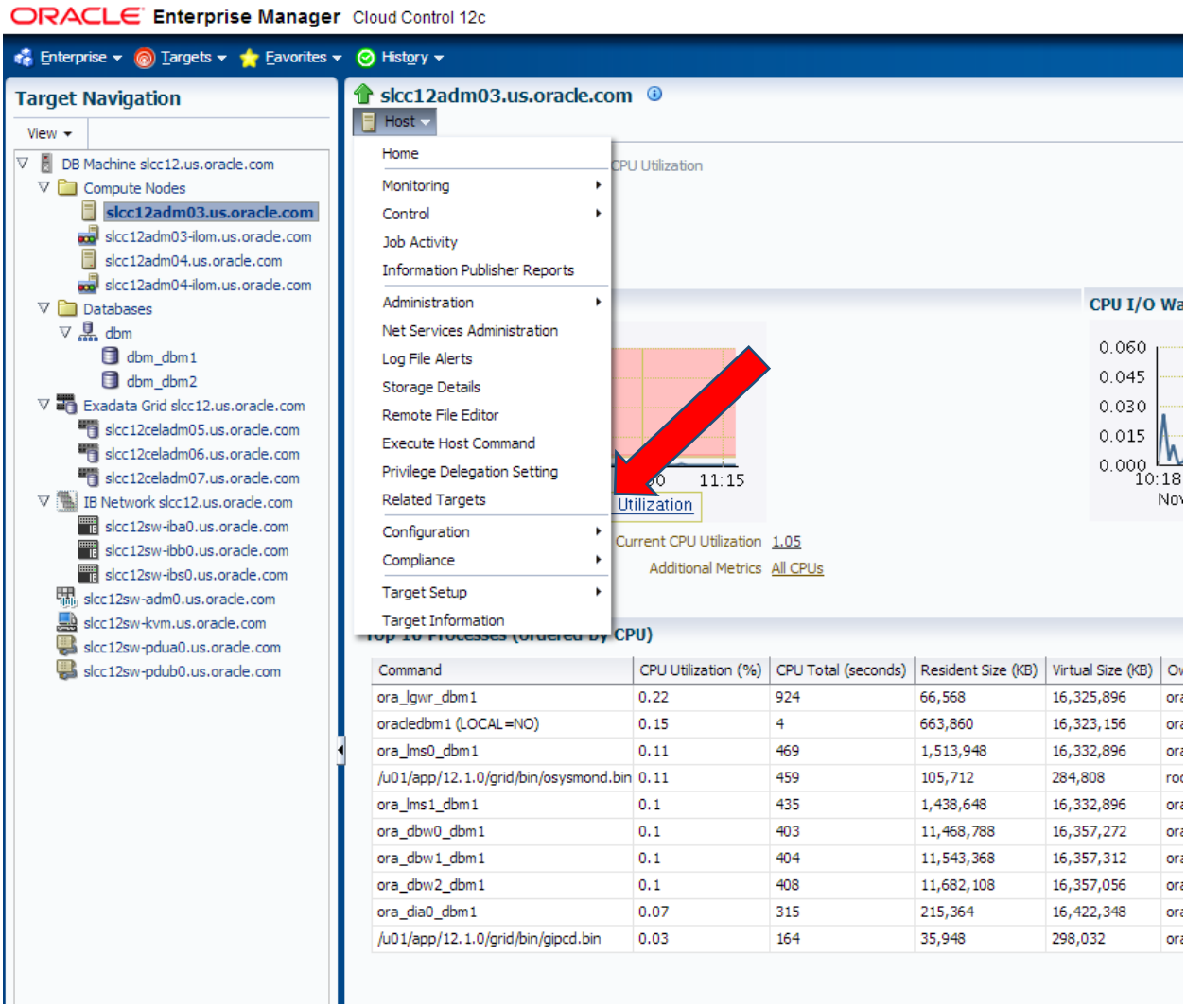


The resulting screen allows access to storage information on the file systems, ASM, local disks and

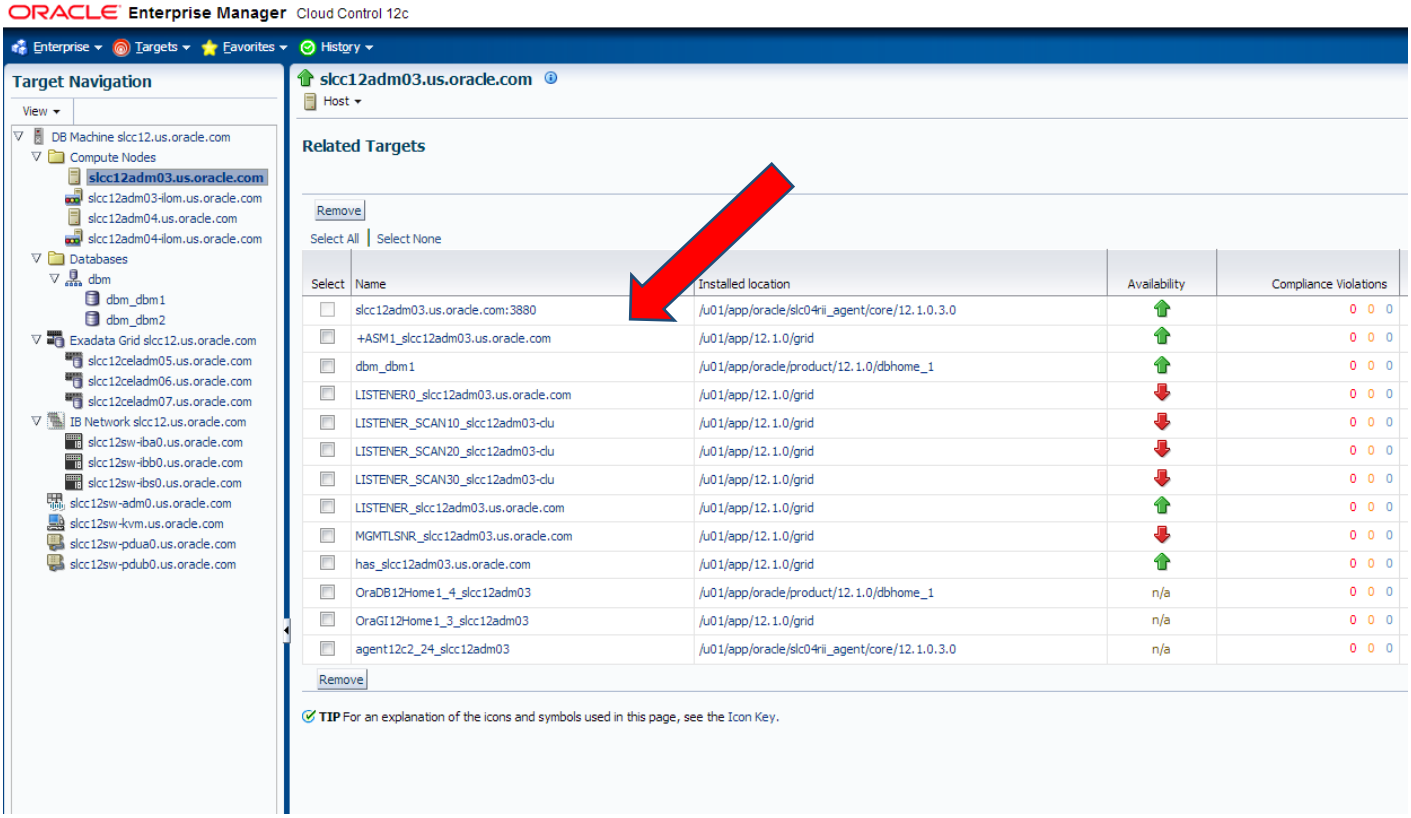


Appendix 5.3

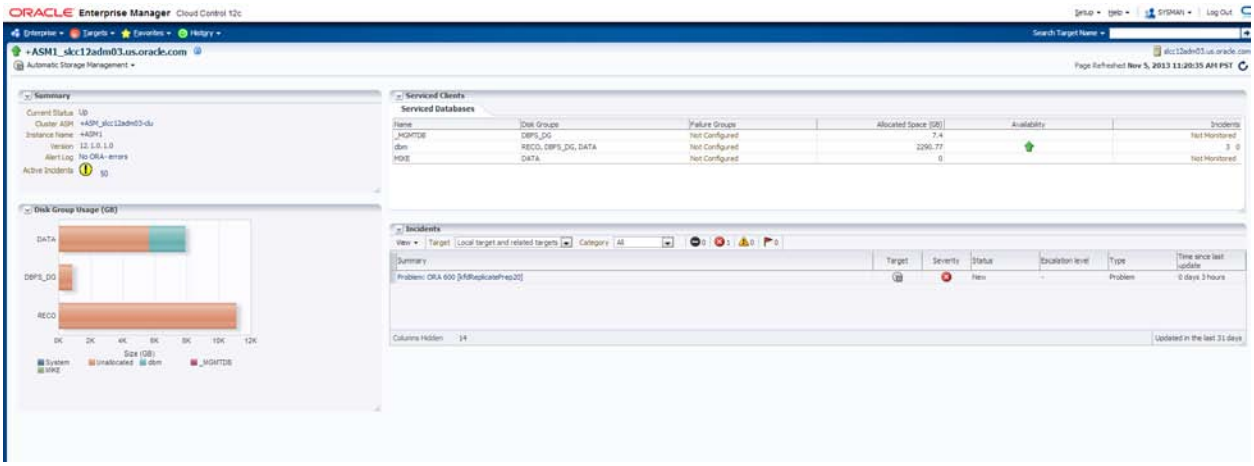
Detailed file system Information

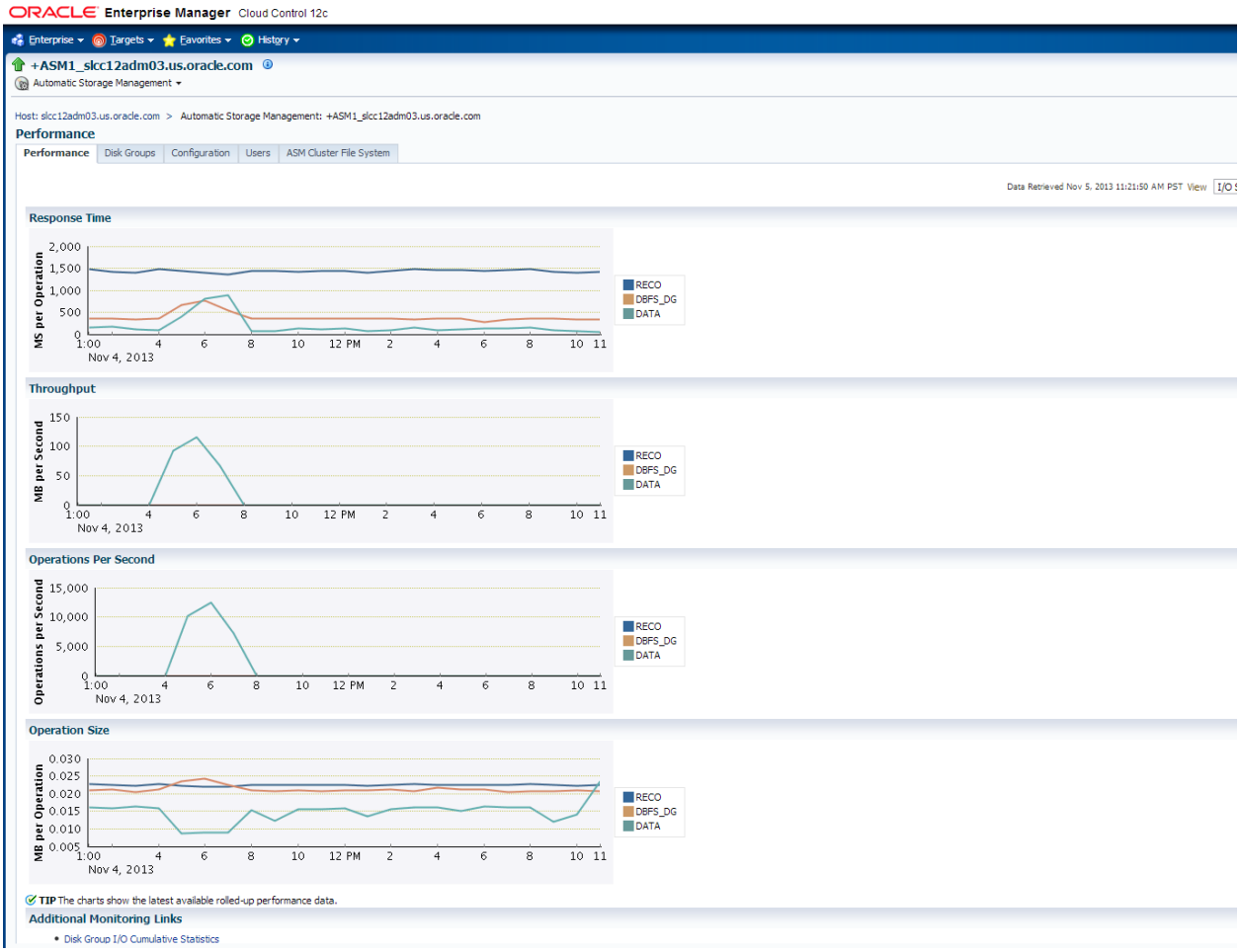


The ASM target should be listed on the resulting screen



Next click on the target and the ASM home page will be displayed




ORACLE Enterprise Manager Cloud Control 12c
Enterprise
Targets
Favorites
History
+ASM1_slcc12adm03.us.oracle.com
Automatic Storage Management
Host: slcc12adm03.us.oracle.com > Automatic Storage Management: +ASM1_slcc12adm03.us.oracle.com
Performance
Performance
Disk Groups
Configuration
Users
ASM Cluster File System
Data Retrieved Nov 5, 2013 11:21:50 AM PST View
Response Time
MS per Operation
2,000
1,500
1,000
500
0
1:00
4
6
8
10
12 PM
2
4
6
8
10
11
Nov 4, 2013
RECO
DBFS_DG
DATA
Throughput
MB per Second
150
100
50
0
Operations Per Second
Operations per Second
15,000
10,000
5,000
Operation Size
MB per Operation
0.030
0.025
0.020
0.015
0.010
0.005
TIP The charts show the latest available rolled-up performance data.
Additional Monitoring Links
Disk Group I/O Cumulative Statistics

Exadata Health and Resource Usage Monitoring

August 2013
Author: Mike Chafin
Contributors: Jia Shi, Shari Yamaguci, Juan Loaiza
Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.
Worldwide Inquiries: