Oracle Exadata

Tales from the X-files

Bryan Grenn

Contents

- Overview of V1 HP Exatata
- Overview of the v2 Sun Exadata
- Overview of the v2-2 Sun Exadata
- Overview of the v2-8 Sun Exadata
- The database layer
- The storage layer
- What is different with the Exadata
- What is not different with the Exadata.
- What to think about when purchasing
- Lessons learned
- Patching

V1 Exadata HP

- 8 database cells (2x4) 64 cores
- 14 storage cells (8 cores, 8g ram, 1tb sata or 450g SAS drives)
- 10g/second Infiniband bandwidth
- Sas or Sata Drives
- No flash cache

V2 Sun Exadata

- 8 database servers (8 cores 2.53 ghz, 72G memory per server).
- 14 storage servers (8 cores, 5.3 Tb flash cache)
- Sas or Sata drives
- 20g/second Infiniband

V2-2 Exadata

- 8 database servers (12 cores 2.93 ghz, 96G memory per server).
- 14 storage servers (12 cores, 378g flash/cell, 5.3 Tb total flash cache)
- SAS or SATA (with a SAS interface)
- 40g/second Infiniband

V2-8 Exadata

- 2 database servers (64 cores 2.26 ghz, 1tb memory per server).
- 14 storage servers (12 cores,378g flashcache/cell, total 5.3 Tb flash cache)
- SAS or SATA (with a SAS interface)
- 40g/second Infiniband

What the 2-2 offers

- More CPU and memory than the V2 Exadata
- Faster CPU’s than the v2 Exadata
- Faster CPU’s than the v2-8 Exadata.
- More parallelization across nodes

What the v2-8 offers

- More CPU’s than both the v2 and v2-2 exadata’s
- Slower CPU’s than the v2 and the v2-2
- More memory per node, and more memory total.

Why choose the v2-2

- COST ! It is about $1,000,000 cheaper (list price)
- Same list price as the v2 Exadata.
- Larger cluster with more redundancy and with more database nodes.
- Faster CPU's
- More parallelization for DW workload
- Can be purchased in ¼ increments.

Why choose the V2-8

- More memory (1tb/node is huge)
- More CPU (by count)
- Less RAC overhead
- More for OLTP processing

- *** Only available in Full RAC.

Storage Tier

- 14 storage cells
- 40gb/second Infiniband
- More CPU’s than the database servers
- 5.3 TB flashcache (write through)
- SAS or SATA.

SAS Vs Sata

SAS (high performance)

- Drives run at 15,000 RPM
- Drives are typically 300g or 600g

SATA (high capacity)

- Drives run at 7,500 RPM
- Drives are typically 2tb

Why choose one or the other

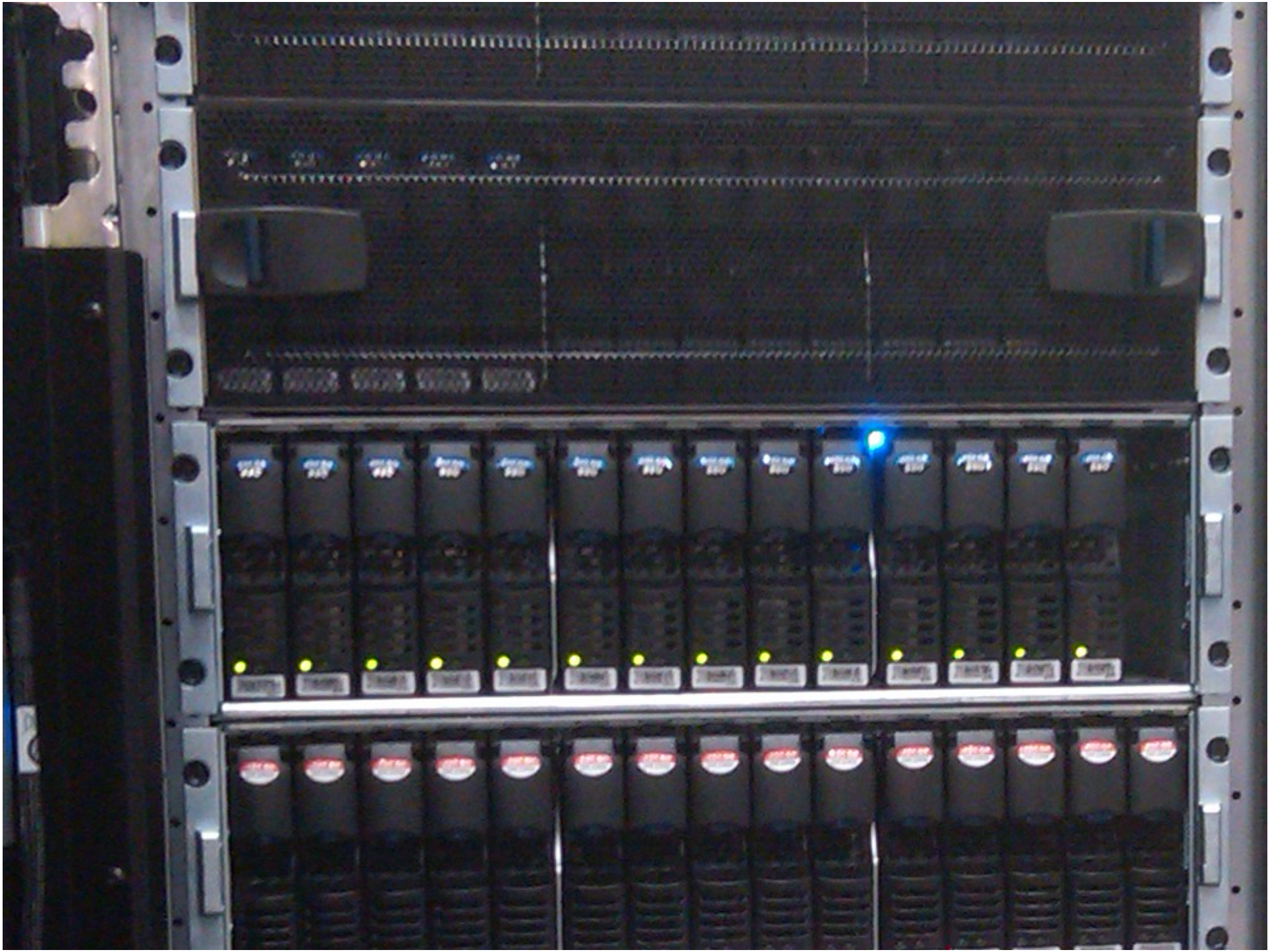
- SAS is 2x the speed for Seeks. (to move the head to find a block). Once a block is found, then sequential reads with read ahead are the same as Sata
- SAS has a much lower capacity (600g vs 2tb)
- SATA typically is 75% the speed of SAS, and will hold 3.5x the data (per Kevin Closson)
- Space vs Speed. Financial Apps typically go for Speed.
- 28tb of SAS vs 100tb of Sata.

How is the Exadata Different?

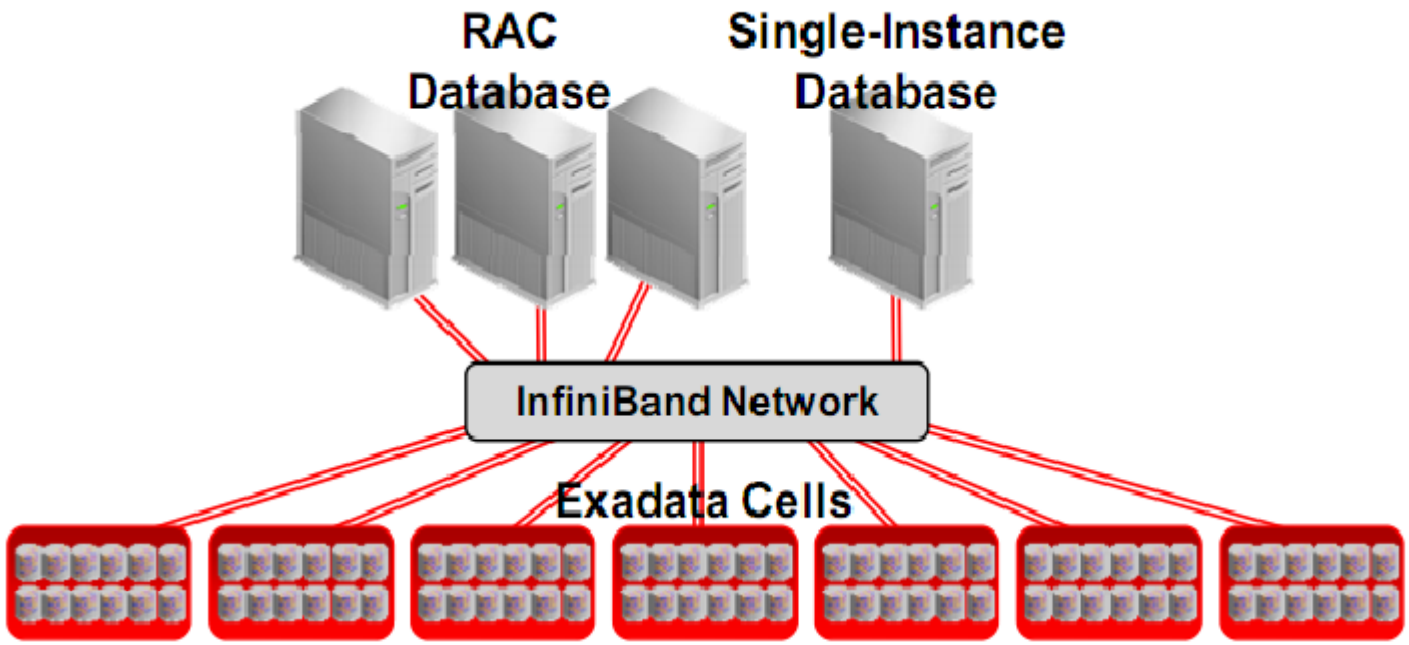
SAN

VS

Exadata

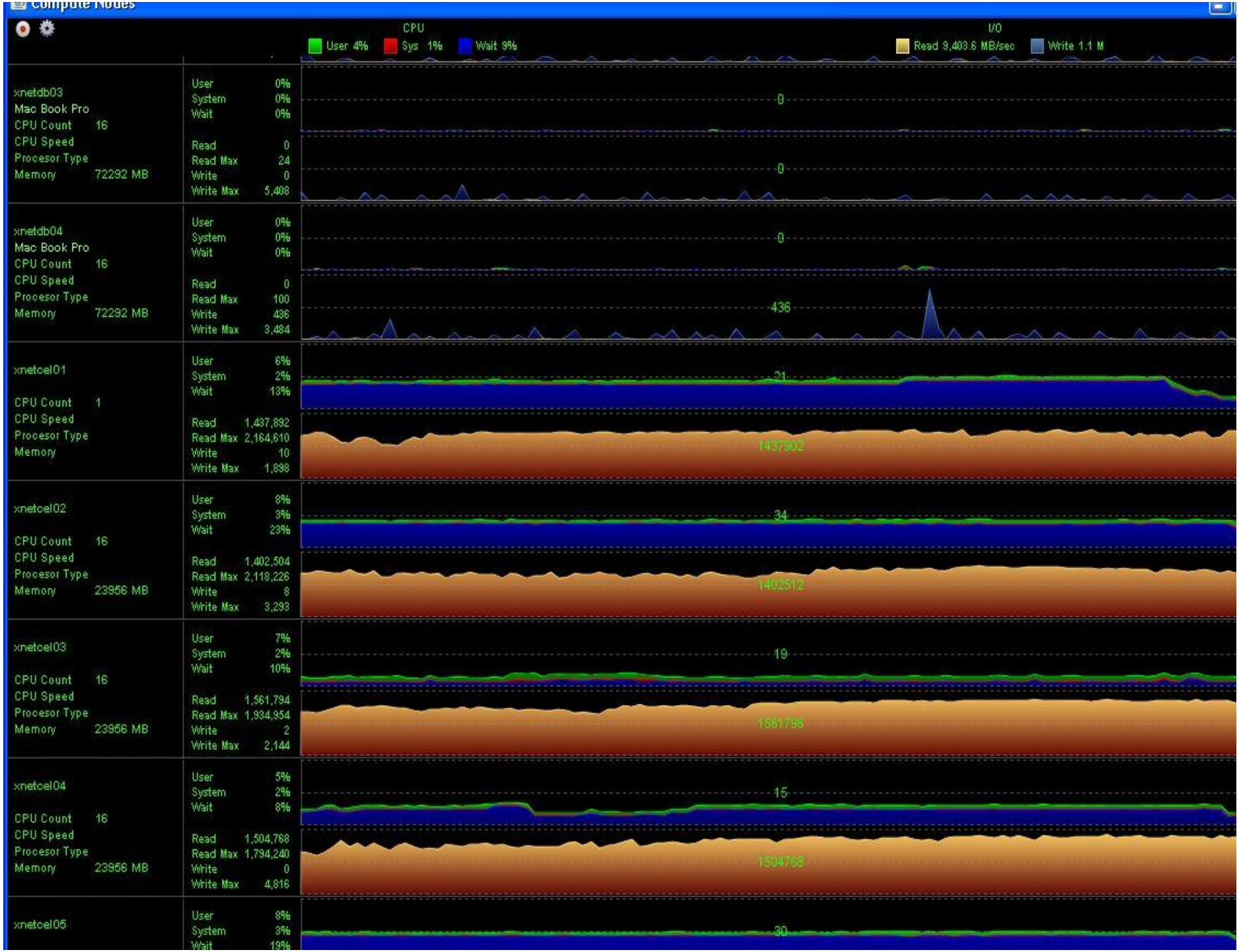


Back of an Exadata



Database Machine Half Rack Deployment

How do you know what it is doing?

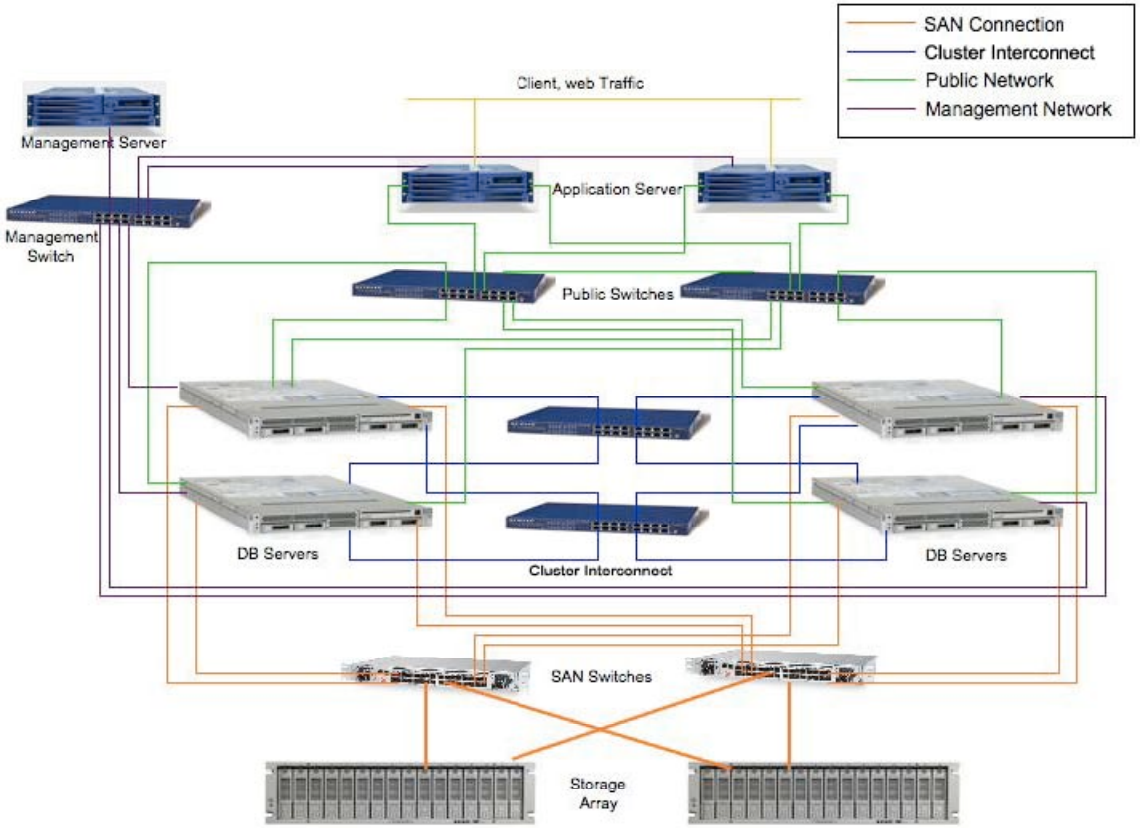

Compute Nodes
CPU
User 4%
Sys 1%
Wait 9%
I/O
Read 9,403.6 MB/sec
Write 1.1 M
xnetdb03
Mac Book Pro
CPU Count 16
CPU Speed
Procesor Type
Memory 72292 MB
User 0%
System 0%
Wait 0%
Read 0
Read Max 24
Write 0
Write Max 5,408
0
0
xnetdb04
Mac Book Pro
CPU Count 16
CPU Speed
Procesor Type
Memory 72292 MB
User 0%
System 0%
Wait 0%
Read 0
Read Max 100
Write 436
Write Max 3,484
0
436
xnetcel01
CPU Count 1
CPU Speed
Procesor Type
Memory
User 6%
System 2%
Wait 13%
Read 1,437,892
Read Max 2,164,610
Write 10
Write Max 1,898
21
1437902
xnetcel02
CPU Count 16
CPU Speed
Procesor Type
Memory 23956 MB
User 8%
System 3%
Wait 23%
Read 1,402,504
Read Max 2,118,226
Write 8
Write Max 3,293
34
1402512
xnetcel03
CPU Count 16
CPU Speed
Procesor Type
Memory 23956 MB
User 7%
System 2%
Wait 10%
Read 1,561,794
Read Max 1,934,954
Write 2
Write Max 2,144
19
1561796
xnetcel04
CPU Count 16
CPU Speed
Procesor Type
Memory 23956 MB
User 5%
System 2%
Wait 8%
Read 1,504,768
Read Max 1,794,240
Write 0
Write Max 4,816
15
1504768
xnetcel05
User 8%
System 3%
Wait 13%
30

The database layer

The Exadata really describes the storage layer. The database layer is truly just a RAC cluster.

This is important when looking at an Exadata. Look at your AWR. Is your workload mostly Logical reads or physical reads ? Are they index lookups, or FTS?

RAC database layer



Exdata Database nodes
VS
Traditional Database nodes

- Exadata utilizes most current Intel X86-64 chipset (westmere EP or Nehalem EX) and Memory architecture. This chipset is 200% the speed of AMD Opteron Chipset (Istanbul).
- Exadata utilizes Infiniband as the interconnect.

What this mean for Testing ?

- You will see speed gains because of the faster CPU's. Be sure to benchmark your systems performance against the CPU's in the Exadata you're comparing.
- You will see speed gains because of the version of Oracle. Oracle introduced an improved optimizer, and Direct path reads.

Direct path Reads

- For large scans (Oracle keeps track of this based on the explain plan), oracle will bypass the SGA, and bring the data back directly to the process memory (PGA). This ensures that large scans will not age out data in the Shared pool, and allows reads to be up to 2x the speed of traditional reads.
- The data read in a Direct path read is NOT sharable between processes, and is not kept for any future executions.

Thinks to keep in Mind when Testing

Try to compare Apples to apples as much as possible.

- Version of Oracle (direct path reads, more efficient explain plans,etc)
- CPU speed
- SGA size.

Why ?

All these items will improve performance and must be used to adjust for performance when doing benchmarking.

How to get a reliable test

- Baseline you current system
- Use DBReplay on your current system, and POC an Exadata
- Do AWR comparisons between current system and Exadata (easier said than done).
- Turn off storage indexes (alter session set "_kcfis_storageidx_disabled"=true -- turn them off)
- Turn off storage software altogethor (ALTER SESSION SET CELL_OFFLOAD_PROCESSING = false)

What is different with the Exadata

- Infiniband interconnect
- Storage cells
- Infiniband to the storage layer
- Flash cache on the storage cells
- Storage software
- Disk throughput is greater than that of Fiber
- Why HCC won’t work on Fiber.
- NO ACFS

What is the same with Exadata

- Multi-node RAC cluster
- Oracle 11.2.0.2
- Partitioning/Advanced compression/dataguard. Etc. Dbreplay
- Intel commodity servers
- ASM

What to watch out for when purchasing.

- Networking (1ge vs 10ge)
- Backup times
- Physical standby
- Fiber connections don’t exist.
- Lots and lots of IP’s (22 servers with 4 1ge ports and 2 10ge ports) along with switches, power supplies, etc. etc
- LOTS of cooling needed. They run very, very hot !!
- Location because Infiniband has a distance limitation.

Lessons Learned

- Oracle will try to compare your current system to an exadata. Try to compare what you would buy to an exadata.
- Getting all the groups to agree on the network configuration and who will support what is a HUGE task for a company.
- Control the testing. Oracle will try control the testing (and the message).
- POC the whole solution if it is more than just the exadata (things like Goldengate, ODI, etc).
- Know everything you can about the building blocks first so you can do your own testing (RAC, 11g, Linux, ASM etc).

More information on size

What it means if you have a ½ full Exadata.

A ½ full exadata is 50 tb of data.

- Data transfers at 100g/hour. That is 500 hours to transfer the data of 1gb network. 50 hours over 10gb.
- Backup times are very long (96 hours for 50tb was our estimate). That is with 4 streams.
- Instantiation times for your physical standby database.
- How do you keep it in hot backup mode for days for backups ?
- Can you run Synchronously ?

Patching

- Cells get patched at once, and some patches are not rolling
- CRS patches are not always rolling
- Database patches are not always rolling
- Exadata mostly gets patched as a whole
- You might want to split an exadata into multiple physical boxes.
- Do you buy patching services from Oracle ?
- How often do you patch ?

Questions ??