

Exadata: Just because it's 10x
faster, we can make 100X Slower

Terry Golden
terrygolden@brandx.net

AGENDA

1. Introductions
2. What is Exadata?
3. Why should I be interested in Exadata?
4. How can I make it 10X faster
5. or stop it being 100X slower?

Terry Golden

31 years of IT experience mostly in clustering
Team Member of Belcore's rewrite of 800 service
on Cluster & SS7

14 years of Oracle experience including 1st
Commercial Version of Oracle

Consultant to Oracle Consulting Services ACS on
RAC, Streams, Golden Gate, CDC, Upgrades
Architect for FiCC dept settlement replacement an
State of Oregon & US Navy RAC Consolidations

Member RAC SIG and numerous Oracle Groups

Source Material for Presentation

1. Experience on building a DW on Exadata V2-2 at handset maker Orangemotkia
2. Tom Kyte's Performance Talk using Exadata
3. Oracle Exadata website, white papers, etc
4. Other vendors' Exadata Experiences

What is Exadata?

- Complete (Servers, Disks, Networks) Rebuilt Configured Environment to run Oracle DBs
- Just Add Power, cooling, NI cables, DNS. DB
- Cluster in cabinet with dual power supplies
- Data Appliance most sql runs in storage server
- Smart Scan + Compression = 10X performance over server + SAN + Network
- V1 was based on HP H/W OLAP Only
- V2 is based on Sun H/W OLAP + OLTP

Exadata Expansion Path

You can chain up to 8 exadata together

Except for qtr rack systems

Patched as an Appliance

One patch for both RDBMS and server OS

Less downtime than san+server+rdbms patches

Half recent patches have required CRS/Grid software updates that meant an exadata outage

Need to buy exadata in pairs or have GG / Streams replicating to another database NOT on same exadata

IB Network Improves OLTP TIPS

- 3X speed of Ethernet
- Major improvement in high volume RACs

What is in Exadata x2-2?

- Cabinet contain Linux DB and Storage Servers
- Contains 2-3 IB + NI switches + HBA
- 4 x 300GB @ 10,000 RPM SAS drives
- Sold in quarter, half, Full RAC
- With 2, 4, 8 DB servers on 2 x 6 Core 2.93G
- 192, 384, 768 GB + 96 GB of Memory
- 1.1, 2.6 5.3 TB of Flash Cache
- 3, 7, 14 storage servers with 12 x 600GB 10K.
- Or 12 X 2TB 7.2K SAS drives

Exadata V2-8

- OCTA Core cpus vs HEX core on x2-6
- 10x/8x TDE encryption / decryption improvement as processing is done in CPU hardware
 - Only in Full rack 2 db nodes 8 x 8 cores
- 2 TB per rack vs. 768 GB on x2-2
 - 16 sockets 128 cores 256 threads on x2-8 vs
- 16 sockets 96 cores 192 threads on x2-6
- 1.5M IOPS vs 1.0M IOPS



Oracle Exadata Database Machine Packaging & Licensing

Complete Family Of Database Machines

Oracle Exadata V2

Oracle Exadata V2



- Quarter, Half, Full and Multi-Racks

- Full and Multi-Racks

Exadata Hardware

Scaleable Grid of industry standard servers for Compute and Storage

- Eliminates long-standing tradeoff between Scalability, Availability, Cost

Database Grid

Software

- Oracle Enterprise Linux OR Solaris 11 Express
- Oracle Database 11g EE R2
- Oracle RAC

Hardware

- 8 Dual-processor x64 servers OR
- 2 Eight-processor x64 servers

InfiniBand Network

- Redundant 40Gb/s switches
- Unified server & storage network



Intelligent Storage Grid

Software

- Exadata Storage Software

Hardware

- 14 High-performance low-cost storage servers
- 100 TB **High Speed** disk OR
- 336 TB **High Capacity** disk

5.3 TB PCI Flash

- Data mirrored across storage servers

Exadata Configurations

		X2-8 Full Rack	X2-2 Full Rack	X2-2 Half Rack	X2-2 Quarter Rack
	No. of Nodes	2	8	4	2
Database Server Nodes	No. of CPUs	16	16	8	4
	No. of Cores	128	96	48	24
	Memory	2 TB	768 GB	384 GB	192 GB
	No. of InfiniBand Switches	3	3	3	2
Exadata Storage Server Cells	No. of Cells	14	14	7	3
	No. of CPUs	28	28	14	6
	No. of Cores	168	168	84	36
	Number of Disk	168	168	84	36

Building blocks: X2-2 nodes = Sun Fire X4170 M2 w/ Intel Xeon X5670 CPUs

X2-8 nodes = Sun Fire X4800 w/Intel Xeon X7560 CPUs

Exadata Storage Server Cells = Sun Fire X4270 M2 w/ Intel Xeon L5640 CPUs

Exadata Product

		X2-8 Full Rack	X2-2 Full Rack	X2-2 Half Rack	X2-2 Quarter Rack
Raw Disk1	High Perf Disk	100 TB	100 TB	50 TB	21 TB
	High Cap Disk	336 TB	336 TB	168 TB	72 TB
Raw Flash1		5.3 TB	5.3 TB	2.6 TB	1.1 TB
User Data2 (assuming no compression)	High Perf Disk	28 TB	28 TB	14 TB	6 TB
	High Cap Disk	100 TB	100 TB	50 TB	21 TB

1 – Raw capacity calculated using 1 GB = 1000 x 1000 x 1000 bytes and 1 TB = 1000 x 1000 x 1000 x 1000 bytes.

2 - User Data: Actual space for end-user data, computed after single mirroring (ASM normal redundancy) and after allowing space for database structures such as temp, logs, undo, and indexes. Actual user data capacity varies by application. User Data capacity calculated using 1 TB = 1024 * 1024 * 1024 * 1024 bytes.

Exadata Product

		X2-8 Full Rack	X2-2 Full Rack	X2-2 Half Rack	X2-2 Quarter Rack
Raw Disk Data Bandwidth ^{1,4}	High Perf Disk	25 GB/s	25 GB/s	12.5 GB/s	5.4 GB/s
	High Cap Disk	14 GB/s	14 GB/s	7 GB/s	3 GB/s
Raw Flash Data Bandwidth ^{1,4}		50 GB/s	50 GB/s	25 GB/s	11 GB/s
Disk IOPS ^{3,4}	High Perf Disk	50,000	50,000	25,000	10,800
	High Cap Disk	25,000	25,000	12,500	5,400
Flash IOPS ^{3,4}		1,000,000	1,000,000	500,000	225,000
Data Load Rate ⁴		5 TB/hr	5 TB/hr	2.5 TB/hr	1 TB/hr

1 – Bandwidth is peak physical disk scan bandwidth, assuming no compression.

2 - Max User Data Bandwidth assumes scanned data is compressed by factor of 10 and is on Flash.

3 – IOPs – Based on IO requests of size 8K

4 - Actual performance will vary by application.

Oracle Exadata Database Machine

Software Licensing



	X2-8 Full Rack	X2-2 Qtr Rack	X2-2 Half Rack	X2-2 Full Rack
# of Database Servers	2	2	4	8
SW License: # of 'Processors'	64	12	24	48
# of Exadata Storage Servers	14	3	7	14
SW Lic: # Software 'Disk Drives'	168	36	84	168

- ✧ Oracle Database 11g R2 and Exadata Storage Server Software are required
 - ✧ Exadata Storage Server Software includes Hybrid Columnar Compression
- ✧ Partitioning Option required for data warehousing
- ✧ RAC required for multiple instance database(s)

Some Notes on Storage Cells

- 7.2/24 TB raw with FRA nets down to 4/14TB
- Default ASM redundancy is Normal could cut usable storage to 25% net FRA B4 Compress
- Data needs to be replicated across storage cells so if a cell fails data lives on another cell
- Think of storage cell as a shelf in SAN.
- Storage cells are more than SAN replacement most of the performance improvement comes from the software running in the cells.
- Performance improvements are automatic no need for special hints configuration

Usable Storage TB

Disk	Qtr	Half	Full
storage svrs	3	7	14
Fast/Small	21.6	50.4	100.8
Slow/Big	72	168	336

Usable Storage with FRA @ 40%

Fast/Small	12.96	30.24	60.48
Slow/Big	43.2	100.8	201.6
Compress 3x			
Fast/Small	38.88	151.2	181.44
Slow/Big	129.6	302.4	604.8

Exadata Performance Features

- I/O Resource Manager auto prioritizes sys I/O
- Flash Cache cuts I/O access to 10%
- Hybrid Columnar Compression / Smart Scans
- Predicate Filtering runs part of sql in storage
- Column Projections
- Bloom Filtering
- Storage Indexes
- Data Mining Scoring

IORM

- System IO (e. g. redo, archive, etc) automatically runs BEFORE other IO
- It's possible to negate by flooding redo with numerous small commits
- Can lower the priority of long running batch operations so OLTP or OLAP queries can complete faster

Flash Cache

- Present SAN's top out at around 120K IO/sec
- Oracle DB can do 1M IO/sec
- Volatile data kept will be lost on reboot
- For data between small hot memory and cold stored on disk blocks
- Great for scratch / worksheet tables
- Alter table cube_worksheet storage (cell_flash_cache keep)
- Start up trigger to populate price lists

HCC

- Compression of 10X on average
- Processing takes place on the storage cell
- Repeating column values are linked to rows in Compression Units in the block
- Need to use Bulk Load (e. g. append hit, parallel dml, sql*ldr direct path, or CTAS)
- Compress for Query | Archive (low | high)

Smart Scan

- Goal reduce the data volume into the SGA as will only match the result set
- Less load on FC and RAC interconnects+CPU
- Predicate Filtering process where in storage
- Column Projections only returns columns = sql
- Bloom filter Joins passes matching values to storage cells to match only requested values
- Storage Indexes High/Low Values @1MB disk
- Data Mining Scoring functions run on storage

Exadata will not cure a bad DB design

But will lower the problem event horizon as more and faster CPU's put more sessions in
Contention

As is often the case in RAC more pressure on single Mutex (e.g. sequences, etc) from more processes results in less through put.

Trick is to schedule parallel operation to run in less time so operations don't end up contending

Tricks to Effect Parallelism

Make non single block reads run in parallel

`SET PARALLEL_DEGREE_LIMIT = 10`

Will let optimizer enable automatic parallelism

Use Parallel Statement Queue

Set `PARALLEL_DEGREE_POLICY = auto` will queue sql waiting for resources

Use Resource Manager to carve your workload

Why Shouldn't I be interested in Exadata?

If can't afford \$300K qtr rack small data size db
unstructured data, performance not an issue,
serialized application that can't be changed,
can't run 11.2, or just need to waste money on
expensive servers and old SAN then
exadata is NOT FOR YOUR ORGANIZATION

Everyone else keep your seats

Budget Constrained Think Used Exadata

- V1 systems are still supported
- Given choice \$100K and choice between buying a V1 system or spending same amount on performance consulting DW V1 is faster less risky buy vs. build option

Who should use Exadata?

Want to save massive money on hardware
staffing overhead by consolidating DB's

DW with near uncontrollable storage growth

OLTP system with poor performance

How Will I Know if Exadata will work?

See if compression will benefit my data and
Profile your workload to see if it will run faster on
exadata then have Oracle or an outside expert do
a proof of concept using your heaviest workload
to prove your application should run faster with
better performance and less cost.

How to calculate Compression Saving

Use `dbms_compression.get_compression_ratio`
Passing `dbms_compression.comp_for_oltp!olap`
Along with `dba_object_table` to sum up present
uncompressed and suggested compressed block
counts. Divide sum of compressed blocks into
uncompressed blocks to lower the number the
higher your saving will be using Exadata
compression

How to Estimate CPU Savings

- Create and add to sqltune set during heavy vol
- Create sqlpa analysis task
- Run `dbms_sqlpa.execute_analysis_task` passing sqltune set and `cell_simulation_enabled` to false
- Rerun above this time with true
- `dbms_sqlpa.execute_analysis_task` with `compare_metric`
- `dbms_sqlpa.report_analysis_task('x','HTML');`

So Exadata is Pre built No Need to Configure the servers right?

- WRONG!!!
- Will have to enable huge pages and set SGA
- Work with network admins to connect net/dns
- DBCA to create your DB(s)

Exadata Doesn't need Indexes Right?

- NO!!!
- You will need to range partition dates and
- Have Primary key indexes
- Or Performance will be uneven
- Use Exadata features for 2nd 3rd dimensions of Data where cardinality trails off

Exadata POC

- Data Pump SQL minus alternative indexes
- Apply above script to Exadata DB
- Use IMPDP to import data
- Alternatively use transportable tablespace / db
- Have startup trigger to refresh flash cache
- Flash back DB & Apply workload from earlier exercise
- Repeat Real Application Testing, AWR until you have indexes you need to meet SLA

Migration to Exadata

- Using POC Exadata
- Impdp truncate from source db('s)
- Or transport tablespace and load table to table
- Remembering you'll need bulk load to get compression

Lessons Learned from Orangemotkia

- Beware of JAVA commit frequently mindset
- JMS feature divides large Messages in 2 txns
- Exadata wants larges batches of data to process in parallel bulk loads
- Application wants to commit after every txn
- Just moving commit from 1 to 50 txns improved performance 4X
- GG, informics etc doesn't do Bulk loads
- Right side index growth contention

Orangemotkia Cont'd

- We found 4 DOP was optimal (20K msgs/sec)
- JAVA developer tried testing with 16 DOP
- EM/GC Performance TAB only needed small star to complete RED FLAG on screen
- It only takes a few contenting processes to bring a DB on a full RAC to it's knees

Orangemotkia Cont'd

AWR will not tell you 'Hay Dumbie you're just archiving that table why didn't you create it with query compress archive (high) and tell the JAVA programmer to include an append hint'. Also why isn't Exadata tab inside performance tab and linked storage cell performance to my sql?

Staff knows what they know, and scrum will not Impart exadata skills that the tools only partially cover

Orangemotkia Cont'd

- I used types and bulk loading got 300X more throughput
- Development can I have you bulk batch messages in GG as multiple pumps start contenting with each?
- Not so much, have you tried writing your own adapter/handler?

Oracle Performance Talk Lessons

- Most DB look like arena scene in last episode of Blood & Sand Gods of the Arena dozens to thousands of processes fighting for resources
- Retro move back to ACMS/CICS where middle tier queued requests to database
- Showed 2000 connections alone can freeze db but limit sessions to 64 → 2000/TPS
- Tune web server to send fewer txns that will run in bulk and in parallel

Exadata DR Realities

- You're going to buy Exadata in pairs
- As while it's possible to DG to non exadata DR
- It will take considerable time to uncompress data so it can be used
- Or performance will be much slower on Logical Standby DB
- So spread the load so when you patch exadata you have another to take over load

Take Exadata move to re-balance load

- Divide load between OLTP and OLAP/DW
- Divide OLTP between Writes and Reads
- Use Active DG to service OLTP reads
- Use Logical DG / Streams / GG for DW
- Moving load will reduce contention in db that can only be good for performance
- Use DB services along with Resource Manager and IORM to craft optimal use plan

Exadata Changes Support Model

- Current Support model is specialized DB, NETWORK, SAN, Sysadmin
- Exadata needs to be managed as it's own island including all of the above skills
- Either have current staff train on Exadata
- Or dedicate staff to be Exadata experts

Exadata Futures

- Just IMHO
- Different types of storage are needed
- Solid State Drives will come as industry is moving in that direction
- FRA specific storage is needed doesn't make sense to store back in smart storage cells
- ExaTape? Oracle owns StorageTek and is pushing Tape Subsystem

Conclusions

- Exadata offers a major cost save in terms of storage and server costs along with massive performance improvement (10X) when compared to transitional SAN + Server environment
- Additional savings in administration costs by consolidating activity into an appliance model
- Appliance model will be future of Oracle DB

Exadata Questions?

- Terrygolden@brandx.net