Oracle Exadata X2/X3-8: A Critical Review

Mike Ault

Oracle Guru

Texas Memory Systems an IBM Company

What Is Exadata X2/X3-8?

- Is it software?
- Is it hardware?
- Is it the Borg?

Exadata as Borg?

Not Really (at least not yet…)

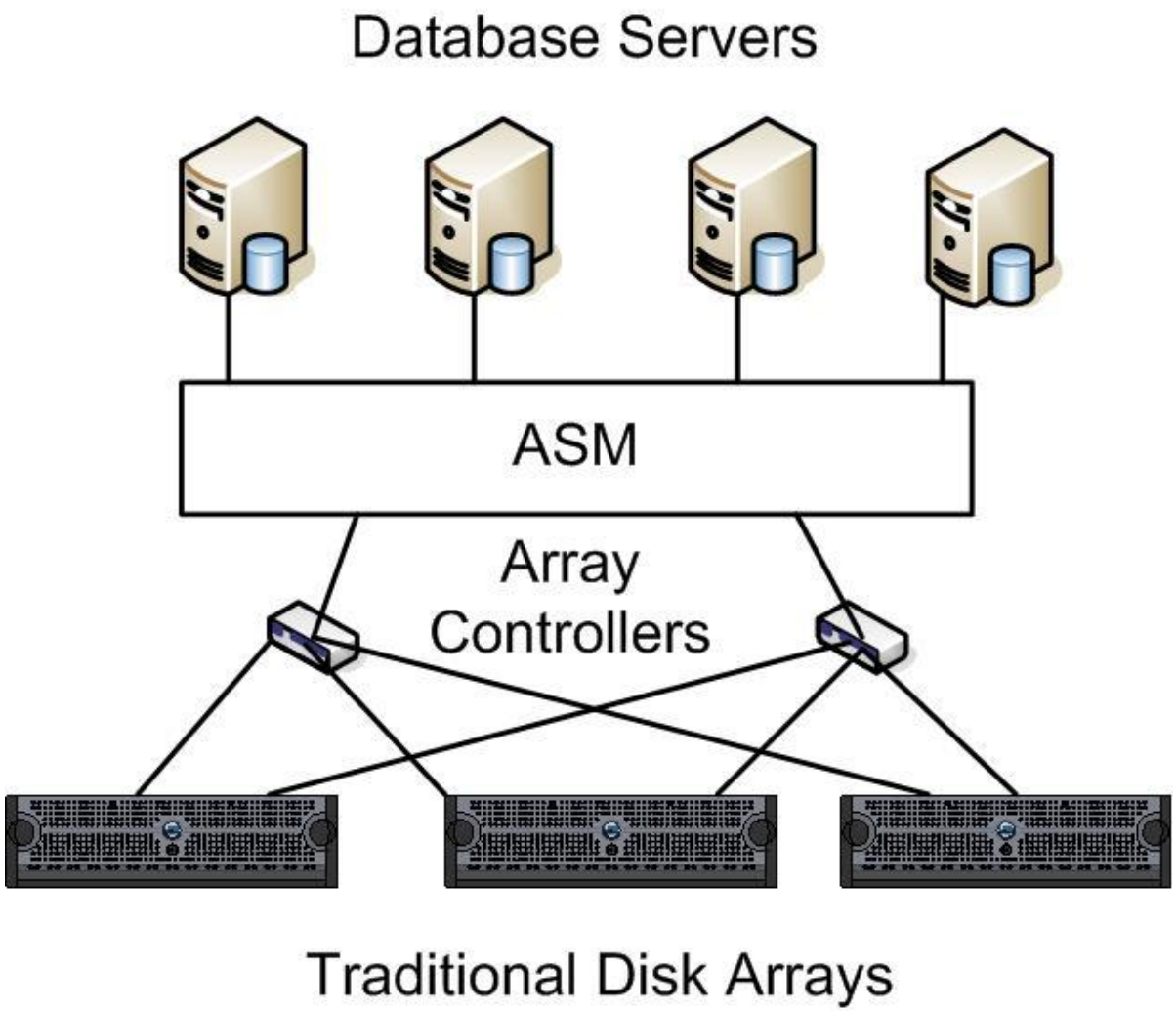
- Exadata is a combination of hardware and software
- Exadata takes state of the art *disk*, computer and flash technology and combines it with specially designed software from Oracle.
- Exadata hardware can only run Oracle software

Exadata X2-8 Hardware

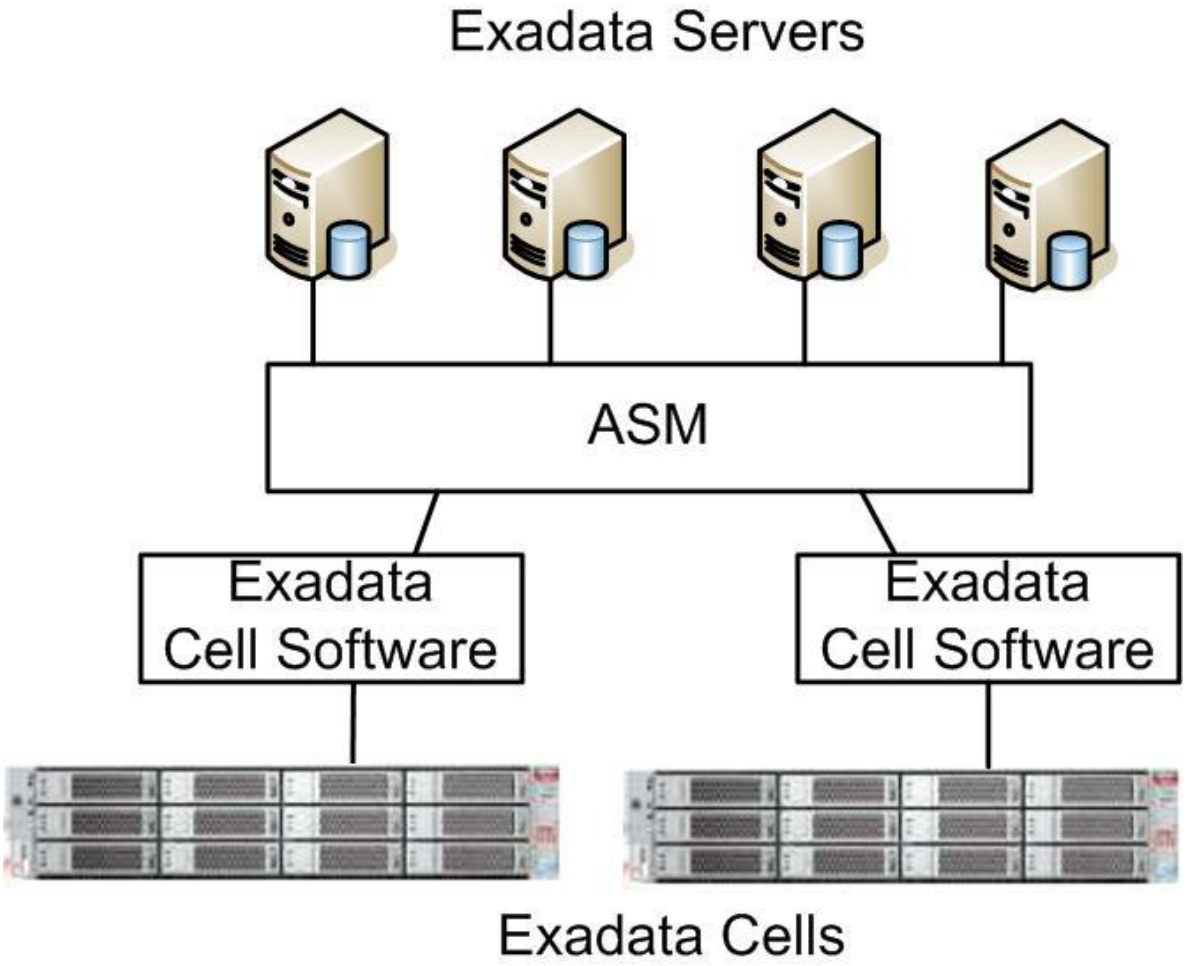
Exadata Database Machine X2-8 Hardware
2 x Database Servers, each with: • 8 x Eight-Core Intel® Xeon® X7560 Processors (2.26 GHz) • 1 TB Memory • Disk Controller HBA with 512MB Battery Backed Write Cache • 8 x 300 GB 10,000 RPM SAS Disks • 8 x InfiniBand QDR (40Gb/s) Ports • 8 x 10 Gb Ethernet Ports based on the Intel 82599 10GbE Controller • 8 x 1 Gb Ethernet Ports • 1 x ILOM Ethernet Port • 4 x Redundant Hot-Swappable Power Supplies
3 x 36 port QDR (40 Gb/sec) InfiniBand Switches
14 x Exadata Storage Servers X2-2 with 12 x 600 GB 15,000 RPM High Performance SAS disks or 12 x 2 TB 7,200 RPM High Capacity SAS disks Includes 5.3 TB Exadata Smart Flash Cache

(Looks to be the SUN X4800)

Traditional Setup



Exadata Setup



Each Exadata Cell has 4-96 GB flash cards

Exadata X2-8 Storage Hardware

- Sun x4270 M2 servers that contain dual six-core Xeon L5640 processors running at 2.26 GHz, with 24GB of memory
- 4-96 MB flash cards for Smart flash cache used to accelerate disk reads
- Disks are either high performance low capacity 15K or low performance high capacity 7.5K SAS drives
- Lose 66% or more of raw space for redundancy options or formatting losses
  - High Perf: 7.2 down to 2 TB per cell, 28 TB full rack
  - Low Perf: 24 down to 7 TB per cell, 98 TB full rack
- You pay license on a per disk basis
  - $10,000.00/disk, $120,000.00 per cell (total cost per cell $180K)
- Get IOPS based on large number of disks (168 to get 50K IOPS)
  - 200 IOPS/DISK*168 DISK=33,600 IOPS so this is doubtful
- Promises of 1,000,000 IOPS from FLASH (read-only)
- Full X2-8 hardware cost: $1,500,000.00 (doesn't include software!)

All prices/data taken from:

Oracle Exadata and Exalogic Pricelist, July 1, 2011

Oracle Technology Global Pricelist, July 1, 2011, Software Investment Guide

Exadata Smart Flash Cache and the Sun Oracle Database Machine, Oracle, Oct 2009

Something to Note:

Oracle Database Software (sold separately)	
For database servers	Oracle Database 11g Release 2 Enterprise Edition (11.2.0.2 or higher required), Oracle Real Application Clusters, Oracle Partitioning, and other Oracle Database options
For storage servers	Oracle Exadata Storage Server Software
Oracle Software (included)	
For database servers	Oracle Linux 5 Update 5 with the Unbreakable Enterprise Kernel or Solaris 11 Express: selectable at install time Zero-loss Zero-copy Datagram Protocol (ZDP) InfiniBand protocol used to communicate between the Exadata Storage Servers and the Oracle Database which is based on the Reliable Datagram Sockets (RDS) OpenFabrics Enterprise Distribution (OFED)

Exadata Configurations/Costs

System	Servers	Cores	Total Memory (GB)	Exadata Stroage Cells	Total Storage (HP) GB	Usable GB	Total Storage (HC) GB	Usable GB	Total Flash (Cell level) GB	Storage Cell Cores	Storage Cell Memory (GB)
X2-2 1/4	2	24	192	3	21,600	7,128	73,728	24,330	1,152	36	1,800
X2-2 1/2	4	48	384	7	50,400	16,632	172,032	56,771	2,688	84	4,200
X2-2 Full	8	96	768	14	100,800	33,264	344,064	113,541	5,376	168	8,400
X2-8	2	128	2,048	14	100,800	33,264	344,064	113,541	5,376	168	8,400

System	HW Cost	Oracle SW cost	Exadata Storage Cell Software	total cost
X2-2 1/4	$ 330,000.00	$ 1,182,000.00	$ 360,000.00	$ 1,872,000.00
X2-2 1/2	$ 625,000.00	$ 2,364,000.00	$ 840,000.00	$ 3,829,000.00
X2-2 Full	$1,100,000.00	$ 4,728,000.00	$ 1,680,000.00	$ 7,508,000.00
X2-8	$1,650,000.00	$ 6,304,000.00	$ 1,680,000.00	$ 9,634,000.00

All prices taken from:

Oracle Exadata and Exalogic Pricelist, July 1, 2011

Oracle Technology Global Pricelist, July 1, 2011, Software Investment Guide

Exadata 3 Year Projected

System	Hardware support/yr	Software Support/yr	Total Support Cost/yr
X2-2 1/4	$ 99,000.00	$ 232,440.00	$ 331,440.00
X2-2 1/2	$ 202,000.00	$ 491,280.00	$ 693,280.00
X2-2 Full	$ 352,000.00	$ 982,560.00	$ 1,334,560.00
X2-8	$ 452,000.00	$ 1,186,880.00	$ 1,638,880.00

System	Total 3 Year cost
X2-2 1/4	$ 2,534,880.00
X2-2 1/2	$ 5,215,560.00
X2-2 Full	$ 10,177,120.00
X2-8	$ 12,911,760.00

All prices taken from:

Oracle Exadata and Exalogic Pricelist, July 1, 2011

Oracle Technology Global Pricelist, July 1, 2011, Software Investment Guide

Exadata Features

Exadata Storage Software Features
• Smart Scan Technology • Smart Flash Cache • IO Resource Manager • Storage Index Technology • Hybrid Columnar Compression • Smart Scans of Data Mining model scoring

Smart Scan

- Based on maps (storage index technology) created at the cell level for each storage extent.
- High and low value for each column is stored
- Cell software uses these storage indexes to pre-process SQL
- Only the cells and extents that have data are searched.
- Hardware based fine grained partitioning of data.
- Each restart causes the storage indexes to be rebuilt.
- Storage indexes don’t work very well for OLTP
- If you have fairly calm data, such as a data warehouse, the smart scan technology (and storage index technology) will work well for you.

Smart Flash Cache

- SANs have caches to speed access.
- SAN caches will probably be DDR (on older machines) or flash (on newer systems)
- Cache frequently accessed blocks that aren't frequently updated.
- For Oracle SAN caches are set to be write-through, essentially making them read-only.
- Smart flash cache, a SAN cache optimized for Oracle.
- Only available on Exadata in the 4-96 GB SUN flash cards in each Exadata cell.
- The Smart Flash Cache is read-only and is for stable non-changing data.
- The flash cards can also be configured as flash LUNs giving high speed (well, as fast as SUN flash goes) access for read and write

SUN Flash Specifications

Performance	
Random Read (4K)	101 K IOPS
Random Write (4K)	88 K IOPS
Sequential Read (1M)	1.1 GB/sec
Sequential Write (1M)	567 MB/sec
IO service time (latency +4K transfer)	0.22 ms
Capacity	
Capacity - User	96 GB
Capacity - Raw[1]	128 GB
Domains (FMods)	4

- 220 microsecond response.
- Smart Flash Cache is a SAN cache optimized for Oracle and is only really useful for non-changing data.
- Only 394 GB of flash per 5 or so terabytes of storage

IO Resource Manager

- An extension to the DBMS_RESOURCE_MANAGER
- Specific to the storage cells
- Allows you to restrict IO resources by database.
- Great feature if you are consolidating many databases
- If you only have a few databases or one, it is a non-starter.
- You can only get 50,000 IOPS from a full X2-8 this can be a critical feature when consolidating.
- Specifications also state that you get 1.5 million IOPS from the flash cache but it is not managed with IO resource manager.
- When consolidating several databases of various IO needs not likely will get that many useful IOPS out of the flash cache.

Hybrid Columnar Compression

- The most fanfare in Exadata seems to be for a feature called Hybrid Columnar Compression (HCC).
- Optimizes data storage requirements while avoiding some of the performance issues associated with compression.
- HCC requires that data be loaded using data warehouse bulk loading techniques, it will not work on data entered from applications.
- Can provide data optimization rates as much as 15X normal capacity requirements but may also cause a noticeable performance loss, especially with volatile data.
- Use HCC on infrequently accessed OLTP data or non-changing data warehouse data and to only set the compression level to “low” or 4X for OLTP.
- HCC data in the flash cache is kept in compressed form, so that if any row is needed, the entire 32 KB compression unit needs to be stored
- HCC increases the size of the cache needed for full scans, but dramatically reduces the available size of the flash cache for random I/O.

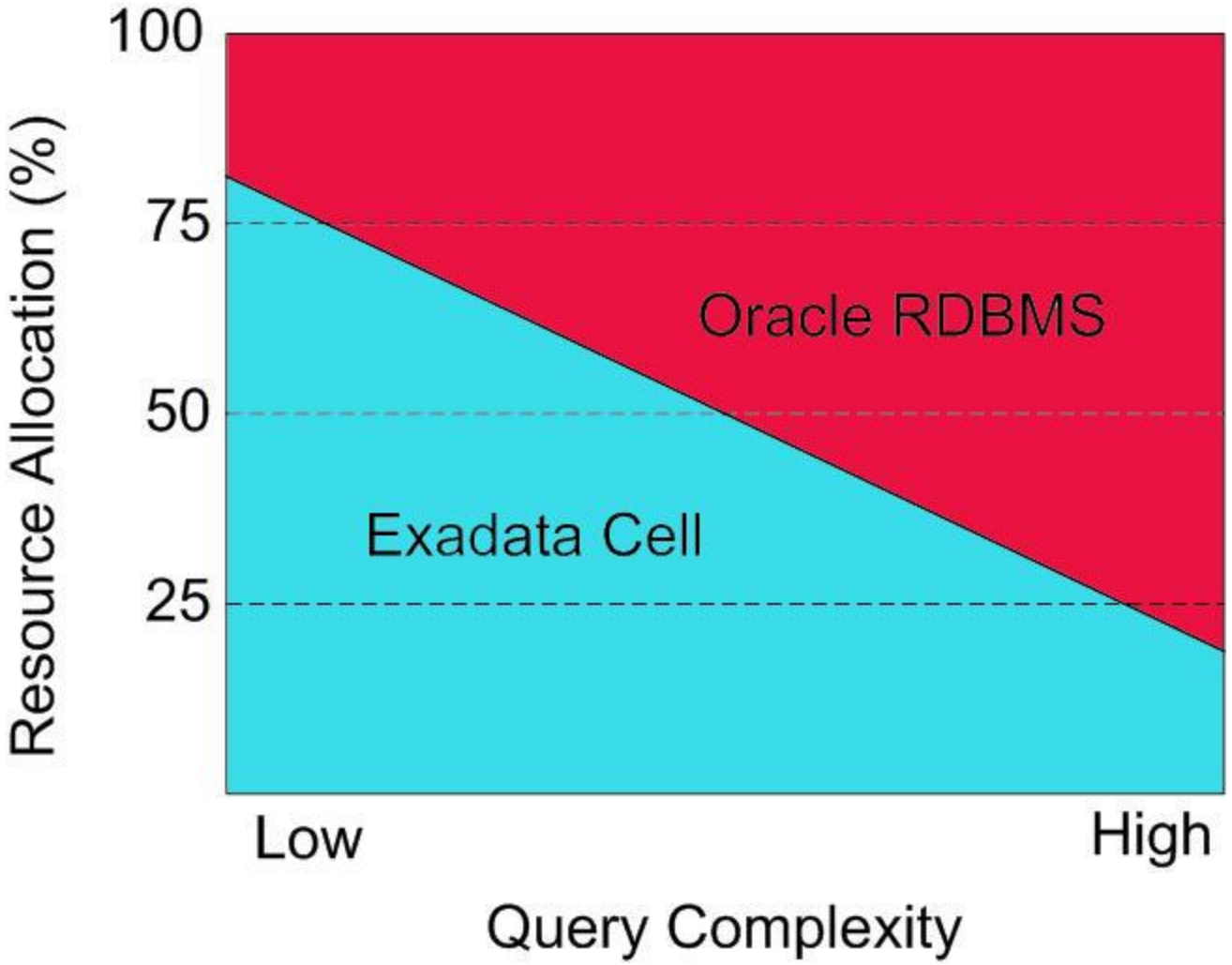
Hybrid Columnar Compression

- Updates to are much more complex since the full compression unit needs to be retrieved, uncompressed, modified, re-compressed and re-stored on the drives.
- Unless you have a poorly designed OLTP database with lots of duplicate entries in each table, HCC will not be effective
- In data warehouse or DSS/OLAP databases HCC can be very effective for non-changing data.
- Note that the highest compression settings on HCC can only be used for archival data and aren't recommended for your in-use data.
- It would seem that HCC would be best used on data that would be stored on cheap SATA or SAS based slow hard drive inexpensive storage.
- The overhead in loading, unloading and creating the HCC storage units is a negative.
- You can achieve high compression on static data with many repeating values but you get poor compression if you have few repeating values.
- You can get as good a result on most data using Oracle's Advanced compression which is available in Enterprise Oracle on all platforms.

Smart Scans of Data Mining model Scoring

- What a mouthful.
- The work of doing advanced models for data mining will be offloaded to the storage cells.
- There have been reports that Exadata has issues with complex summaries and queries.
- Oracle11GR2 has added many new statistical analysis features that can be built into models.
- Building of the models from these PL/SQL procedures can then be pushed down to the storage cells for a speed increase according to Oracle of anywhere from 2-26X.
- Have not seen many reviews of this feature and haven't the experience in BI and analytics to review it properly.

Overall



From: Teradata, Exadata is Still Oracle, March 2011

Overall

- Low
  - Simple star schemas
  - Simple joins
  - Fixed data with many duplicate entries
- High
  - Hybrid or snowflake schemas
  - Complex joins, summaries, etc.
  - Changing, non-duplicated data

Exadata Strengths

- One vendor
- High bandwidth data path
- Moves processing (in some cases) closer to the data
- Offers advanced compression for archival/fixed data
- Offers 5-20X acceleration for simple queries against relatively calm data
- Within the Exadata X2-2 family, easy upgrades

Exadata Weaknesses

- Limited to one vendor for hardware and software (limited flexibility)
- Not good for rapidly changing data
- Expensive
- Complex hardware and software
- Daisy cutter approach to upgrade from existing system
- Heavy metal approach to performance
- For DP if you go Dataguard, must use Exadata

X3-8

- 160 CPUs
- 4 TB of memory
- 168 cores in storage
- 22.4 TB of flash cache
- Starts at 1/8 rack (turns off CPUS)
- Enhanced HCC compression
- Write back caching

Summary

- If you intend on keeping your licenses where they are and just getting an Exadata, it is really expensive
- Uses “old” technology (disks) and charges you for them
- May have issues with rapidly changing data (frequent IUD) negating the new performance features
- Can get similar or better performance improvements without getting rid of existing technologies.
- Question: If all that processing is moved to the Cells, what do you need 128 to 160 CPUs for?

If Not Exadata?

If Not Exadata?

- So you may be asking the question: If not an Exadata, then what should we buy?
- Let's examine one of the alternatives to the Exadata X2-8 or X3-8, since they are the top of the line.
- An Exadata X2-8 will cost around $12M dollars over a three year period considering initial cost, hardware support and software licensing.
- Didn't include the required installation and consulting fees that go with that
- Let's look at performance

Exadata Performance

Exadata Product Performance

		X2-8 Full Rack	X2-2 Full Rack	X2-2 Half Rack	X2-2 Quarter Rack
Raw Disk Data Bandwidth[1,3]	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[1,3]	High Perf Disk	75 GB/s	75 GB/s	37.5 GB/s	16 GB/s
	High Cap Disk	64 GB/s	64 GB/s	32 GB/s	13.5 GB/s
Disk IOPS[2,3]	High Perf Disk	50,000	50,000	25,000	10,800
	High Cap Disk	25,000	25,000	12,500	5,400
Flash IOPS[2,3]		1,500,000	1,500,000	750,000	375,000
Data Load Rate[4]		12 TB/hr	12 TB/hr	6 TB/hr	3 TB/hr

1 - Bandwidth is peak physical disk scan bandwidth achieved running SQL, assuming no compression.

2 - IOPs – Based on peak IO requests of size 8K running SQL. Note that other products quote IOPs based on 2K, 4K or smaller IO sizes that are not relevant for databases.

3 - Actual performance will vary by application.

4 - Load rates are typically limited by CPU, not IO. Rates vary based on load method, indexes, data types, compression, and partitioning

ORACLE

Taken from a presentation given by Greg Walters, Senior Technology Sales Consultant, Oracle, Inc. to the Indiana Oracle Users Group on April 11, 2011

Exadata Performance

- We are primarily concerned with the numbers in the first column for the Exadata X2-8 Full Rack.
- Most will be buying the high performance disks so if we look at those specifications and meet or beat them, then we will also beat the low performance values as well.
- **Raw Disk Data Bandwidth: 25 GB/s**
- **Raw Flash Data Bandwidth: 75 GB/s**
- **Disk IOPS: 50,000**
- **Flash IOPS: 1,500,000**
- **Data Load Rates: 12 TB/hr**

Exadata Performance

- **Note 2 says:**
  - **IOPS- based on peak IO requests of size 8K running SQL. Note that other products quote IOPS based on 2K, 4K or smaller IO sizes that are not relevant for databases.**
- The actual value for IOPS is based on peak not steady state values.
- The system cannot sustain the peak value except for very short periods of time.
- When the IO is passed to the OS the request is broken down into either 512 byte or 4K byte IO requests since most OS can only handle 512 byte or 4K IOs.
- Modern disks (like those in the storage cells in Exadata) will only support 4K IO size so arguing that testing at 8K is more realistic is rather simplistic.
- In addition most flash IO is usually done at 4K

Exadata Performance

- **Note 3 says:**
  - **Actual performance will vary by application.**
- This is similar to mileage may vary and simply means that the numbers are based on ideal situations and the actual performance will probably be much less.

Injecting Some Reality

- Are these based on measurement or on what the interface will provide?
- At 50K IOPS with an 8K block size You only get 0.38 GB/s do the math: 50,000*8192/1024^3=0.3814.
- On the 1,500,000 IOPS from the flash: 1,500,000*8192/1024^3=11.44 GB/s
- Highest bandwidth that can actually be attained at peak IOPS for both disk and Flash would be 11.82 GB/s. Note 1 says that are not including any credit for either advanced or HCC compression.
- They don't tell you if the IOPS are based on 100% read, 80/20 read/write or 50/50 read/write, a key parameter is the mix of reads and writes if it is not specified the number given is useless.
- The Flash cache is at the Cell level and is actually used as an Oracle optimized SAN cache. This is read-only.
- Unless the data is relatively stable (non-changing) the actual useful IOPS from the cache could be quite a bit lower than advertised.
- In a DWH with unchanging data they may get read numbers that high at peak.

Injecting Some Reality

- Ok, so now we have some performance numbers to compare to:

- **Disk IO bandwidth:** **0.38 GB/s**
- **Flash IO Bandwidth:** **11.44 GB/s**
- **Disk IOPS:** **50,000 (read/write ratio unknown)**
- **Flash IOPS:** **1,500,000 (since this is cache, read-only)**
- **Total IOPS:** **1,550,000 (high estimate, unlikely you will get this IOPS)**

- The total IOPS for the system is 1,550,000 IOPS and the total bandwidth is 11.82 GB/s.
- They quote a loading bandwidth of 12 TB/s but make the claim it is based on the CPUs more than the IO capabilities.
- If we provide adequate bandwidth and CPUs we should be able to match that easily.

Disk performance Realities

- A high-performance disk will be lucky to achieve 250 random IOPS.
- 14 Cells X 12 Disk/cell X 250= 42,000,
- 300 IOPS for non-random IO then you get 50,400.
- In a test to achieve 100,000 IOPS from disks, EMC needed 496 disks yielding a value of 202 IOPS/disk,
- Exadata X2-8 disk farm can only achieve close to 34,000 IOPS

How About Some CPUs?

- 2-SUN Fire X4800 (Same as Exadata)
- Each with:
  - 8-8 core 2.26 GHz 7560 CPUs
  - 1 TB memory
  - 8 PCIe modules
- Cost: $268,392.00
- $240K for X3 (approx)
- May not need 128-160 C

Main Storage-SSD

- 1-U 10-20 TB HA eMLC Flash
- 450,000 IOPS per unit
- 4 QDR Infiniband ports per unit
- 110 microsecond read latency worst case (4k)
- (HP 28 TB) 3 - 10 TB - $450,000.00 30 TB
  - 45 TB X3 – 3 – 20 TB - $900,000.00 60 TB
- (HC 224 TB) 12 – $3,600,000.00 240 TB
- User can choose from 10-800 TB in 10 TB Chunks for a full rack

Some Licenses (Same base Oracle licenses as Exadata)

- 128 CPUS*$51,500.00=$6,592,000
- 160 CPUs*$51,500.00=$8,240,000
- No need for Cell licenses since no Cells!

Misc
(Support and connection)

- Switches
- Rack
- Cables, etc
- $40,000.00

Total

3 SSDs	$900,000.00
Servers	$240,000.00
Oracle	$8,240,000.00
Misc	$40,000.00
Total	$9,420,000.00*

(Plus shipping, handling, installation, support)

Save over $369K in ongoing license costs!

***Increases by $2,700,000.00 with HC option (100TB)**

New Specifications

- What would the specifications for this configuration look like?
- **Total servers: 2**
- **Total cores: 160**
- **Total memory: 4 TB**
- **Interface for IO: Infiniband**
- **Bandwidth: 12 GB/s from the interfaces, 5 GB/s sustained (by IOPS)**
- **Total Storage: 60 TB**
- **Total IOPS: 1,350,000 IOPS 80/20 read/write ratio doing 4K IOs (which by the way, map nicely to the standard IOs on the system). Peak rates would be much higher.**
- **Total cost with Oracle licenses and support for three years: Base: $9,420,000.00* + Support and licenses 2 additional years: $2,230,560.00=$11,650,560.00 for a savings of $2,618,808.00 over the three years.**

** Close to $8.2m of this cost is for Oracle core based licenses due to the 128 cores*

Support

- You would also get a savings in support and license costs of $523,600.00 for each year after the first three in addition to the savings in power and AC costs.
- Unless you are really consolidating a load of databases you will not need the full 128 CPUs
- Save license fees by reducing the number of cores (approximately $49K/core)
- In addition the X2-8 servers are configured with several terabytes of disk, another unneeded expense.
- You can do similar comparisons to the various X2-2 quarter, half and full racks and get similar savings.

All-In-One-Place

	Exadata X2-8,X3-8	SSD	Difference
IOPS (Storage)	50,000	1,350,000-4,900,000	1-3x
IOPS (Cache)	1,500,000	N/A	N/A
CPUs	128/160	128/160	0
Memory	2-4 TB	2-4 TB	0
Storage Capacity	28-224 TB Disk*	60-240 TB Flash**	3.9 to 1.02X larger
Storage Latency	1-5 ms	0.110 ms	10-45X smaller
Flash Cache	5.2-22 TB	N/A	N/A
Cost	$9,634,000.00-$12,672,200.00	**$9,420,000.00-$12,120,000.00**	less

*** Using Oracle numbers**

**** User incremented 10-200 in 10 TB increments**

What do you lose?

- Technology used to fix disk issues:
  - Smart Scan/storage indexes
  - HCC
  - IORM

Everything else is included with Oracle11gR2 Enterprise and the listed licenses!

Oracle Performance Claims

- I have no doubt that the claims made by Oracle for the various clients that have bought the Exadata are correct.
- They never show what the configuration of the previous system was in comparison to the new Exadata they have purchased.
- You can easily see 10X or even a 100X improvement in performance if the before configuration was severely under configured and the after configuration is severely over configured.
- It would be impossible to not get significantly better performance with almost any properly configured replacement system in many of the user cases used by Oracle.
- Oracle should show the complete configuration they replaced and then tell us how much performance improved with the new hardware and software.
- Without knowledge of both the before and after configurations any performance comparison is invalid.

Summary

- Exadata uses old technology (disk), write back flash caches and new software to brute force performance gains
- New technology such as Flash and SSD as storage can get the same benefits, cheaper
- Exadata locks you into Oracle technology and hardware *and license fees*
- Using newer technology keeps your options open with better performance

Thank You!

Mike Ault

mrault@us.ibm.com

http://www.statspackanalyzer.com/