CONFIO SOFTWARE

Ignite IT Performance™

Exadata Performance, Yes You Still Need to Tune

Kathy Gibbs
Senior Database Administrator, CONFIO Software

Who Am I?

- Over 18 years in IT and 12+ Years in Oracle & SQL Server
  - DBA and Developer
  - Worked for various industries (Telecom, Retail, Finance)
  - Oracle, SQL Server, Sybase, DB2 on VMware
- Sr DBA for Confio Software
  - KathyGibbs@confio.com
  - Makers of Ignite8 Response Time Analysis Tools
  - IgniteVM for Oracle/SQL/Sybase/DB2 on Vmware
  - Alarm VM for VM Admins

Agenda

- What is Exadata
- How Exadata Solves Performance Problems
- How is Performance Tuning different on an Exadata Machine
- How Exadata Can Create Performance Problems
- Questions

What are other Exa Products?

- First, the different ‘Exa’ Products
  - Exalogic and the Elastic Cloud
    - Think ‘WebLogic’ this is the application server solution
    - The Elastic Cloud is factory assembled and installed
    - Exabus is the defining architectural feature. It is basically the I/O subsystem
  - Exalytics In-Memory Machine
    - Contains ‘optimized’ TimesTen db
    - Includes BI Enterprise Edition and Essbase
    - The rumor is this is Oracle competing with SAP’s HANA hardware
  - Database Appliance
    - This is just a preconfigured 11gr2 database on OEL ‘Fully Redundant Integrated Database Appliance in a single box’
    - RAC and RAC One Licensing is included in the price.

What is Exadata?

- It is a preconfigured combination of hardware and software that provides a platform for running Oracle Db.
- Since Exadata includes a storage subsystem, new software has been developed to run at the storage layer.
  - This has allowed the developers to do some things that are just not possible on other platforms.

History of Exadata

- In 2008 Oracle introduced its 1st 'Database Machine'.
  - This Exadata V1 was based on HP hardware
  - This version was really marketed to big Data Warehouse shops.
- In 2009 they came out with Exadata V2 with Sun hardware, added Smart Flash Cache and now OLTP recommended
- In 2010 Oracle completed the acquisition of Sun and they came out with X2-2 and X2-8

What is Exadata – V2

8 Compute Servers

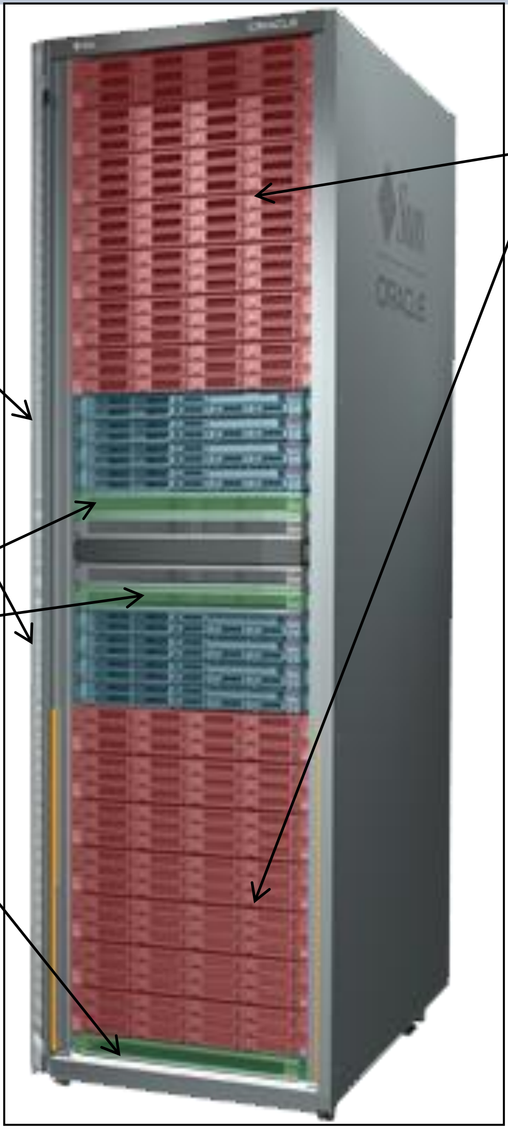
- 8 x 2 sockets x 4 cores = 64 cores
- 576 GB DRAM

InfiniBand Network

- 40 Gb/sec each direction
- Fault Tolerant

I/O Capacity and Performance

15K RPM 600GB SAS or 2TB SATA 7.2K RPM disks



14 Storage Servers

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

5TB+ flash storage!

What is Exadata – X2-2

8 Compute Servers

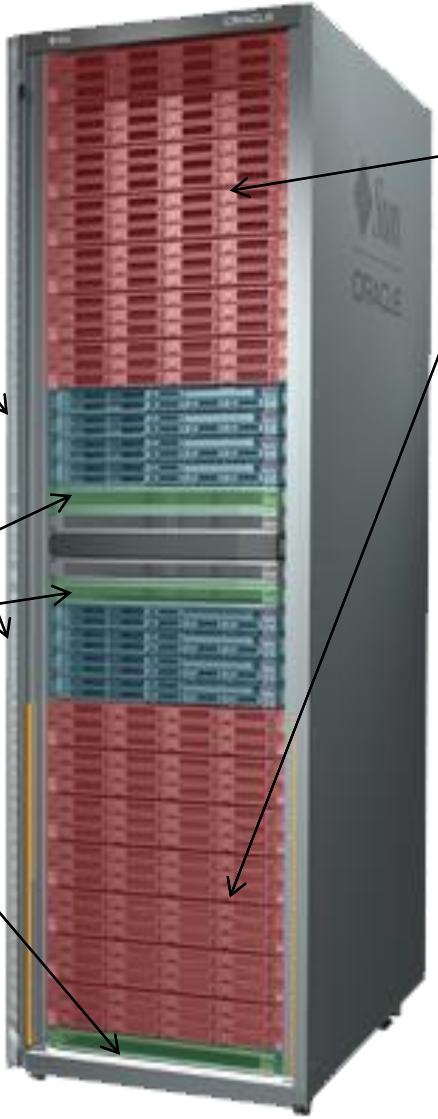
- 8 x 2 sockets x 6 cores = 96 cores
- 768 GB DRAM

InfiniBand Network

- 40 Gb/sec each direction
- Fault Tolerant

I/O Capacity and Performance

15K RPM 600GB SAS (HP model – high performance) or 2TB **SAS** 7.2K RPM disks (HC model – high capacity)
Note that 2TB SAS are the same old 2 TB drives with new SAS electronics.



14 Storage Servers

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

5TB+ flash storage!

What is Exadata – X2-8

2 Compute Servers

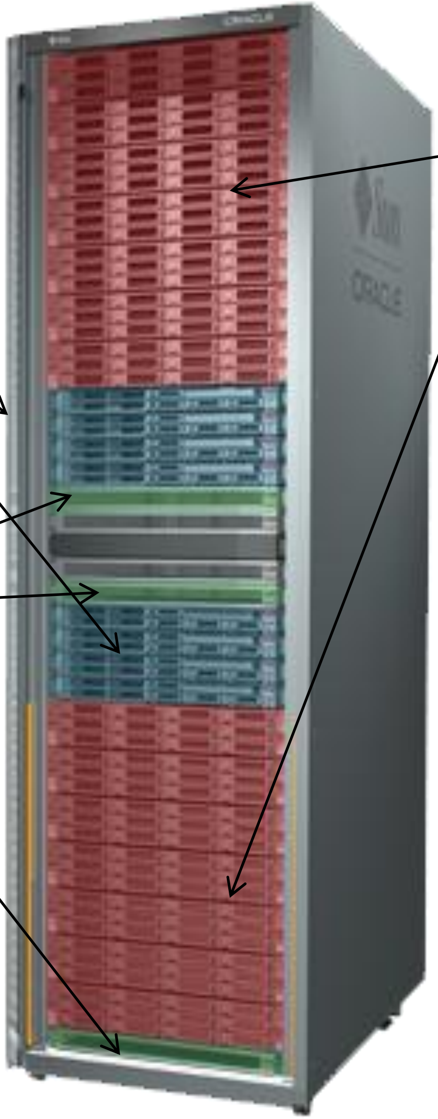
- 8 x 2 sockets x 8 cores = 128cores
- 2 TB DRAM

InfiniBand Network

- 40 Gb/sec each direction
- Fault Tolerant

I/O Capacity and Performance

15K RPM 600GB SAS (HP model – high performance) or 2TB **SAS** 7.2K RPM disks (HC model – high capacity)
Note that 2TB SAS are the same old 2 TB drives with new SAS electronics.



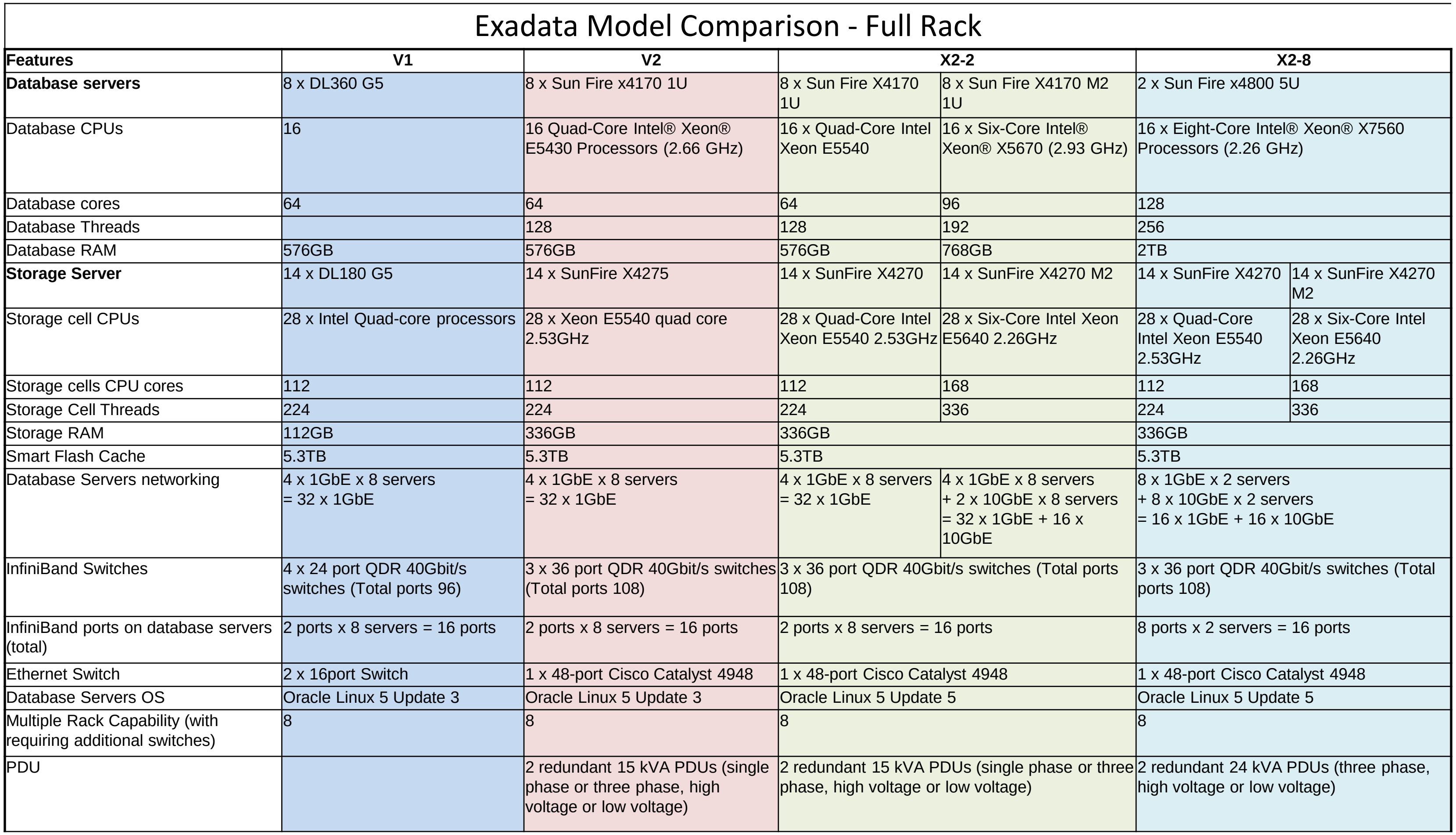
14 Storage Servers

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

Oracle Linux or Solaris x86

5TB+ flash storage!

What is Exadata – Comparison



Exadata Model Comparison - Full Rack

Features	V1	V2	X2-2		X2-8	
Database servers	8 x DL360 G5	8 x Sun Fire x4170 1U	8 x Sun Fire X4170 1U	8 x Sun Fire X4170 M2 1U	2 x Sun Fire x4800 5U	
Database CPUs	16	16 Quad-Core Intel® Xeon® E5430 Processors (2.66 GHz)	16 x Quad-Core Intel Xeon E5540	16 x Six-Core Intel® Xeon® X5670 (2.93 GHz)	16 x Eight-Core Intel® Xeon® X7560 Processors (2.26 GHz)	
Database cores	64	64	64	96	128	
Database Threads		128	128	192	256	
Database RAM	576GB	576GB	576GB	768GB	2TB	
Storage Server	14 x DL180 G5	14 x SunFire X4275	14 x SunFire X4270	14 x SunFire X4270 M2	14 x SunFire X4270	14 x SunFire X4270 M2
Storage cell CPUs	28 x Intel Quad-core processors	28 x Xeon E5540 quad core 2.53GHz	28 x Quad-Core Intel Xeon E5540 2.53GHz	28 x Six-Core Intel Xeon E5640 2.26GHz	28 x Quad-Core Intel Xeon E5540 2.53GHz	28 x Six-Core Intel Xeon E5640 2.26GHz
Storage cells CPU cores	112	112	112	168	112	168
Storage Cell Threads	224	224	224	336	224	336
Storage RAM	112GB	336GB	336GB		336GB	
Smart Flash Cache	5.3TB	5.3TB	5.3TB		5.3TB	
Database Servers networking	4 x 1GbE x 8 servers = 32 x 1GbE	4 x 1GbE x 8 servers = 32 x 1GbE	4 x 1GbE x 8 servers = 32 x 1GbE	4 x 1GbE x 8 servers + 2 x 10GbE x 8 servers = 32 x 1GbE + 16 x 10GbE	8 x 1GbE x 2 servers + 8 x 10GbE x 2 servers = 16 x 1GbE + 16 x 10GbE	
InfiniBand Switches	4 x 24 port QDR 40Gbit/s switches (Total ports 96)	3 x 36 port QDR 40Gbit/s switches (Total ports 108)	3 x 36 port QDR 40Gbit/s switches (Total ports 108)		3 x 36 port QDR 40Gbit/s switches (Total ports 108)	
InfiniBand ports on database servers (total)	2 ports x 8 servers = 16 ports	2 ports x 8 servers = 16 ports	2 ports x 8 servers = 16 ports		8 ports x 2 servers = 16 ports	
Ethernet Switch	2 x 16port Switch	1 x 48-port Cisco Catalyst 4948	1 x 48-port Cisco Catalyst 4948		1 x 48-port Cisco Catalyst 4948	
Database Servers OS	Oracle Linux 5 Update 3	Oracle Linux 5 Update 3	Oracle Linux 5 Update 5		Oracle Linux 5 Update 5	
Multiple Rack Capability (with requiring additional switches)	8	8	8		8	
PDU		2 redundant 15 kVA PDUs (single phase or three phase, high voltage or low voltage)	2 redundant 15 kVA PDUs (single phase or three phase, high voltage or low voltage)		2 redundant 24 kVA PDUs (three phase, high voltage or low voltage)	

Thanks to Oracle and http://blog.vishalgupta.com/

What is Exadata – Key Capabilities

Exadata Model Comparison - Full Rack

Key Capabilities	High Performance SAS Disks	High Capacity SATA Disk	High Performance SAS Disks	High Capacity SATA Disk	High Performance SAS Disks	High Capacity SAS Disk	High Performance SAS Disks	High Capacity SAS Disk
Hard Disk Type	450GB 15k RPM SAS	1TB 10k RPM SATA	600GB 15k RPM SAS	2TB 7.2k RPM SATA	600GB 15k RPM SAS	2TB 7.2k RPM SATA	600GB 15k RPM SAS	2TB 7.2k RPM SATA
Uncompressed raw disk bandwidth			21 GB/sec	14 GB/sec	25 GB/sec	14 GB/sec	25 GB/sec	14 GB/sec
Uncompressed Flash data bandwidth			50 GB/sec	50 GB/sec	50 GB/sec	50 GB/sec	50 GB/sec	50 GB/sec
Disk IOPS			50,000 IOPS	25,000 IOPS	50,000 IOPS	25,000 IOPS	50,000 IOPS	25,000 IOPS
Flash IOPS			1,000,000 IOPS	1,000,000 IOPS	1,000,000 IOPS	1,000,000 IOPS	1,000,000 IOPS	1,000,000 IOPS
Raw disk data capacity			100 TB	336 TB	100 TB	336 TB	100 TB	336 TB
Uncompessed user data			28 TB	100 TB	28 TB	100 TB	28 TB	100 TB
Smart Flash Cache			5.3TB		5.3TB		5.3TB	
Data Load Rate					5 TB/hour	5 TB/hour	5 TB/hour	5 TB/hour

Thanks to Oracle and http://blog.vishalgupta.com/

What is Exadata – Database Features

Exadata Model Comparison - Full Rack

Database Server	8 x DL360 G5	8 x Sun Fire x4170 1U	8 x Sun Fire X4170 1U	8 x Sun Fire X4170 M2 1U	2 x Sun Fire X4800 5U
CPU	2 x Intel Quad-core processors	2 x Quad-Core Intel Xeon E5540 2.53GHz	2 x Quad-Core Intel Xeon E5540 2.53GHz	2 x Six-Core Intel Xeon X5670 2.93GHz	8 x Eight-Core Intel® Xeon® X7560 Processors (2.26 GHz)
Memory	32 GB	72 GB	72 GB	96 GB	1 TB
Disk Controller		HBA with 512MB Battery Backed Write Cache	HBA with 512MB Battery Backed Write Cache		HBA with 512MB Battery Backed Write Cache
Local Disks	4 x 146 GB SAS disks	4 x 146GB 10K RPM SAS Disks	4 x 146 GB 10K RPM SAS	4 x 300 GB 10K RPM SAS	8 x 300GB 10K RPM SAS Disks
Local Storage	292GB (RAID1)	292GB (RAID1)	292GB (RAID1)	600GB (RAID1)	1.2 TB (RAID1)
Infiniband Ports		2 x QDR (40Gb/s) Ports	2 x QDR (40Gb/s) Ports		4 x Dual-port 4X QDR PCIe 2.0 (40Gb/s)
Ethernet Ports		4 Embedded Gigabit Ethernet Ports	4x1Gb	4x1Gb + 2x10Gb (Intel 82599 Controller)	8x1GbE and 8x10GbE using SFP+ connectors (Intel 82599 Controller)
ILOM Ethernet Port		1 Ethernet port (iLO2 with Advanced Pack)	1		1
Power Supplies		2 x Redundant Hot-Swappable	2 x Redundant Hot-Swappable		4 x Redundant Hot-Swappable

Thanks to Oracle and http://blog.vishalgupta.com/

What is Exadata – Storage Server

Exadata Model Comparison - Full Rack

Storage Server	14 x DL180 G5	14 x SunFire X4275 2U	14 x SunFire X4275 2U	14 x SunFire X4270 M2 2U	14 x SunFire X4275 2U	14 x SunFire X4270 M2 2U
CPU	2 x Intel Quad-core processors	2 x Quad-Core Intel Xeon E5540 2.53GHz	2 x Quad-Core Intel Xeon E5540 2.53GHz	2 x Six-Core Intel Xeon E5640 2.26GHz	2 x Quad-Core Intel Xeon E5540 2.53GHz	2 x Six-Core Intel Xeon E5640 2.26GHz
Memory	8GB	24 GB	24 GB		24 GB	
Flash Card		4 x 96 GB Sun Flash Accelerator F20 PCIe Cards	4 x 96 GB Sun Flash Accelerator F20 PCIe Cards		4 x 96 GB Sun Flash Accelerator F20 PCIe Cards	
Smart Flash Cache		384 GB	384 GB		384 GB	
Local Disks Count	12	12	12 (Non M2 model has 2TB SATA disks)		12	
Disk Controller		HBA with 512MB Battery Backed Write Cache	HBA with 512MB Battery Backed Write Cache		HBA with 512MB Battery Backed Write Cache	
Infiniband Ports		Dual-port 4X QDR (40Gb/s)	Dual-port 4X QDR (40Gb/s)		Dual-port 4X QDR (40Gb/s)	
Ethernet Ports		1 Embedded Gigabit Ethernet Port	1 Embedded Gigabit Ethernet Port		1 Embedded Gigabit Ethernet Port	
ILOM Ethernet Port		1 Ethernet port (LO100c)	1 Ethernet port (LO100c)		1 Ethernet port (LO100c)	
Power Supplies		2 x Redundant Hot-Swappable	2 x Redundant Hot-Swappable		2 x Redundant Hot-Swappable	
OS			OEL 5.5		OEL 5.5	

Thanks to Oracle and http://blog.vishalgupta.com/

What is Exadata – Environmental Specifications

Exadata Model Comparison - Full Rack

Environmental Specifications				
Height		42U, 78.66” - 1998 mm	42U, 78.66” - 1998 mm	42U, 78.66” - 1998 mm
Width		23.62” (600mm)	23.62” – 600 mm	23.62” – 600 mm
Depth		47.24” – 1200 mm	47.24” – 1200 mm	47.24” – 1200 mm
Weight		2171 lbs (986.8 kg)	2,131 lbs. (966.6 kg)	2,080 lbs. (943.5 kg)
Power - Maximum power usage		13.2 kW (13.6 kVA)	14.0 kW (14.3 kVA)	14.0 kW (14.3 kVA)
Power - Typical power usage		9.6 kW (9.9 kVA)	9.8 kW (10.0 kVA)	9.8 kW (10.0 kVA)
Cooling - At max usage		44,800 BTU/hr	47,800 BTU/hour (50,400 kJ/hour)	48,600 BTU/hour (51,280 kJ/hour)
Cooling - At typical usage		32,800 BTU/hr	33,400 BTU/hour (35,300 kJ/hour)	34,020 BTU/hour (35,890 kJ/hour)
Airflow - At max usage (front-to-back)		1680 CFM	2,200 CFM	2,200 CFM
Airflow - At typical usage (front-to-back)		950 CFM	1,560 CFM	1,560 CFM
Operating temperature		41° to 95° F (5° to 35° C) at sea level	5 ºC to 32 ºC (41 ºF to 89.6 ºF)	5 ºC to 32 ºC (41 ºF to 89.6 ºF)
Operating humidity		10% to 90% relative humidity	10% to 90% relative humidity	10% to 90% relative humidity

Thanks to Oracle and http://blog.vishalgupta.com/

Exadata Performance

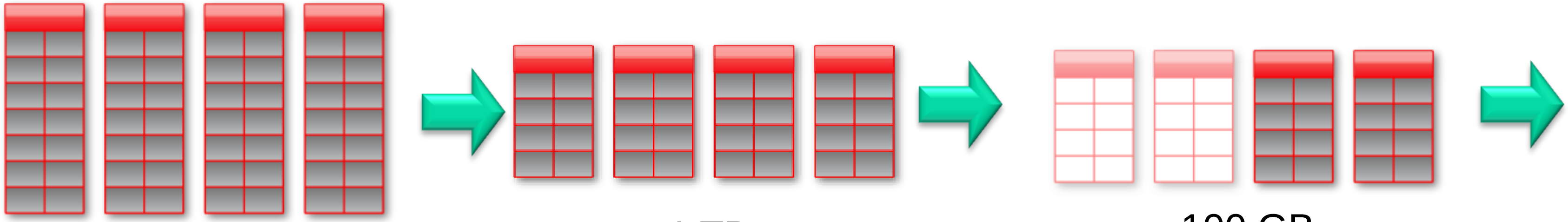
- Cell Offloading
  - Smart Scan
  - Storage indexes
  - Hybrid Columnar Compression
- DBRM/IORM
- Smart flash cache
- Additional Views

Exadata Performance - Cell Offloading

- Offloading is the generic term to describe what is now done at the storage tier that would have been done at the db tier.
  - Smart Scans are the access mechanism used to off load the tasks
  - DataFile Initialization – Through ASM speeds up initialization
  - RMAN Offload - When Block Change Tracking is used with RMAN and Exadata the Cell does the incremental backups at a granularity of the individual block, rather than at the granularity of a group of blocks as is done without Exadata.
  - HCC – will talk to in a later slide.

Why Exadata is so Efficient – Smart Scan

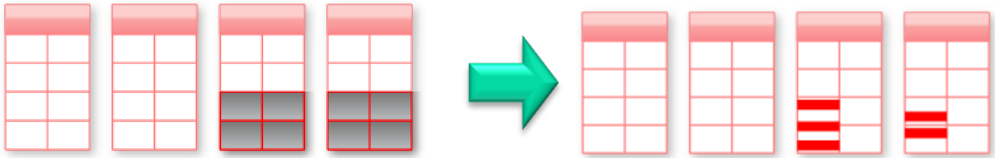
```
Select a.account_num, c.customer_name
From all_accounts a, all_customers c
Where a.acct_id = c.acct_id
And state = 'PA'
```



10 TB of user data
Requires 10 TB of IO

1 TB
with compression

100 GB
with partition pruning



20 GB
with Storage Indexes

5 GB
with Smart Scans

Sub second
On Database
Machine

Smart Scan

- Two Main Wait Events
  - Cell Smart Table Scan
  - Cell Smart Index Scan
- The flow of data from Smart Scan can’t be buffered in SGA buffer pool. (Think PGA (heap)

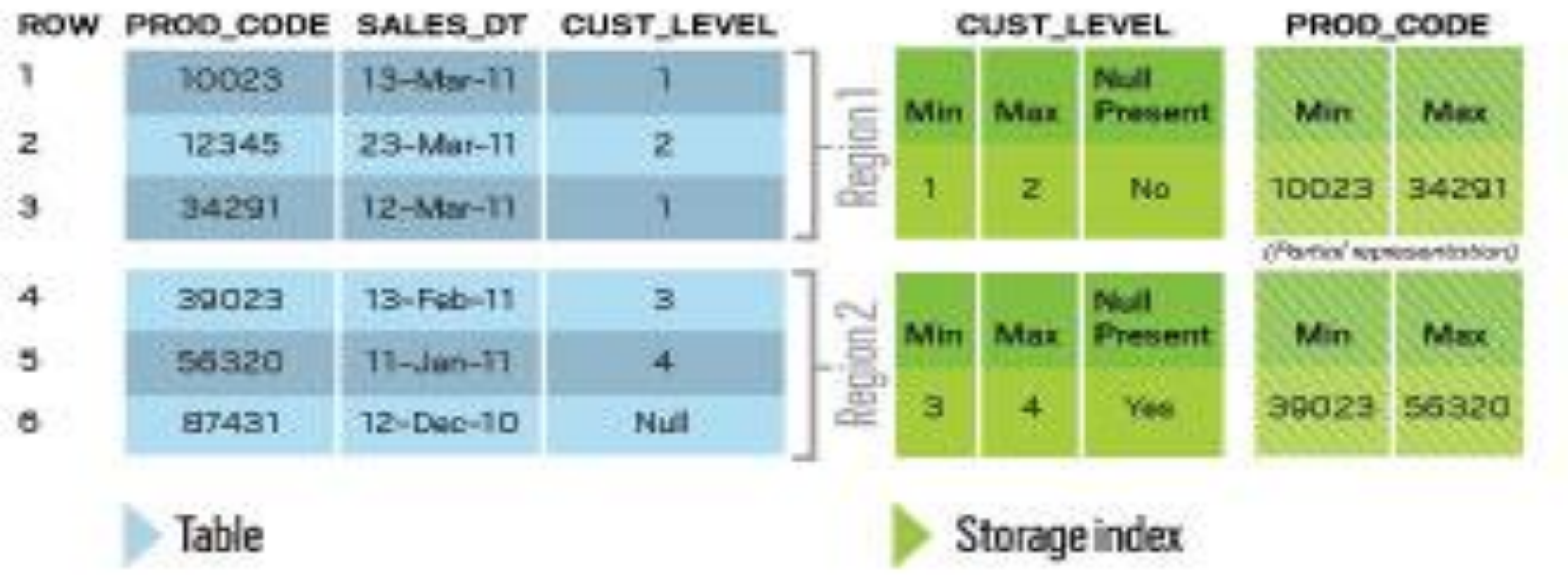
Exadata Performance – Storage Index

Storage Index

- A storage index is an in-memory structure that holds some information about the data inside specified regions of physical storage
- More importantly, it knows what is NOT located in that region.
- Think of this feature as a pre-filter

Exadata Performance – Storage Index

- Storage indexes actually work in the negative, find where it doesn't exist and therefore eliminate cells needed to be looked at
- To use them, the queries must use smart scans which means not all queries will benefit. Typically they are utilized for queries using predicates, full table scans or fast fill scans of indexes



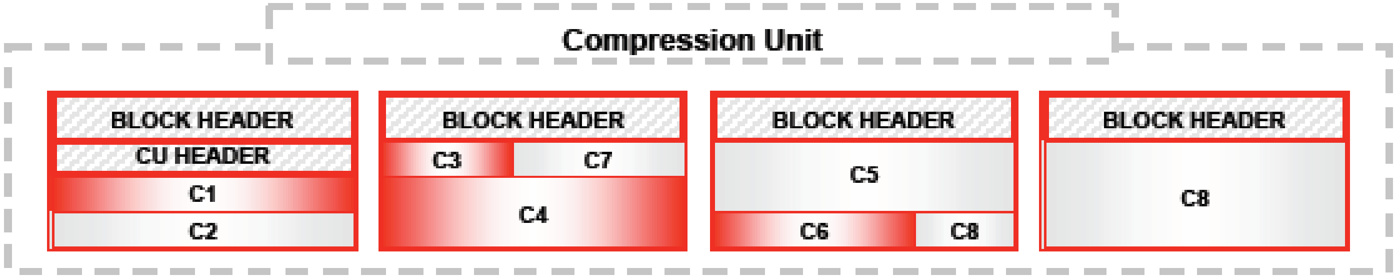
ROW	PROD_CODE	SALES_DT	CUST_LEVEL
1	10023	13-Mar-11	1
2	12345	23-Mar-11	2
3	34291	12-Mar-11	1
4	39023	13-Feb-11	3
5	56320	11-Jan-11	4
6	87431	12-Dec-10	Null

	CUST_LEVEL			PROD_CODE	
	Min	Max	Null Present	Min	Max
Region 1	1	2	No	10023	34291
Region 2	3	4	Yes	39023	56320

(Partial representation)

Exadata Performance – Hybrid Columnar Compression

Conceptual Illustration of a Logical Compression Unit



- This technology utilizes a combination of both row and columnar methods for storing data.
- A logical construct called the compression unit is used to store a set of HCC rows. When data is loaded, column values for a set of rows are grouped together and compressed. After the column data for a set of rows has been compressed, it is stored in a compression unit.

Exadata Performance – Hybrid Columnar Compression

- To maximize loading, use below ‘DW’ options. However you can also ‘regular’ DML
  - Insert statements with the APPEND hint
  - Parallel DML
  - Direct Path SQL*LDR
  - CTAS
- Types
  - Query Low
  - Query High
  - Archive Low

DBRM/IORM

- DBRM – Database Resource Monitor
  - This is the resource manager you are familiar with.
  - Without all sessions are given equal priority
  - The main reason is for use with consolidation
- IORM - I/O Resource Monitor
  - Added with V2 and beyond
  - Can prioritize I/O across dbs
  - For first time can virtually guarentee I/O service levels within and among dbs

DBRM/IORM

- DBRM
  - CPU Quantum – wait event. Resmgr: cpu quantum. Is the unit of CPU time that the DBRM uses for allocating CPU to consumer groups. Occurs when DBRM is actively throttling Cpu Consumpton.
  - Check DBRM Metrics V$RSRC_Consumer_group. Also v_$RSRCMGRMETRIC and V_$RSRCMGRMETRIC_HISTORY for monitoring effect of DBRM resource allocations have on sessions
  - Instance Caging – Provisions CPU at db instance level.

DBRM/IORM

- IORM
  - Interdatabase IORM – Manages priority among multiple dbs by db name
  - IORM Categories – This is a new attribute. Still by dbname. Ex oltp_category batch_category
  - Intradatabase IORM – On exadata when a DBRM plan is activated the db transmits a desc of this plan to all cells in the storage grid. So in a way this is a bit of the 'default'

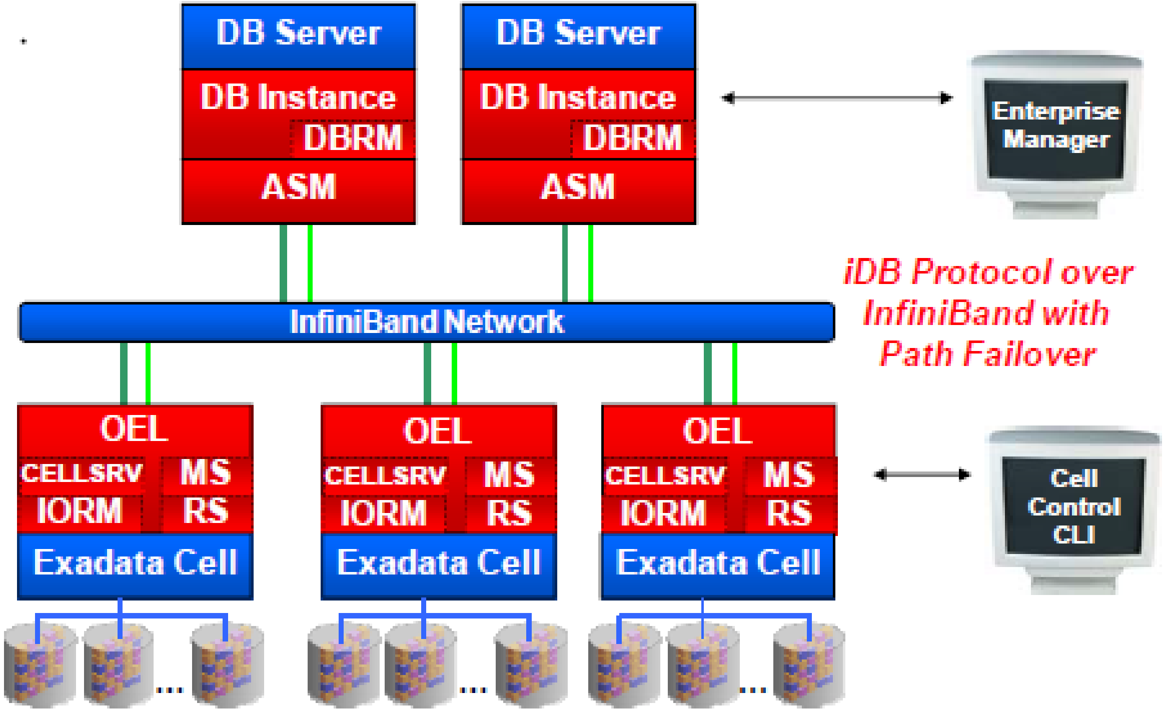
DBRM/IORM

- IORM
  - IORM manages at the storage cell
  - IORM distinguishes between small and large io request (<128k in size or > 128K)
  - For each cell disk cellsrv maintains an IORM queue for each consumer group and each backgrou pprocess. For each db accessing the cell.

```
Cellcli> alter iormplan objective = low_latency
Cellcli> list iormplan attributes object
 low_latency
```

Exadata Performance

DBRM/IORM



Or SQLPlus

Figure 4: Exadata Software Architecture

Exadata Performance – Smart Flash Cache

- This feature is what allowed V2 to be good for OLTP databases. 5TB on a full rack.
  - ~ 366 G usable flash storage per storage cell
- Can be used in two ways (or both)
  - Configured as cache - recommended
  - Carved out as solid state disks for ASM
- Write through cache. Writes bypass the cache and go directly to disk. However can copy data into cache if likely to be used again

Exadata Performance – Smart Flash Cache

- Follow by querying v$sysstat (and related v$views)
  - Cell flash cache read hits
- CELL_FLASH_CACHE=KEEP
  - None, default, keep
  - Alter table oltp.busy_table (cell_flash_cache=keep);
- Cellcli CELL_FLASH_CACHE=KEEP
  - Create flashcache all size=300g
  - List flashcache detail
  - List celldisk attributes name, diskType, size where name like 'FD.*'
  - List metriccurent where objecttype = 'FLASHCACHE'
  - List flashcachecontent – shows objects in cache

Exadata Performance – Smart Flash Cache

- Follow by querying v$sysstat (and related v$views)
    - Can be overestimated if all in ‘KEEP’ state
    - Cell flash cache read hits

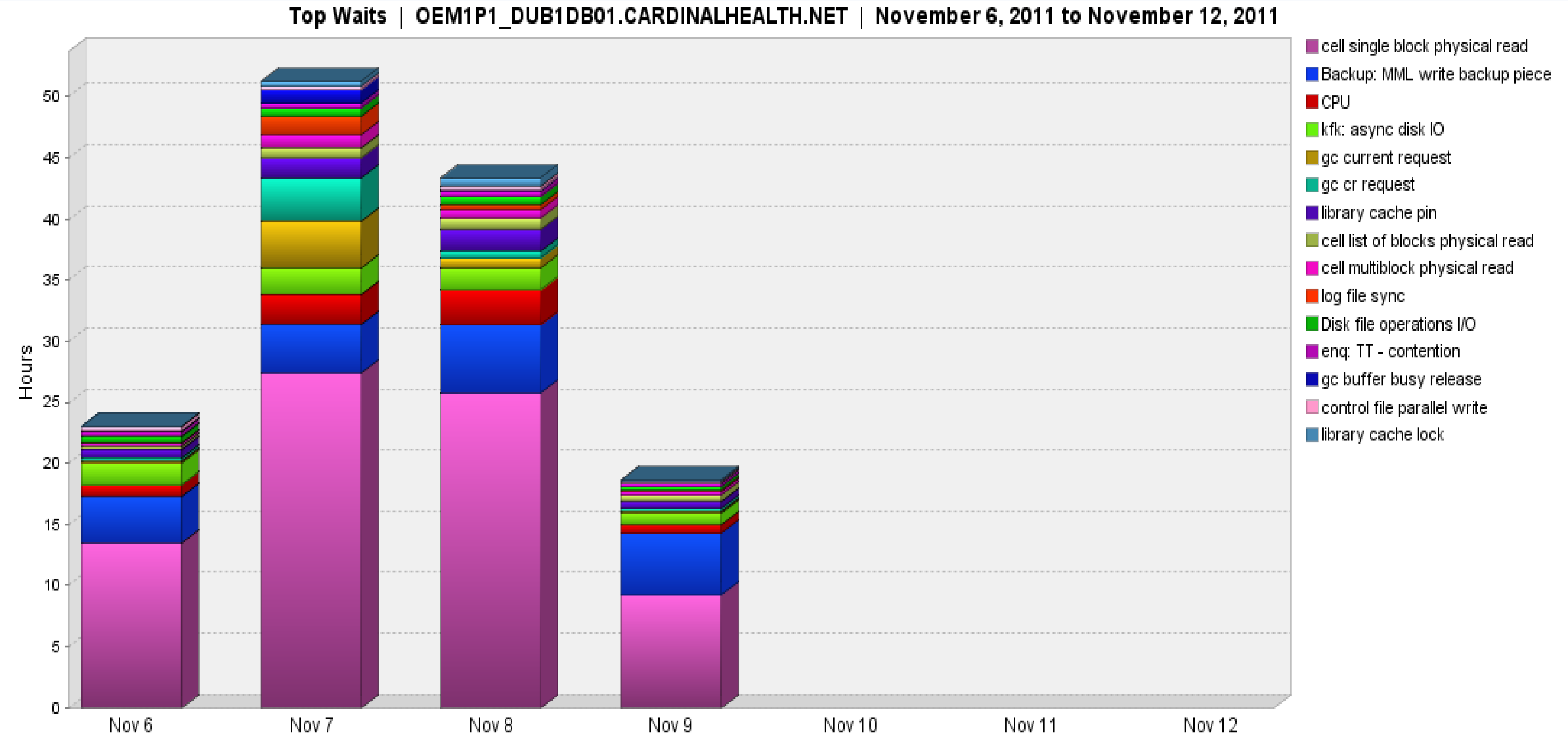
```
Select ‘cell single + multiblock reads’ c1, c2, c3, c5, c6
C6/decode(nvl(c2,0),0,1,c2) hit_ratio
From(
Select sum(total_waits) c2,
Avg(value) c6,
Sum(time_waited /100) c3
Avg((average_wait/100)*1000) c5
From v_$system_event, v$sysstat ss
Where event in (‘cell single block physical read’, ‘cell multiblock physical read’)
And name like ‘cell flash cache read hits’
And event not like ‘%Idle%’)
Order by c3
```

Exadata Performance

What to look for when tuning

- Sql statement response Time Monitoring
- DB Layer utilization and efficiency
- Storage Cell layer utilization and efficiency
- Advanced metrics and monitoring for Exadata
- Will be looking at
  - Sql statements v$views
  - Cellcli
  - OSWatcher

Exadata Performance



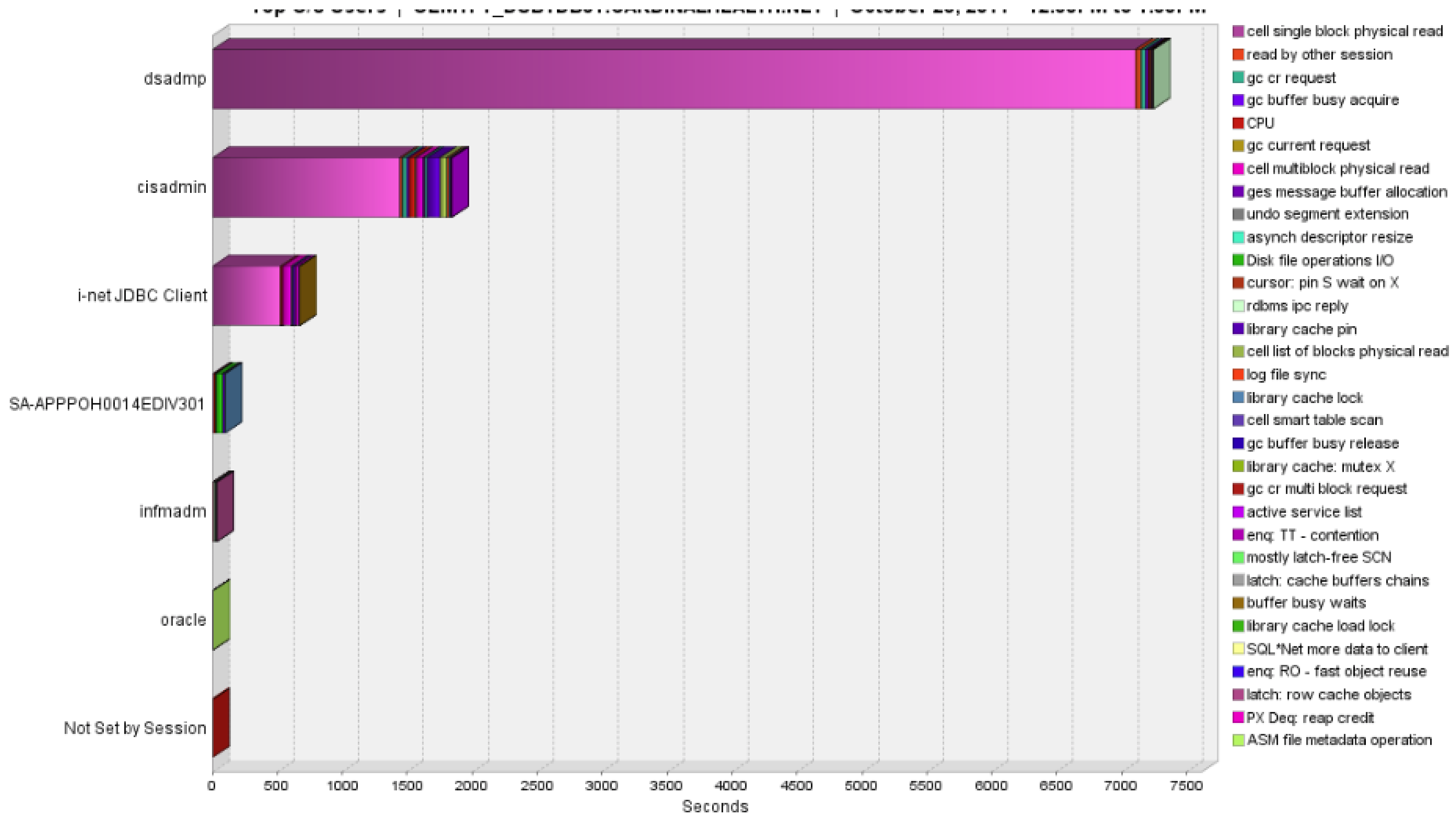
Exadata Performance - Exadata Wait Events

- Actually none that only exist on exadata, built into db code. No time is allocated on other platforms for Exadata specific
- Wont list all here.

```
Select event operation, count(*) from (
 select sql_id, event, sql_plan_operation||”||sql_plan_options operation
  from DBA_HIST_ACTIVE_SESS_HISTORY
  where event like ‘cell %’
  group by operation, event
  order by 1,2,3
```

- Cell smart table scan event - see offloading
- Cell smart index scan – fast full index scans that are offloaded

Exadata Performance



Exadata Causes Problems?

- This is not a reasoning not to get Exadata, just a note that Exadata is not the end all be all
- Oracle states to drop all indexes, if you follow this you could have problems.
- If you have a very finely tuned query especially OLTP you may not see any benefit on this query
- This is the exception but everyone I have talked to has that 1 query.

Conclusion

- At the root, Exadata houses an Oracle db.
- Always start with Cell Offloading
- Besides understanding roles, especially around patching, our biggest problems came from the upgrade to 11gr2
- Be aware, Exadata is not the be all end all, testing must be done and tuning must be done
- DBRM/IORM and Flash cache are your friends, don’t be afraid to use them even if you have had issues in the past

Confio Software

CONFIO SOFTWARE
Ignite IT Performance™

- Award Winning Performance Tools
- Ignite8 for Oracle, SQL Server, DB2, Sybase
- IgniteVM for Databases on VMware
  - Download at www.confio.com
- Provides Answers for
  - What changed recently that affected end users
  - What layer (VM or DB) is causing the problem
  - Who and How should we fix the problem

Download free trial at

www.confio.com