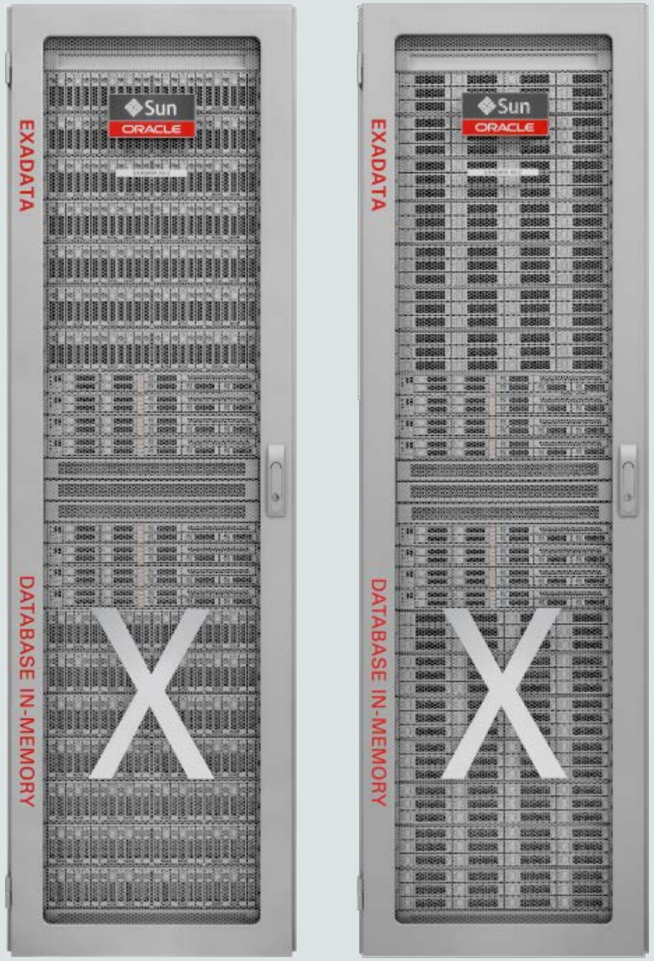
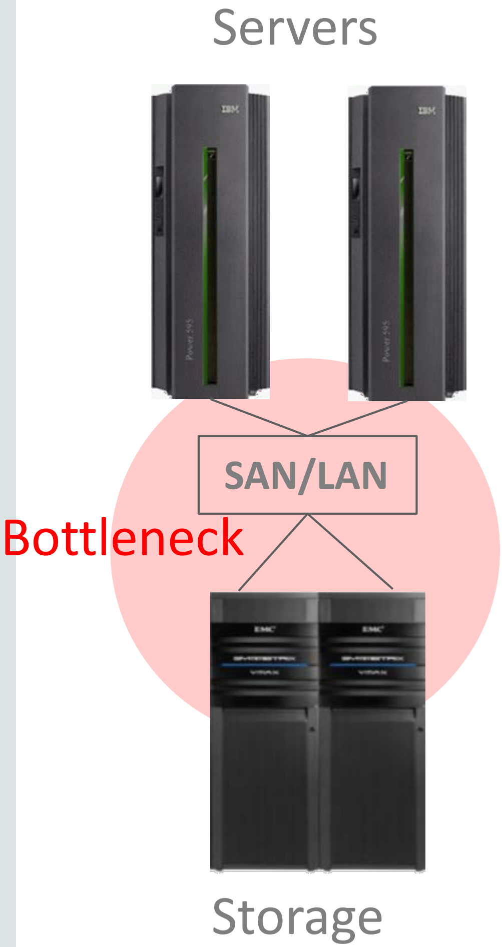
Databases Aware NVMe Flash: Pushing Application Performance



Gurmeet Goindi

Group Product Manger - Exadata, Oracle

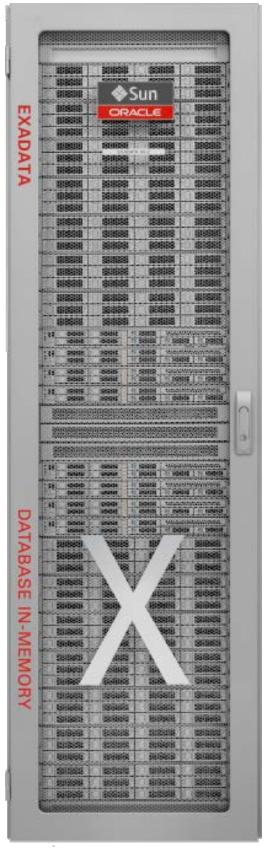
Traditional Database Deployment Issues



- Separation of servers and storage bottlenecks database performance
  - Flash produces data much faster than LANs and SANs can transport it
- Storage dominates the costs of database deployments and yet is limited to simple block serving
- Deployments are unique, complex
- Database runs on top of generic protocols and algorithms
  - Huge performance gains are squandered

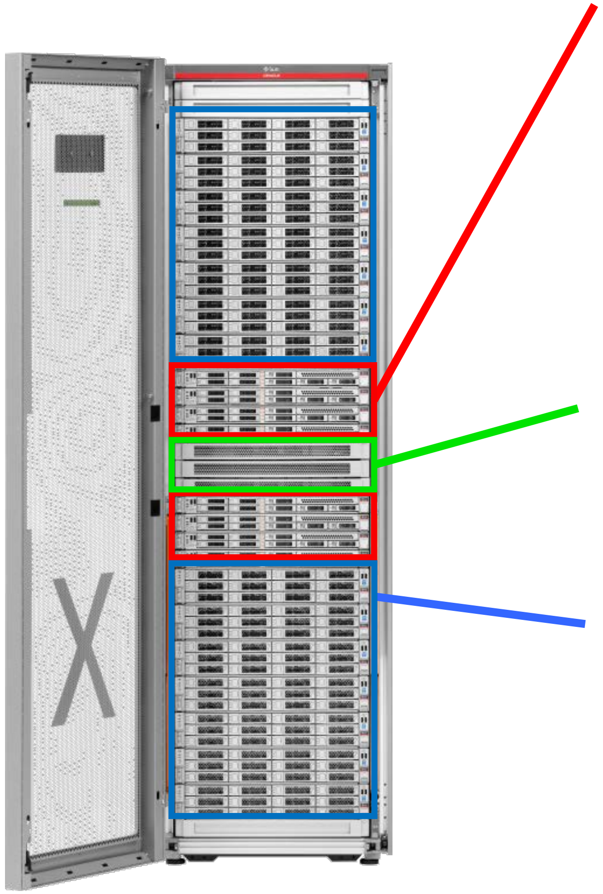
Oracle Exadata Database Machine

The Best Oracle Database Platform



- **Pre-Integrated Hardware and Software** – The latest hardware - sized, tuned and tested for **Oracle Database** workloads.
- **Unique Software and Protocols** – database, networking and storage software collaborate to power ***fastest*** and **most efficient Oracle Database** processing
- **End-to-End Support** – one integrated support team to **reduce complexity** and **lower operations costs.** All technologies owned and supported by Oracle

Exadata X5-2 Product Components



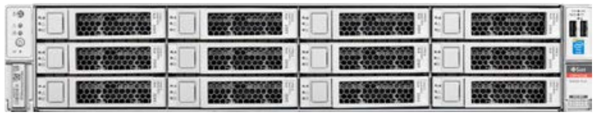
- **Scale-Out Database Servers**
  - **Two 18-core x86 Processors (36 cores)**
  - Oracle Linux 6
  - Oracle Database Enterprise Edition
  - Oracle VM (optional)
  - Oracle Database options (optional)
- **Fastest Internal Fabric**
  - **40 Gb/s InfiniBand**
  - Ethernet External Connectivity
- **Scale-Out Intelligent Storage**
  - **High-Capacity Storage Server**
  - **Extreme Flash Storage Server**
  - **Exadata Storage Server Software**

X5-2 Database Server

36 cores per server

256 – 768 GB DRAM

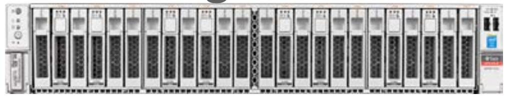
High-Capacity Storage Server



Extreme Flash Storage Server

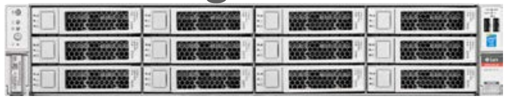
Exadata X5 Storage Servers

Extreme Flash Storage Server



All-Flash

High-Capacity Storage Server



Disk + Flash Cache

State-of-the-art NVMe PCIe flash
Consistently Low Response Times
Optimized InfiniBand I/O Protocols

Exadata Storage Server Software

Smart Scan (SQL Offload)
Smart Flash Cache
I/O Resource Management
Hybrid Columnar Compression

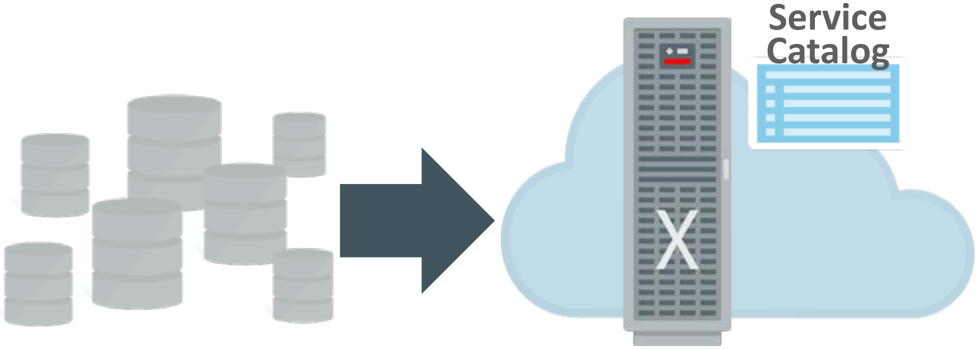
Performance	Extreme Flash	High-Capacity
Analytic Scans	263 GB/s	140 GB/s
OLTP Reads (8K)	4.14 M IOPS	4.14 M IOPS
OLTP Writes (8K)	4.14 M IOPS	2.69 M IOPS
Flash Latency	0.25 ms @ 2M IOPS	0.25 ms @ 1M IOPS

Capacity	Extreme Flash	High-Capacity
Cores (for SQL offload)	16	16
Disk (per server)	-	48 TB
Flash (per server)	12.8 TB	6.4 TB
Disk (full rack)*	-	672 TB
Flash (full rack)*	179.2 TB	89.6 TB

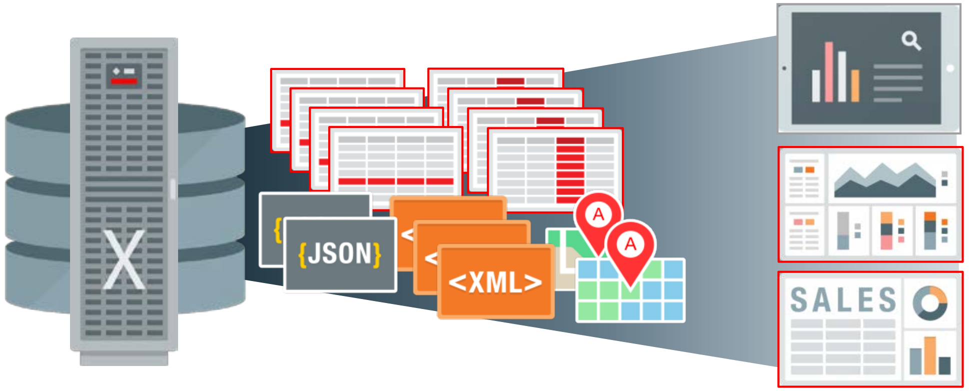
* Full Rack : 8 DB servers, 14 storage servers

Exadata Use Cases

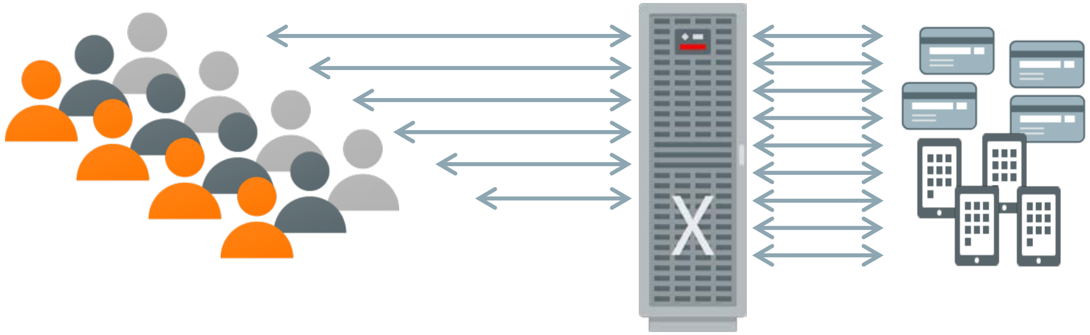
- DATABASE CONSOLIDATION / DBaaS



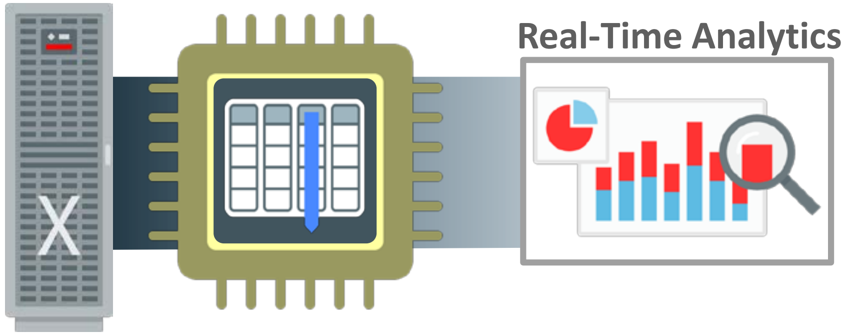
- DATA WAREHOUSING



- ONLINE TRANSACTION PROCESSING

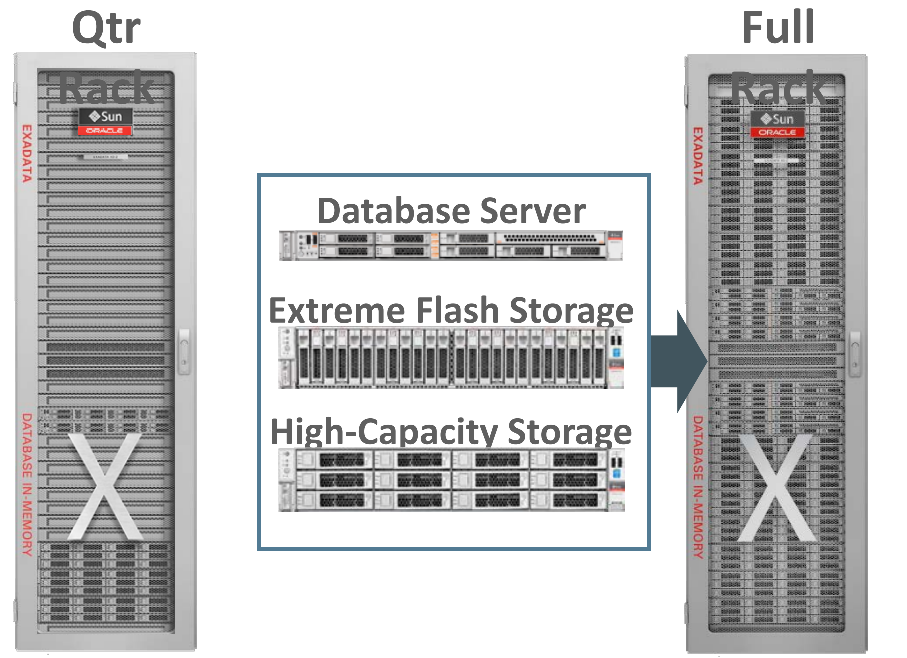


- IN-MEMORY DATABASE



Exadata Elastic Configurations

Optimize Exadata for any Workload



Start with

2 Database Servers

3 Storage Servers

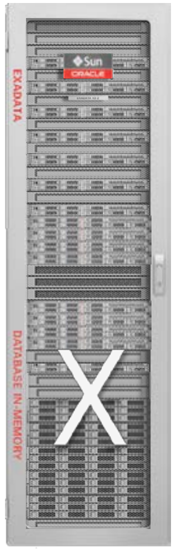
Add Servers

Any Kind

Any Quantity

Configuration Examples

DB **In-Memory** Machine



15 DB Servers
5 Storage Servers

576 DB Cores
13.3 TB RAM
192 TB Disk

Extreme Flash OLTP Machine

11 DB Servers
11 Storage Servers

396 DB Cores
8 TB RAM
140 TB Flash

Data Warehousing Machine

8 DB Servers
14 Storage Servers

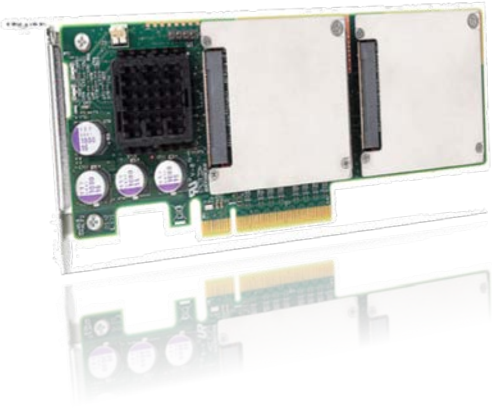
512 Cores
90 TB Flash Cache
672 TB Storage

ORACLE

Oracle’s Flash Architecture

- Scale out architecture
  - adds flash capacity and performance by adding storage servers
  - adds networking and CPU needed to process flash in one unit
- Database Aware Storage
  - Metadata about IO present on the cell
- Flash on the Storage Server enables sharing
  - A block on disk is stored in only one flash cache

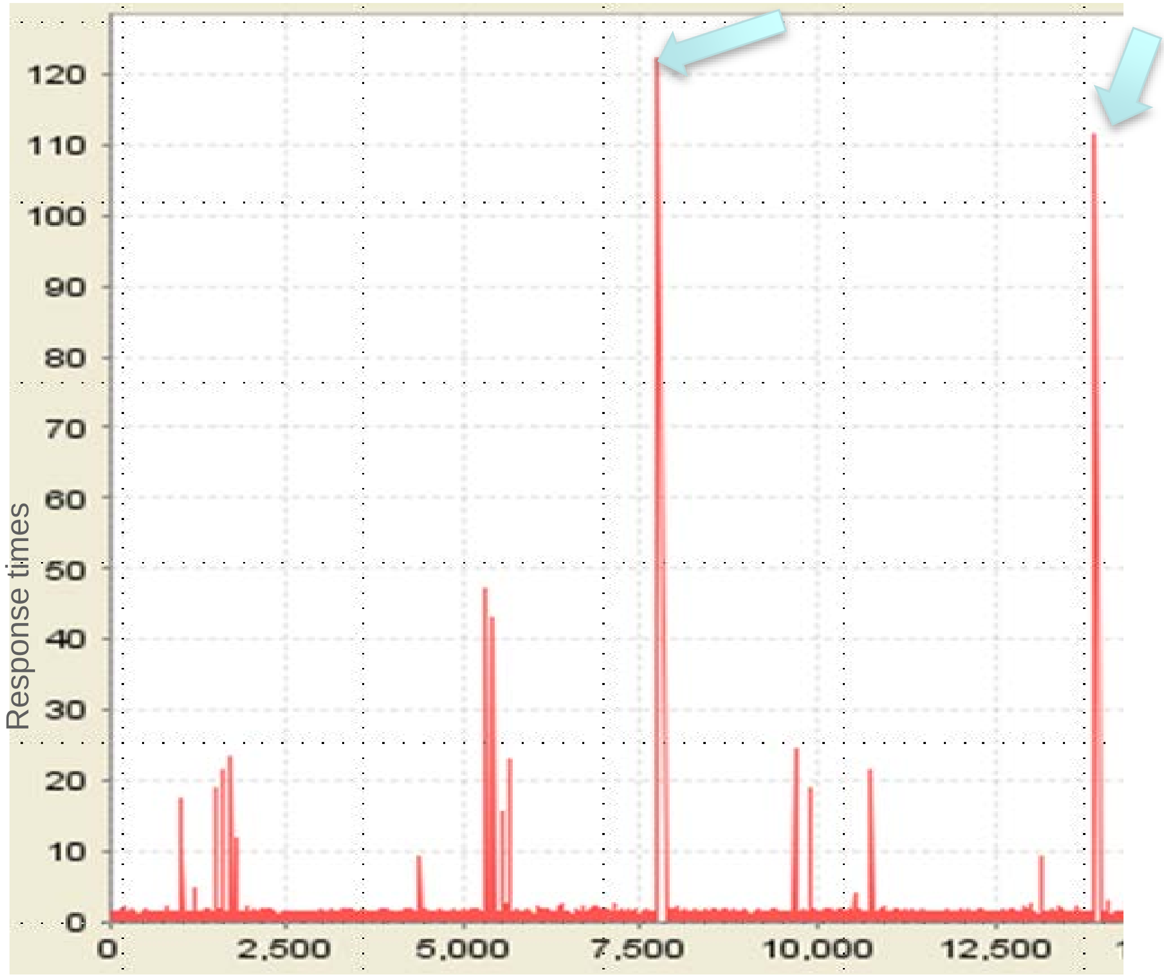
Exadata Smart Flash Cache



- Understands different types of I/Os from database
  - Skips caching I/Os to backups, data pump I/O, archive logs, tablespace formatting
  - Caches Control File Reads and Writes, file headers, data and index blocks
- Write-back flash cache
  - Caches writes from the database not just reads
- RAC-aware from day one

ORACLE

Flash And Database Logs

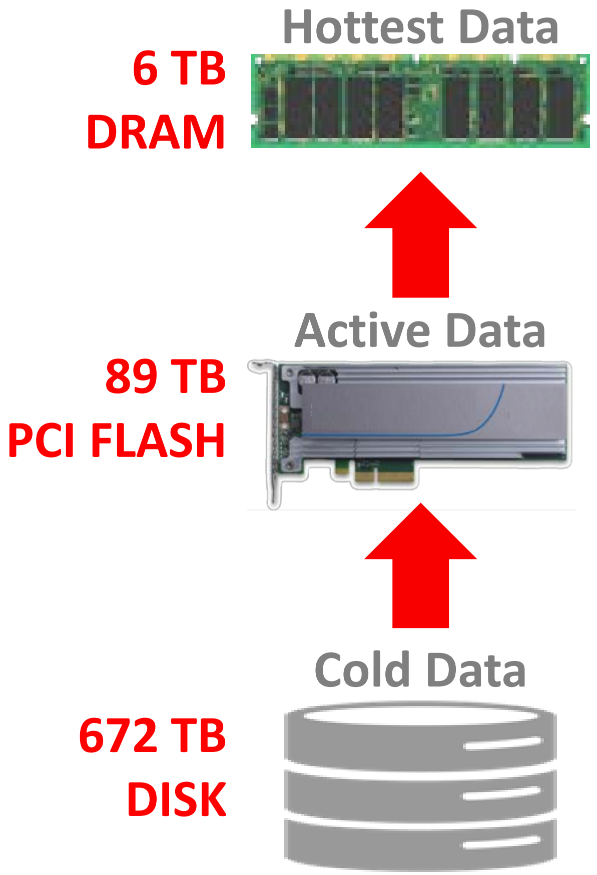


Number of transactions

- Flash has very good *average* write latency
- Greatly improves user transaction response time
- Flash occasional outliers, one or two orders of magnitude slower
- OLTP workloads dislike such large variations
- *Oracle's Approach:* Write to Flash and the DRAM cache in the disk controller simultaneously to even out the impact of outliers
  - the first to complete "wins" so that outliers are avoided (on either medium)

Most Cost Effective Database Storage

- **Exadata software transparently gives best of memory, flash, disk**
  - **Cost and Capacity** of SAS Disk Storage
  - **I/Os** of Scale-Out PCI Flash
  - **Speed** of In-Memory DB
- **Hybrid Columnar Compression (HCC)**
  - **Industry best data compression (10x average) for analytics & archive**
  - Data remains compressed in flash, memory, backups, standbys



Hottest Data

6 TB DRAM

Active Data

89 TB PCI FLASH

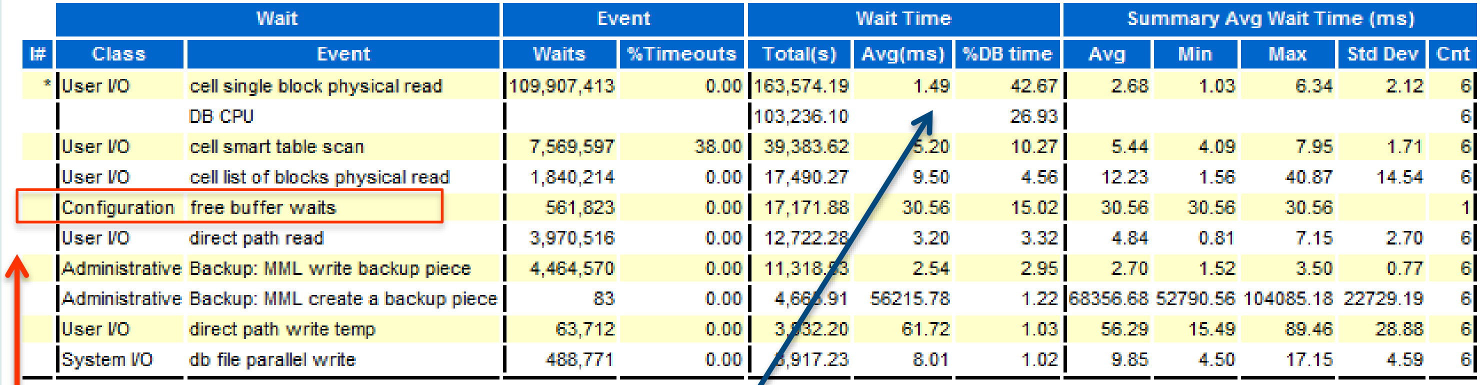
Cold Data

672 TB DISK

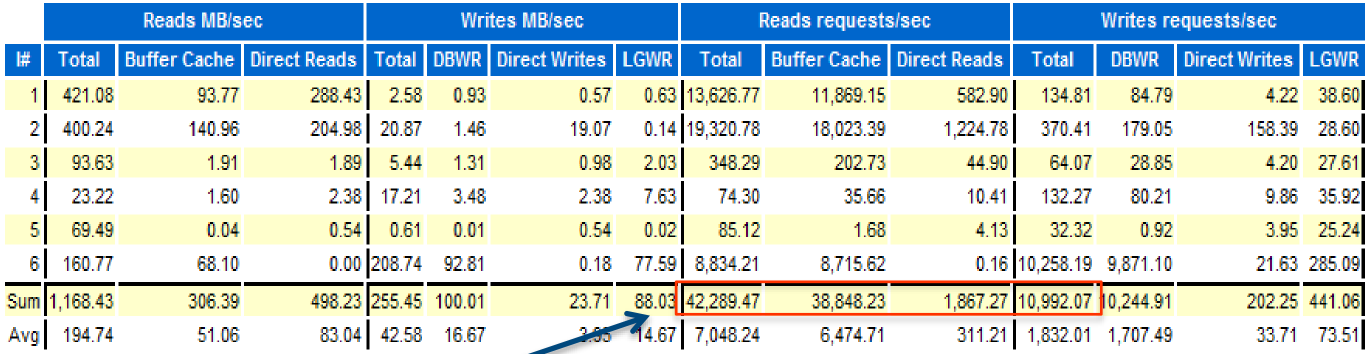
Per standard DB Machine full rack
8 DB, 14 HC storage servers

Customer Case Study

What Did We See - Exadata ODS



	Wait		Event		Wait Time			Summary Avg Wait Time (ms)				
I#	Class	Event	Waits	%Timeouts	Total(s)	Avg(ms)	%DB time	Avg	Min	Max	Std Dev	Cnt
*	User I/O	cell single block physical read	109,907,413	0.00	163,574.19	1.49	42.67	2.68	1.03	6.34	2.12	6
		DB CPU			103,236.10		26.93					6
	User I/O	cell smart table scan	7,569,597	38.00	39,383.62	5.20	10.27	5.44	4.09	7.95	1.71	6
	User I/O	cell list of blocks physical read	1,840,214	0.00	17,490.27	9.50	4.56	12.23	1.56	40.87	14.54	6
	Configuration	free buffer waits	561,823	0.00	17,171.88	30.56	15.02	30.56	30.56	30.56		1
	User I/O	direct path read	3,970,516	0.00	12,722.28	3.20	3.32	4.84	0.81	7.15	2.70	6
	Administrative	Backup: MML write backup piece	4,464,570	0.00	11,318.53	2.54	2.95	2.70	1.52	3.50	0.77	6
	Administrative	Backup: MML create a backup piece	83	0.00	4,665.91	56215.78	1.22	68356.68	52790.56	104085.18	22729.19	6
	User I/O	direct path write temp	63,712	0.00	3,932.20	61.72	1.03	56.29	15.49	89.46	28.88	6
	System I/O	db file parallel write	488,771	0.00	3,917.23	8.01	1.02	9.85	4.50	17.15	4.59	6



	Reads MB/sec			Writes MB/sec				Reads requests/sec			Writes requests/sec			
I#	Total	Buffer Cache	Direct Reads	Total	DBWR	Direct Writes	LGWR	Total	Buffer Cache	Direct Reads	Total	DBWR	Direct Writes	LGWR
1	421.08	93.77	288.43	2.58	0.93	0.57	0.63	13,626.77	11,869.15	582.90	134.81	84.79	4.22	38.60
2	400.24	140.96	204.98	20.87	1.46	19.07	0.14	19,320.78	18,023.39	1,224.78	370.41	179.05	158.39	28.60
3	93.63	1.91	1.89	5.44	1.31	0.98	2.03	348.29	202.73	44.90	64.07	28.85	4.20	27.61
4	23.22	1.60	2.38	17.21	3.48	2.38	7.63	74.30	35.66	10.41	132.27	80.21	9.86	35.92
5	69.49	0.04	0.54	0.61	0.01	0.54	0.02	85.12	1.68	4.13	32.32	0.92	3.95	25.24
6	160.77	68.10	0.00	208.74	92.81	0.18	77.59	8,834.21	8,715.62	0.16	10,258.19	9,871.10	21.63	285.09
Sum	1,168.43	306.39	498.23	255.45	100.01	23.71	88.03	42,289.47	38,848.23	1,867.27	10,992.07	10,244.91	202.25	441.06
Avg	194.74	51.06	83.04	42.58	16.67	3.95	14.67	7,048.24	6,474.71	311.21	1,832.01	1,707.49	33.71	73.51

What? Writes are supposed to be fast! Wait until later slides.

1.49 ms single block reads

While doing 42K read IOPS and 11K write iops over an hour period.

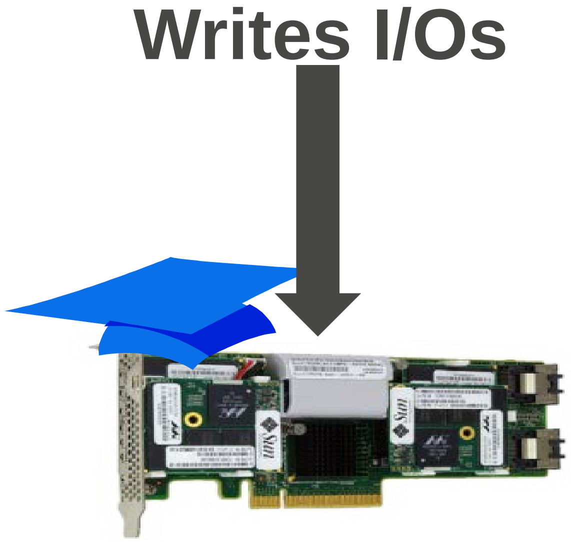
Note: The other databases were active on the Exadata System during this time.

Comparison to Old system

Metric	Exadata ODS	Monolithic Hardware ODS	Comparison
Single Block Reads	1.5 ms	3.8 ms	> 2x
Log File Synch Waits	.85 ms	5.7 ms	> 6x

Note: The Exadata ODS is over twice the workload as the previous version. In addition, the Exadata system is shared with several databases, while the Monolithic Hardware was dedicated.

Write Back Flash Enablement

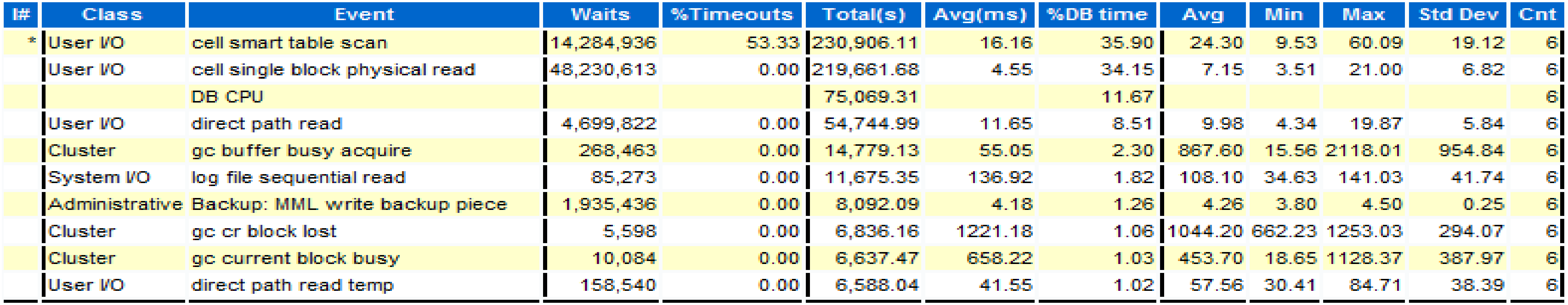


Design to accelerate write intensive workloads.

From previous slide, we had lots of “free buffer waits”.

Enabled this feature on X2-2.

Result: No more “free buffer waits”.



I#	Class	Event	Waits	%Timeouts	Total(s)	Avg(ms)	%DB time	Avg	Min	Max	Std Dev	Cnt
*	User I/O	cell smart table scan	14,284,936	53.33	230,906.11	16.16	35.90	24.30	9.53	60.09	19.12	6
	User I/O	cell single block physical read	48,230,613	0.00	219,661.68	4.55	34.15	7.15	3.51	21.00	6.82	6
		DB CPU			75,069.31		11.67					6
	User I/O	direct path read	4,699,822	0.00	54,744.99	11.65	8.51	9.98	4.34	19.87	5.84	6
	Cluster	gc buffer busy acquire	268,463	0.00	14,779.13	55.05	2.30	867.60	15.56	2118.01	954.84	6
	System I/O	log file sequential read	85,273	0.00	11,675.35	136.92	1.82	108.10	34.63	141.03	41.74	6
	Administrative	Backup: MML write backup piece	1,935,436	0.00	8,092.09	4.18	1.26	4.26	3.80	4.50	0.25	6
	Cluster	gc cr block lost	5,598	0.00	6,836.16	1221.18	1.06	1044.20	662.23	1253.03	294.07	6
	Cluster	gc current block busy	10,084	0.00	6,637.47	658.22	1.03	453.70	18.65	1128.37	387.97	6
	User I/O	direct path read temp	158,540	0.00	6,588.04	41.55	1.02	57.56	30.41	84.71	38.39	6

What This Means to Us

More Flexibility in System Use

- We are less concern about unplanned activities on the system. The users can go after the system when they need to, not during certain windows.
- Maintenance activities have less impact on system availability.

More Use of the Data

- Exadata's Flash reduces the i/o contention of the mixed workloads within the database and between competing databases
- More concurrent users mean more business questions being answered.

Faster Access to the Data

- Faster I/O means less time waiting for queries to return, more time to analyze the results

Integrated Cloud

Applications & Platform Services

ORACLE®