
ORACLE®

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Dramatic Reduction in Oracle Database I/O Wait Times with Oracle FS Flash Storage Systems

Simon Towers
Architect
Flash Storage Systems Group
Oracle Corporation

All Flash FS | All-Flash Performance, Simplicity, Security

ORACLE
FLASH STORAGE SYSTEMS

Flash Performance

- Scales to 912 TB Flash, Up to 360K IOPS

QoS: Storage Quality of Service

- Allocation of storage resources according to business priority

Engineered for Oracle Database and Apps.

- Engineered to leverage Oracle's Hybrid Columnar Compression (HCC) , Automatic Storage Manager (ASM), Automatic Data Optimization (ADO), Oracle Linux, Solaris, Cloud Services, and Fusion Application Profiles

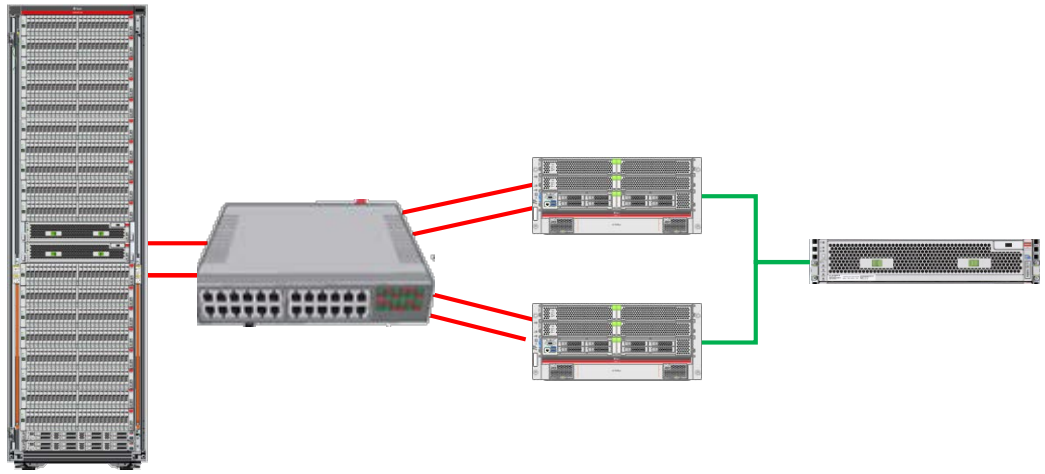
Enterprise-Grade

- < 1 second failover, Warmstart technology, No SPOF, Pre-emptive Copy, SSD gauges, T10-PI, Replication and Copy Services, Ships fully tested and racked, Business Critical Service for Systems

ORACLE

Hardware

Physical Setup



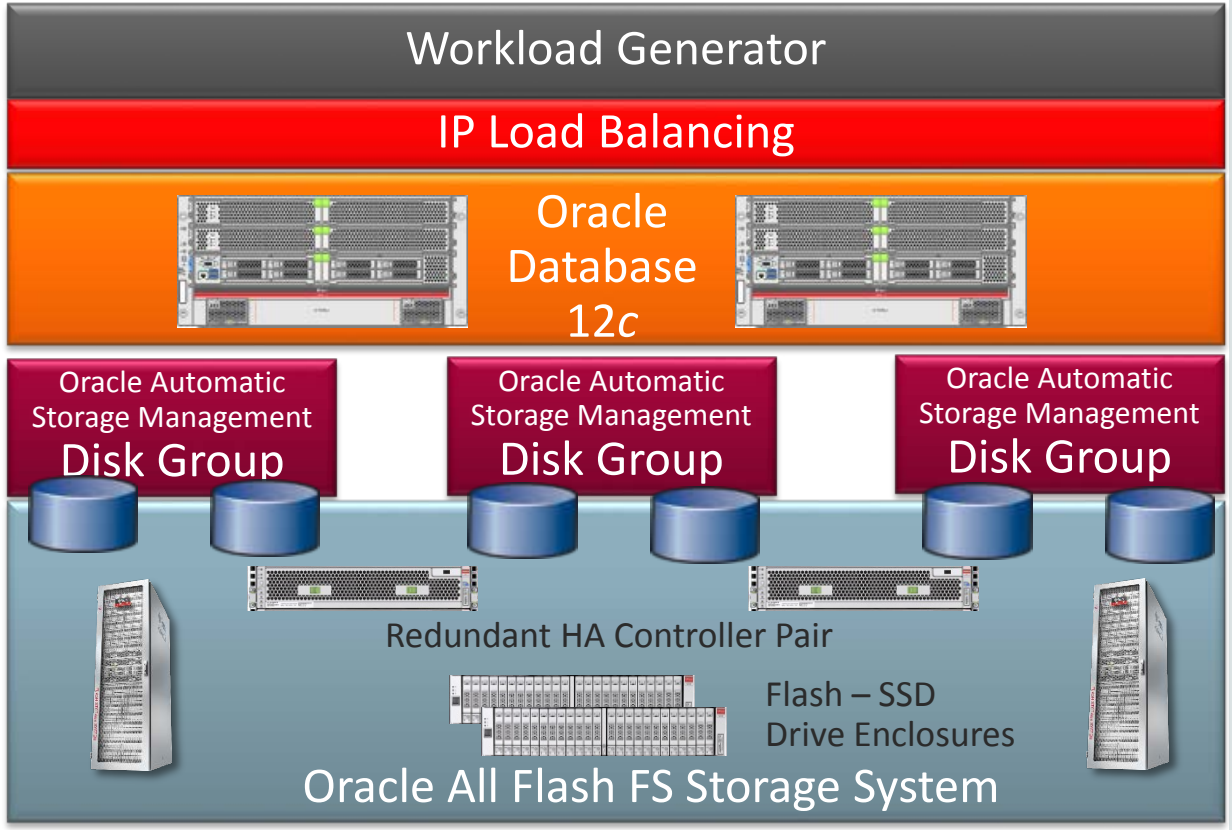
Oracle All Flash FS or FS1-2

16 Gb/sec FC Switch

Sun Server X5-2

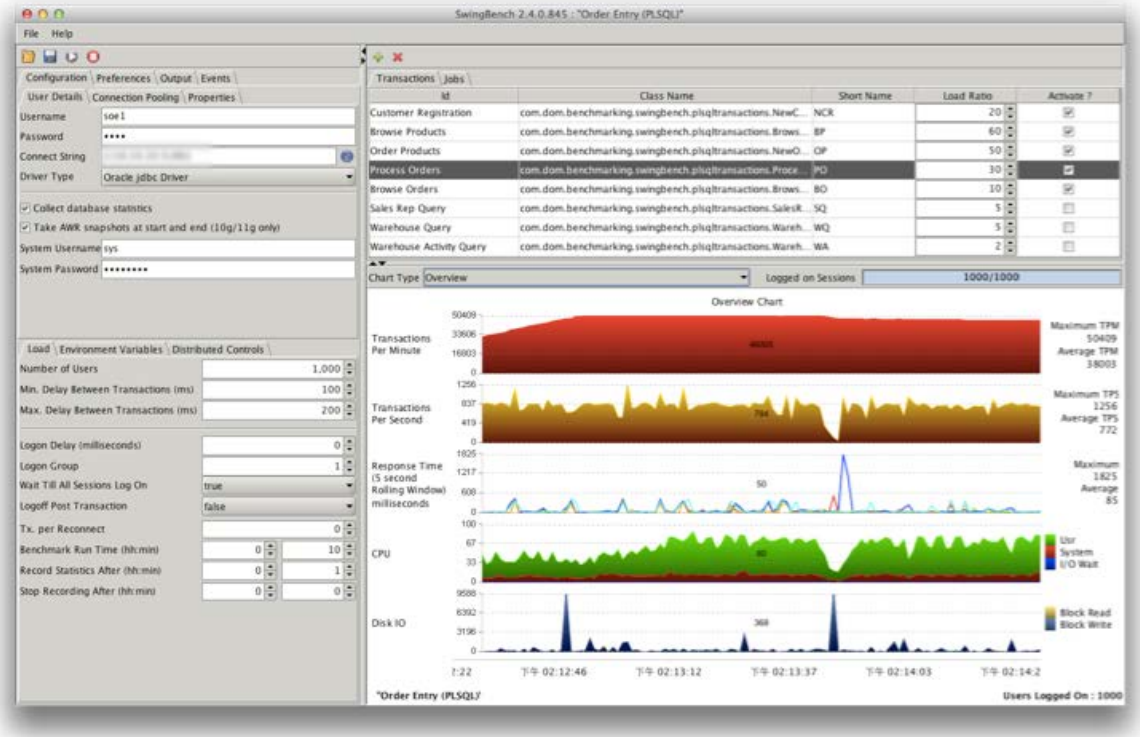
Workload Generator Server

Logical Setup



Software: Swingbench

Load Generator:	• Designed to stress test Oracle Databases
Consists of:	• Load generator • Coordinator • Cluster Overview
Includes 4 Benchmarks:	• OrderEntry • SalesHistory • CallingCircle • StressTest



Transaction Processing Example

Swingbench Order Entry (PLSQL) V2 workload

Varied from 100 to 3000 simulated users

9.4%
Customer registration

6.3%
Update customer details

31.3%
Browse products

25.0%
Order products

25.0%
Process orders

3.1%
Browse orders

Scenario 1 - All HDD Configuration

2-node RAC Cluster

- Oracle X4 servers:
  - 128 GB memory
  - 2 CPU x 12 cores
  - 2-port QLE8362 16 Gbit FC HBA
  - 2-port QDR Infiniband HCA for RAC interconnect

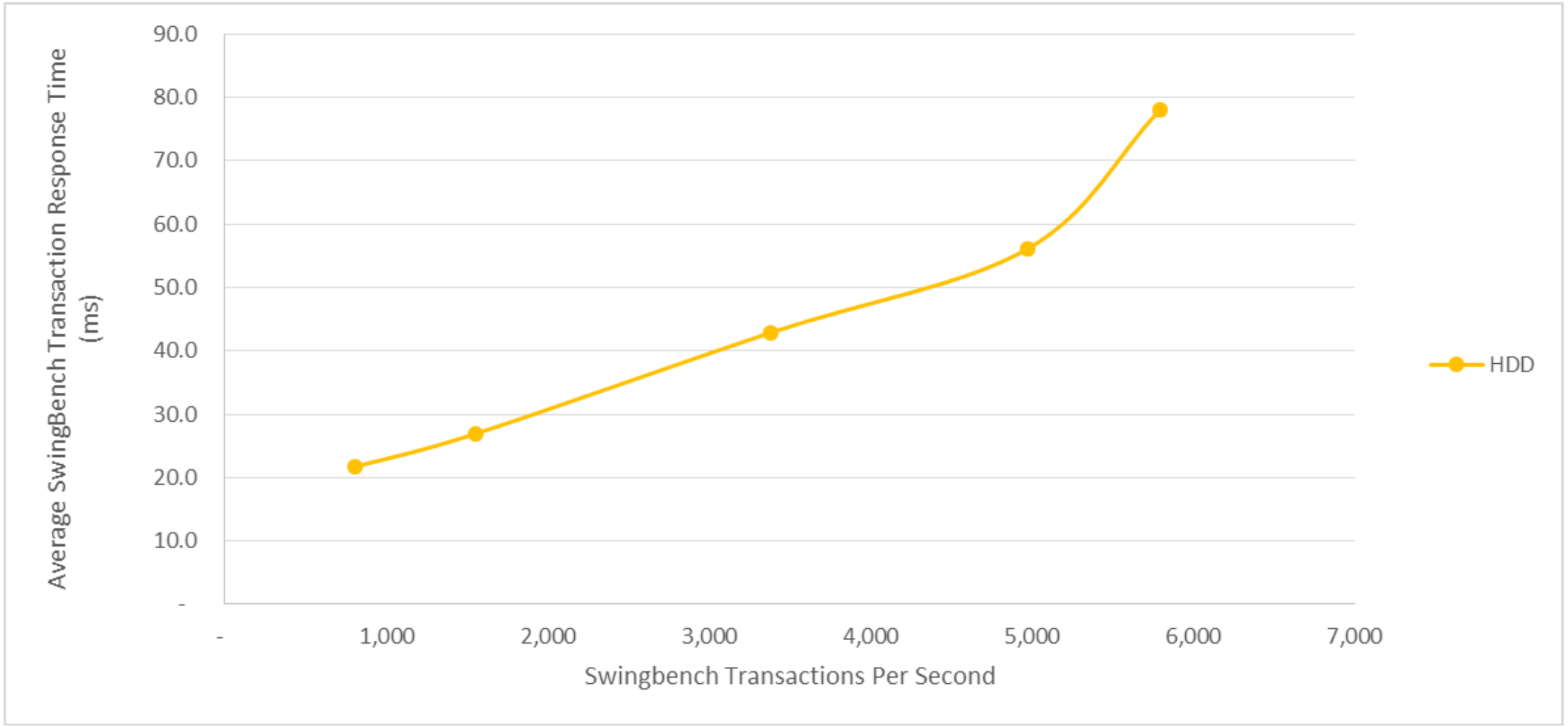
Oracle Database 12c

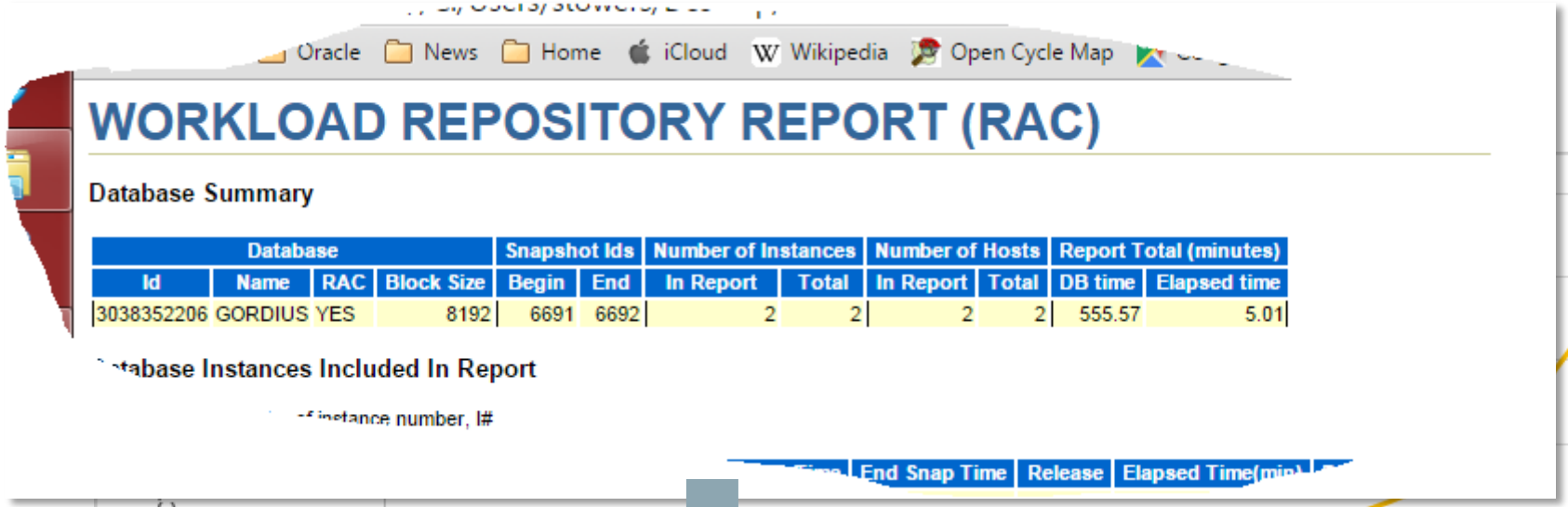
- 100 GB Aggregate SGA
- 2 TB database (much larger than the SGA)

HDD Storage

- 3 x 24-HDD enclosures
- 300 GB 10K RPM SAS HDDs
- Database, indexes, redo logs striped over all the drives

Scenario 1 Results





Scenario 1: AWR Wait Events

Top Timed Events

- Instance '*' - cluster wide summary
- '*' Waits, %Timeouts, Wait Time Total(s) : Cluster-wide total for the wait event
- '*' 'Wait Time Avg (ms)' : Cluster-wide average computed as (Wait Time Total / Event Waits) in ms
- '*' Summary 'Avg Wait Time (ms)' : Per-instance 'Wait Time Avg (ms)' used to compute the following statistics
- '*' [Avg/Min/Max/Std Dev] : average/minimum/maximum/standard deviation of per-instance 'Wait Time Avg(ms)'
- '*' Cnt : count of instances with wait times for the event

	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	db file sequential read	4,397,286	0.00	38,613.77	8.78	115.84	8.81	8.34	9.28	0.67	2
		DB CPU			2,789.55		8.37					2
	Commit	log file sync	994,776	0.00	825.16	0.83	2.48	0.83	0.82	0.84	0.01	2
	System I/O	log file parallel write	767,610	0.00	380.38	0.50	1.14	0.50	0.49	0.51	0.02	2
	Cluster	gc current block 2-way	1,798,267	0.00	218.25	0.12	0.65	0.12	0.12	0.12	0.00	2
	Cluster	gc cr block 2-way	1,512,667	0.00	175.82	0.12	0.53	0.12	0.11	0.12	0.00	2
	Other	target log write size	343,167	0.00	172.07	0.50	0.52	0.50	0.50	0.50		2
	Cluster	gc current block busy	22,746	0.00	135.80	5.97	0.41	5.98	5.87	6.08	0.15	2
	Cluster	gc cr grant 2-way	961,064	0.00	108.62	0.11	0.33	0.11	0.11	0.12	0.01	2
	User I/O	read by other session	5,478	0.00	92.87	16.95	0.28	16.96	16.79	17.13	0.24	2

Top 3 wait events are I/O related

Log file sync & log file parallel write are good (< 1ms)

But db file sequential read is terrible – it's the bottleneck by a large margin

Let's try this on SSDs!

Scenario 1 - All HDD Configuration

2-node RAC Cluster

- Oracle X4 servers:
  - 128 GB memory
  - 2 CPU x 12 cores
  - 2-port QLE8362 16 Gbit FC HBA
  - 2-port QDR Infiniband HCA for RAC interconnect

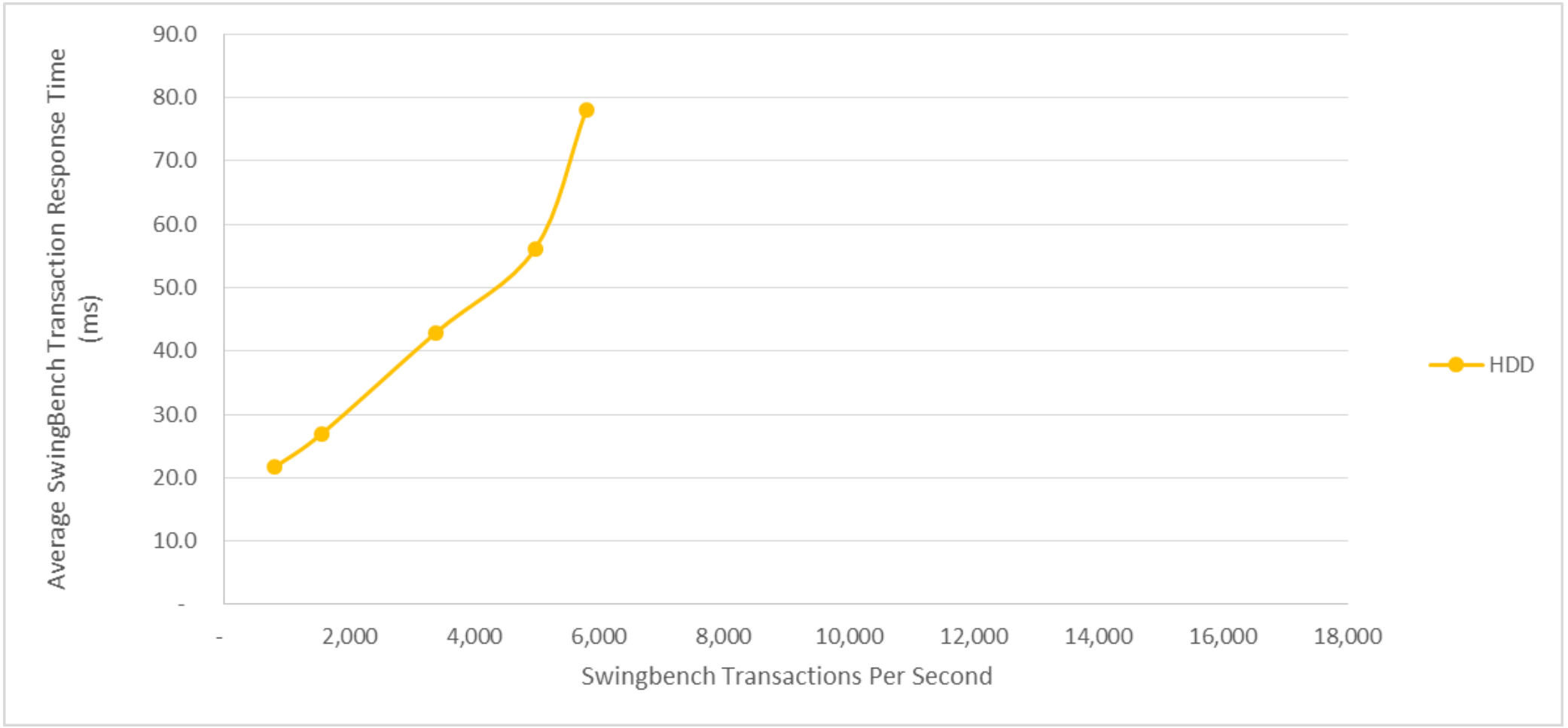
Oracle Database 12c

- 100 GB Aggregate SGA
- 2 TB database (much larger than the SGA)

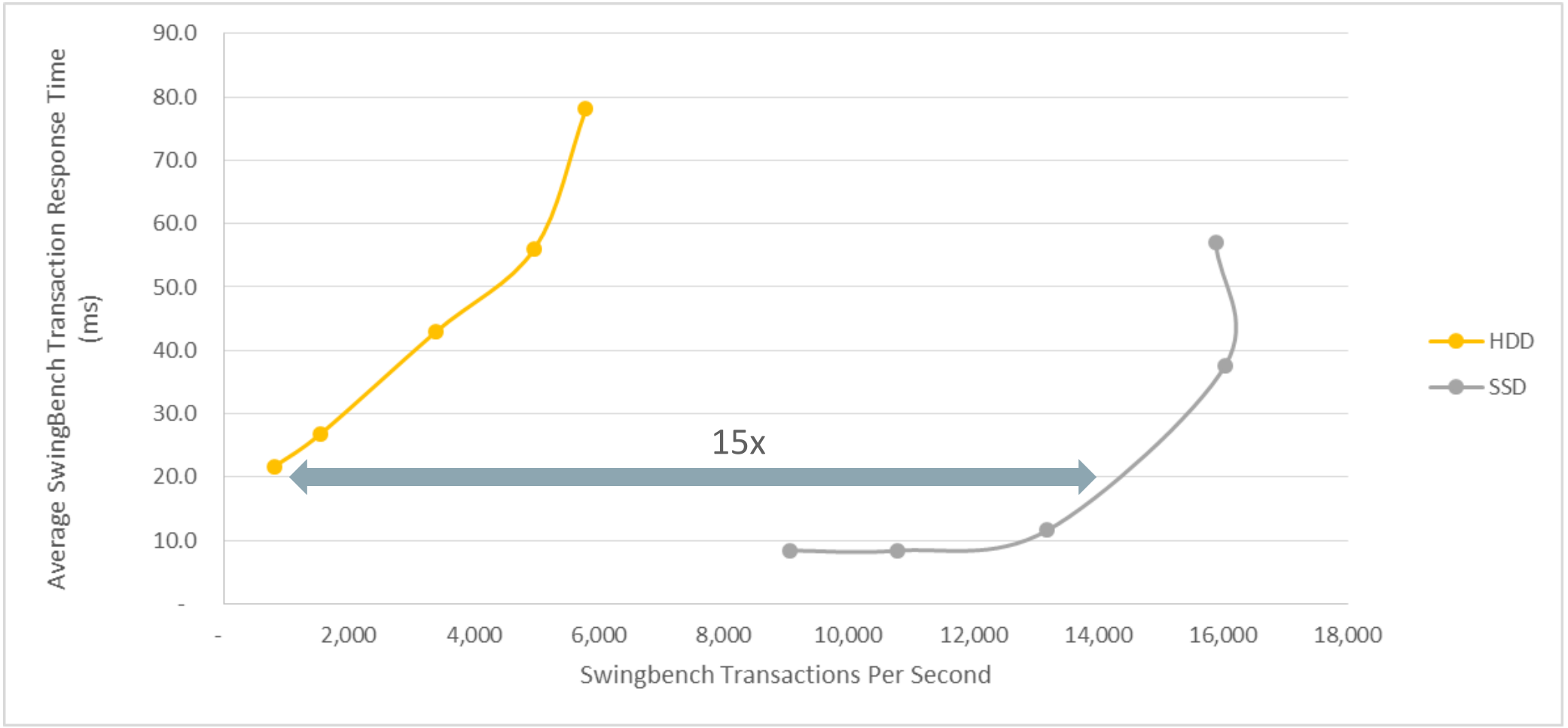
SSD Storage

- 3 x 13-SSD enclosures
- 400 GB SAS-attached SSDs
- Database, indexes, redo logs striped over all the drives

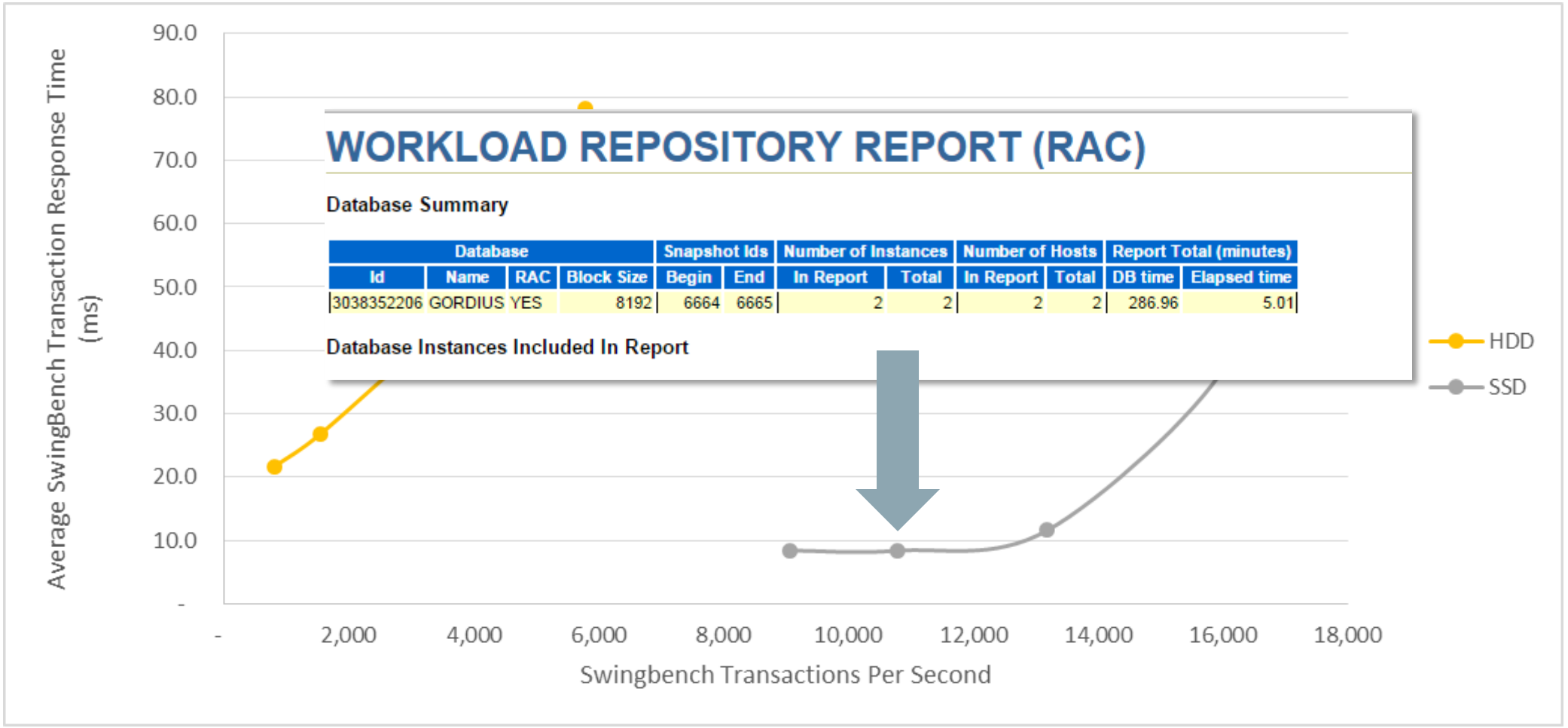
Scenario 2 Results



Scenario 2 Results



Scenario 2 Results



WORKLOAD REPOSITORY REPORT (RAC)

Database Summary

Database				Snapshot Ids		Number of Instances		Number of Hosts		Report Total (minutes)	
Id	Name	RAC	Block Size	Begin	End	In Report	Total	In Report	Total	DB time	Elapsed time
3038352206	GORDIUS	YES	8192	6664	6665	2	2	2	2	286.96	5.01

Database Instances Included In Report

Scenario 2: AWR Wait Events

Top Timed Events

- Instance '*' - cluster wide summary
- '*' Waits, %Timeouts, Wait Time Total(s) : Cluster-wide total for the wait event
- '*' 'Wait Time Avg (ms)' : Cluster-wide average computed as (Wait Time Total / Event Waits) in ms
- '*' Summary 'Avg Wait Time (ms)' : Per-instance 'Wait Time Avg (ms)' used to compute the following statistics
- '*' [Avg/Min/Max/Std Dev] : average/minimum/maximum/standard deviation of per-instance 'Wait Time Avg(ms)'
- '*' Cnt : count of instances with wait times for the event

	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
*		DB CPU			9,802.05		56.93					2
	User I/O	db file sequential read	11,607,406	0.00	6,757.51	0.58	39.25	0.58	0.58	0.59	0.01	2
	Commit	log file sync	3,223,931	0.00	3,283.35	1.02	19.07	1.02	1.01	1.03	0.02	2
	Cluster	gc current block 2-way	5,397,126	0.00	1,524.55	0.28	8.85	0.28	0.25	0.32	0.05	2
	Cluster	gc cr block 2-way	5,111,598	0.00	1,449.93	0.28	8.42	0.28	0.25	0.32	0.05	2
	Cluster	gc current grant busy	985,119	0.00	583.36	0.59	3.39	0.68	0.46	0.90	0.31	2
	Concurrency	library cache: mutex X	289,497	0.00	579.23	2.00	3.36	1.99	1.78	2.20	0.30	2
	System I/O	log file parallel write	948,149	0.00	483.34	0.51	2.81	0.51	0.51	0.51	0.00	2
	Cluster	gc current grant 2-way	1,594,965	0.00	424.88	0.27	2.47	0.27	0.23	0.30	0.05	2
	Other	Failed Logon Delay	373	100.00	373.03	1000.07	2.17	1000.07	1000.07	1000.07	0.00	2

Substantial improvement in db file sequential read

From: 8.81ms
To: 0.58ms

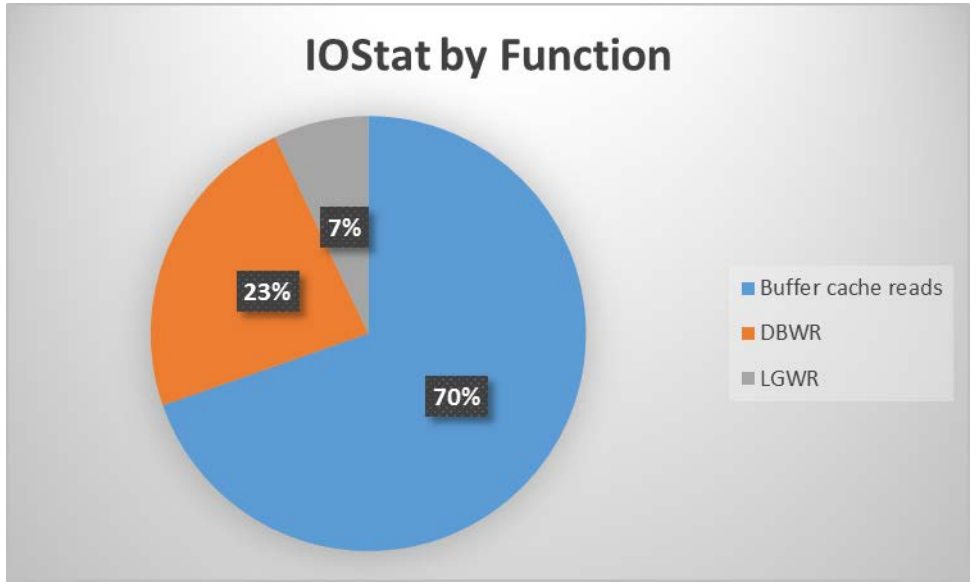
Putting your Oracle Database on SSD radically improves transaction throughput and response time

More AWR Statistics

IOStat by Function (per Second)

- Total Reads includes all Functions: Buffer Cache, Direct Reads, ARCH, Data Pump, Others, RMAN, Recovery, Streams/AQ and XDB
- Total Writes includes all Functions: DBWR, Direct Writes, LGWR, ARCH, Data Pump, Others, RMAN, Recovery, Streams/AQ and XDB

	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	258.86	258.76	0.00	141.98	112.82	0.01	29.14	33,118.28	33,110.27	0.00	13,226.78	10,876.97	0.03	2,349.30
2	143.87	143.76	0.00	80.44	64.24	0.01	16.19	18,400.27	18,393.08	0.00	9,232.20	6,363.99	0.04	2,867.66
Sum	402.73	402.52	0.00	222.42	177.06	0.02	45.33	51,518.56	51,503.35	0.00	22,458.98	17,240.96	0.07	5,216.95
Avg	201.37	201.26	0.00	111.21	88.53	0.01	22.66	25,759.28	25,751.67	0.00	11,229.49	8,620.48	0.03	2,608.48



Data & Index read/write are 93% of IOPS

Log Writer:
- Small fraction of IOPS
- Sequential writes
- Array write cache

⇒ HDD on array are just fine for redo logs

Putting ALL database files on SSD is best

But if you only have limited SSD, put Data & Index on SSD first

Order Entry on SSD Wrap-Up

Running the order entry application on SSD takes I/O time out of the picture

As a DBA, now you can focus purely on database design issues

Storage is taken care of!

Other Reasons to Love SSDs

High density disk drives have become complex

- Alternate track interference refresh
- Bad sector scrubbing
- Very long data recovery scenarios
- These issues cause unpredictable and long I/O delays
- **Flash drives don’t have these issues**

Additional demands on drive performance

- Rebuild to recover from failed drives
- Zeroing after LUN deletion
- **Having the performance headroom in flash reduces the impact of these activities**

Summary

- Oracle Database performance comparison between HDD and SSD
- SSD solution scales to 15X the throughput
- SSD solution response time > 10X faster
- With SSD, your performance is all in the database. *I/O is taken care of*

Storage Tiering with Oracle 12c ADO and Oracle FS Flash Storage

Thursday, April 14th @ 8:30am

Session ID: 1248

Extraordinary Data Warehousing on Oracle SuperCluster with Oracle FS1 Flash Storage System

Thursday, April 14th @ 11am

Session ID: 4739

ORACLE

Database Protection Options from the Experts

Wednesday, April 13th @ 3pm

Session ID: 4749

Zero Data Loss Recovery Appliance Deep Dive: Direct from Development

Thursday, April 14th @ 11am

Session ID: 4105

ORACLE

Integrated Cloud

Applications & Platform Services


ORACLE®