Understanding Storage's Impact on Oracle Performance

Jamon Bowen
Texas Memory Systems

Why Understanding Hardware is Important for a DBA

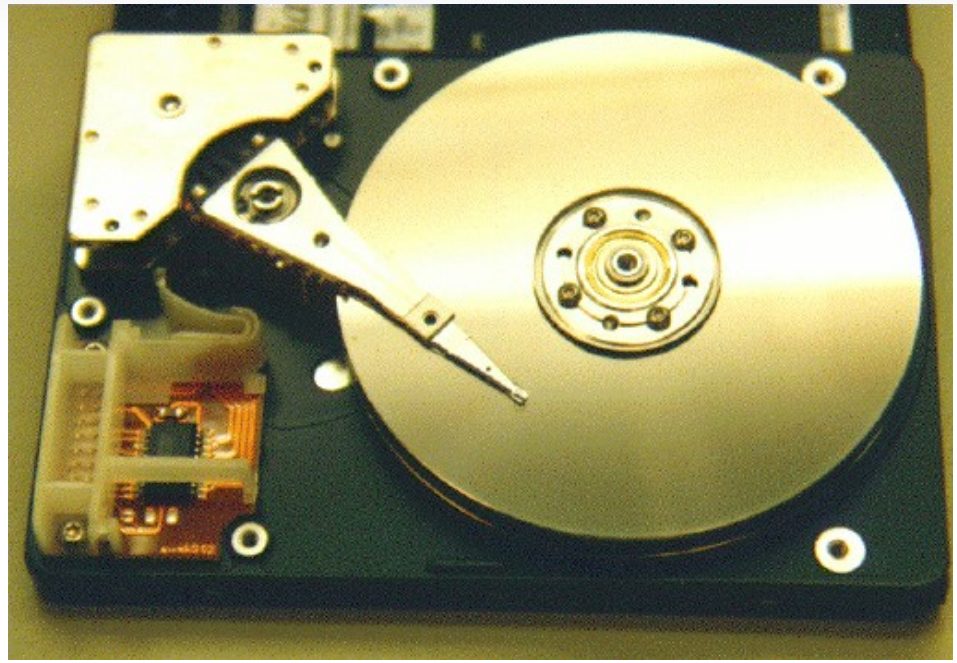
- One a of main features of a production system is its performance.
  - Query Tuning or Hardware selection are regularly used to address performance issues.
- Query Tuning is much like selecting the route for a cross country trip.
  - In this analogy Hardware selection is the same as selecting the mode of transport.
  - Both can have dramatic impacts on performance and work best when they are combined.
- Understanding Hardware is key to requesting the best equipment for an application.

Terms to Understand

- Capacity – amount of space used for data storage
- Bandwidth – How much data is passed through at a given time
- IOPS – How many reads and writes occurred on the storage at a given time (Inputs & Outputs Per Sec.)
- Latency – How quick each command finished. Often referred to as response time.

Single Disk Data

- Bandwidth –
  - Depends of the drive capacity, RPM, and interface.
  - 15K RPM drives support from 40-80 MB/s
- Latency -
  - Depends on the required side to side movement of the disk head and the RPM of the drive.
  - 15K RPM drives with true Random IO support 3-10 MS latency
- IOPS
  - Depends on the latency
  - 15K RPM drives under Random IO support 100-300 IOPS



How do disks Arrays increase performance?

- Massive Arrays of Disks:

- This can ensure that access time doesn't degrade below the 5 - 10 ms Access time, and that parallel operations can be handled.
- Can increase IOPS, Bandwidth, does not impact latency.

TMS
TEXAS MEMORY SYSTEMS

Using Statspack/ AWR Reports To identify IO Bottlenecks.

Statspack / AWR

Statspack is a comparison of various counters Oracle tracks over a set interval to gauge how much time was spent on various things in relation to on another. It also gives a snapshot of configuration parameters. The next few slides step through IO related sections of a single statspack

```
STATSPACK report for

DB Name         DB Id    Instance     Inst Num Release     Cluster Host
------------ ----------- ------------ -------- ----------- ------- ------------
MTR           3056795493 MTR                 1 9.2.0.6.0   NO      XXXXX

             Snap Id     Snap Time      Sessions Curs/Sess Comment
             ------- ------------------ -------- --------- -------------------
Begin Snap:   25867 13-Dec-06 11:45:01       31        .9
  End Snap:   25868 13-Dec-06 12:00:01      127       7.5
   Elapsed:               15.00 (mins)

Cache Sizes (end)
~~~~~~~~~~~~~~~~~
               Buffer Cache:     7,168M      Std Block Size:         8K
           Shared Pool Size:       400M          Log Buffer:     2,048K
```

900 seconds Wall time

World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

Top 5 Timed events

The Top 5 timed events is the most important section of a statspack/AWR report for determining if an application is IO bound.

```
Top 5 Timed Events
~~~~~~~~~~~~~~~~~~                                                 % Total
Event                                             Waits    Time (s) Ela Time
--------------------------------------------- ------------ ----------- --------
db file sequential read                        8,587,142      45,110    83.20
CPU time                                                       4,981     9.19
latch free                                       109,044       1,420     2.62
buffer busy waits                                 46,525       1,305     2.41
db file parallel read                             23,687         744     1.37
          ---------------------------------------------------------------
```

The YAPP method.

Single Block Read, 5.25 mS per wait.

9,541 Waits per second.

These Correspond to physical read IOs

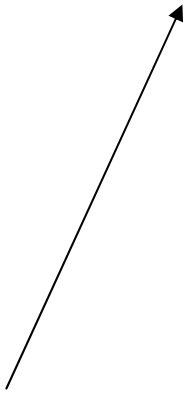
Major IO Related Waits

Event	Description
db file sequential read	The sequential read event is caused by reads of single blocks by the Oracle Database of a table or index.
db file scattered read	The scattered read event is caused by reads of multiple blocks by the Oracle Database of a table or index.
CPU time	This is the amount of time that the Oracle database spent processing SQL statements, parsing statements, or managing the buffer case. Tuning the SQL statements and procedures, or increasing the server's CPU resources generally best reduce this event.
log file parallel write	This event is caused by waiting for the writes of the redo records to the redo log files.
log file sync	This event is caused by waiting for the LGWR to post after a session performs a commit. This can be tuned by reducing the number of commits.
free buffer wait	This wait occurs when a session needs a free buffer and cannot find one. A slow DBWR process that cannot quickly flush dirty blocks from the buffer cache can cause this. This wait can also occur is one session requests a buffer that another session has requested from disk.

Determining the Database IO load

Load Profile

	Per Second	Per Transaction
Redo size:	17,007.41	16,619.62
Logical reads:	351,501.17	343,486.49
Block changes:	125.08	122.23
Physical reads:	**11,140.07**	**10,886.06**
Physical writes:	**1,309.27**	**1,279.41**
User calls:	7,665.49	7,490.70
Parses:	14.34	14.02
Hard parses:	4.36	4.26
Sorts:	2.85	2.78
Logons:	0.17	0.17
Executes:	22.41	21.90
Transactions:	**1.02**	



Note these are Blocks per second (not reads per second). 8k Block size = 87 MB/s read, 10 MB/s write.

World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

Tablespace IO Stats

Tablespace IO Stats for DB: MTR Instance: MTR Snaps: 25867 -25868
->ordered by IOs (Reads + Writes) desc

Tablespace

	Reads	Av Reads/s	Av Rd(ms)	Av Blks/Rd	Writes	Av Writes/s	Buffer Waits	Av Buf Wt(ms)
MSERVERTAB	7,861,586	8,735	5.5	1.0	59,214	66	45,542	28.9
MSERVERIND	884,275	983	5.0	1.0	24,261	27	925	19.1
TEMP	122,465	136	7.7	8.9	121,028	134	0	0.0
TOOLS	1,166	1	1.3	1.5	452	1	0	0.0
UNDOTBS1	66	0	5.6	1.0	353	0	2	0.0
SYSTEM	51	0	9.0	1.0	9	0	0	0.0

These are read/write IOs per second. 9,855 Read IOPS, 227 write IOPS.
The average read response time is ~5.5 ms.

Instance Activity Stats

Instance Activity Stats for DB: MTR Instance: MTR Snaps: 25867 -25868

Statistic	Total	per Second	per Trans
...			
physical reads	**10,026,061**	**11,140.1**	**10,886.1**
physical reads direct	**1,087,774**	**1,208.6**	**1,181.1**
physical writes	**1,178,340**	**1,309.3**	**1,279.4**
physical writes direct	**1,093,945**	**1,215.5**	**1,187.8**
physical writes non checkpoint	**1,178,292**	**1,309.2**	**1,279.4**
..			
redo blocks written	15,984	17.8	17.4
redo buffer allocation retries	0	0.0	0.0
redo entries	104,742	116.4	113.7
redo log space requests	0	0.0	0.0
redo log space wait time	0	0.0	0.0
redo ordering marks	0	0.0	0.0
redo size	15,306,672	17,007.4	16,619.6
redo synch time	421	0.5	0.5
redo synch writes	**1,056**	**1.2**	**1.2**
redo wastage	697,864	775.4	757.7
redo write time	638	0.7	0.7
redo writer latching time	12	0.0	0.0
redo writes	1,267	1.4	1.4

Overall IO analysis

- Single Block waits are a significant Database Wait event.
- Read IO load is high, 87 MB/s, 9,855 Read IOPS
- Storage Response time is good for an array under this load: 5.5 mS per read.

Conclusion:

Storage with a lower response time can improve the database performance. Adding additional spindles to the array is likely to only have minimal results.

AWR/ Oracle 10g New IO Counters

Instance Activity Stats DB/Inst: RAMSAN/ramsan Snaps: 22-23

Statistic	Total	per Second	per Trans
physical read IO requests	302,759	**4,805.7**	20,183.9
physical read bytes	**35,364,380,672**	**561,339,375.8**	###########
physical read total IO requests	**302,945**	**4,808.7**	**20,196.3**
physical read total bytes	**35,367,449,600**	**561,388,088.9**	###########
physical read total multi block r	**292,958**	**4,650.1**	**19,530.5**
physical reads	**4,316,960**	**68,523.2**	**287,797.3**
physical reads cache	4,316,941	68,522.9	287,796.1
physical reads cache prefetch	4,014,197	63,717.4	267,613.1
physical reads direct	0	0.0	0.0
physical reads direct temporary t	0	0.0	0.0
physical reads prefetch warmup	0	0.0	0.0
physical write IO requests	484	7.7	32.3
physical write bytes	5,398,528	85,690.9	359,901.9
physical write total IO requests	615	9.8	41.0
physical write total bytes	7,723,520	122,595.6	514,901.3
physical write total multi block	124	2.0	8.3
physical writes	659	10.5	43.9
physical writes direct	0	0.0	0.0
physical writes from cache	659	10.5	43.9
physical writes non checkpoint	582	9.2	38.8

World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

ORioN – Oracle's Storage Benchmarking Tool

ORioN – Oracle I/O Numbers

- Tool designed to predict the performance of an Oracle database without having to install Oracle or create a database.
- Designed expressly to simulate Oracle IO by using the same software stack.
- Can simulate the effect of striping on performance done by ASM.
- Simulates both OLTP and data warehouse loads.

Where can I get ORioN

- Available here:
  http://www.oracle.com/technology/software/tech/orion/
  - Requires free Oracle Technology Network login
- Binaries available for:
  - Linux/Solaris
  - Windows
  - EM64 Linux

Using ORioN

- Command line driven utility, however, a configuration file is also required.
- EX: orion –run simple –testname <Configuation File> -num_disks 8
- Configuration File
  - Contains the path to the physical LUNs to test
  - Windows EX: “\\.\e:” for the E: drive
  - Linux /dev/sda
- Allows multiple drives for ASM striping

Command Line Options for predefined tests

- -run <simple, normal, advanced>
  - Simple: 30-50 minute test for baseline IOPS, Bandwidth and Latency.
  - Normal: full day test of different combinations of 8K and 1024 K random reads. Fully maps out read performance.
- -num_disks <#>
  - Increases maximum load and range for storage systems with higher disk counts.
- -testname <testname.lun>
  - The name of the configuration file that identifies the test LUNs
- (optional) –cache_size <#>
  - Specifics the size of the cache on the LUN in MB. Based on this number ORioN will warm the cache with random 1 MB reads to prevent cache masking true performance.

Output

- 5 files
  - <testname>_iops.csv
    - Matrix of the IO per second as the number of concurrent 8K and 1024 KB random reads varies.
  - <testname>_lat.csv
    - Matrix of the average latency as the number of concurrent 8K and 1024 KB random reads varies.
  - <testname>_mbps.csv
    - Matrix of the average MB per second as the number of concurrent 8K and 1024 KB random reads varies.
  - <testname>_summary.txt
    - List of all input parameters, information on the LUNs tested
    - Maximum MBPS, IOPS, and minimum latency recorded.
  - <testname>_trace.txt
    - Verbose test output.

Mytest_iops.csv Example

Large/Small	1	2	4	6	8	10	12	14
0	10360	18692	34592	37556	40038	40990	41041	41285
1								
2								
3								

TMS
TEXAS MEMORY SYSTEMS

RamSan Solid State Disks

Texas Memory Systems, Inc.
Some RamSan customers …

- World's Fastest Storage®
- Over 30 years of experience with high bandwidth and low latency architectures
- Delivering tenth generation SSD
- Privately owned with no debt/venture capital
- Repeat customers demonstrate high customer satisfaction:

World's Fastest Storage®

TMS
TEXAS MEMORY SYSTEMS

RamSan Solid State Disks: Performance

Why Oracle Performance Increases with the RamSan

- Many Oracle deployments spend a significant amount of time waiting on a large number of IOs
- The RamSan does nothing to reduce the number of times that Oracle waits on IO, it just significantly reduces the duration of each wait
- The degree to which the database is waiting on IO before the RamSan is deployed determines how large the performance gain.

SPC-1 Report

SPC-1 IOPS™ Results

SPC-1 IOPS: ***291,208.58***

$/SPC-1 IOPS: ***$0.67***

*In **2008**, SPC-1 Ranked the RamSan-400 as:*

#1** for **Performance

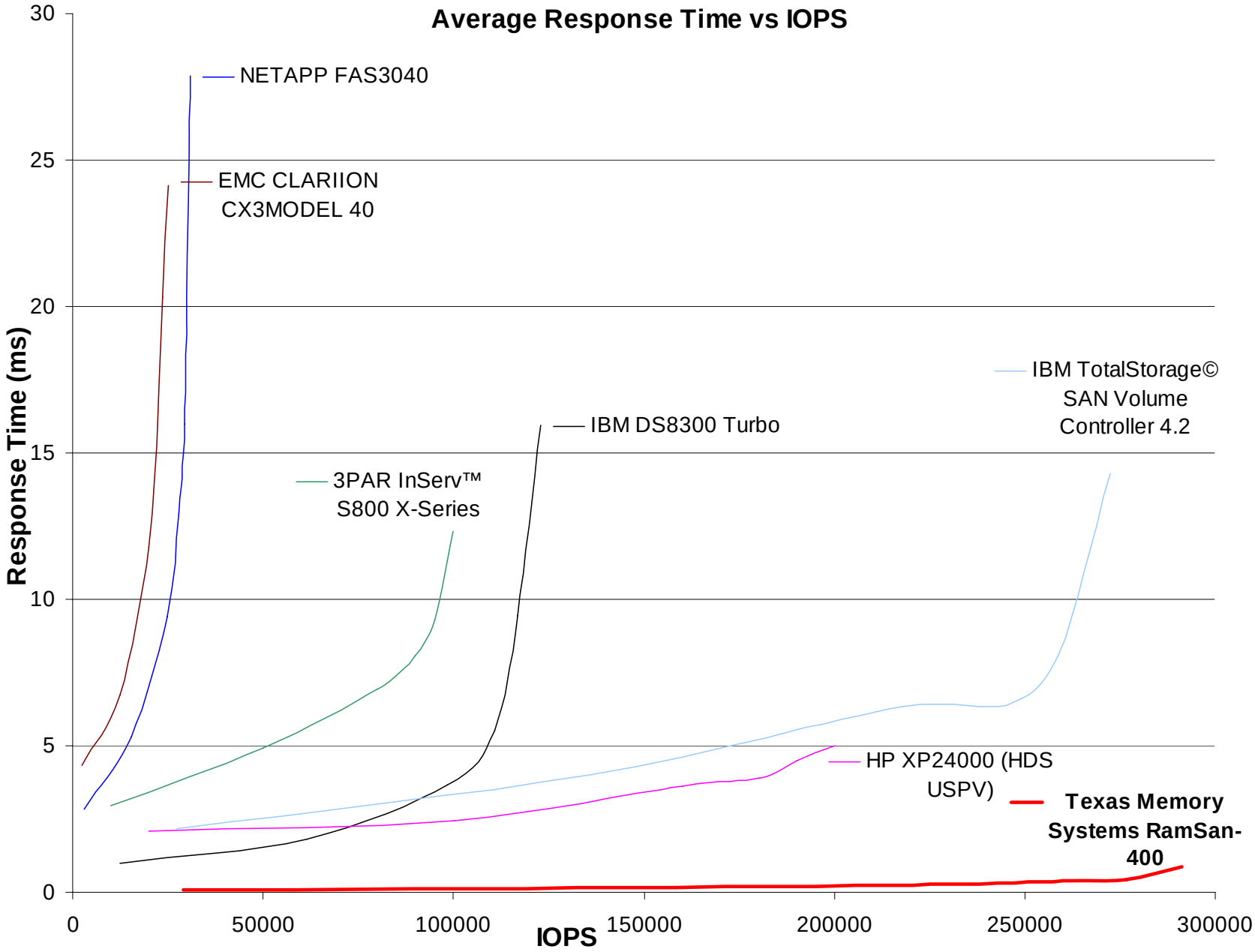
AND

#1** for **Price/Performance.

World's Fastest Storage®

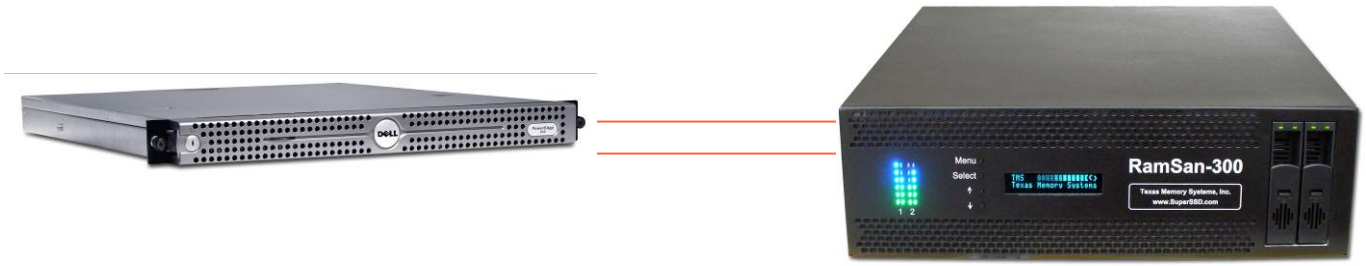
TMS TEXAS MEMORY SYSTEMS

SPC-1: Comparing Results (www.storageperformance.org)



ORioN Tests

- Setup
  - Pentium-D (3.0 GHz) Dell PowerEdge 850
  - Qlogic QLE2462 4 Gbps HBA
  - 32 GB RamSan
  - 2 Fibre channel connection

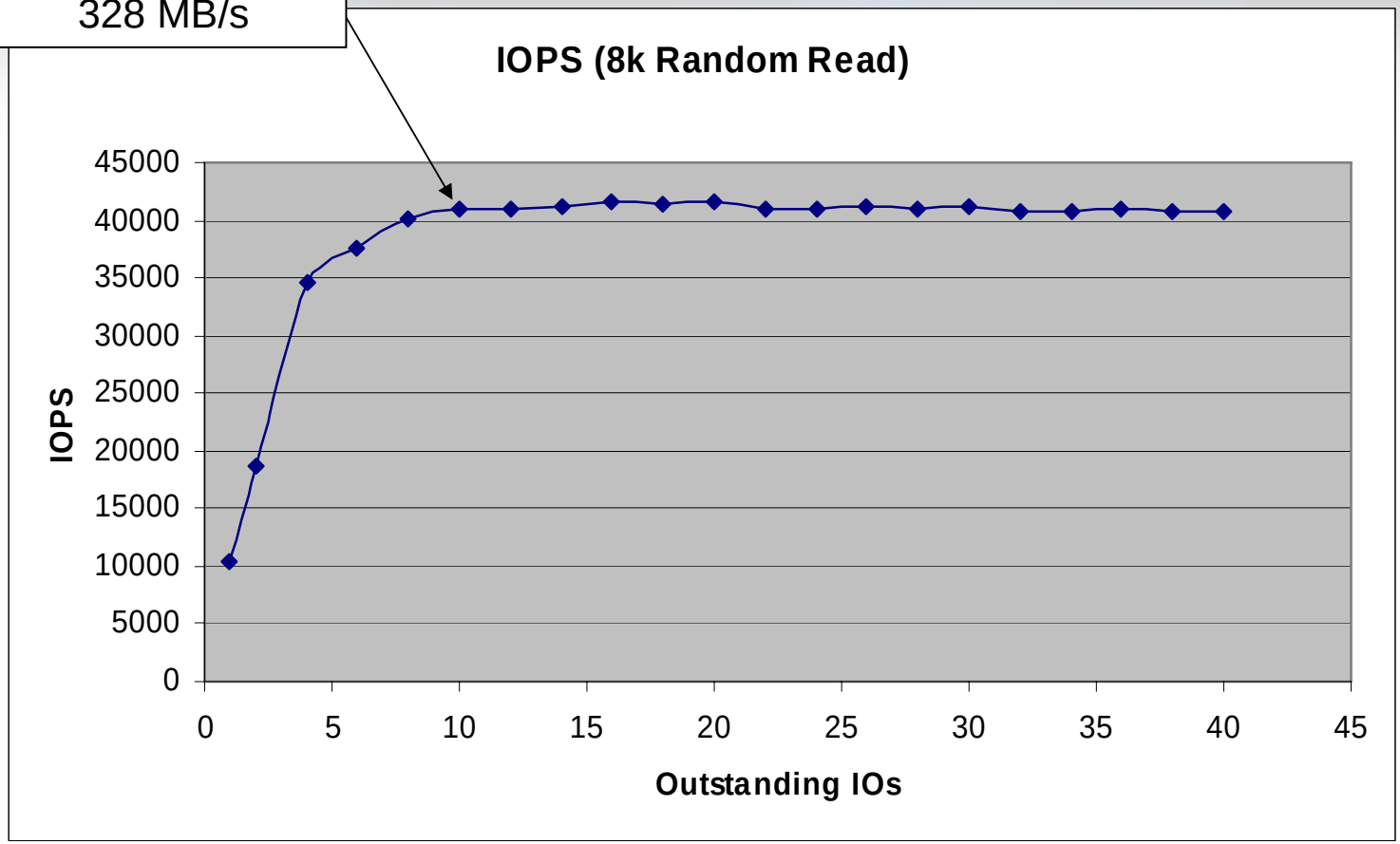


World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

IOPS Results

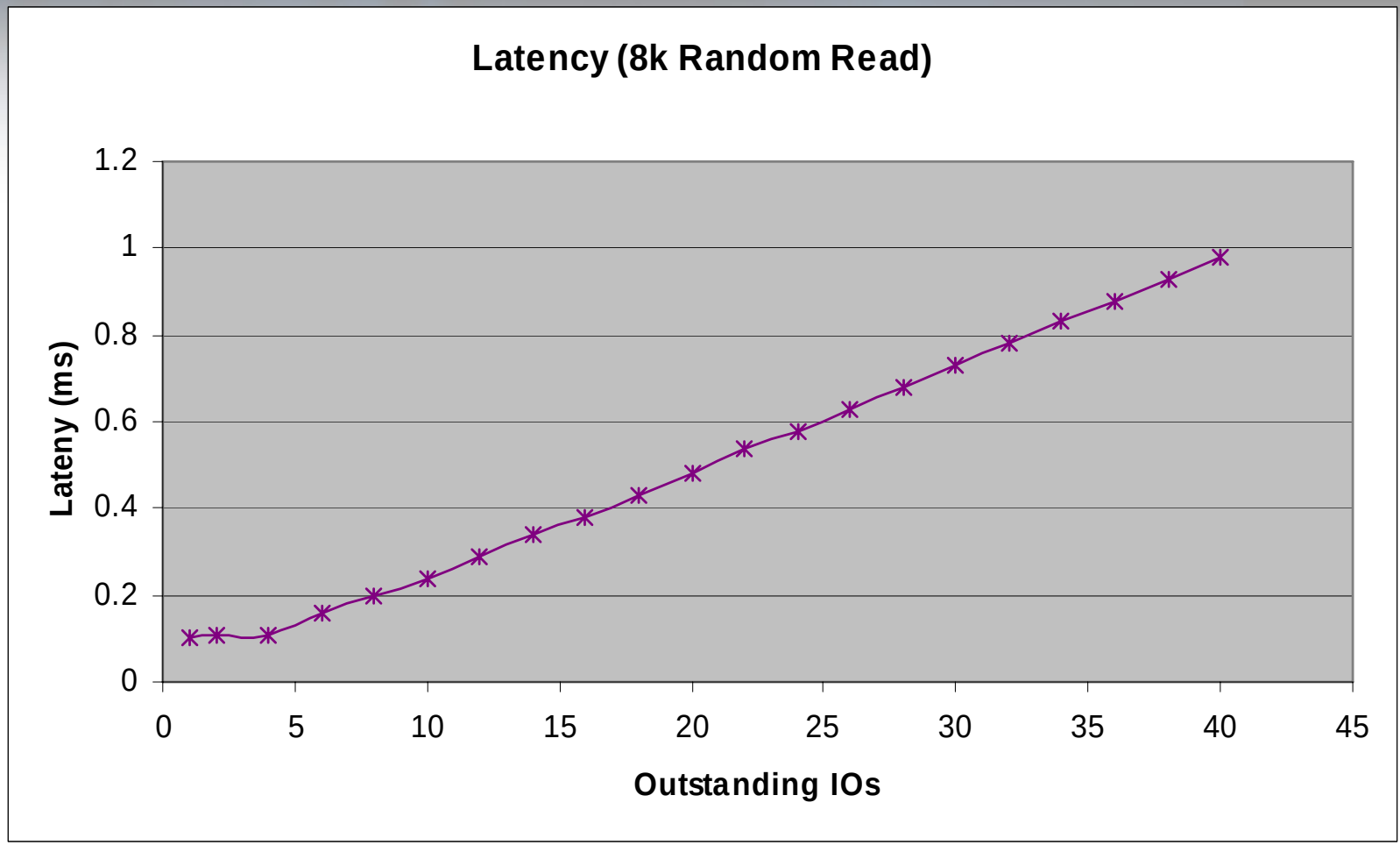
41000 IOPS with 8K block size, 328 MB/s



World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

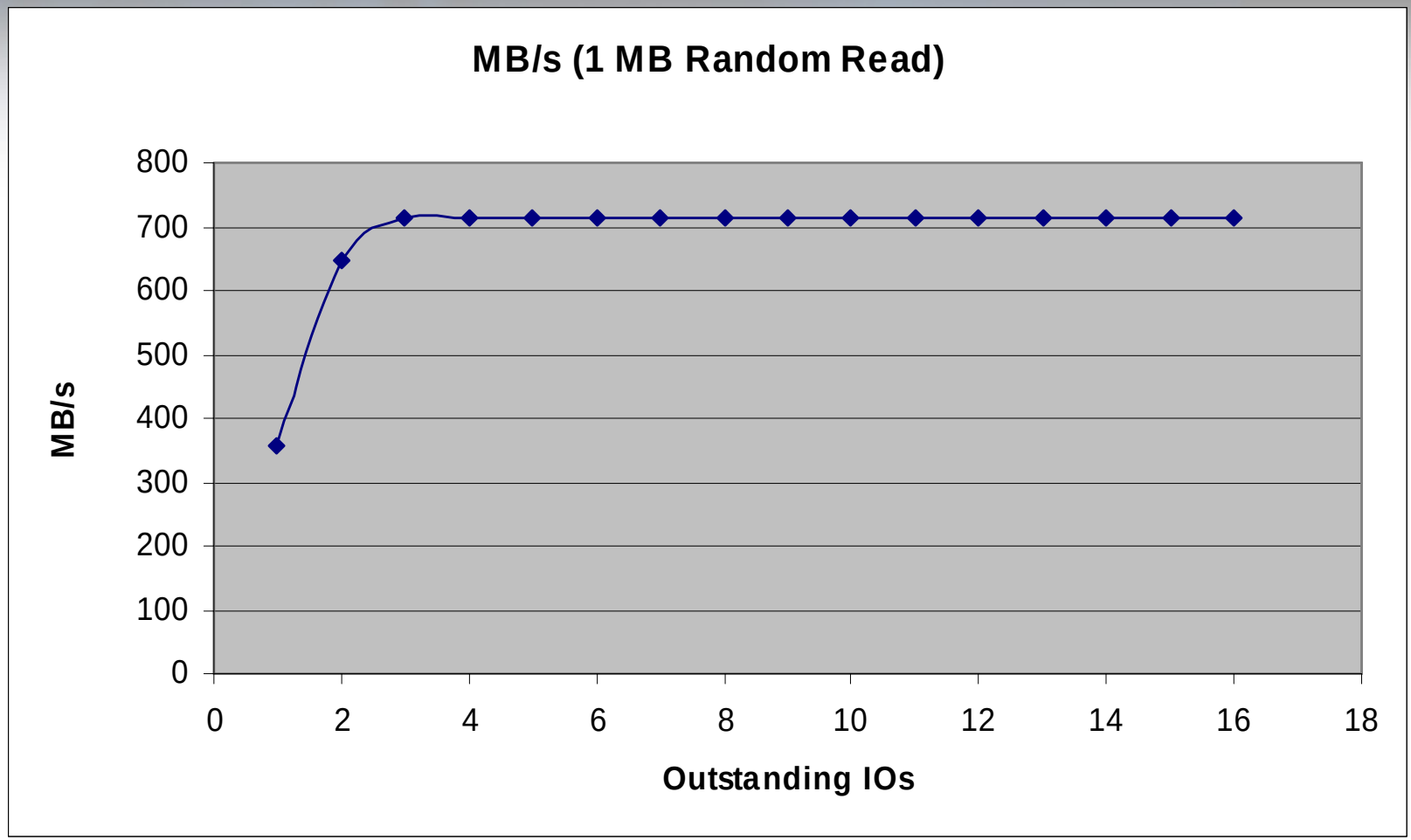
Latency Result



World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

Bandwidth Results



World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

The RamSan's Effect On Oracle

- Has no direct impact on the Number of IO related waits that occur, only on the amount of time that each wait takes. For a batch job the IO load will increase quite a bit with faster storage. For an OLTP load the IO load will not necessarily increase, but the end user response time will decrease.
- If the % of time that the database or the user spends waiting on IO is significant, the RamSan can offer considerable gains.

After the RamSan for the Same Database Examined Earlier

Before the RamSan

Top 5 Timed Events

Event	Waits	Time (s)	% Total Ela Time
db file sequential read	8,587,142	45,110	83.20
CPU time		4,981	9.19
latch free	109,044	1,420	2.62
buffer busy waits	46,525	1,305	2.41
db file parallel read	23,687	744	1.37

After the RamSan – *Note that even thought the # of waits increased significantly, the total time for waits decreased.*

Top 5 Timed Events

Event	Waits	Time (s)	% Total Ela Time
db file sequential read	20,159,505	22,973	64.32
CPU time		9,887	27.68
db file scattered read	97,723	992	2.78
buffer busy waits	63,767	855	2.39
db file parallel read	85,300	657	1.84

World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

TMS
TEXAS MEMORY SYSTEMS

RamSan Solid State Disks: Product Line

Overview of RamSan Solid State Disks

Feature	**RamSan-300**	**RamSan-400**	**RamSan-500**
Media	**DDR RAM**	**DDR RAM**	**Cached Flash**
Size	3U x 17”	3U x 25”	4U x 20”
Latency	15 microsecond	15 microsecond	200 microseconds
Capacity	16 to 32GB	32 to 128GB	1024 GB to 2048GB
Bandwidth	1.5 GB/sec	3 GB/sec	2 GB/sec
Random IOPS	200,000	400,000	100,000 4 Gbit FC
Interface	4Gbit FC or IB	4Gbit FC or IB	2 to 8
Ports	2 to 4	2 to 8	Up to 1024 LUNs
LUN Mapping	Up to1024 LUNs	Up to 1024 LUNs	Yes
LUN Masking	Yes	Yes	Yes
Hot Swap Power	Yes	Yes	Yes
Non-Volatile	Yes	Yes	Chipkill™, RAID
Error Protection	Chipkill™	Chipkill™	

World’s Fastest Storage®

TMS TEXAS MEMORY SYSTEMS

Statspackanalyzer.com

Free Statspack/AWR analysis

-Looks for IO bottlenecks and other configuration issues.

-Straightforward Tuning advice

World's Fastest Storage®

TMS TEXAS MEMORY SYSTEMS