

Pump-up Array Performance

Ray Lucchesi, President
Silverton Consulting, Inc.

<http://www.SilvertonConsulting.com>

INTEROP[®]

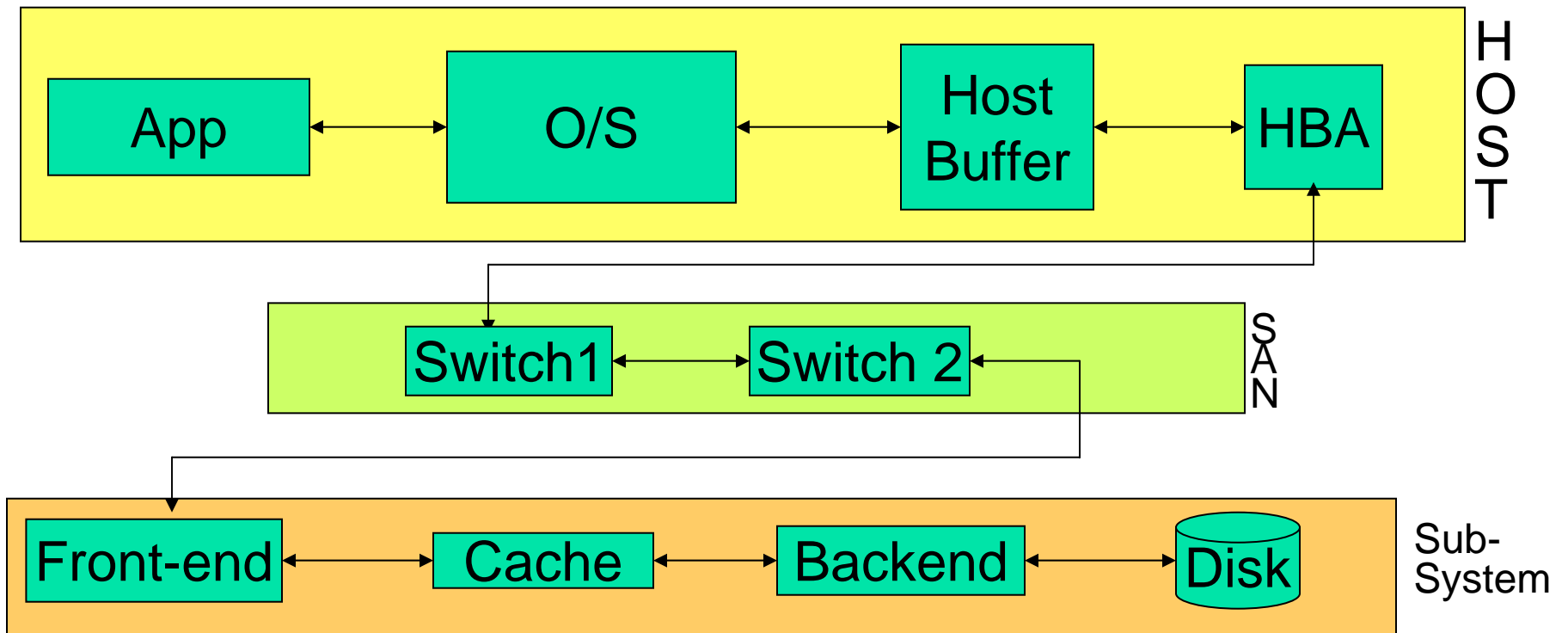
BUSINESS. TECHNOLOGY.
ONE WEEK. ONE PLACE.



Agenda

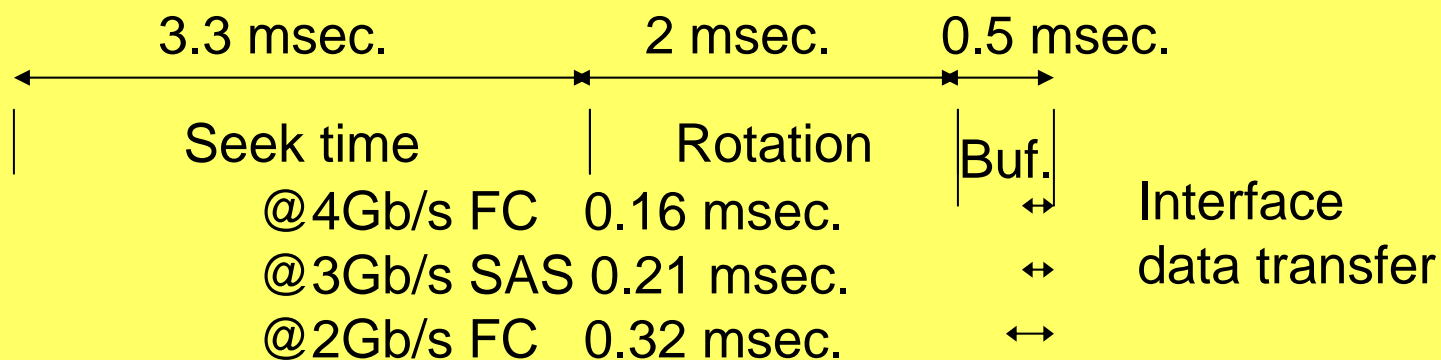
- Performance Fundamentals
- Performance Impacts
- Side Issues
- Final Thoughts

IO Journey



Fast Disk I/O

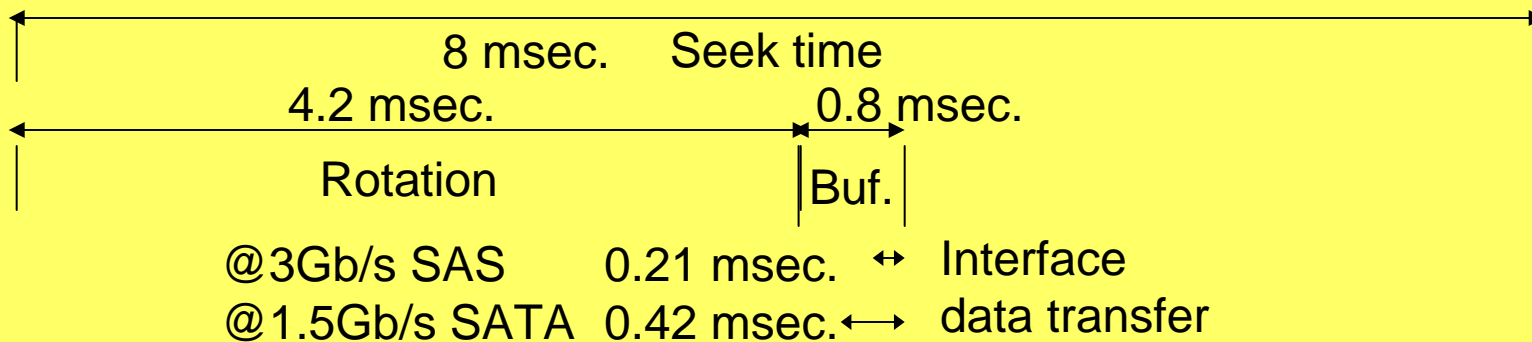
64K byte block high end disk I/O - 5.8 msec.



	Seagate Cheetah 15K.5	HitachiGST Ultrastar 15K	Fuji Max3147 15K
Read seek (msec)	3.5	3.6	3.3
Write seek (msec)	4.0	?	3.8
Rotational speed (KRPM)	15	15	15
Sustained transfer (MB/s)	125	123	?
Capacity (GB)	300,146, or 73	300,147, or 73	147, 73, or 36

Slow Disk I/O

64K byte block high capacity/slow disk I/O - 13 msec.



	Seagate Barracuda ES	HitachiGST Deskstar 7.2K
Read seek (msec)	8.5	8.5
Write seek (msec)	9.5	9.2
Rotational speed (KRPM)	7.2	7.2
Sustained transfer (MB/s)	78	?
Capacity (GB)	750, 500, 400, 320, or 250	1000 or 750

Cache I/O

64K byte block cache I/O 0.2 msec.



- Add subsystem overhead 2.25msec, for I/O add 1/2 in front and 1/2 at end
- Must add overhead to disk times above.



Transfer speed

- Burst data rates <> sustained transfer rates
- Fibre channel 4Gb/s, 2Gb/s, 1Gb/s - front-end or back-end
- SCSI Ultra 320 (3.2Gb/s) - front-end or back-end
- Ethernet 10Gb/s, 1Gb/s, 100Mb/s - front-end only
- SAS/SATA 3Gb/s, 1.5Gb/s - back-end only or direct attached storage

Enterprise Class

- Also called monolithic arrays
- Larger and better cache, more front-end & back-end interfaces, but fewer drive options
- Local and remote replication options
- High availability
- Typically better throughput

	HDS USP-V	EMC DMX-4	IBM DS8300 Turbo
Front-end/backend interfaces	224/?	64/64	128/64
Cache size (GB)	256	256	256
Drive options (GB)	73, 147, 300	73, 147, 300, 500	73, 147, 300, 500

Midrange Class

- Also called modular arrays
- More drive types but cache, front-end, and back-end limited
- Less replication options
- Less availability options
- Typically better latency
- SAS/SATA backend interfaces

	Dot Hill 2730T	LSI 6998	EMC CX3 model 80	HDS AMS1000	IBM DS4800
Backend	SAS/SATA II	FC/8-ports	FC/8-ports or SATA	FC or SATA	FC/8-ports
Front-end	4	8	8	8	8
Cache size (GB)	1	16	16	16	16
Drive options (GB)	73,146, 250, 300, 500, 750	73, 146, 250, 300	73, 146, 300, 500s, 750s	73,146, 250s, 300, 500s	146, 250, 300, 400, 500



JBODs

- Direct attached storage -
SATA/SAS, SCSI Ultra 320, or
FC/AL
- RAID either S/W or HBA based
- Only disk and host buffer for
cache I/O



Fundamentals Summary

- I/O journey
- Fast, medium, & slow drive
- Cache
- Transfer speeds
- Enterprise, midrange, and JBOD subsystems



Agenda

- Performance Fundamentals
- Performance Impacts
- Side Issues
- Final Thoughts

Cache

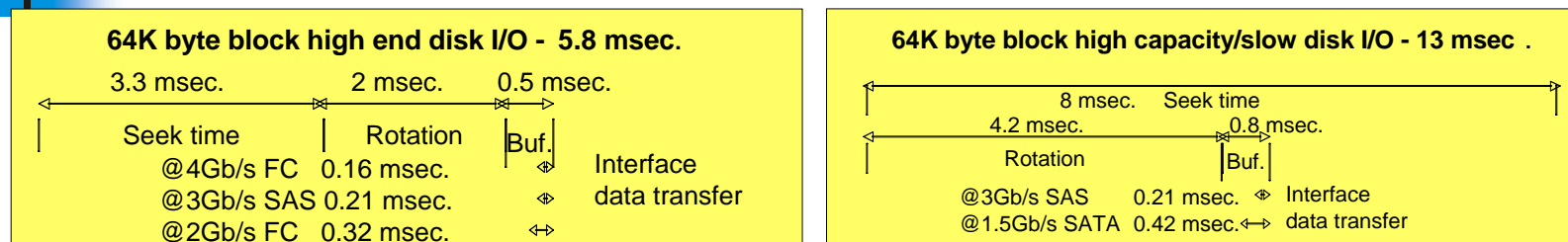
64K byte block cache I/O 0.2 msec.

1.1 msec. 1.1 msec. SubSys Overhead

 @4Gb/s FC 0.16 msec.

- Cache hit, the fastest way to do I/O 2.4 vs. 8.1 msec.
- Larger cache helps, but
 - Write hit ultimately needs to be destaged and written to disk, may impact throughput - writing directly to disk may be faster than a write hit followed by destage
- Sophistication matters

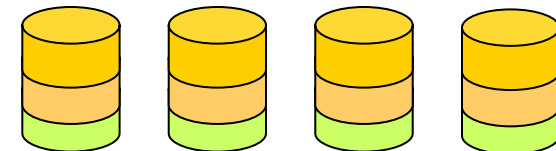
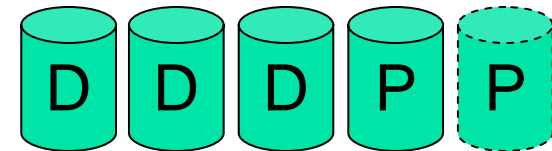
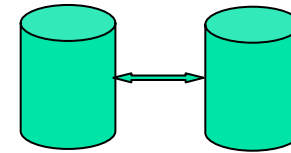
Drive speed



- Fast drives = big difference for miss and destage
8.1 vs. 15.3 msec.
 - Assuming no other bottlenecks
- High capacity/slow drives degrade miss/destage
 - Response time concern, subsystem sophistication masks throughput impacts for non-busy drives

RAID and Striping

- RAID-1 - mirrored data
 - Reads use closest seek
 - Writes both, 2nd destaged later
 - Reads split across 2X drives
- RAID-4, 5, 6, DP - parity + data blocks
 - Parity block write penalty
 - RAID 5, 6, & DP distributed parity
 - RAID 4 single parity drive (potential hot drive)
 - RAID 6 & DP two parity drives, RAID 5 has one
- LUN striping - LUNs stripped across RAID groups (same type)
 - Eliminates hot RAID groups





I/O Balance

LUN I/O activity spread

- Across RAID groups - no hot RAID groups, drives
- Across front-end interfaces/controllers - no hot controllers, front-end interfaces
- Across back-end interfaces - no hot back-end interfaces
- Application/workload mix - toxic workloads reduce cache hits



Cache Revisited

- Cache read-ahead insures follow-on I/O in cache
 - Sophisticated subsystems compute in real-time
 - Others specify (consider cache demand at time of I/O)
- Cache read to write boundary
 - Some subsystems have hard boundary
 - Others have soft boundary - sized based on average or peak write workload



Remote Replication

- Remote replication - duplicates data written on subsystem to remote subsystem
- Synchronous - write degradation
- Semi-Synchronous - remote data at 1- to N-I/Os behind primary
- Asynchronous - data duplication scheduled and only guaranteed at end of activity
- Midrange and enterprise differences
 - use of backend disk vs. cache for holding data

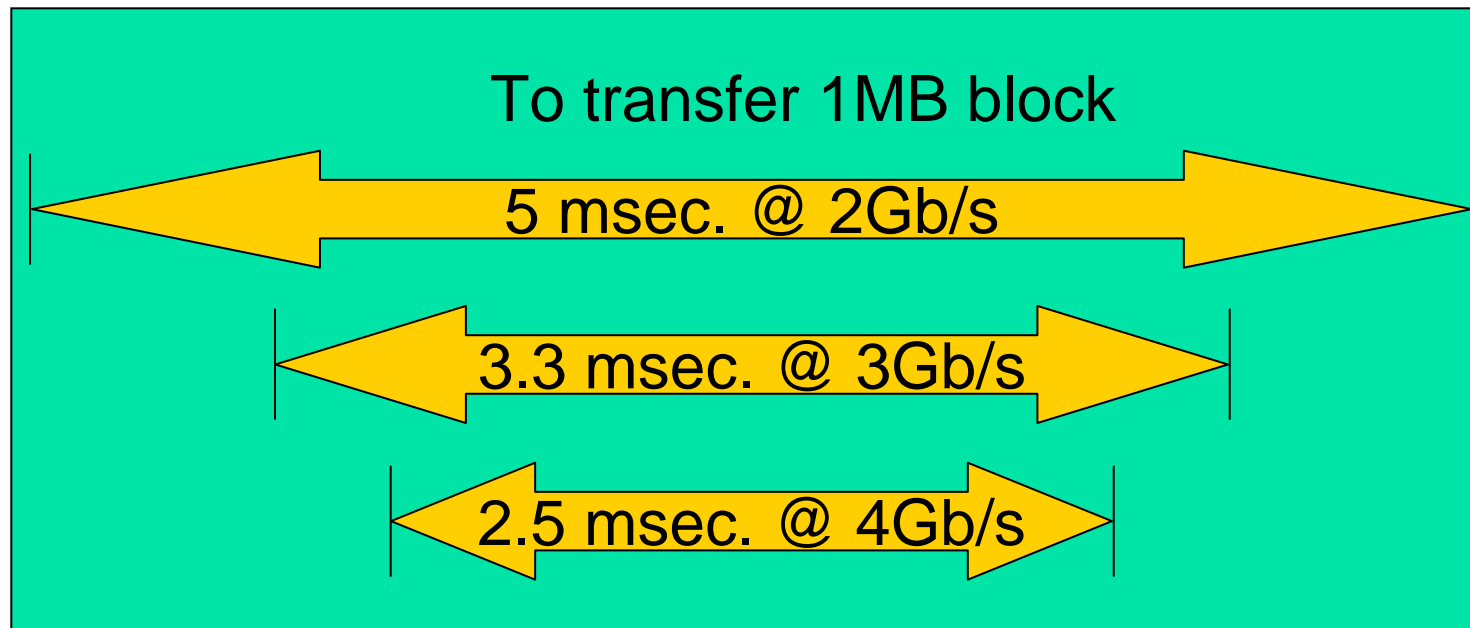


Transfer size

- For sequential - the larger the better
 - Most transfers generate full I/O (seek+rotation+transfer), bigger transfers \Rightarrow less seeks+rotations for same file size.
 - Each transfer invokes 2.3msec overhead, less transfers \Rightarrow less overhead for same file size
- For random I/O - larger transfers don't work
 - Each random request typically processes only small amount of data, large transfers \Rightarrow wasted data
- Real workloads mixed, seldom pure sequential or random

Transfer Speed

- Transfer speed impacts performance for large transfer sizes at backend as well as front-end





Midrange Cache Mirroring

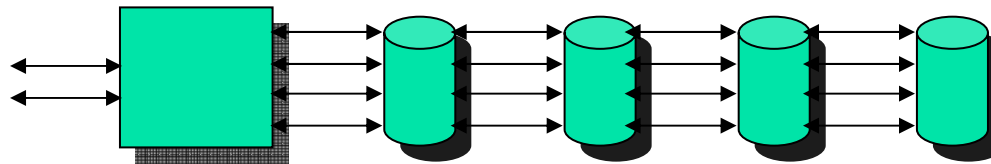
- Adds additional transfer (between controllers) for each write
- Performance impact depends on transfer size and speed between controllers



Point-in-time (P-I-T) Copy

- P-I-T copy - used to replicate data locally for backup and test purposes
- Copy-on-write technology
 - Takes added cache, disk, and/or other resources for each write

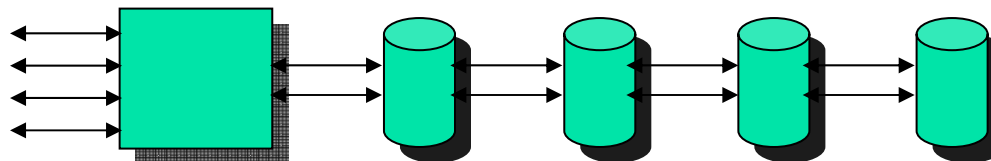
Front-end Limits



Number of front-end interfaces can limit performance

- FC sustains ~90% of rated speed
 - for 4Gb/s= \sim 360MB/s per FC link
- iSCSI sustains 50-80% of rated speed
 - 1Gb/s=50 to 80MB/s per GigE link
- Connectivity often dictates number of front-ends but performance requirements should be considered

Back-end Limits



Back-end number of FC or SAS/SATA links also limits I/O

- Cache miss activity translates into backend I/O
 - Write hits also
- FC Switched vs. FC/Arbitrated Loop (FC/AL)
 - Switched provides more throughput per drive
 - Loop provides less throughput per drive sharing link
- SAS backend is point-to-point



Drive Limits

Drive count and speed limit
subsystem I/O rate

- Single drive has upper limit of I/Os
 - Faster drives do more
- Max subsystem drive I/O compared to peak miss/destage workload



Impacts Summary

- Cache, drive speed, RAID & LUN striping, and I/O balance
- Cache parameters, remote replication, transfer size & speed, cache mirroring, and P-I-T copy
- Front-end, back-end, and drive I/O limits



Agenda

- Performance Fundamentals
- Performance Impacts
- Side Issues
- Final Thoughts



Pre-purchase decisions

- Drives (count and performance)
 - Performance cost 50% more (\$/GB)
- Interfaces front-end and back-end (type/speed and number)
- Cache size and sophistication
 - 2X cache ~10% more readhits



Configuration Time

- RAID type for LUNs
- LUN striping or not
- Fixed cache parameters - cache mirroring, look ahead, read to write boundary
- I/O balance - across LUNs, RAID groups, controllers, front-end & back-end interfaces
- Subsystem partitioning - cache, interfaces, drives (RAID groups)



Host Side

- Multi-path I/O for performance
- HBA configuration matches subsystem
 - Host transfer size > or = subsystem
- Host buffer cache for file system I/O
 - Write-Back vs. Write-Thru
 - Sync's for write back
 - May use all available memory
 - Database cache, buffer cache, and subsystem cache interaction



SAN Performance

- Fan-in ratio 5:1 to 15:1 server to storage ports
- Hop counts
- ISL and FC link oversubscription
- Locality



Exchange Server

- Three files per exchange storage group
 - Jet DB (.edb file) data from MAPI clients
 - Stream DB (.stm file) attachments (ptrs from .edb)
 - Transaction Log (.log files)
- Isolate each storage group on own set of LUNs
 - Separate log file LUN from Jet and Stream DB LUNs
- Other exchange I/O besides reading & writing MS mail
 - Beware of BlackBerry/Trea users



Database I/O

For Oracle, DB2/UDB, MS SQL, etc.

- Separate log files from table spaces
- Typically separate indices from the table spaces they index
- Tailor transfer size to use
 - For sequential use larger transfer sizes
 - For random use smaller transfer sizes



Ongoing Workload Monitoring

What to look for

- Overall I/O activity to subsystem LUNs
- I/O balance over controllers, front-end interfaces, RAID groups, LUNs
- Read and write hit rates
- Sequential vs. random workload
 - Workload mix toxicity

Monitoring Tools 1

IOSTAT (Solaris example)

```

iostat -xtc 5 2          extended disk statistics tty    cpu
disk                    r/s  w/s  Kr/s  Kw/s  wait  actv  svc_t  %w  %b  tin tout us sy
  wt id
sd0          2.6  3.0  20.7  22.7  0.1   0.2  59.2   6   19  0  84  3  85 11 0
sd1          4.2  1.0  33.5   8.0  0.0   0.2  47.2   2   23
sd2          0.0  0.0   0.0   0.0  0.0   0.0   0.0   0    0
sd3         10.2  1.6  51.4  12.8  0.1   0.3  31.2   3   31
  
```

Monitoring Tools 2

SAR (HP-UX example)

```
/usr/bin/sar -d 15 4
```

```
HP-UX gummo A.08.06 E 9000/??? 02/04/92
```

Time	device	%busy	avque	r+w/s	blks/s	await	avserv
17:20:36	disc2-1	33	1.1	16	103	1.4	20.7
	disc2-2	56	1.1	42	85	2.0	13.2
17:21:06	disc2-0	2	1.0	1	4	0.0	24.5
	disc2-1	33	2.2	16	83	24.4	20.5
	disc2-2	54	1.2	42	84	2.1	12.8
Average	disc2-0	2	1.0	1	4	0.0	29.3
	disc2-1	44	1.8	21	130	16.9	21.3
	disc2-2	45	1.2	34	68	2.0	13.2



Monitoring Tools 3

- OS specific performance monitoring tools

AIX	Performance monitor
HP-UX	Disk performance monitor
Linux	lostat
Solaris	Dtrace
Windows	Performance monitor



Monitoring Tools 4

- DB specific performance monitoring tools

DB2/UDB	DB2 performance monitor
MS SQL	SQL performance monitor
Oracle	STATSPACK



Monitoring Tools 5

- Subsystem specific monitoring tools

EMC	ControlCenter Performance Manager
LSI	Storage Performance Analyzer
HDS	Hi-Command Tuning Manager
IBM	TotalStorage productivity center



Side Issues Summary

- Pre-purchase and configuration
- Host issues
- SAN design
- Exchange I/O
- Database I/O
- Workload monitoring



Agenda

- Performance Fundamentals
- Performance Impacts
- Side Issues
- Final Thoughts



Performance Automation

Some enterprise subsystems automate performance tuning

- LUN balancing
 - Across RAID groups
 - Across controllers/front-end interfaces
- Cache hit maximization
 - Read ahead amount
 - Read:write boundary partitioning
- Others



iSCSI vs. FC

- Ethernet at 50-80% vs. FC at 90% of sustained rated capacity
- Ethernet 1Gb/s vs. FC 2-4Gb/s
- Processor overhead for TCP/IP stack vs. HBA handling FC protocol overhead
- iSCSI continuum from desktop NIC to iSCSI HBA
 - iSCSI HBA costs \approx FC HBA
 - For iSCSI use server level NIC



NFS/CIFS vs. Block I/O

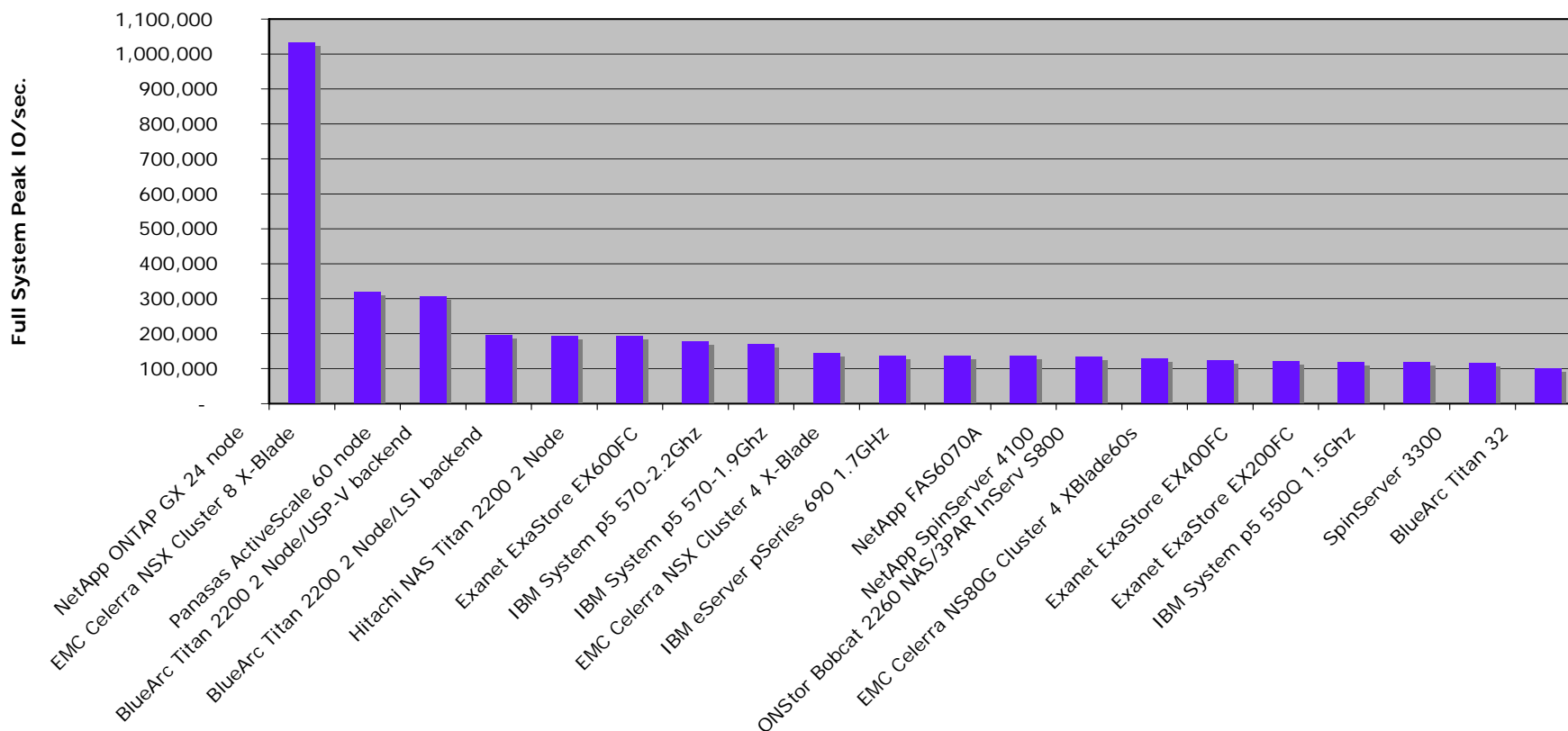
- NFS/CIFS Performance \approx block I/O
- # directory entries/mount point
- Gateway vs. integrated system
- Single vs. parallel vs. cluster vs global file systems
- No central repository for NetBench CIFS benchmarks

What Price Performance?

- Drive cost differential 50% for faster drives
- Enterprise - midrange cost differential
 - Subsystem sophistication cost differential, Enterprise class subsystems ~\$30/GB, Midrange = ~\$20/GB, Entry = ~\$10/GB
 - Cache size differential 100GB's or more for Enterprise class, 10GB's for midrange.

SPEC* SFS NFS full

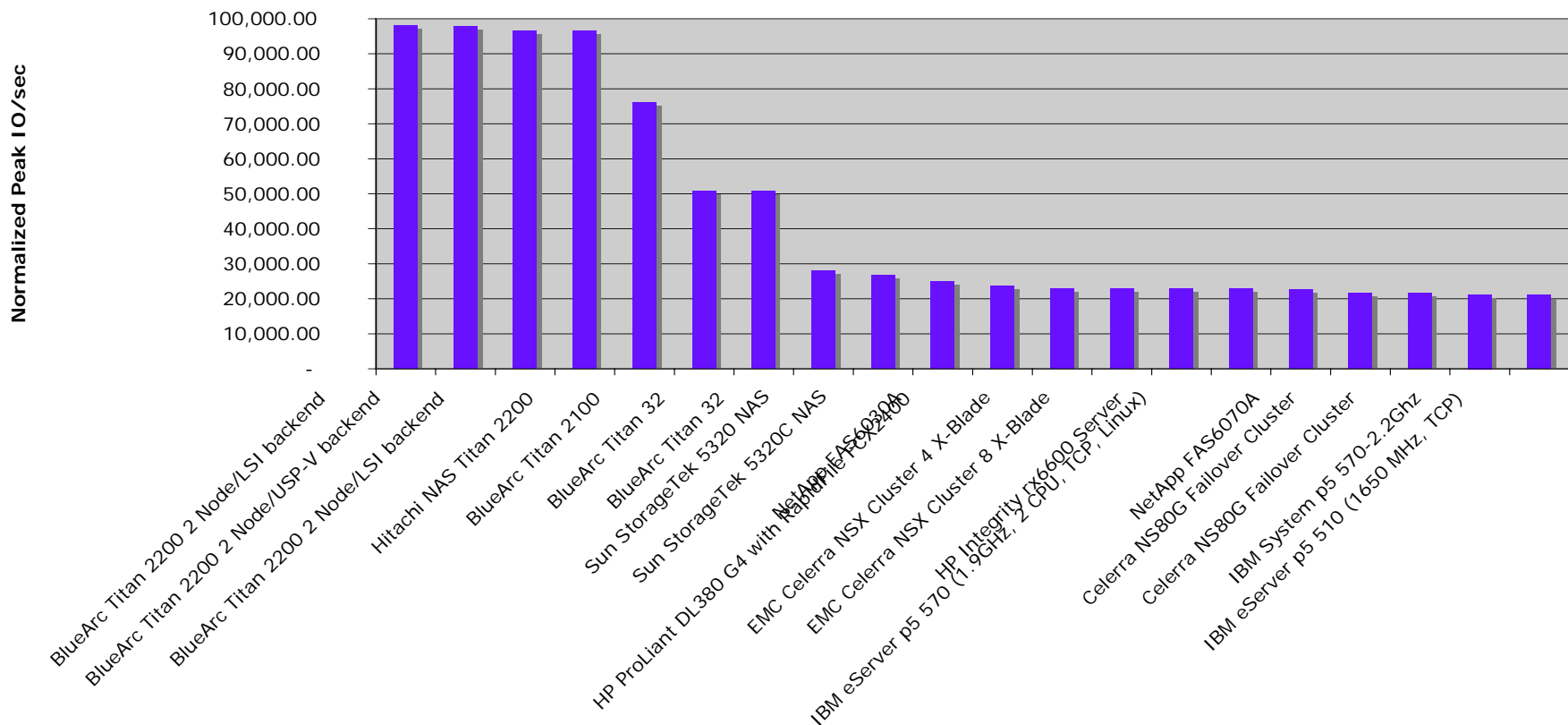
SPEC* SFS97_R1 NFS V3 Full System Performance as of 03 August 2007, TCP results only, Un-normalized performance, Top 20 systems



*All SPEC SFS results Copyright © 1995-2007 Standard Performance Evaluation Corporation (SPEC). All rights reserved, permission granted for use, data from <http://www.spec.org> as of 03 August 2007

SPEC* SFS NFS normalized

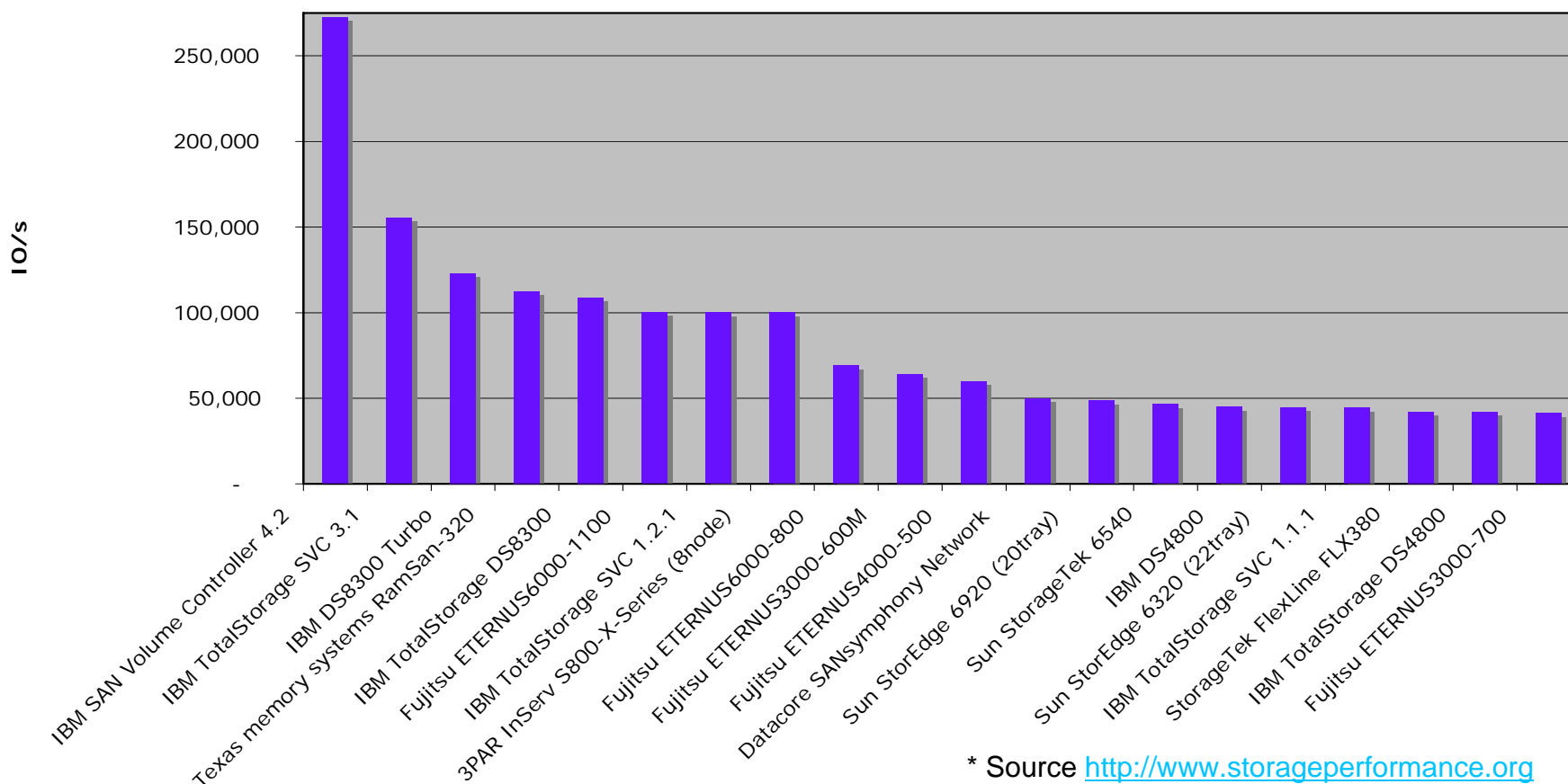
SPEC* SFS97_R1 NFS V3 Normalized Results as of 03 August 2007, TCP results only, Normalized by processors (chips or cores), Top 20 overall



*All SPEC SFS results Copyright © 1995-2007 Standard Performance Evaluation Corporation (SPEC). All rights reserved, permission granted for use, data from <http://www.spec.org> as of 03 August 2007

SPC-1* IOPS™ Top 20

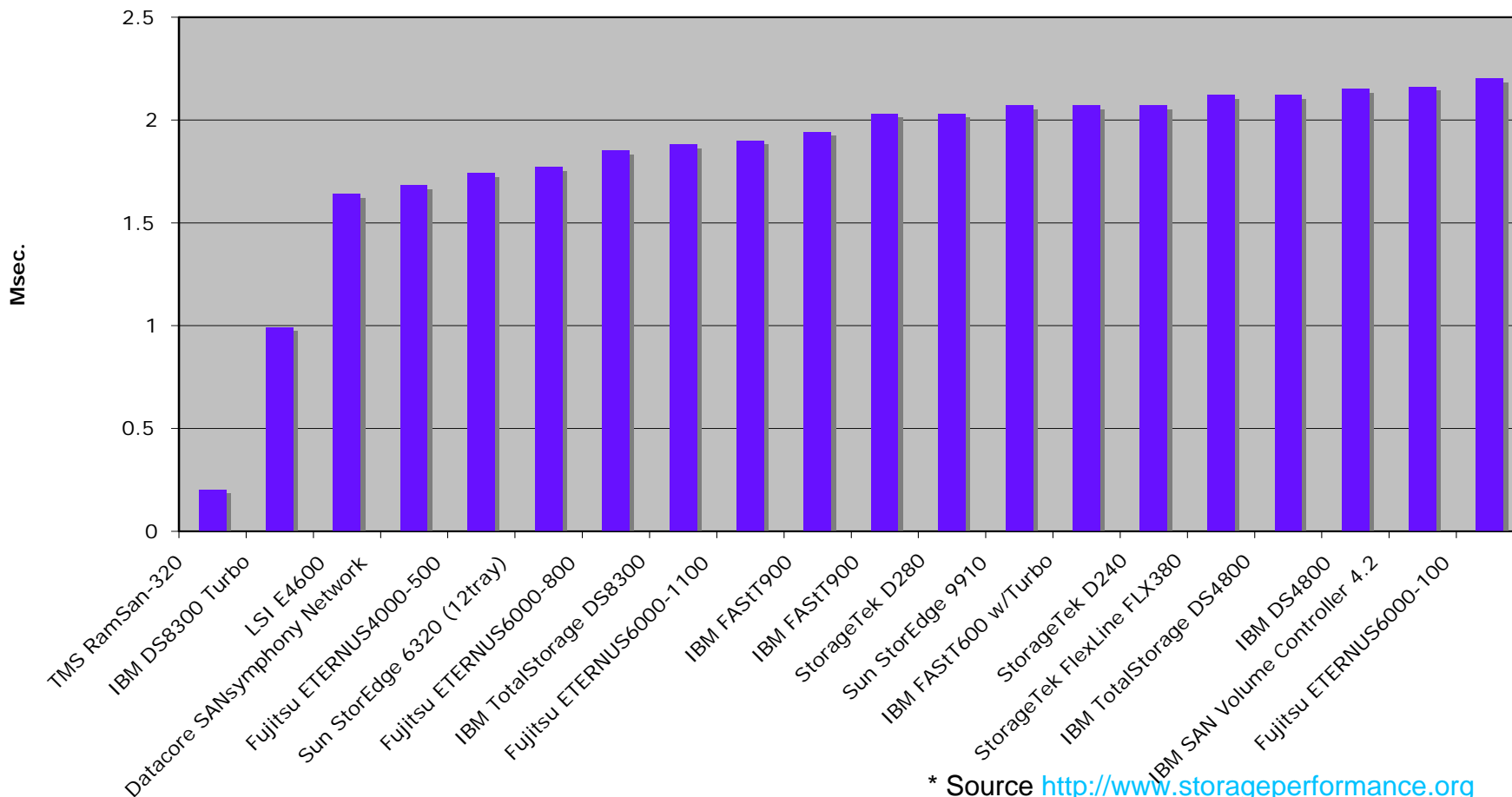
SPC-1* IOPS™ performance as of 31 July 2007 - top 20



* Source <http://www.storageperformance.org>

SPC-1* LRT™ Top 20

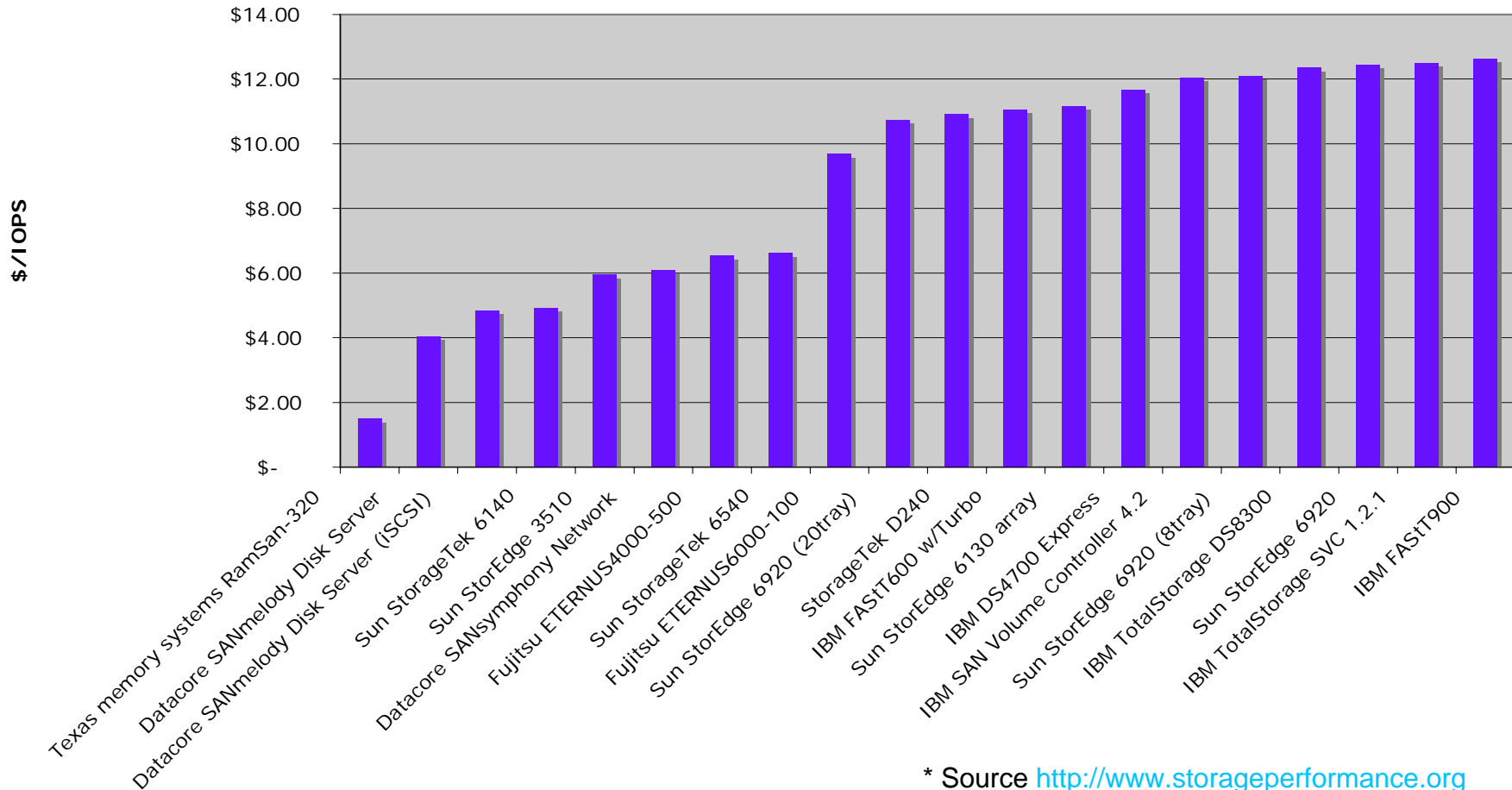
SPC-1* LRT™ avg. resp. time as of 31 July 2007 - top 20



* Source <http://www.storageperformance.org>

SPC-1* \$/IOPS™ Top 20

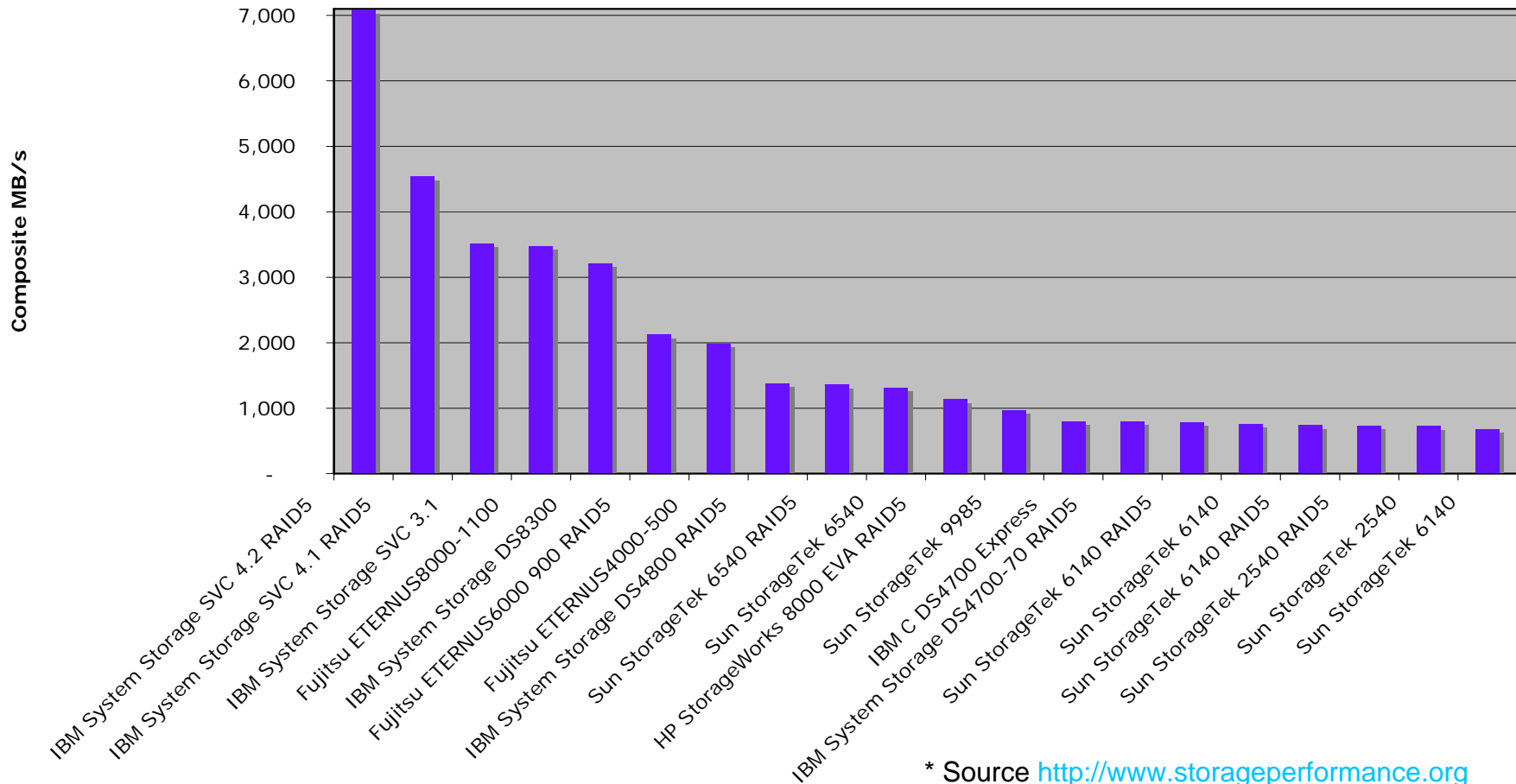
SPC-1* \$/IOPS(TM) as of 31 July 2007, Top 20



* Source <http://www.storageperformance.org>

SPC-2* MPBS™ Top 20

SPC-2* MBPS™ Performance data as of 31Jul07, Top 20



* Source <http://www.storageperformance.org>



Final Thoughts Summary

- Automated performance tuning
- iSCSI vs. FC
- NFS/CIFS vs. block I/O
- Cost of performance
- Benchmark results



For More Information

- Storage Performance Council (SPC) block I/O benchmarks www.storageperformance.org
- Standard Performance Evaluation Corp. (SPEC) SFS NFS I/O benchmarks www.spec.org
- Computer Measurement Group - more than just storage performance www.cmg.org
- Storage Networking Industry Association - standards with some performance info www.snia.org
- Silverton Consulting - StorInt™ Briefings & Dispatches, articles, presentations and pod casts from Silverton Consulting www.SilvertonConsulting.com



For More Information

Contact: Ray Lucchesi,
Info@SilvertonConsulting.com
+1-720-221-7270



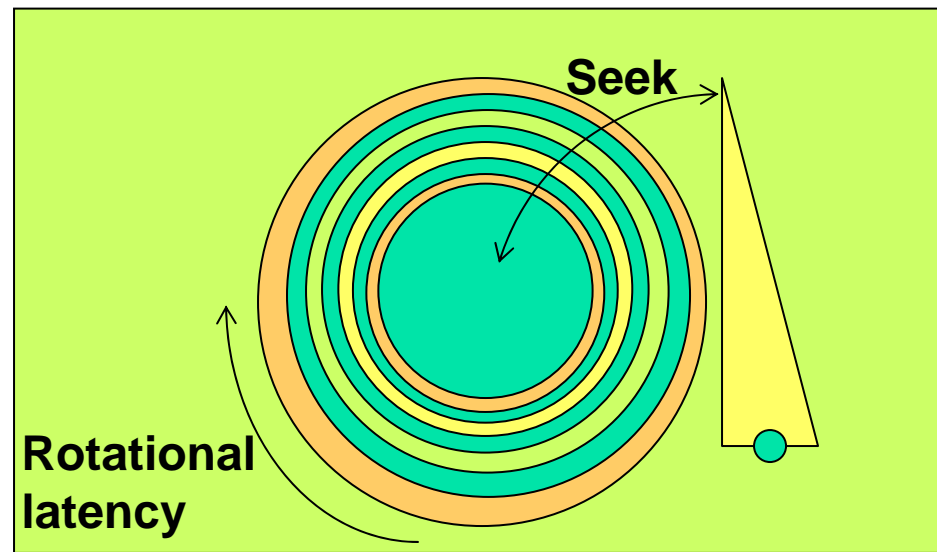
Background Information

Disk Array Terminology

- Direct attached or SAN attached disk arrays
 - Enterprise class - big subsystems with cache, multiple front-end interfaces and 10 to 100s of TB of disk
 - Mid-range and entry level have smaller amounts of each of these
- Just a bunch of disks (JBODs) internal attached disks

Disk Terminology

- Disk seek in milliseconds (msec.)
- Disk rotational latency
- Disk data transfer
- Disk buffer





Cache Terminology

- Cache read hit - when a read request finds its data in cache
- Cache write hit - when a write request writes to cache instead of disk, data is later destaged to backend disk
- Cache miss - when either a read or write have to use disk to perform the I/O request
- Cache read ahead - during sequential read requests, reading ahead of where the I/O is requesting data



IO Performance Terminology

- Throughput - data transferred per time unit (MB/s or GB/s)
- Response time - average time to do I/O (msec.)
- Sequential workload - multi-block accesses in block number sequence
- Random workload - no discernible pattern to block accesses



Acronyms

FC	Fibre channel	LUN	Logical unit number
FC/AL	Fibre channel arbitrated loop	MB/s	Mega-bytes per second
Gb/s	Giga-bits per second	Msec	1/1000 of a second
GB/s	Giga-bytes per second	P-I-T copy	Point-in-time copy
HBA	Host bus adapter	RAID	Redundant array of inexpensive d
I/O	Input/output request	SAN	Storage area network
iSCSI	IP SCSI	SAS	Serial attached SCSI
JBOD	Just a bunch of disks	SATA	Serial ATA
KRPM	1000 revolutions per minute	Xfer	Transfer