



Pump-up Array Performance

Ray Lucchesi, President
Silverton Consulting

<http://www.SilvertonConsulting.com>

INTEROP[®]

THE LEADING BUSINESS TECHNOLOGY EVENT

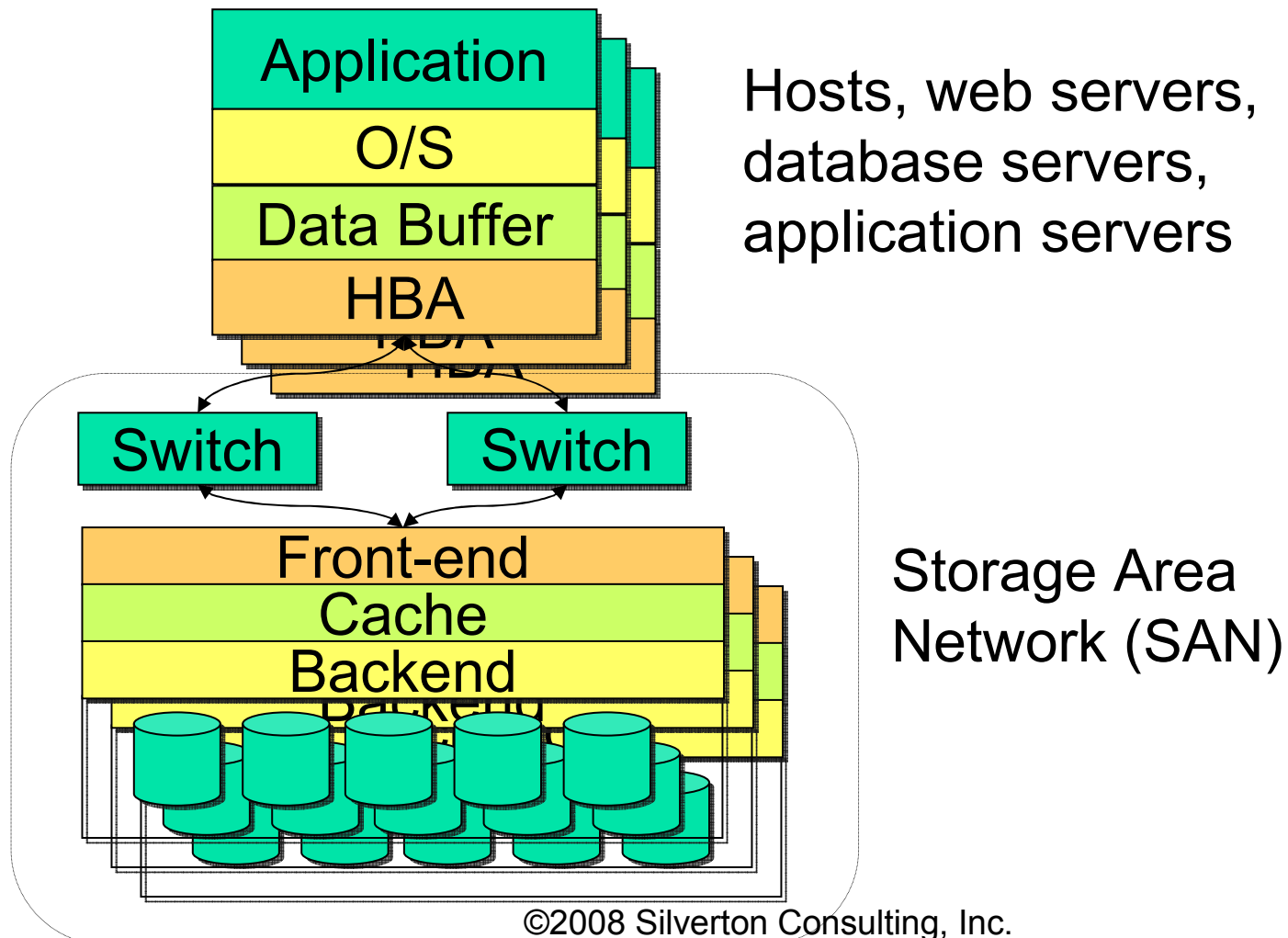


Pump-up Array Performance

Ray Lucchesi, President
Silverton Consulting

<http://www.SilvertonConsulting.com>

I/O journey



©2008 Silverton Consulting, Inc.

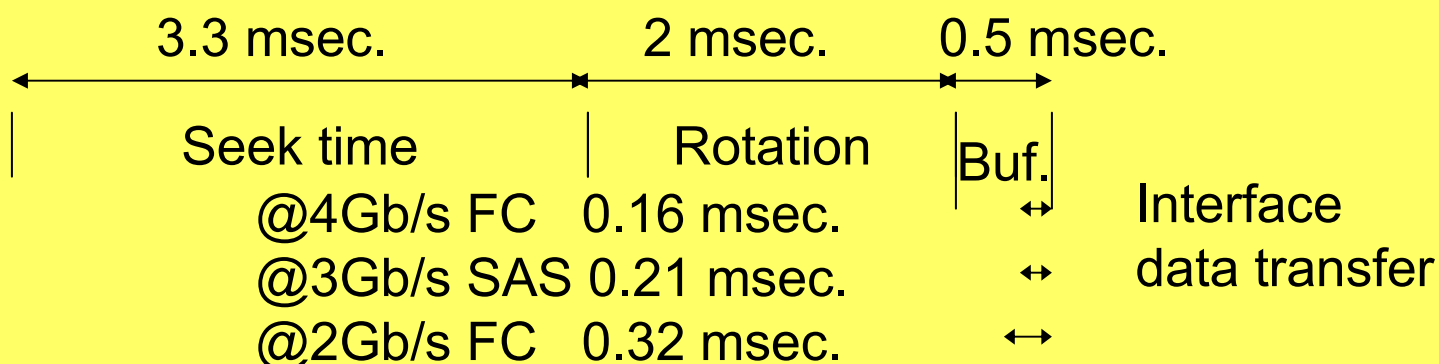
All Rights Reserved

April 08

3

Fast Disk I/O

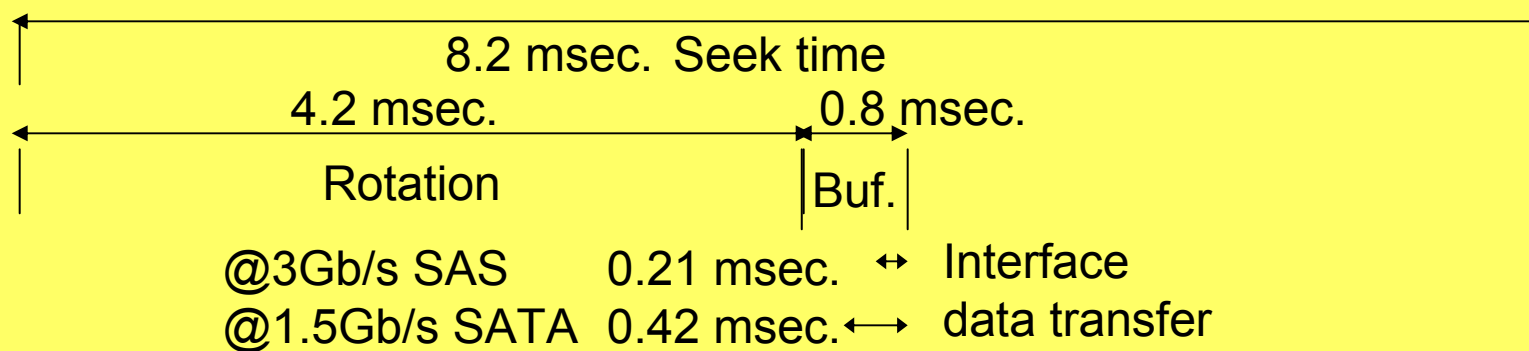
64K byte block high end disk I/O - 5.8 msec.



	Seagate Cheetah 15K.6	HitachiGST Ultrastar 15K300	Fuji MBA3300 15K
Read seek (msec)	3.4	3.6	3.4
Write seek (msec)	3.9	?	3.9
Rotational speed (KRPM)	15	15	15
Sustained transfer (MB/s)	164	123	179
Capacity (GB)	450, 300,146	300,147, 73	300, 147, 73

Slow Disk I/O

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



	Seagate Barracuda ES	HitachiGST Ultrastar 7.2K
Read seek (msec)	8.5	8.2
Write seek (msec)	9.5	?
Rotational speed (KRPM)	7.2	7.2
Sustained transfer (MB/s)	78	85
Capacity (GB)	750, 500, 400, 320, 250	1000, 750, 500

Cache I/O

64K byte block cache I/O 0.2 msec.

1.1 msec. 1.1 msec. SubSys Overhead

 @4Gb/s FC 0.16 msec.

- Add overhead ~2.2msec for I/O
 - Must add overhead to disk times above
- Larger cache helps, but
 - Write hits later destaged to disk causing 2 transfers
 - Write direct to disk only causes 1 transfer
- Sophistication matters.

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
- FC backend may include SATA
- Higher 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, 146, 300, 750	73, 146, 300, 500, 1000	73, 146, 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
- FC and/or SAS/SATA backend
- Better latency

	HP EVA8100	LSI 6998	EMC CX3 model 80	HDS AMS1000	IBM DS4800
Backend	FC/8-ports	FC/8-ports	FC/8-ports or SATA	FC or SATA	FC/8-ports
Front-end	8	8	8	8	8
Cache size (GB)	8	16	16	16	16
Drive options (GB)	146, 300, 450, 500f, 1000f	73, 146, 300, 500s	73, 146, 300, 500s, 750s	73,146, 300, 500s, 750s	36, 73, 146, 300, 500f, 750f



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



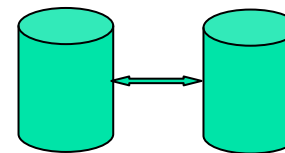
Take-aways

- Drive performance function of
 - Seek time
 - Rotation rate=rotational latency
 - Sustained transfer rate
- Subsystem performance function of
 - Class of subsystem
 - Drives, interfaces, cache size
 - Sophistication

RAID levels

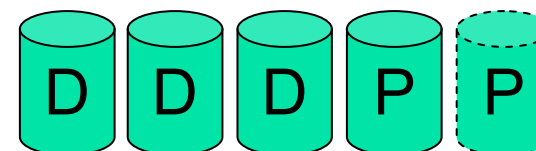
■ RAID-1 - mirrored data

- Reads use closest seek
- Writes both, 2nd destaged later
- Reads split across 2X drives
- Fastest response time



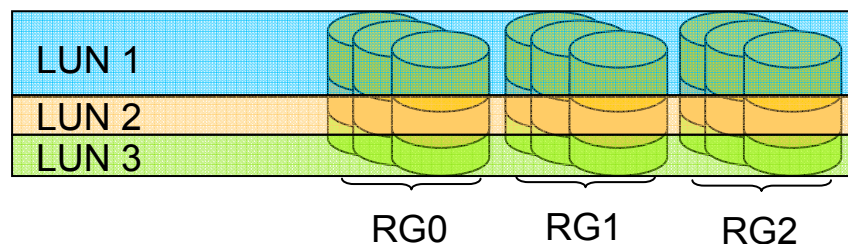
■ RAID-3, 4, 5, 6, DP - parity + data blocks

- Parity write penalty
- RAID 5, 6, & DP distributed/rotated parity
- RAID 3& 4 single parity drive (potential hot drive)
- RAID 6 & DP two parity drives, RAID 5 has one
- Throughput ok



LUN striping

- LUN striping - LUNs stripped across RAID groups (same type)
 - Eliminates hot RAID groups
- Called metaLUN striping, Vraid-0, -1, -5, RAID 0+1, 1+0, 3+0, 5+0, 10, 30, 50, 60
- Also comes with thin provisioning



I/O Balance

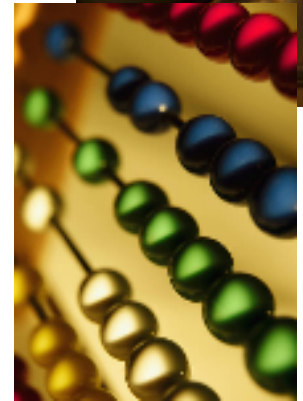
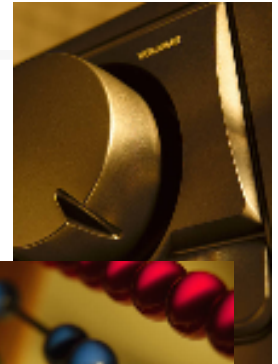
LUN I/O activity spread

- 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



Cache Parameters

- 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



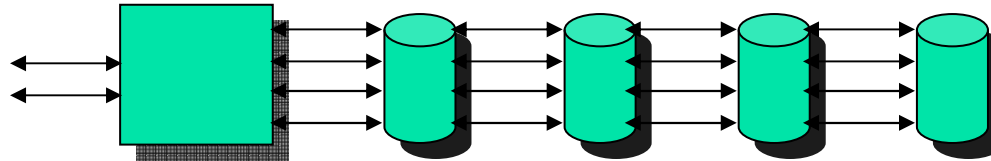
Drive Limits

Drive count and speed limit
subsystem I/O rate



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

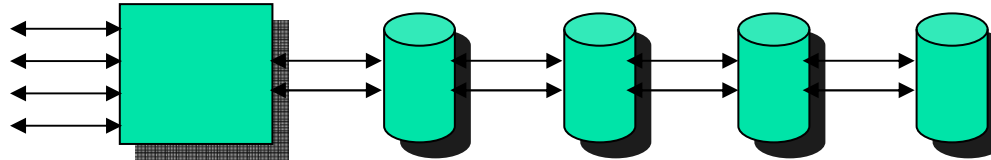
Front-end Limits



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&availability dictates front-end links
 - performance should be considered

Back-end Limits



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

- Cache miss is backend I/O
 - Write hits/destage 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

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 stink
 - Each random I/O processes only small amount of data, large transfers \Rightarrow wasted data
- Real workloads mixed
 - seldom pure sequential or random




Take-aways

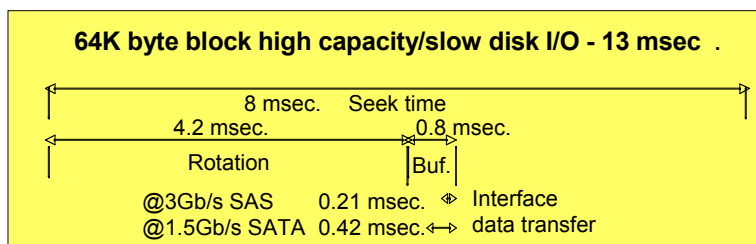
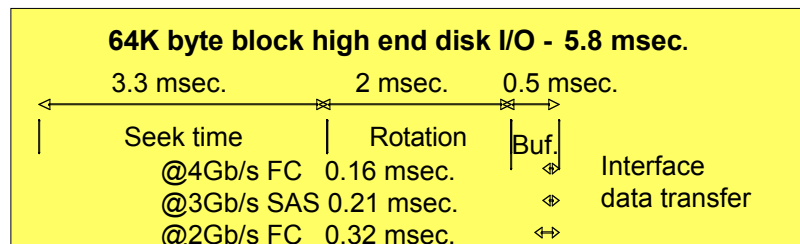
- Subsystem configuration affects performance
 - RAID level
 - LUN striping
 - I/O balance
 - Transfer size
- Subsystem performance is limited by
 - Drive speed
 - Number and speed of front- and back-end interfaces

Pre-purchase decisions



- Drives (count and performance)
 - Performance cost 50% more (\$/GB)
 - Enterprise - midrange cost differential
 - Subsystem sophistication cost differential, Enterprise class subsystems ~\$30/GB, Midrange = ~\$20/GB, Entry = ~\$10/GB
 - Cache size and sophistication
 - 2X cache ~10% more readhits
 - Interfaces front-end and back-end (type/speed and number)
- 

Drive speed

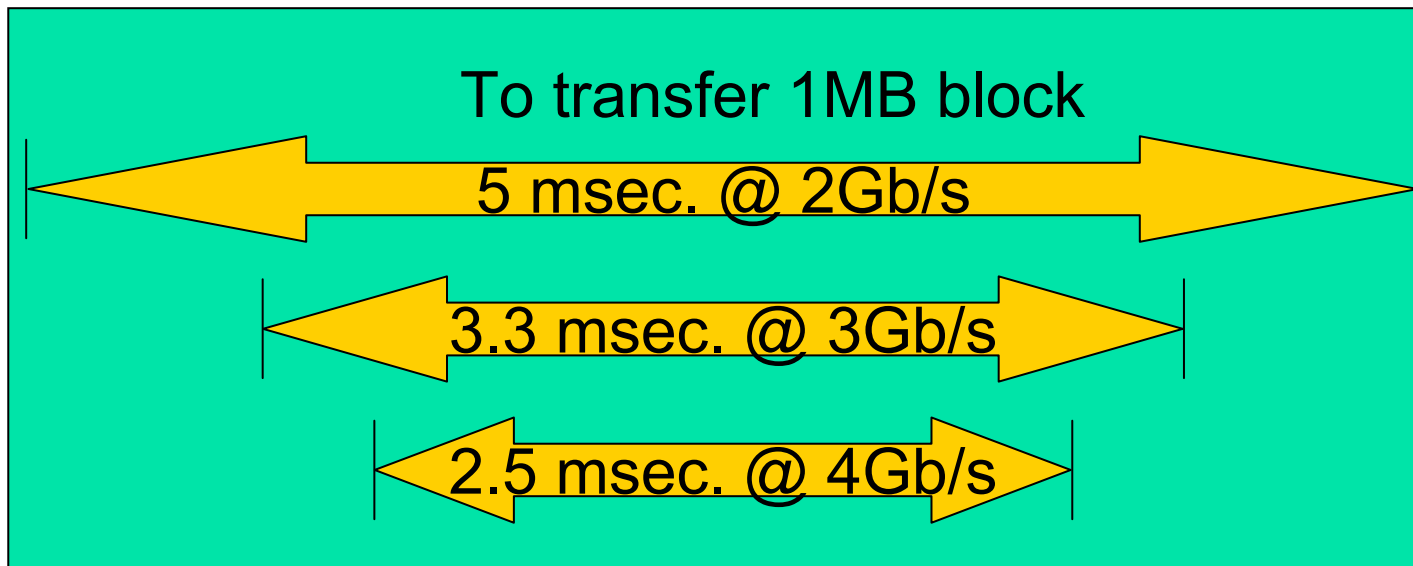


- Fast drive improves miss and destage 8.1 vs. 15.3 msec.
 - Assuming no other bottlenecks
- High capacity/slow drives degrade miss/destage
 - Response time concern, subsystem sophistication masks throughput impacts for non-busy drives

Transfer speed

Burst data rates <> sustained transfer rates

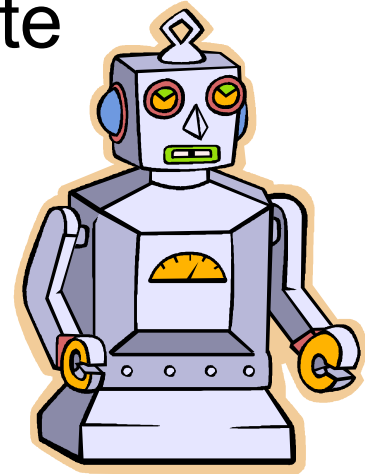
- Ethernet 0.1 to 10Gb/s - front-end only
- Fibre channel 1 to 8Gb/s front or back-end
- SCSI Ultra 320 3.2Gb/s front or back-end
- SAS/SATA 1.5 to 6Gb/s - back-end or DAS



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

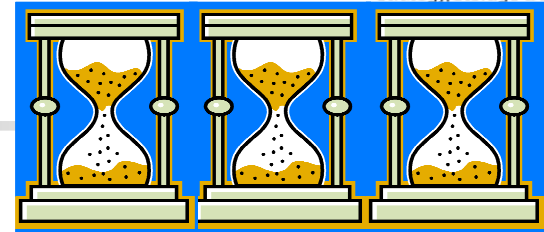
- Ethernet at 50-85% vs. FC at 90% of sustained rated capacity
- Ethernet 1Gb/s vs. FC 4Gb/s
- TCP/IP stack overhead for vs. HBA FC h/w
- iSCSI continuum from desktop NIC to iSCSI HBA
 - iSCSI HBA costs ~= FC HBA
 - Use server class NICs
 - Jumbo frames, q-depth level, separate storage LAN/VLAN
 - More hints on iSCSI storage deployment at <http://www.demartek.com/>



NFS/CIFS vs. block I/O

- NFS/CIFS Performance \approx block I/O
 - Latency/response time not same
- # directory entries/mount point
- Gateway vs. integrated system
- Single vs. parallel vs. cluster vs global file systems
- No central repository for NetBench CIFS benchmarks
- JetStress benchmarks at ESRP

SAN performance



- ISL and FC link oversubscription
 - Fan-in ratio 5:1 to 15:1 server to storage ports
 - Virtualization 40:1
- Hop counts
- Locality

Configuration Time

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





■ Server side

- Multi-path for performance and availability
- HBA configuration matches subsystem
 - Host transfer size > or = subsystem
 - Qdepth
- 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

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



Free monitoring tools



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
  
```

Free 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
17:20:51	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





O/S monitoring tools

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



Database monitoring tools

DB2/UDB	DB2 performance monitor
Informix	onpladmin
MS SQL	SQL performance monitor
MySQL	mysqladmin
Oracle	STATSPACK
PostgreSQL	Pq_statio

Subsystem monitoring tools

Dot Hill	SANscape
EMC	ControlCenter Performance Manager
LSI	Storage Performance Analyzer
HDS	Hi-Command Tuning Manager
HP	Storage Essentials SRM Performance Pack
IBM	TotalStorage productivity center

Midrange cache mirroring

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





Remote replication

Also called Remote Mirroring - duplicates data written to local on remote subsystem

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



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

Also called Snapshot

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

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&Treo users





Database I/O

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

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

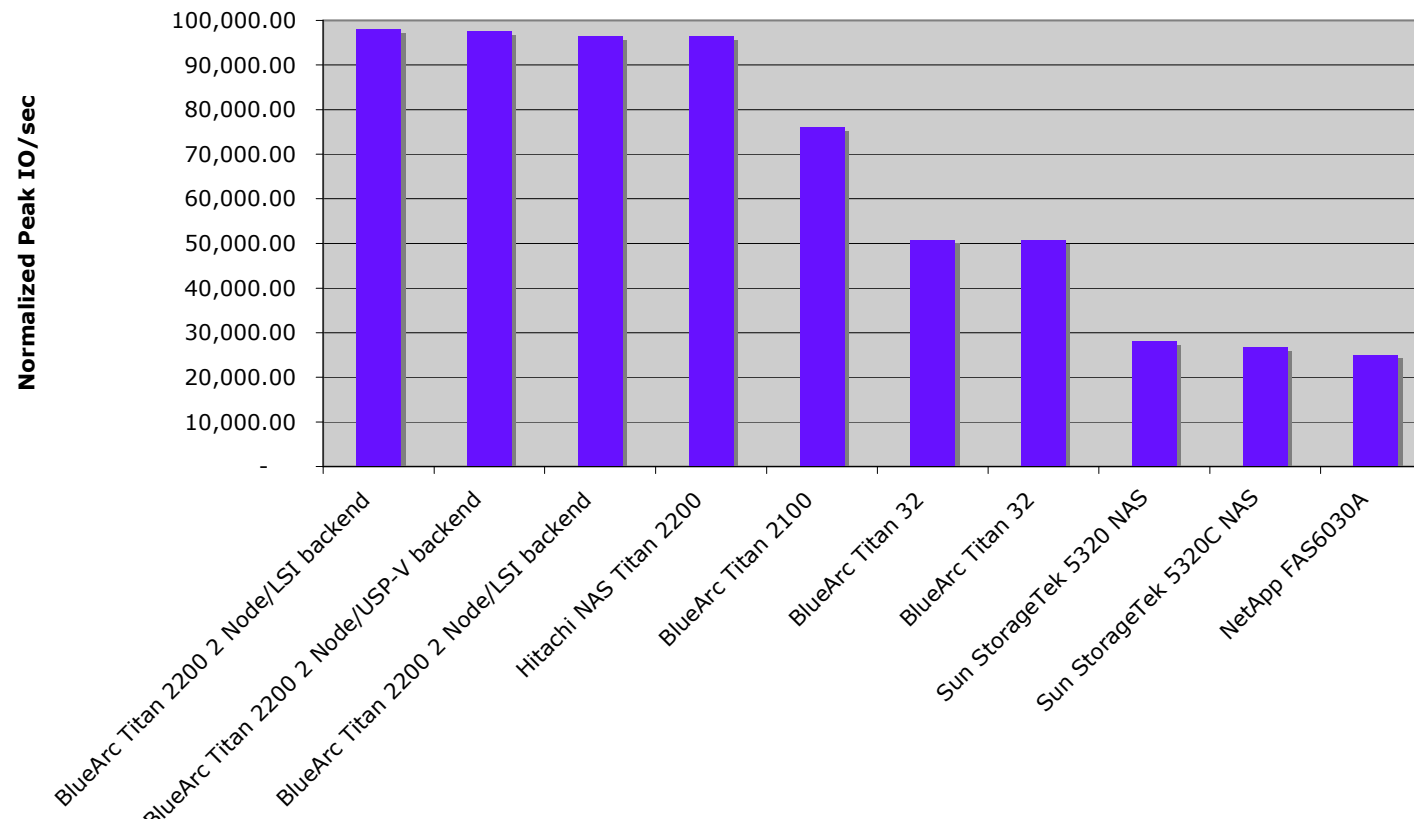


Take-aways

- Time to consider performance is early and often
- Purchase establishes I/O performance limits
 - Configuration impacts how well subsystem performs within limits
 - Monitor subsystem for performance problems
 - Application specific configurations help I/O performance

SPEC* NFS normalized

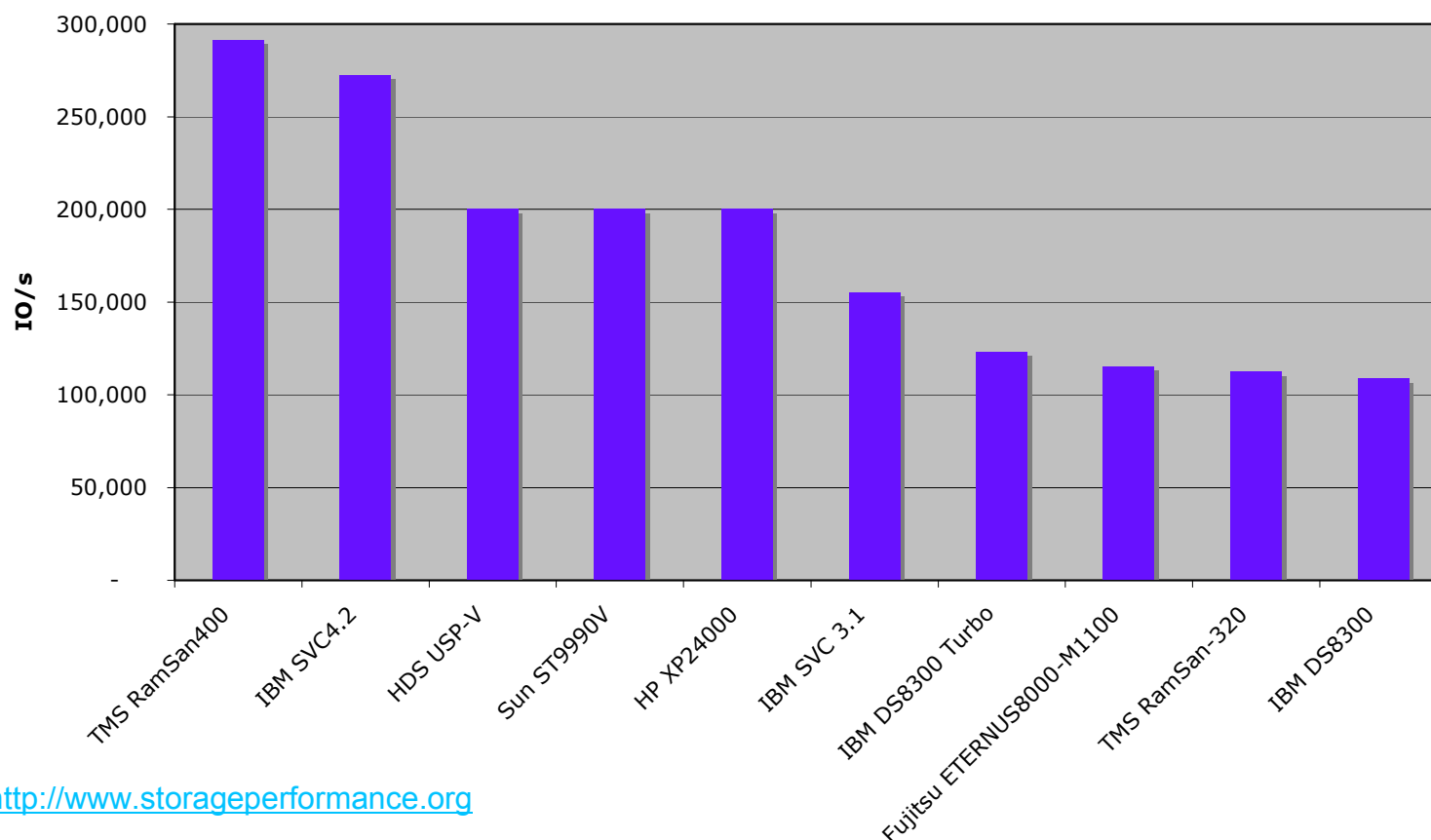
SPEC* SFS97_R1 NFS V3 Normalized Results as of 05 March 2008, TCP results only, Normalized by processors (chips or cores), Top 10 overall performers



*All SPEC SFS results Copyright © 1995-2008 Standard Performance Evaluation Corporation (SPEC). All rights reserved, permission granted for use, data from <http://www.spec.org> as of 05 March 2008

SPC-1* IOPS™

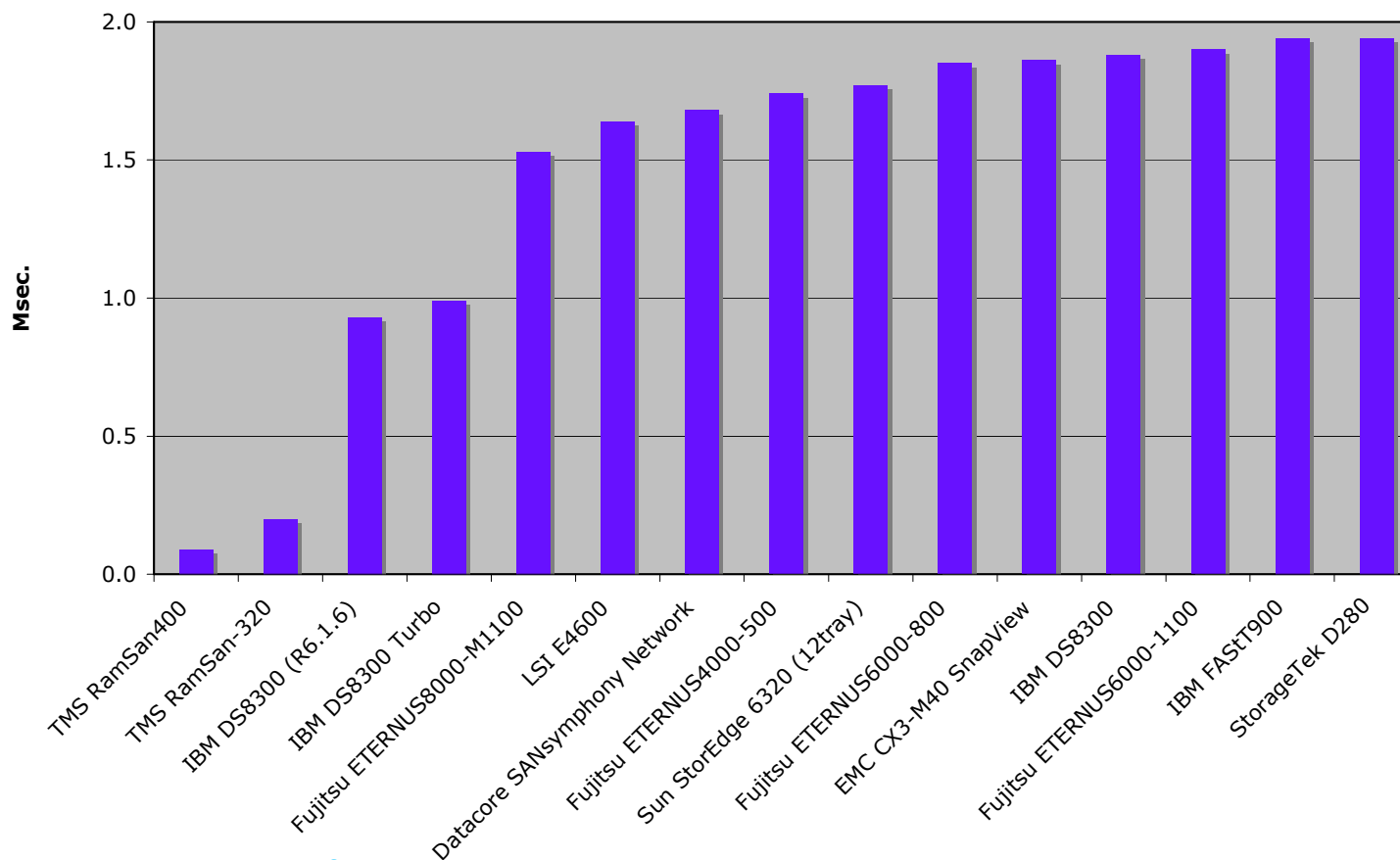
Top 10 SPC-1* IOPS™ performance as of 14 Feb 2008



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

SPC-1* LRT™

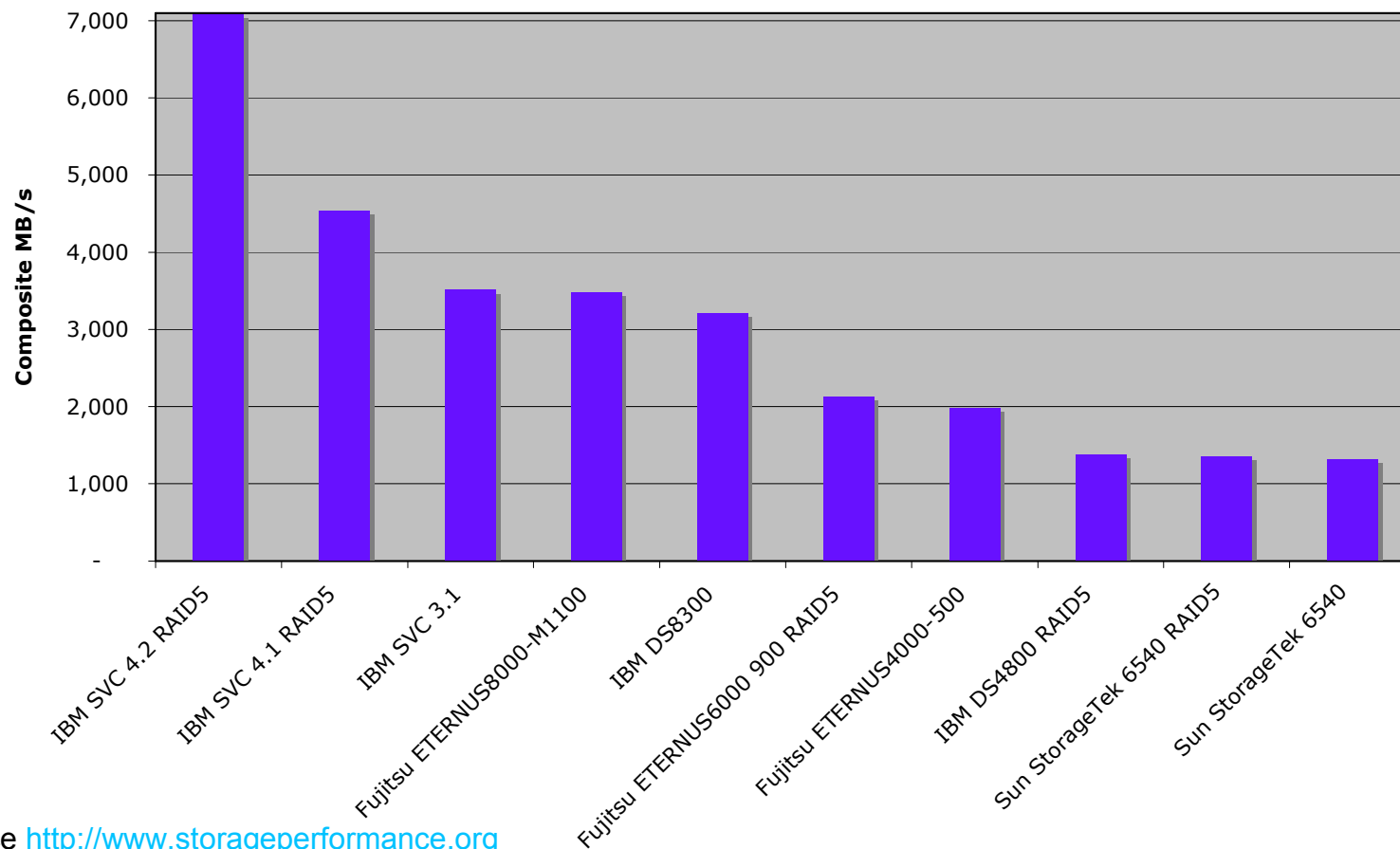
Top 15 SPC-1* LRT™ performance as of 14 Feb 2008 -



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

SPC-2* MPBS™

Top 10 SPC-2* MBPS™ performance as of 14 Feb 2008



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

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



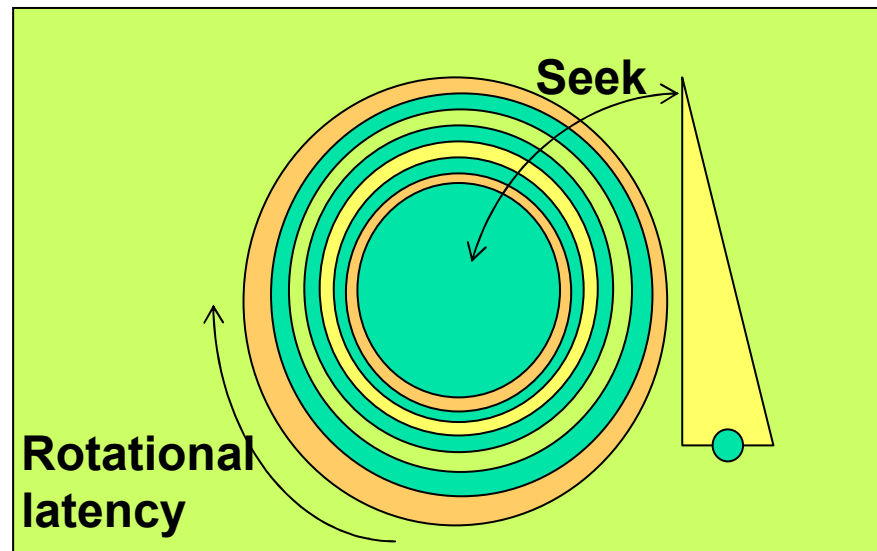
Background Information

Disk array terminology

- **Direct attached storage (DAS)**
- **Storage Area Network (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
- **Destage** - data written from cache 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