



Education

Storage Performance 101

Ray Lucchesi, President
Silverton Consulting, Inc.

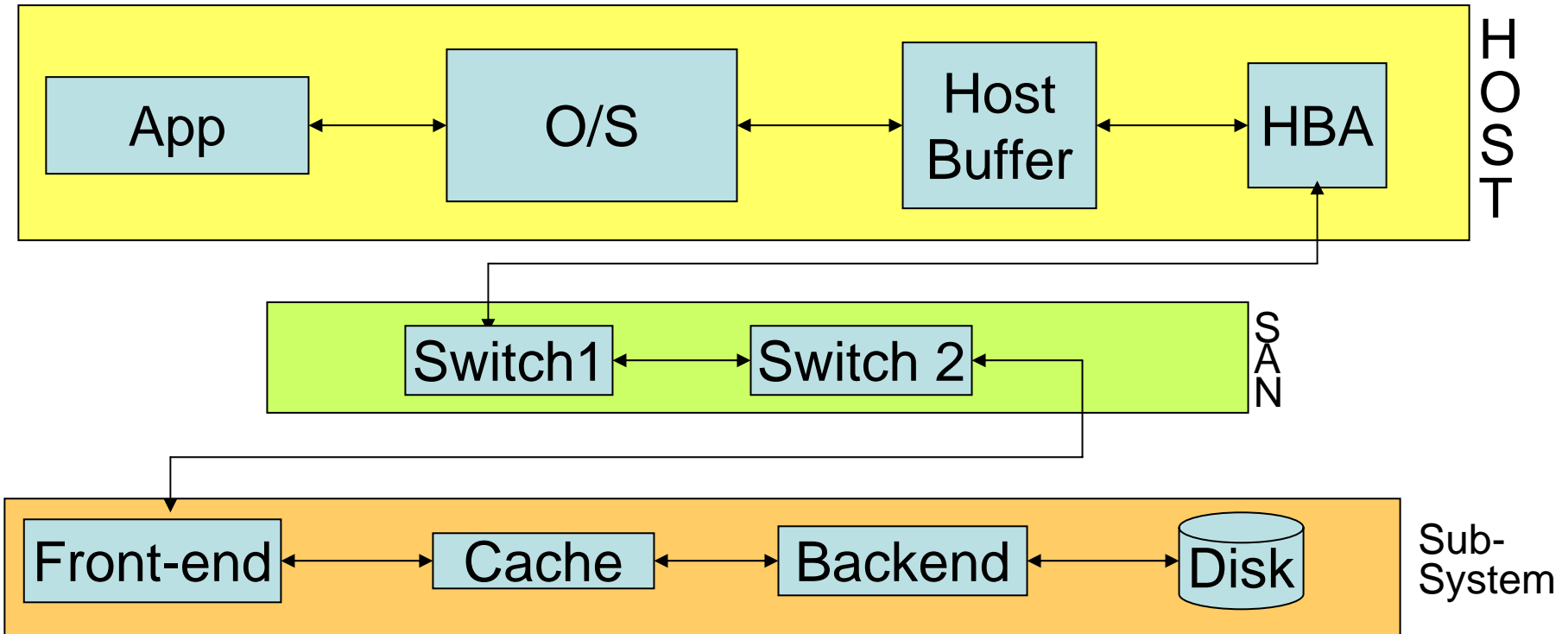
- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

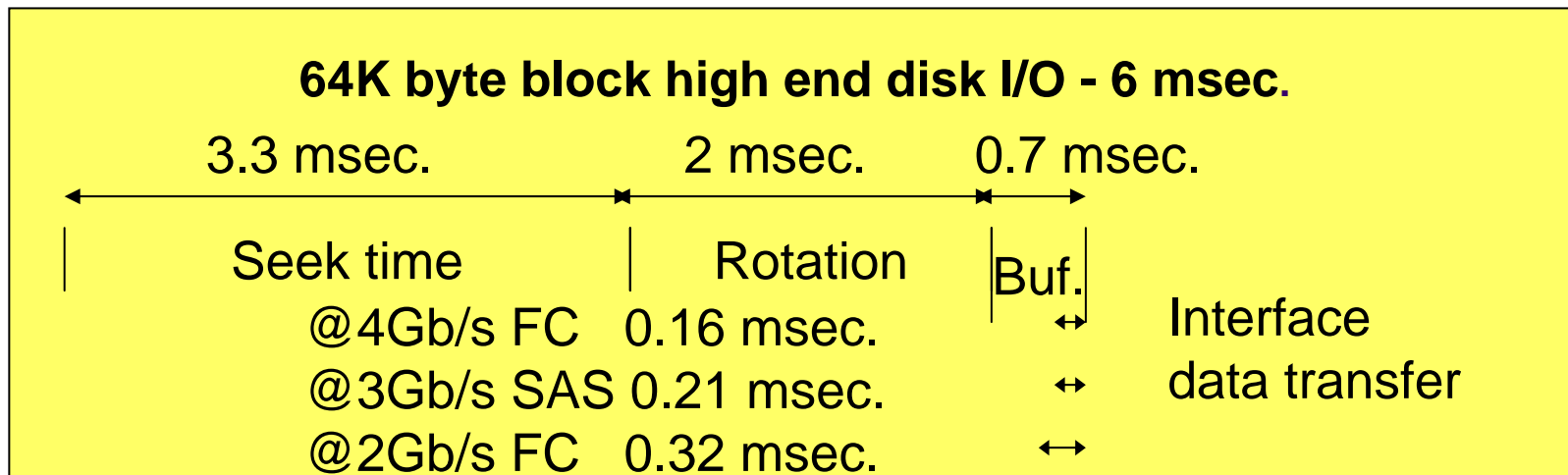
Storage Performance 101

This tutorial is an introduction to storage array performance tuning and should be a first step for anyone working to improve data storage performance.

Most data centers are dealing with ever increasing amounts of data storage. Although the need for storage seems insatiable array performance typically plateaus or worse, degrades post installation. Storage performance tuning is a teachable, ongoing activity. In addition, the vocabulary and activities are something any administrator should be able to master within a short time.

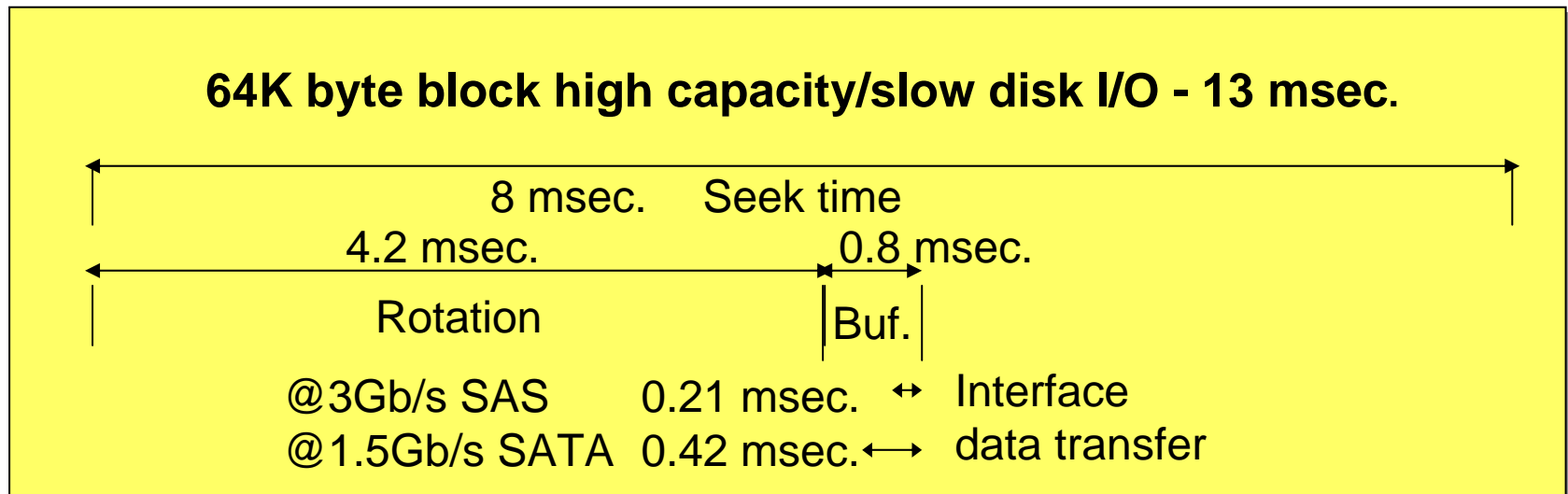
I/O Journey





- Read seek times from 3.3 to 3.6msec
- Write seek times from 3.8 to 4.0msec
- Rotational speed 15KRPM
- Sustained data xfer from 93 to 125MB/s
- Capacity from 36 to 300GB

Slow Disk I/O



- Read seek times ~8.5msec
- Write seek times ~9.5msec
- Rotational speed 7.2KRPM
- Sustained data xfer from 65 to 72MB/s
- Capacity from 250 to 750GB

64K byte block cache I/O 0.2 msec.



- Need to add subsystem overhead
~2.4msec - shown 1/2 in front and 1/2 at end
- Overhead must be added to bare disk I/O times above.

Transfer rates

- Fibre channel 4Gb/s, 2Gb/s, 1Gb/s - front-end or backend
- SCSI Ultra 320 (3.2Gb/s) - front-end or backend
- Ethernet 10Gb/s, 1Gb/s, 100Mb/s - front-end only
- SAS/SATA 3Gb/s, 1.5Gb/s - backend only or direct attached storage

Enterprise Class Subsystem

Larger and better cache, more front-end & backend interfaces, more drive options

- Local and remote replication options
- High availability
- Better throughput
- Cache size ~256GB
- Front-end from 64 to 224 FC interfaces
- Drive options from 73 to 500GB

Midrange Class Subsystem

More drive options but cache size, front-end, and back-end limitations

- Typically less replication options
- Typically less availability options
- Typically better response time
- Backend SAS/SATA II and/or FC interfaces
- Front-end 8 FC interfaces per dual controllers
- Cache from 1 to 16GB
- Drives from 73 to 750GB, 7.2 to 15KRPM

Just a bunch of disks

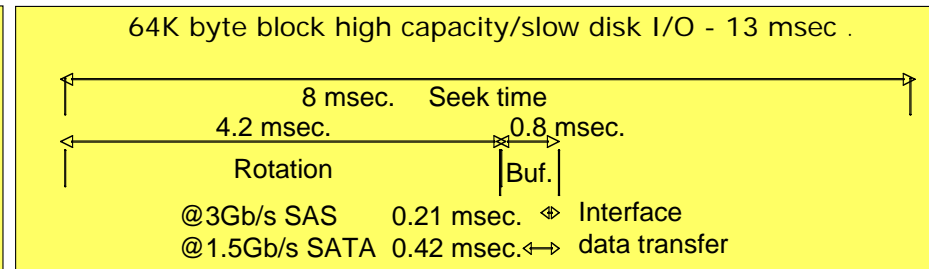
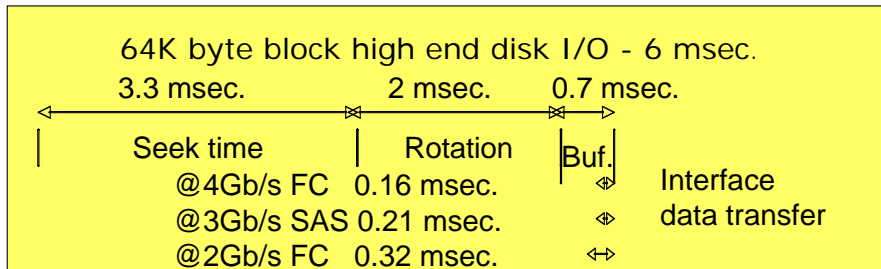
- Direct attached storage - SATA/SAS, SCSI Ultra 320, or FC/AL
- RAID either S/W or HBA based
- Only disk buffer and host buffer cache for cache-like I/O

64K byte block cache I/O 0.2 msec.



- Cache hit, the fastest way to do I/O 2.6 vs. 8.4 msec.
- Larger cache helps, but
 - ◆ Write hit de-staged to disk, impacts sequential throughput
 - ◆ Writing disk directly faster for sequential
- Sophistication matters

Drive speed

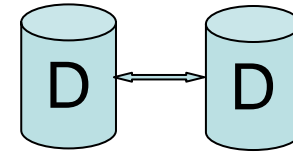


- Fast drives helps miss and destage activity 8.4 vs. 15.4 msec.
 - ◆ Assuming no other bottlenecks
- High capacity/slow drives degrades miss/destage performance seriously
 - ◆ Response time concern, subsystem can mask throughput impacts with non-busy drives

RAID and Striping

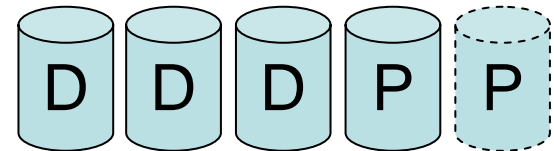
➤ RAID-1 - mirrored data

- ◆ Reads use closest seek
- ◆ Writes both, 2nd destaged later



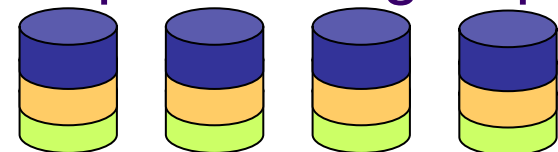
➤ RAID-4, 5, 6, DP - parity + data blocks

- ◆ Parity block write penalty
- ◆ RAID 5, 6, & DP parity block(s) distributed
- ◆ RAID 4 lone parity drive (hot drive)
- ◆ RAID 6 & DP two parity drives, RAID 5 has one



➤ LUN striping - LUN striped across multiple RAID groups (of same type)

- ◆ Eliminate hot RAID groups



LUN I/O activity spread

- Across RAID groups
- Across Front-end interfaces/controllers
- Across Back-end interfaces
- Application/workload mix - toxic workloads reduce cache hits

Cache Revisited

- Cache read-ahead - insures follow-on sequential requests from cache.
 - ◆ Some subsystems compute value in real-time
 - ◆ Others specify (consider cache demand at time of I/O)
- Cache read to write boundary -
 - ◆ Some have a hardware boundary
 - ◆ Others specify boundary (sized on average or peak write workload)

Remote Replication

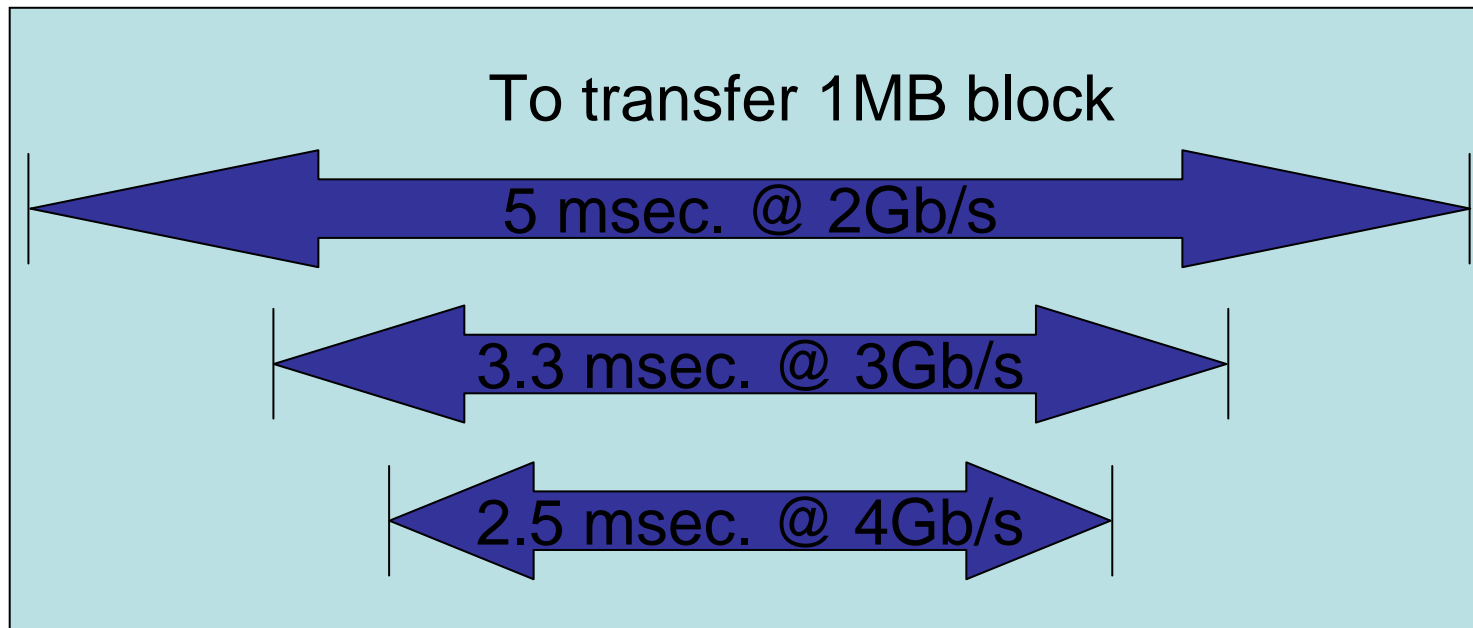
Remote replication - mirrors data written on one subsystem to remote subsystem

- Synchronous - write performance degrades
- Semi-synchronous - remote site data behind primary site
- Asynchronous - data duplication scheduled only correct at end of activity
- Enterprise vs. midrange - cache use vs. backend use

Transfer size

- For sequential, the larger the better
 - ◆ Most requests generate seek, rotation & transfer, bigger transfers, less I/Os per file
 - ◆ Each transfer adds 2.4 msec overhead, less I/O, less overhead, better throughput
- For random I/O, larger transfers don't work as well
 - ◆ Each random I/O processes only small amount of data
- Real workloads always mixed

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



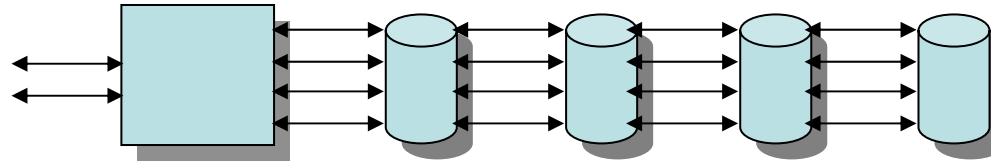
Midrange Cache Mirroring

- Each write adds transfer (between controllers)
- Performance depends on transfer size and link speed between controllers

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

- P-I-T copy - copies data locally for backup and test purposes
- Copy-on-write needs cache, disk, and/or other resources for write

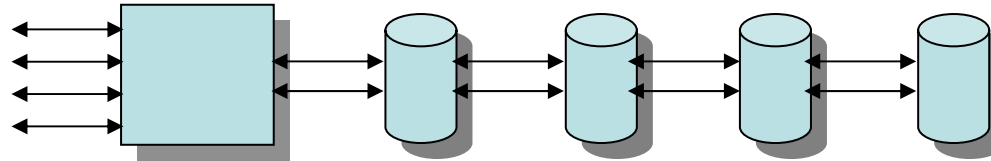
Front-end Limits



Front-end interfaces can limit performance

- FC achieves ~90% of rated speed, 2Gb/s= \sim 180MB/s per FC link
- iSCSI achieves ~50-85% of rated speed, 1Gb/s=50 to 85MB/s per iSCSI link
- Network connectivity often dictates number of front-end links but performance requirements should be considered

Backend Limits



Backend 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 has more throughput per drive vs sharing across FC/AL
- SAS backend - point-to-point

Drive Limits

Number and speed of drives can limit I/O performance

- Upper limit to the number of I/Os one drive can do
 - ◆ Faster drives do more
- Compute max drive I/O and know peak miss/destage workload to check limits

Pre-purchase decisions

- Drives (number and performance level)
 - ◆ Performance drives cost 50% more (\$/GB)
- Interfaces front-end and possibly backend (type, number, and transfer speed)
- Cache size and sophistication
 - ◆ 2X cache size for 10% readhit improvement

Configuration Time

- RAID level
- LUN Striping
- Fixed cache parameters - look ahead, cache mirroring, read to write partition boundary
- I/O balance - across LUNs, RAID groups, front-end & back-end interfaces, controllers
- Subsystem partitioning - cache, interfaces, drives (RAID groups)

Host Side

- HBA configuration matches subsystem
 - ◆ Host transfer size should be $>$ or $=$ subsystem
- Host buffer cache for file system I/O
 - ◆ Write-back vs. write-thru
 - ◆ Sync's for write-back
 - ◆ May use all free host memory
 - ◆ Database cache, host buffer cache, and subsystem cache interaction
- Multi-path I/O

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

- Many email servers use multiple databases
 - ◆ One database stores MAPI clients data
 - ◆ One database stores attachments (ptrs from above)
 - ◆ One database is a transaction Log
- Isolate each database to own set of LUNs
 - ◆ Transaction log should be separate from other two
- Email I/O besides reading&writing mail
 - ◆ Beware of BlackBerry and other push users

Database I/O

- Separate log files from table spaces
- Indices from the table spaces
- For heavy sequential DB access use larger transfer sizes
- For heavy random DB access use smaller transfer sizes

Ongoing Workload Monitoring

What to look for

- OS, database, or subsystem specific tools
- Overall I/O activity to subsystem LUNs
- I/O balance over controllers, RAID groups, LUNs
- Read and write hit rates
- Sequentiality vs. random workload mix toxicity

Workload Monitoring Tools

➤ IOSTAT

```
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
```

SAR

```
/usr/bin/sar -d 15 4
AA-BB gummo A.08.06 E 9000/??? 02/04/92
17:20:36      device  %busy  avque r+w/s blks/s avwait 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
```

Performance Automation

Some enterprise subsystems can automate performance tuning for you

➤ LUN balancing

- ◆ Across RAID groups
- ◆ Across controllers/front-end interfaces

➤ Cache hit maximization

- ◆ Read ahead amount
- ◆ Read:write boundary partitioning

➤ Others

- Ethernet at typically at 50-85% vs. FC at 90% of sustained rated capacity
- Ethernet 1Gb/s vs. FC 2-4Gb/s
- Processor overhead for TCP/IP stack, TOE vs. HBA handling FC protocol overhead
- iSCSI hints
 - ◆ iSCSI HBAs, Server class NICs or Desktop 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 ~same as block I/O
- # Directory entries/Mount point
- Gateway vs. integrated system
- Parallel vs. cluster vs global file systems
- No central repository for NetBench CIFS benchmarks

What Price Performance?

- Drive cost differential 50% more (\$/GB) for faster drives
- Enterprise - midrange cost differential
 - ◆ Enterprise ~\$30/GB,
 - ◆ Midrange ~\$20/GB
 - ◆ Entry ~\$10/GB
- Cache size differential 100GB's or more for Enterprise vs. 10GB or less for midrange.

Performance Benchmarks

- For NFS results raw data available at <http://www.spec.org/osg/sfs97r1/results/>
- For block storage results raw data available at http://www.storageperformance.org/results/benchmark_results_all
- For summary charts and analysis of both NFS and block performance see my Performance Results StorInt™ Dispatch available at <http://www.SilvertonConsulting.com/>

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 (CMG) - more than just storage performance www.cmg.org
- Storage Networking Industry Association (SNIA) - standards with performance info www.snia.org
- Silverton Consulting - StorInt™ Briefings & Dispatches, articles, presentations and pod casts www.SilvertonConsulting.com

- Please send any questions or comments on this presentation to SNIA: track-storage@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.**

SNIA Education Committee

Ray Lucchesi

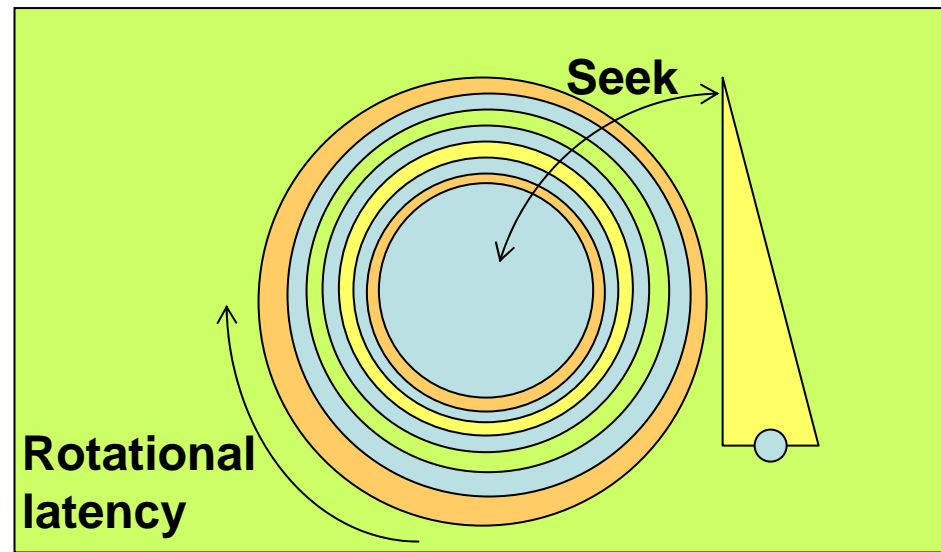
Background Information

Disk Array Terminology

- 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) internally attached disks

Disk Terminology

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



Cache Terminology

- Cache read hit - read request finds data in cache
- Cache write hit - write request writes to cache instead of disk, later destaged to backend
- Cache miss - either a read or write that uses disk to perform I/O
- Cache read ahead - during sequential reads, reading ahead of where I/O requests data

IO Performance Terminology

- ▶ Throughput - bytes transferred over time (MB/s or GB/s), also measured in I/O operations per second
- ▶ Response time - average time to do I/O (msec.) includes seek, rotation, and transfer
- ▶ Sequential workload - multi-block accesses in block number sequence
- ▶ Random workload - no discernible pattern to block number accesses

FC	Fibre channel	LUN	Logical unit number
FC/AE	Fibre channel arbitrated Ethernet	MBps	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 disks
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