Storage Performance 101

Ray Lucchesi, President Silverton Consulting, Inc.
Ray@SilvertonConsulting.com
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.

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 necessary and ongoing activity. In addition, the vocabulary and activities are something any administrator should be able to master within a short time.
I/O Journey

Application, web, or database servers

Storage subsystems:

Application
O/S
Data Buffer
HBA

Switch

Storage Area Network (SAN)

Front-end
Cache
Backend

Switch

Storage Performance 101
© 2008 Storage Networking Industry Association. All Rights Reserved.
Storage performance terms

- **Throughput** - bytes transferred over time (MB/s or GB/s),
- **IOP/s** - maximum I/O operations per second
- **Response time** - average time to do I/O (millisecond) for drive I/O includes seek, rotation, and transfer
- **Cache read/write hit** - read request finds data in cache, write directed to cache.
- **Destage** - write hit data later flushed from cache and written to backend disk
- **Cache miss** - either a read or write that goes directly to disk to perform I/O
Fast Disk I/O

Seek time | Rotation | Buf. | Interface | data transfer
---|---|---|---|---
3.3 msec. | 2 msec. | 0.7 msec. | @4Gb/s FC 0.16 msec. | @3Gb/s SAS 0.21 msec. | @2Gb/s FC 0.32 msec.

- Read seek times from 3.3 to 3.6msec
- Write seek times from 3.8 to 4.0msec
- Rotational speed 15KRPM
- Sustained data transfer from 93 to 125MB/s
- Capacity from 36 to 300GB
- Good for miss and destage activity
## Slow Disk I/O

- **64K byte block high capacity/slow disk I/O - 13 msec.**
  - Seek time: 8 msec.
  - Rotation: 4.2 msec.
  - Buffer: 0.8 msec.
  - Interface data transfer:
    - @3Gb/s SAS: 0.21 msec.
    - @1.5Gb/s SATA: 0.42 msec.

- Read seek times ~8.5msec, Write seek times ~9.5msec
- Rotational speed 7.2KRPM
- Sustained data transfer from 65 to 72MB/s
- Capacity from 250GB to 1TB
- Ok for throughput with subsystem sophistication
Cache I/O

- Add subsystem overhead ~2.4msec to transfer
- Must add overhead to disk I/O times above
- Larger cache helps
- Algorithm sophistication matters

64K byte block cache I/O 0.2 msec.

1.2 msec. 1.2 msec. SubSys Overhead

@4Gb/s FC 0.16 msec.
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
- Back-end typically FC, with lots of links
- Drive options from 73 to 500GB
More drive options but cache size, front-end, and back-end limitations

- Less front-end links, up to 8 FC links per dual controller
- Less backend links
  - Backend SAS/SATA II, with occasional FC interfaces
- Less cache from 1 to 16GB
- Typically
  - Less replication options
  - Less availability options
- Controller to controller interconnect usually FC but some Infiniband
- Lots of Drive options from 73 to 750GB, 7.2 to 15KRPM
- Usually better response time
RAID Levels

- **RAID-1 - mirrored data**
  - Reads use closest seek
  - Writes both, 2nd destaged later
  - Fastest response time but costs

- **RAID-4, 5, 6, DP - parity + data blocks**
  - Parity block write penalty
  - RAID 4 lone parity drive (hot drive on writes)
  - RAID 5, 6, & DP parity block(s) distributed
  - RAID 6 & DP two parity drives, RAID 5 has one
  - Ok throughput
LUN Striping

Logical Unit Number (LUN) = host’s drive #/letter

- LUNs striped across multiple RAID groups (of same type)
  - Eliminate hot RAID groups, hot drives, hot backend links
- Called RAID 0+1, 1+0, 5+0, 10, or 50
- Also available with thin provisioning
I/O activity should be spread equally

- Across LUNs - no hot LUNs
- Across RAID groups - no hot drives
  - Across Back-end interfaces
- Across Front-end interfaces/controllers - no hot interfaces/controllers
- Typically 35-55% of all subsystem I/O on a single LUN
Cache Revisited

- **Cache read-ahead** - insures follow-on sequential requests come out of 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)
For sequential I/O, the larger the transfer size the better

- Many IO requests generate seek, rotation & transfer, bigger transfers cause less I/Os per file
- Each transfer adds 2.4 millisecond overhead, less I/O means less overhead provides better throughput

For random I/O, larger transfers stink

- Each random I/O processes only small amounts of data

Real workloads always mixed

- Beware toxic workload mixes
Transfer speed

- Fibre channel 1 to 8Gb/s - front-end or backend
- Ethernet 0.1 to 10Gb/s - front-end only
- SAS/SATA 1.5 to 6Gb/s - backend only or direct attached storage
- SCSI Ultra 320 (3.2Gb/s) - front-end or backend

To transfer 1MB block:
- 5 msec. @ 2Gb/s
- 3.3 msec. @ 3Gb/s
- 2.5 msec. @ 4Gb/s
Number and speed of drives can limit subsystem performance

- Upper limit to the number of I/Os one drive can do
  - Faster drives do more

- Compute max drive I/O, multiply by number of data drives to determine peak miss/destage workload limit
Front-end interfaces can limit performance

- FC achieves ~90% of rated speed, 2Gb/s=\(\sim 180\text{MB/s}\) per FC link
- iSCSI achieves ~50-85% of rated speed, 1Gb/s=50 to 85MB/s per iSCSI link
- Availability and connectivity often dictate front-end links but performance requirements should be considered
Backend number of FC or SAS/SATA links also limits I/O

- Cache miss activity translates into backend I/O
  - Write hits followed by destage 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
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

➤ Enterprise - midrange cost differential
  ✷ Enterprise ~$30/GB,
  ✷ Midrange ~$20/GB
  ✷ Entry ~$10/GB
RAID level

LUN striping vs. manual I/O balancing

Fixed cache parameters - look ahead, cache mirroring, read to write partition boundary
  - For mid-range - cache mirroring, adds overhead to each write, but more fault tolerant

Transfer size

Subsystem partitioning - cache, interfaces, drives (RAID groups)
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 for performance and availability

Qdepth settings on LUN vs. RAID group basis

Partition - RAID group/stripe alignment
SAN Considerations

- ISL and FC link oversubscription
  - Fan-in ratio 5:1 to 15:1 server to storage ports
  - Even higher for VMware/Zen servers
- Hop counts
- Locality
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

Some free 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  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  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
```
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**
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

Local (point-in-time) replication

- Copy-on-write needs cache, disk, and other resources for each update to replicated data, persists until replica terminated
- Cloning does complete copy uses more resources, persists indefinitely
Most 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 push users
Databases typically have multiple files/LUNs for transaction logs, indices, and tables

- Isolate table spaces from log files and indices
  - Indices from log files
- For heavy sequential DB access use larger transfer sizes
- For heavy random DB access use smaller transfer sizes
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
NFS/CIFS vs. block I/O

- NFS/CIFS Performance ~same as block I/O
  - NFS/CIFS response time > block I/O response time
- # Directory entries/Mount point
- Gateway vs. integrated system
- Standard vs. parallel vs. cluster file systems
- Global vs. local name space considerations
- CIFS vs. NFS?
Performance Benchmarks

- For NFS results SpecSFS data available at [http://www.spec.org/osg/sfs97r1/results/](http://www.spec.org/osg/sfs97r1/results/)
- For summary charts and analysis of both NFS and block performance see my Performance Results StorInt™ Dispatch available at [http://www.SilvertonConsulting.com/](http://www.SilvertonConsulting.com/)

- Others without central repository of results
  - IOMETER - good for basic response time
  - VDBench - good for throughput
  - NETbench - good for CIFS
For More Information

- Storage Performance Council (SPC) block I/O benchmarks
  [www.storageperformance.org](http://www.storageperformance.org)
- Standard Performance Evaluation Corp. (SPEC) SFS NFS I/O benchmarks
  [www.spec.org](http://www.spec.org)
- Computer Measurement Group (CMG) - more than just storage performance
  [www.cmg.org](http://www.cmg.org)
- Storage Networking Industry Association (SNIA) - standards with performance info
  [www.snia.org](http://www.snia.org)
- Silverton Consulting - StorInt™ Briefings & Dispatches, articles, presentations and podcasts
Q&A / Feedback

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

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

SNIA Education Committee

Ray Lucchesi
Background Information
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

Sequential workload - multi-block accesses in block number sequence

Random workload - no discernible pattern to block number accesses
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,
  - Destage - process of writing cache data to disk
- 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
Acronyms

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