

SQLskills Immersion Event IE2: Performance Tuning

Module 3: IO – Storage Area Networks for DBAs

Jonathan Kehayias
Jonathan@SQLskills.com



Overview

- SAN components
- Multipathing
- iSCSI vs. fibre channel (FC)
- Shared vs. dedicated arrays
- Cache (on the drives, SAN controller, RAID controller, and Windows)
- Clustering
- Appendix: snapshot backups (SAN, not SQL)



What Do You Think of as a SAN?



Scale out IBM DS4800

OR



Basic Dell MD3000i

- Reality is: both are SANs and have very different capabilities!

What is a SAN?

- "A storage area network (SAN) is any high-performance network whose primary purpose is to enable storage devices to communicate with computer systems and with each other"
- (Storage Networking Industry Association)
 - Does not specify interconnect technology
 - Does not specify types of storage devices

SAN Advantages

- **Shared storage**
 - Increases disk utilization
 - Reduces management by making it easier to create new volumes and dynamically allocate storage
 - Create diskless servers that boot from SAN only
- **Advanced features**
 - Mirroring, snapshots, continuous data protection, clustering and geo-clustering only offered by SANs
- **Performance**
 - Almost unlimited number of spindles, controllers, and cache can be put together to meet the requirements

SAN Disadvantages

- **Unpredictable performance**
 - When you share your disks, controllers, and fibre switches between a dozen of servers is very difficult to have predictable performance
- **Higher latency**
 - Distance the I/Os have to travel; added layers of switches, cabling and ports
 - PCI Bus → HBA → FC switches → FC ports → array processors → disks
- **Limited bandwidth**
 - FC now 16Gb/s max
 - iSCSI 10Gb/s max
 - Note: 1 Gb/s = 128 MB/s
- **Cost**

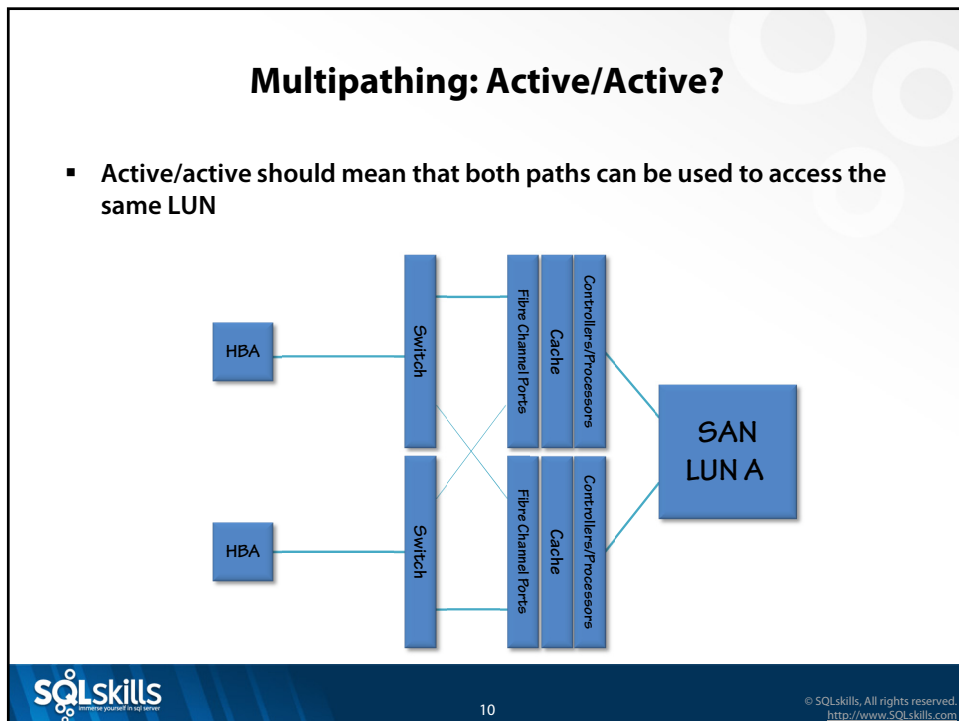
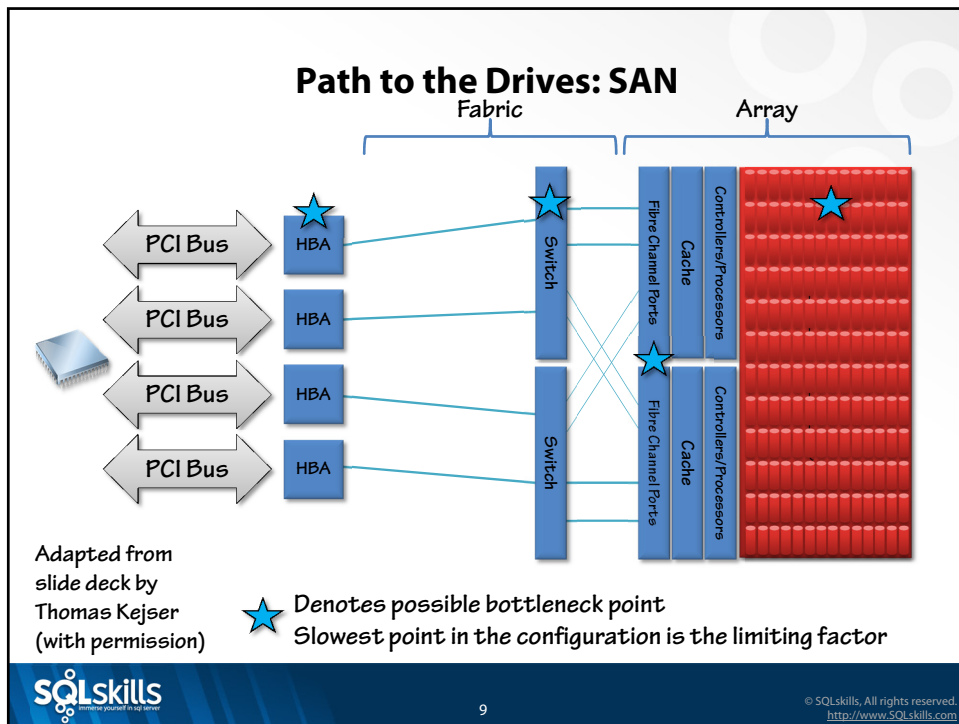
SAN Components

Basics

- **Drives**
 - Physical interface, size, rotation speed, and RAID configuration level
- **SAN storage controllers**
 - FA port/iSCSI port count and speeds
 - Cache size and configuration
- **SAN fabric**
 - Number of switches, their port speeds and path configurations
- **Host bus adapter**
 - Queue depth, port speeds, pathing/multipathing

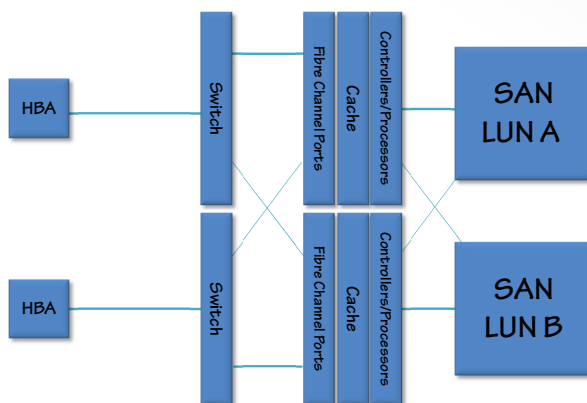
Multipathing

- **Primary purpose of multipathing configurations is to provide redundancy and protection**
- **Secondary to redundancy is to improve performance**
 - Just having multiple paths doesn't guarantee a performance improvement
 - Some SANs only allow one path per LUN
 - iSCSI allows one connection per target
 - Achieving performance improvements may require changes to the physical database layout



Pathing: Active/Active?

- However, in some configurations active/active actually means you have two paths to separate LUNs



Understanding SAN I/O Capacity

- The available bandwidth of a SAN is limited to the lowest available bandwidth bottleneck in the configuration
- Example scenario
 - HBAs: 2x 8Gb/s = ~1,600MB/s (max)
 - Switch: 2x 8Gb/s = ~1,600MB/s (max)
 - SAN has 4x 2Gb/s FA Ports = ~800MB/s (max)

Understanding SAN I/O Capacity

Scenario 1

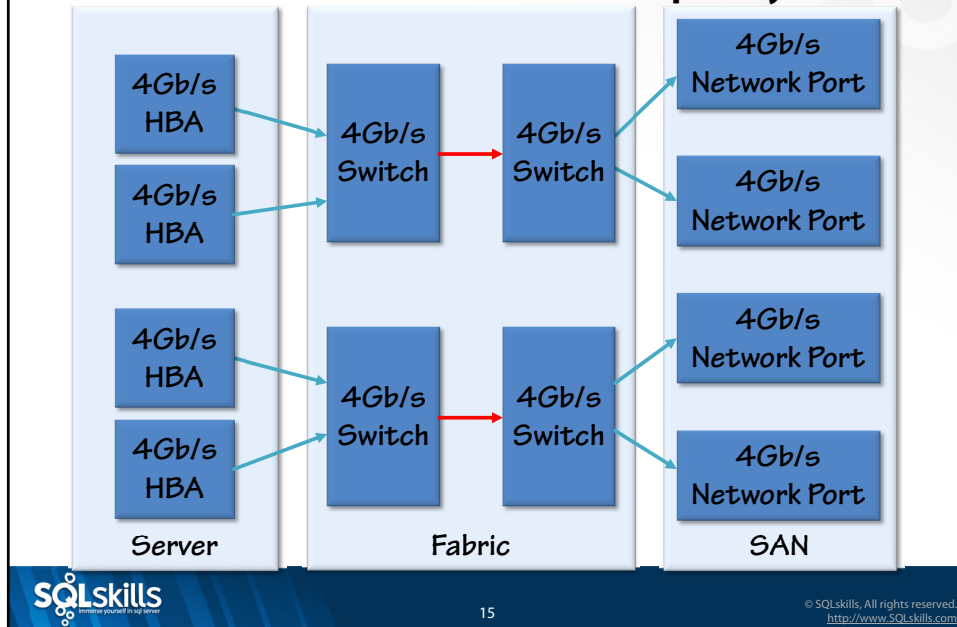
- HBAs: $2 \times 8\text{Gb/s} = \sim 1,600\text{MB/s}$ (max)
 - Switch: $2 \times 8\text{Gb/s} = \sim 1,600\text{MB/s}$ (max)
 - SAN ports: $4 \times 2\text{Gb/s}$ FA ports = $\sim 800\text{MB/s}$ (max)
 - Cache: $1 \times 1,600\text{MB/s} = \sim 1,600\text{MB/s}$ (max)
 - Disks: $80 \times 20\text{MB/s} = \sim 1,600\text{MB/s}$ (max)
-
- Maximum bandwidth is limited by the SAN ports

Understanding SAN I/O Capacity

Scenario 2

- HBAs: $2 \times 8\text{Gb/s} = \sim 1,600\text{MB/s}$ (max)
 - Switch: $2 \times 8\text{Gb/s} = \sim 1,600\text{MB/s}$ (max)
 - SAN ports: $4 \times 8\text{Gb/s}$ FA ports = $\sim 3,200\text{MB/s}$ (max)
 - Cache: $1 \times 1,600\text{MB/s} = \sim 1,600\text{MB/s}$ (max)
 - Disks: $40 \times 20\text{MB/s} = \sim 800\text{MB/s}$ (max)
-
- Maximum bandwidth is limited by the disks

Understand Bottlenecks Completely



General iSCSI vs. Fibre Channel

- **iSCSI**
 - Transfer rates
 - 10Gb/s or 1Gb/s
 - Disk connections
 - SAS limited to 1.5Gb/s, 3Gb/s, or on newer configurations 6Gb/s
- **FC**
 - Transfer rates
 - 16Gb/s, 8Gb/s, 4Gb/s, 2Gb/s and 1Gb/s
 - Disk connections
 - FC limited to 8Gb/s, 4Gb/s, 2Gb/s and 1Gb/s
 - SAS limited to 1.5Gb/s, 3Gb/s, or on newer configurations 6Gb/s

iSCSI vs. Fibre Channel

Best Practices

- **Use the fastest interface you can:**
 - 16Gb/s FC, 10Gb/s iSCSI, 8Gb/s FC, 4Gb/s FC, 2Gb/s FC, 1Gb/s iSCSI or FC
- **The slower the interface the more important appropriate multipathing becomes to achieving the needed throughput**
- **Know the limitations of the controller's connections**
 - You can probably connect more than you are currently using

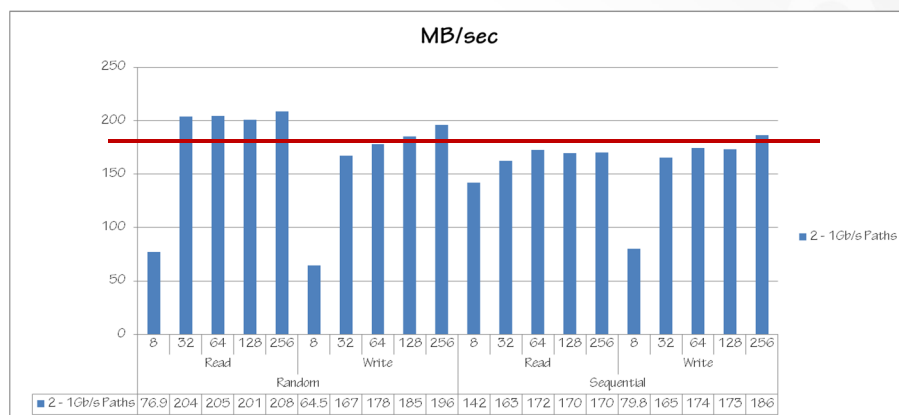
iSCSI Considerations

- **Network design**
 - Dedicated hardware including HBAs, switches, and cabling
 - TCP Offload Engine (TCOE) enabled NICs or iSCSI HBAs for servers
 - Use Jumbo Frames end to end on iSCSI network to improve throughput for larger packet sizes
- **Single path per target may require considerations for LUN and database file layout to achieve the optimal I/O configuration**
- **May require intervention to bring multiple paths back online after failure**

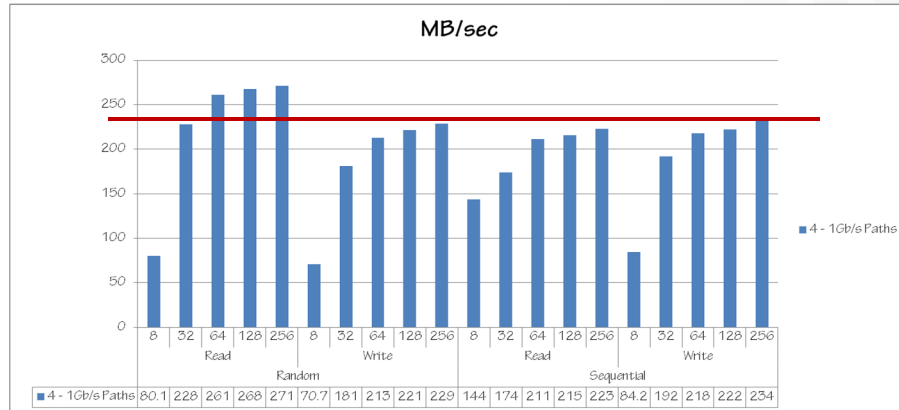
iSCSI Performance Characteristics

- **Equallogic PS-6100XS**
 - 7x 400GB SSD + 17x 600GB 10K RPM
 - RAID 6
 - Two controllers - 4GB cache, 4x 1Gb/sec per controller
- **Equallogic PS-6110XS**
 - 7x 400GB SSD + 17x 600GB 10K RPM
 - RAID 6
 - Two controllers - 4GB cache, 4x 10Gb/sec per controller

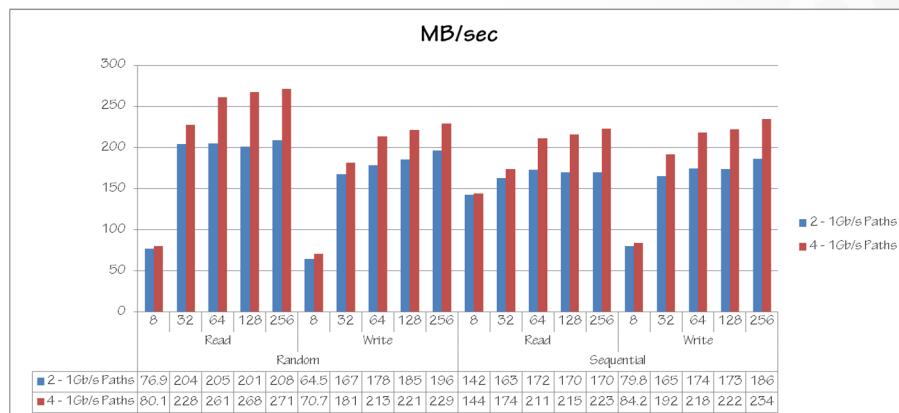
PS-6100XS - Two 1Gb/s Paths



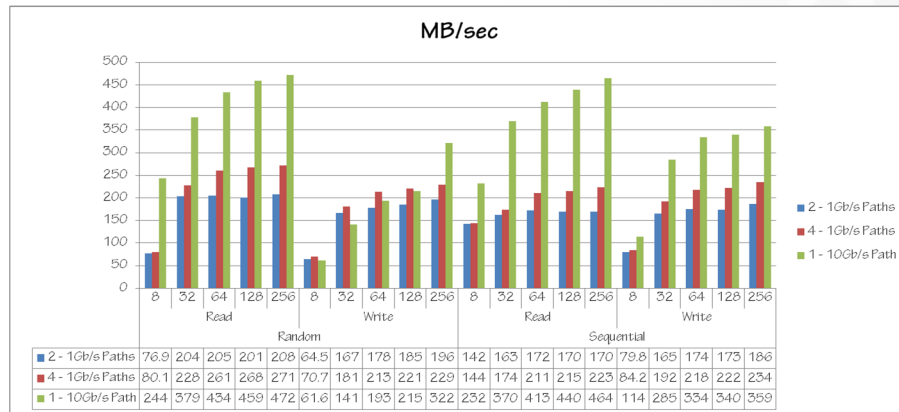
PS-6100XS - Four 1Gb/s Paths



PS-6100XS - Side-by-side Comparison



Comparison to PS-6110XS - 10Gb/s iSCSI



Storage Design Shared vs. Dedicated Spindles

- **Shared spindles allows for higher utilization of the underlying storage capacity**
 - Requires significant consideration up front to minimize the performance impact of sharing spindles
 - One application (often SQL Server, Exchange, or other OLTP solution) can cause the majority of I/O demand
 - I/O demanding applications should never share the same physical spindles
- **Dedicated spindles have lower utilization of storage capacity but predictable performance**

Storage Design (Shared Spindles)

- **Small databases vs. VLDBs**
 - Depends on utilization and workload
 - Size is not the significant consideration
- **Virtualization**
 - Underlying VHD/VMDK appears dedicated but may be on shared LUN physically
 - Hypervisor overheads
- **Can minimize TCO with acceptable performance if SAN Admin's skillset is weak**
 - Easy to extend volumes and add more spindles to balance load

Storage Design (Dedicated Spindles)

- **May provide predictable performance**
 - Only if the remaining infrastructure exceeds the throughput of the spindle configuration
- **Requires significantly more administration and monitoring of performance**
 - Extending volumes and adding spindles may require LUN migrations and/or downtimes
 - Generally have less disks in the RAID configuration based on the amount of space needed

The Effect of Cache

- **Cache exists at almost every layer in the I/O stack**
 - On the disks, the RAID controllers, SAN controllers, inside Windows
 - Must be battery backed to preserve writes in the event of a power failure
- **SQL Server uses FILE_FLAG_WRITE_THROUGH option to bypass on-disk cache that is not battery-backed**
- **Battery learning cycles can impact performance significantly**
 - Maintains the battery by draining it and running through multiple recharge/drain cycles to ensure it is healthy
- <http://support.microsoft.com/kb/234656>

How Should You Use Cache?

- **It Depends....**
 - Is the SAN dedicated to SQL Server or shared across the enterprise?
- **Dedicated SAN for SQL: limit read cache size**
 - Read from disk and cache data in the RAID controller, SAN controller, and the SQL Server buffer pool
 - For high I/O configurations cache can be slower than the potential disk throughput
- **Shared SAN: follow vendor recommendations**

SAN Cache Options

- **Multi-level caching**
 - Use write-only cache on the RAID controller and a shared read/write 80/20 cache on the SAN
 - Dedicate a portion non-shared cache for one demanding LUN with its own read/write ratio
- **Boosted (SSD cached) LUNs**
 - Use a small number of SSD disks in the SAN for caching read and writes of hot spots
 - Provides the benefits of a non-volatile memory and allows larger cache sizes than controller RAM

Review

- SAN components
- Multipathing
- iSCSI vs. fibre channel (FC)
- Shared vs. dedicated arrays
- Cache (on the drives, SAN controller, RAID controller, and Windows)
- Clustering
- Appendix: snapshot backups (SAN, not SQL)

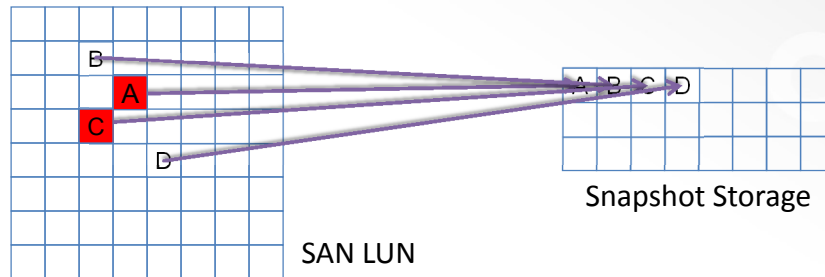
Appendix

- **How a Snapshot Works**

Snapshots

- **LUN level snapshots can provide a read only version of data**
 - SQL Server VSS integration allows storage vendors to quiesce or freeze the disk I/O from SQL Server to create a snapshot
 - Needs to be consistent across all LUNs to avoid corruption
- **Provide a fast “rollback” to a known working state**
 - Protection from hotfix and service pack installations as well as upgrades
- **Can be used in lieu of backups for VLDBs**
 - SAN is a single point of failure and data loss if it fails catastrophically
 - Should be used in conjunction with block level replication of the SAN to another location
- **Snapshots can be attached to a secondary server to perform CHECKDB operations and SQL native backups**
 - Still incurs I/O against the SAN and may impact performance if the disks are already at peak capacity

How a Snapshot Works



- Changes to data blocks write the original data to the snapshot storage (Copy On Write technology – specifically, Copy Before Write)
- Subsequent changes only affect the SAN LUN and are not written to the snapshot storage
 - Only the original data is maintained in the snapshot to provide a point in time view of the data
 - Only pages that are changed are written to the snapshot – unchanged pages remain in the database (this is NOT a backup)

Snapshot Considerations

- **Data and Log files on different LUNs**
 - The underlying SAN must be able to quiesce and snapshot all of the LUNs associated with a single database at the exact same point in time for a snapshot to be considered crash consistent
- **Snapshot space allocations**
 - Depending on the SAN vendor you may need to plan your snapshot space allocations to account for all potential changes during the snapshot existence
- **Virtual Machines**
 - If the virtual hard disks are on the same SAN LUN a single LUN snapshot will work
 - If the virtual hard disks are on different SAN LUNs the same considerations must be made for physical implementations
- **Recovery**
 - Not for crash recovery; it is not a replacement for a solid backup strategy

Questions?

