



PASS SUMMIT 2014

Glenn Berry, Principal Consultant
SQLskills.com

Analyzing I/O Subsystem Performance

○ DBA-211



Please silence cell phones



NOVEMBER 4-7 | SEATTLE, WA
The Conference for SQL Server Professionals.

Session Evaluations

3

ways to access

Your feedback is important and valuable.

Submit by 11:59 PM EST
Friday Nov. 7 to
WIN prizes

Evaluation Deadline:
11:59 PM EST, Sunday Nov. 16



Go to
passsummit.com/evals



Download the GuideBook App
and search: **PASS Summit 2014**



Follow the QR code link displayed
on session signage throughout the
conference venue and in the
program guide

Explore Everything PASS Has to Offer



Free SQL Server and BI Web Events



Free 1-day Training Events



Regional Event



This is Community



Business Analytics Training



Local User Groups Around the World



Session Recordings



CommunityCONNECTOR


PASS Newsletter



Free Online Technical Training

Glenn Berry



- Consultant/Trainer/Speaker/Author
- Principal Consultant, SQLskills.com
- Email: Glenn@SQLskills.com
- Blog: <http://www.SQLskills.com/blogs/Glenn>
- Twitter: @GlennAlanBerry
- Regular presenter at worldwide conferences on hardware, scalability, and DMV queries
- Author of SQL Server Hardware
- Chapter author of Pro SQL Server Practices
- Chapter author of Professional SQL Server 2012 Internals and Troubleshooting
- Chapter author of MVP Deep Dives Volumes 1 and 2
- Instructor-led training: Immersion Events
- Online training:  <http://pluralsight.com/>
- Consulting: health checks, hardware, performance, upgrades
- Become a SQLskills Insider: <http://www.sqlskills.com/Insider>



Agenda

- Three main metrics for storage performance
- SQL Server I/O workload metrics
- Tools for testing storage subsystems
- Primary storage types for SQL Server
- Choosing storage for different workloads and file types
- RAID levels and SQL Server workloads
- Some comparative storage metrics

Three Main Metrics for Storage Performance

1. Latency (ms)

2. Input/output operations per second (IOPS)

3. Sequential throughput (MB/sec or GB/sec)

- These three measurements are all related, so you can't just look at one of them in isolation, without knowing the others
- Storage vendors tend to show their best-case numbers in isolation

Latency

Latency

- The time it takes for an I/O to complete
- Sometimes called response time or service time

Measurement starts when the OS sends a request the drive or controller and ends when the drive finishes processing the request

- Reads are complete when the operating system receives the data
- Writes are complete when the drive informs the OS it has received the data
 - The data may still be in a DRAM cache on the drive or controller
- Write-back caching vs. write-through caching
 - Write-back caching is much faster, but requires a battery backup for the cache

Input/Output Operations per Second

Input/output operations per second (IOPS)

- This metric is directly related to latency
- Constant latency of 1ms means a drive can process 1,000 IOs per second with a queue depth of 1
- As more IOs are added to the queue, latency will increase
- Flash storage can read/write to multiple NAND channels in parallel

$$\text{IOPS} = \text{Queue Depth} / \text{Latency}$$

IOPS by itself does not consider transfer size

- You need to know the transfer size when looking at an IOPS measurement
- You can translate IOPS to MB/s and MB/s to latency as long as you know the queue depth and transfer size

Sequential Throughput

Sequential throughput (MB/sec or GB/sec)

$\text{MB/sec} = \text{IOPS} * \text{Transfer Size}$

- Example: 292 MB/sec = 71,271 IOPS * 4096 bytes transfer size

Sequential throughput often gets short-changed in enterprise storage

- Bandwidth limitations from the storage interface directly affect this
- 1Gbps iSCSI limited to about 100 MB/sec
- 4Gbps FC limited to about 400 MB/sec

Sequential throughput is extremely important for SQL Server

- Database backups and restores
- Index creation and maintenance
- Large sequential reads with reporting workloads

The Importance of Sequential Throughput

Sequential throughput is critical for many database server activities

- Full database backups and restores
 - Make sure to grant “Perform volume maintenance tasks” right to SQL Server Service Account
 - Make sure to use backup compression
 - Make sure to keep your VLF counts under control
- Index creation and rebuilds
 - Use MAXDOP option to indirectly throttle index create or rebuild I/O workload
 - Use data compression where appropriate to reduce I/O workload
- DW-type large sequential scans
 - When your database does not fit into the buffer pool
 - Buffer Pool Extensions (BPE) does not help much for large sequential reads

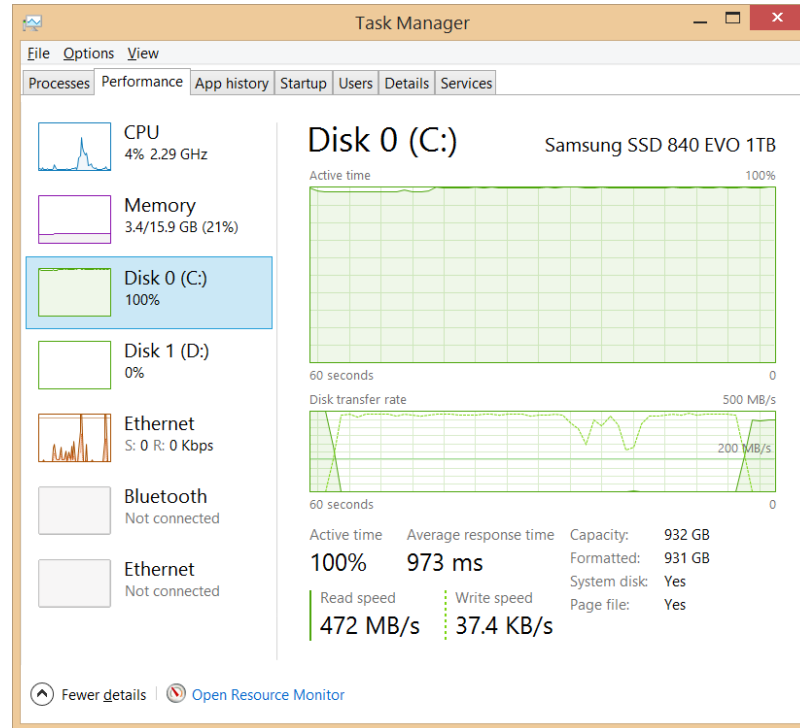
SQL Server I/O Workload Metrics

- What is the read vs. write ratio of the workload?
 - You can use my DMV Diagnostic Queries to determine this
 - Ratios will be different for different SQL Server file types and workloads
- What are the typical I/O rates (IOPS and throughput)?
 - Reads/sec, writes/sec (PerfMon) is IOPS
 - Disk read bytes/sec, disk writes/sec (PerfMon) is throughput
- What is the average logical disk-level latency?
 - Average disk sec/read, average disk sec/write (PerfMon) is latency
- What is the average disk-level and file-level latency for SQL Server database files?
 - You can use my DMV Diagnostic Queries to determine this for every SQL Server database file on your instance

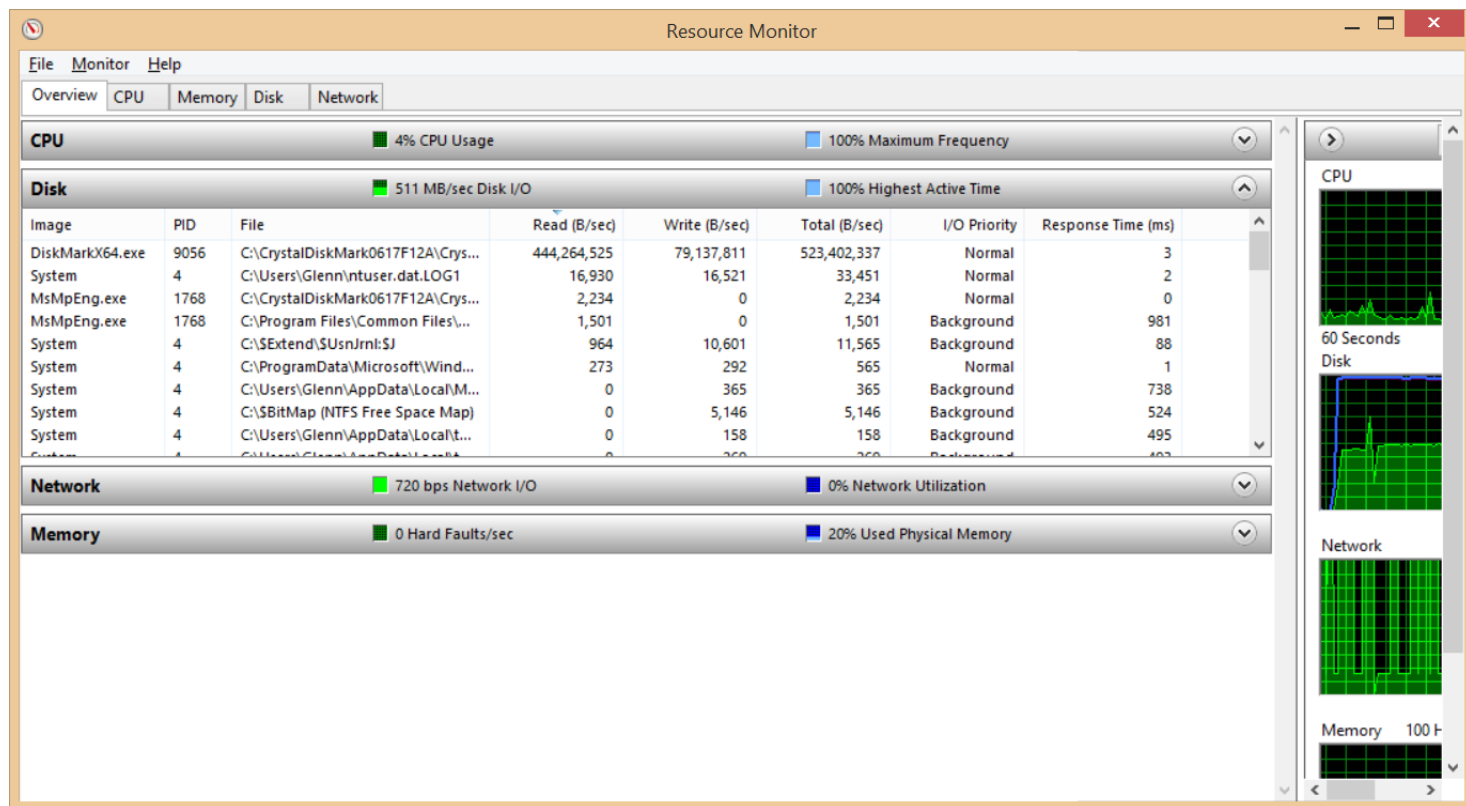
Methods for Measuring I/O Performance

- Task Manager in Windows Server 2012 and newer
 - Depending on what kind of storage you are using
- Disk section in Windows Resource Monitor
- LogicalDisk counters in Performance Monitor
- Disk Benchmark Tools
 - CrystalDiskMark (<http://bit.ly/1vm5dPe>)
 - SQLIO (<http://bit.ly/1obVdIV>)
 - DiskSpd (<http://bit.ly/1whNzQL>)
- SQL Server DMV Diagnostic Queries
 - <http://bit.ly/Q5GAJU>

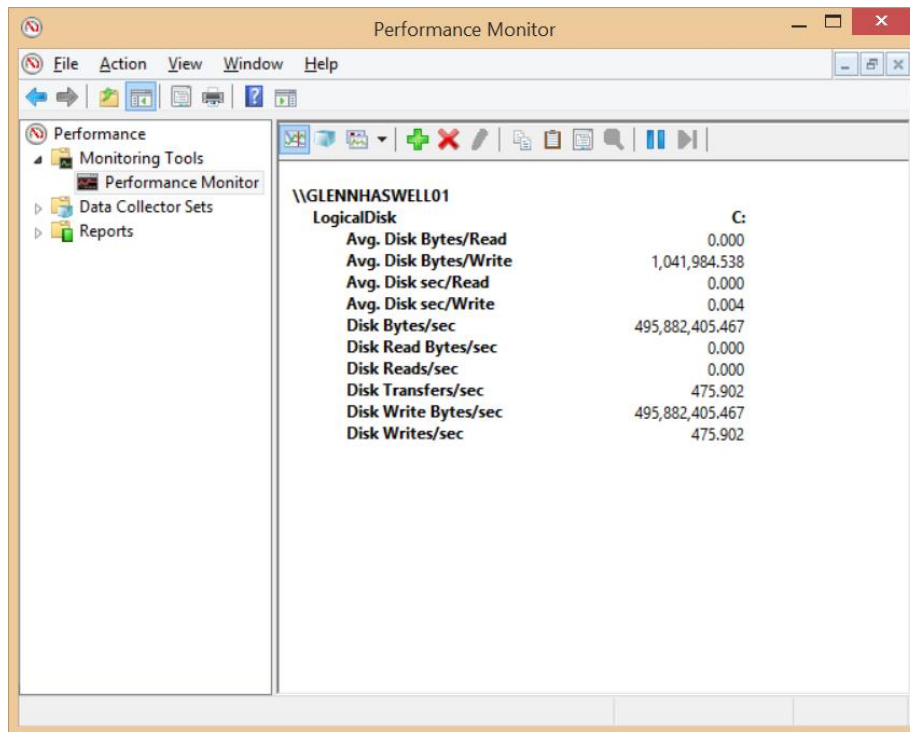
Disk Performance in Windows Task Manager



Disk Performance in Windows Resource Monitor



LogicalDisk Counters in Performance Monitor



Demo

- I/O Diagnostic Queries



NOVEMBER 4-7 | SEATTLE, WA
The Conference for SQL Server Professionals.

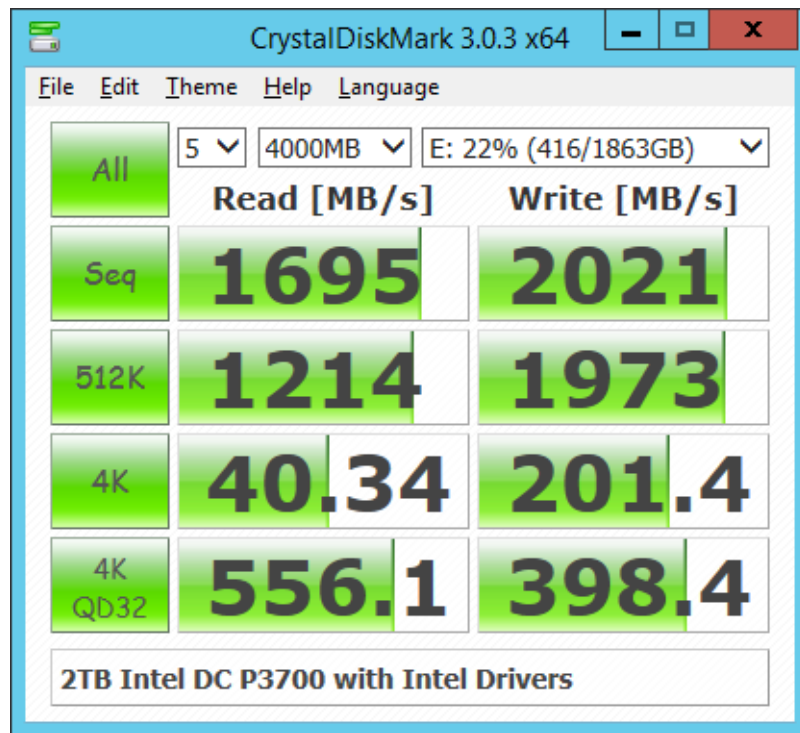
Common DMV I/O Query Result Patterns

- Very common to see high write latency to tempdb data files
 - Make sure you have multiple data files (start with 4-8) that are all the same size (follow Bob Ward's guidance). Make sure you are using TF 1118
 - Consider using local flash-based storage for tempdb
- Common to see high read latency from user database data files
 - Look for signs of memory pressure, consider adding more RAM and doing standard workload and index tuning
 - Consider using SQL Server 2014 BPE (especially for Standard Edition)
- Gather as much evidence as possible to show your SAN administrator
 - Overall SAN metrics may look great, so you need to be prepared with as much data as possible
 - `sys.dm_io_virtual_file_stats` are cumulative since instance was started
 - They include all file activity against your database files

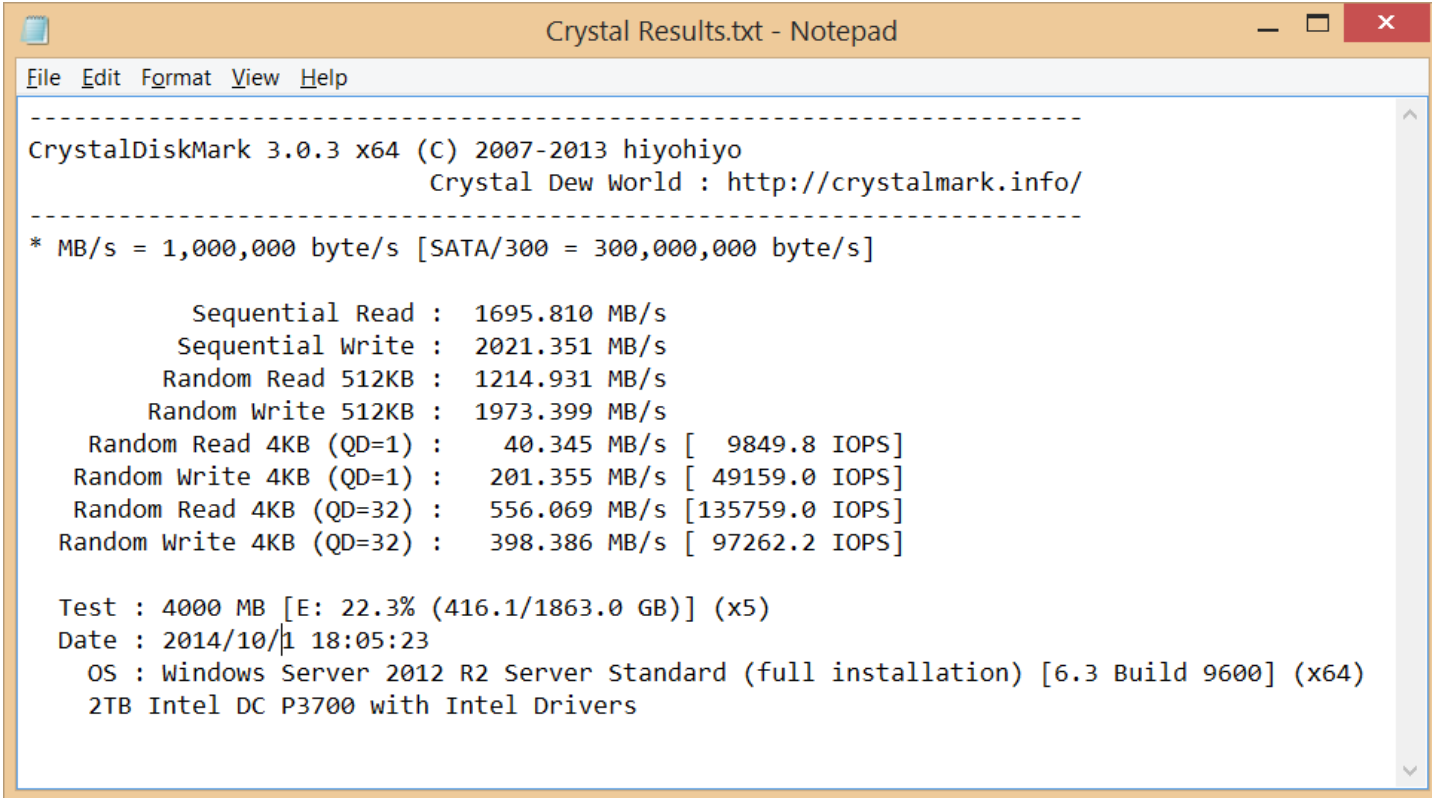
Using CrystalDiskMark To Test Your Storage

- CrystalDiskMark is quick and easy to use
 - It does have a few configuration options
- Make sure to test with a 4000MB test file size
 - This will minimize the influence of any hardware cache
- Make sure to select at least five (or more) test runs
 - This reduces the chances of outliers skewing the results
- Make sure to test with both random and non-random test file types
 - Random data is not compressible, some flash controllers use write compression
- Multiply 4K random I/O results (MB/s) by 244 to get IOPS result
 - Hover tooltip over 4K and 4K QD32 random results to see IOPS result
 - You can also Cntrl-C from GUI into Notepad to get full, detailed results for a test run

CrystalDiskMark Results (Graphical)



CrystalDiskMark Results (Text)



```
Crystal Results.txt - Notepad
File Edit Format View Help
-----
CrystalDiskMark 3.0.3 x64 (C) 2007-2013 hiyohiyo
                        Crystal Dew World : http://crystalmark.info/
-----
* MB/s = 1,000,000 byte/s [SATA/300 = 300,000,000 byte/s]

    Sequential Read : 1695.810 MB/s
    Sequential Write : 2021.351 MB/s
    Random Read 512KB : 1214.931 MB/s
    Random Write 512KB : 1973.399 MB/s
    Random Read 4KB (QD=1) : 40.345 MB/s [ 9849.8 IOPS]
    Random Write 4KB (QD=1) : 201.355 MB/s [ 49159.0 IOPS]
    Random Read 4KB (QD=32) : 556.069 MB/s [135759.0 IOPS]
    Random Write 4KB (QD=32) : 398.386 MB/s [ 97262.2 IOPS]

Test : 4000 MB [E: 22.3% (416.1/1863.0 GB)] (x5)
Date : 2014/10/1 18:05:23
    OS : Windows Server 2012 R2 Server Standard (full installation) [6.3 Build 9600] (x64)
    2TB Intel DC P3700 with Intel Drivers
```

Using SQLIO To Test Your Storage

- SQLIO does not require or use SQL Server for its testing
 - It simply allows you to stress your I/O system in a fairly controlled manner
- SQLIO has many configuration options
 - Can be time consuming and resource consuming to run full suite of tests
- You can use an old style command prompt or PowerShell to run your tests
- Reference:
 - SQLIO, PowerShell and storage performance: measuring IOPs, throughput and latency for both local disks and SMB file shares
 - <http://bit.ly/1n7jm0M>

Using DiskSpd To Test Your Storage

- DiskSpd is a new tool from Microsoft
 - It is far more flexible and powerful than SQLIO
- You can use an old style command prompt or PowerShell to run your tests
- Example commands:
 - PS C:\DiskSpd> C:\DiskSpd\diskspd.exe -c1000G -d10 -r -w0 -t8 -o8 -b8K -h -L X:\testfile.dat
 - Command Line: C:\DiskSpd\diskspd.exe -c1000G -d10 -r -w0 -t8 -o8 -b8K -h -L X:\testfile.dat
- Reference:
 - DiskSpd, PowerShell and storage performance: measuring IOPs, throughput and latency for both local disks and SMB file shares
 - <http://bit.ly/1CeQauw>

Primary Storage Types for SQL Server

Several different storage types are commonly used

- Internal drives (3.5", 2.5", or 1.8")
- Direct-attached storage (DAS)
- Storage area networks (SAN)
- PCIe flash-based storage cards
- Server Message Block (SMB) 3.0/3.02 file shares
 - SQL Server 2012/2014 have full support for using SMB 3.0/3.02 file shares for both user and system databases

Internal Drives

- Internal drives can be adequate for many workloads
 - Possible to have up to 28 2.5" drives in some two-socket servers
- Rack-mount server vertical size affects number of drive bays
 - 1U server might have (8-16) 1.8" or (10) 2.5" drive bays
 - 2U server might have up to (28) 2.5" drive bays
- Use the best hardware RAID controller(s) available for your server
 - Premium RAID controllers have faster processors and larger on-board cache sizes
 - They are less likely to be a bottleneck with SSDs

PCIe Flash Storage

- Flash-based storage on a PCIe expansion card
 - Uses very high bandwidth PCIe slot instead of SAS/SATA port
 - New products using NVM Express (NVMe) have excellent performance
 - Type and speed of PCIe slot can be a limiting factor
- PCIe storage cards can deliver extremely high I/O performance
 - Very high sequential throughput (up to 6.7GB/sec)
 - Extremely high random I/O performance (up to 1.3 million IOPS)
- Capital costs range from low to extremely high
 - Anywhere from \$1000 to \$125K for one PCIe storage card
 - It is common to use two, with software RAID 1 for redundancy

PCIe Slot Bandwidth Limits

- PCIe 1.0 Bus (one-way)
 - x4 slot: 750MB/sec
 - x8 slot: (1.5GB/sec)
- PCIe 2.0 Bus (one-way)
 - x4 slot: 1.5-1.8GB/sec
 - x8 slot: 3.0-3.6GB/sec
- PCIe 3.0 Bus (one-way)
 - x4 slot: 3.0-3.6GB/sec
 - x8 slot: 6.0-7.2GB/sec
- Only Intel Xeon E5, E5 v2, E5 v3 and E7 v2 families have PCIe 3.0

Direct-Attached Storage (DAS)

- External storage enclosure with multiple drive bays
 - Typically (14-24) 2.5" drives in a single external storage enclosure
 - Best practice is to dedicate at least one RAID controller to each storage enclosure
 - Storage enclosures should have dual power supplies
- DAS is easy to configure and manage
 - Does not require special training or expertise (or a cranky SAN admin)
 - Does require planning and common sense
- Can provide excellent sequential read/write performance
 - Limited by PCIe slot bandwidth and RAID controller performance

DAS Considerations

- Use one dedicated RAID controller per storage enclosure
 - You may even want two RAID controllers per enclosure
 - Make sure the hardware cache is enabled
 - Use write-back caching if the RAID controller has a backup battery
- Pay attention to the PCIe slot throughput limits
 - You want to make sure you are not throttled by PCIe slot bandwidth
- Try to dedicate the hardware RAID controller cache to writes
 - Disable read-ahead caching
 - The SQL Server buffer pool is a better read cache than the hardware RAID cache

Storage Area Networks (SAN)

- Shared external storage device with multiple components
 - Large number of drive bays, storage processors, large cache, operating system
 - Much higher initial capital cost, requires some expertise to setup and manage
 - Cranky SAN administrator is often included free of charge!
- Two main types of SANs
 - Fiber-channel, using host bus adapter (HBA)
 - iSCSI, using dedicated Ethernet cards
- SANs are usually optimized for IOPs
 - Sequential throughput can be limited by the interface
 - Example: 1Gbps iSCSI limited to about 100MB/sec

SAN Considerations

- Make an effort to really communicate with your SAN administrator
 - Let the SAN administrator know the type of workload that you have
 - Don't just give the SAN administrator a space requirement
- Your SAN administrator probably has different priorities than you
 - Has to worry about multiple servers with different workloads
 - Has to worry about space, and DBAs complaining about performance
- Consider the complete data path to the SAN
 - HBA/NIC, switches, SAN ports, etc.
 - Be prepared for inconsistent performance with a shared SAN
 - SANs are not magic. The hardware details still matter!

SMB 3.0/3.02 File Shares

- Server Message Block (SMB) 3.0/3.02
 - SQL Server 2012/2014 can store user/system databases on SMB 3.0 file shares
 - SQL Server 2012/2014 can use SMB 3.0 for traditional FCI instances (w/o a SAN)
- Windows Server 2012 has SMB Direct
 - Network adapters should have Remote Direct Memory Access (RDMA)
 - RDMA requires SMB Multichannel in order to be enabled
 - Enables a remote file server to perform like local storage with appropriate NIC hardware (such as 56Gbps Infiniband)
- Microsoft's Jose Barreto is a great resource about SMB file shares
 - <http://blogs.technet.com/b/josebda/>

Negotiated Versions of SMB

Operating System	Windows 8.1 WS 2012 R2	Windows 8 WS 2012	Windows 7 WS 2008 R2	Windows Vista WS 2008	Previous Versions
Windows 8.1 WS 2012 R2	SMB 3.02	SMB 3.0	SMB 2.1	SMB 2.0	SMB 1.0
Windows 8.0 WS 2012	SMB 3.0	SMB 3.0	SMB 2.1	SMB 2.0	SMB 1.0
Windows 7 WS 2008 R2	SMB 2.1	SMB 2.1	SMB 2.1	SMB 2.0	SMB 1.0
Windows Vista WS 2008	SMB 2.0	SMB 2.0	SMB 2.0	SMB 2.0	SMB 1.0
Previous Versions	SMB 1.0	SMB 1.0	SMB 1.0	SMB 1.0	SMB 1.0

Considering Your Workload for Storage

- SQL Server can have several different common workload types
 - Online Transaction Processing (OLTP)
 - Relational Data Warehouse (DW)
 - Online Analytical Processing (OLAP)
- These workload types have different I/O access patterns
 - OLTP workload has frequent writes to data files and log file
 - Also has random reads from data files if database does not fit into buffer pool
 - DW workload has large sequential reads from data files
 - Sequential I/O performance is very important
 - OLAP workload has lots of random reads from cube files
 - Random I/O performance is very important

RAID Levels and SQL Server Workloads

- You need to consider your SQL Server workload type
 - It directly affects your desired RAID level
 - RAID 10 is better for write-intensive workloads
- You also need to consider your availability requirements
 - Some RAID levels are more robust than others
 - RAID 10 > RAID 50 > RAID 5
- Different SQL Server file types have different I/O patterns
 - Data files, log files, tempdb files, backup files, etc.
 - Percentage of reads/writes, sequential vs. random I/O

Selecting a RAID Level For Your SLA

- RAID is not a substitute for a good backup/restore plan!
 - No matter what anyone in your organization tells you...
- RAID is not a substitute for an effective HA/DR strategy
 - No matter what any vendor tells you.
- An appropriate RAID level reduces the chance of unplanned downtime
 - It also reduces the chance of data loss due to disk failure(s)
- RAID 10 and 50 are the most robust common RAID levels
 - RAID 5 can only lose one disk in an array before the array is lost
 - More disks in a RAID 5 array increases the statistical chances that one will fail
 - Consider using hot spares in your arrays and having a cold spare available

Choosing Storage Types Based on Workload

- Flash-based storage gives great random I/O performance
 - It also gives better sequential performance than magnetic storage
 - Flash-based storage is more expensive per GB, but price is declining rapidly
- Magnetic storage gives fair sequential performance
 - Magnetic storage gives quite poor random I/O performance
 - Large controller caches can help mask poor random I/O writes
 - Flash-based caching can improve magnetic storage performance
 - Flash-based storage tiering is not as effective for most SQL Server workloads
- Flash-based storage is the best choice if you have the budget
 - Use where you have heavy random I/O activity or any I/O bottlenecks

Configuring Storage for SQL Server File Types

- SQL Server data files
 - Still common to use magnetic storage (flash gaining in popularity as cost declines)
 - Most common to use RAID 5, 50, or 10
- SQL Server log files
 - Magnetic storage or flash-based storage (flash when you have many databases)
 - Most common to use RAID 10 because of better write performance
- SQL Server tempdb data and log files
 - Magnetic storage or flash-based storage (flash becoming more popular)
 - Most common to use RAID 10 because of better write performance

HA/DR Effects on Storage Choices

- Traditional FCI requires some form of shared storage
 - Usually a SAN, but SMB 3.0/3.02 can be used with SQL Server 2012 or newer
 - SQL Server 2012/2014 can use local storage for tempdb files with FCI
 - Often a very good use for flash-based storage, also reducing load on SAN
- AlwaysOn AGs must use Windows Clustering feature
 - Can use shared storage (SAN or SMB 3.0/3.02)
 - Can also use any type of non-shared storage
- Other native HA/DR technologies can use any type of storage
 - Consider using non-shared storage to eliminate the single point of failure
 - Consider shared storage combined with non-shared storage for different parts of architecture

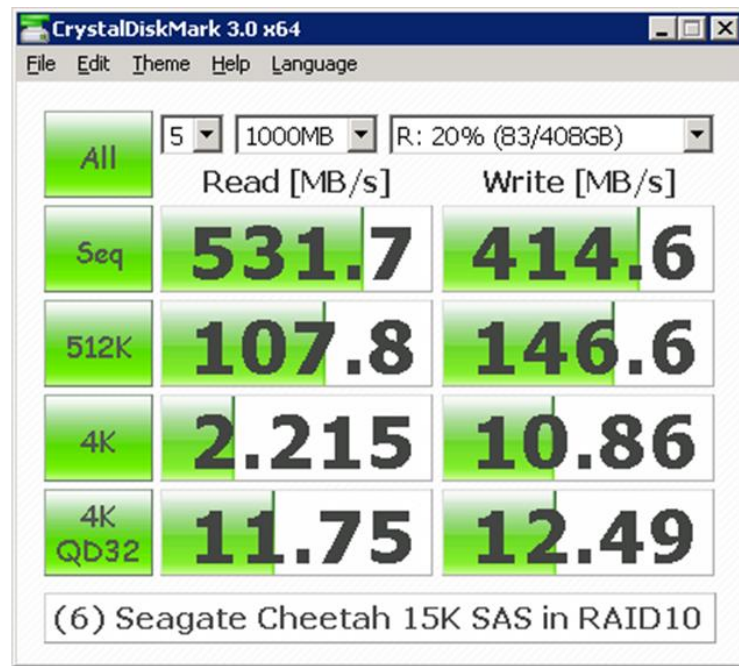
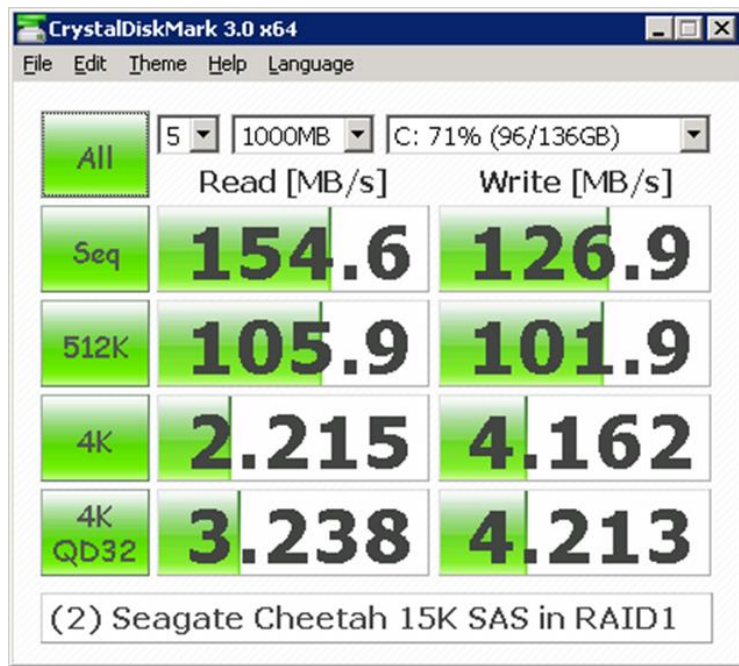
Sizing Your Storage Subsystem

- Use a RAID calculator to ensure you have more than enough space
 - Keep in mind the performance advantages of “short-stroking”
 - Flash-based storage also benefits from ample free space
- After you have enough space, concentrate on performance
 - Don’t negotiate with yourself! Ask for flash-based storage, ask for RAID 10
 - Consider your workload as you make budget-driven compromises
- Suggested I/O performance goals
 - 5,000-10,000 or more IOPS on all LUNs
 - 1 GB/sec or more of sequential throughput on all LUNs

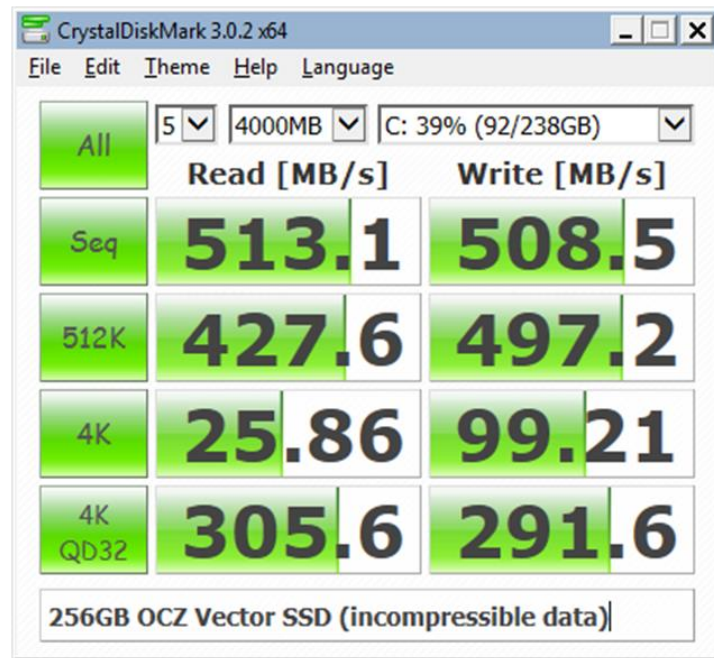
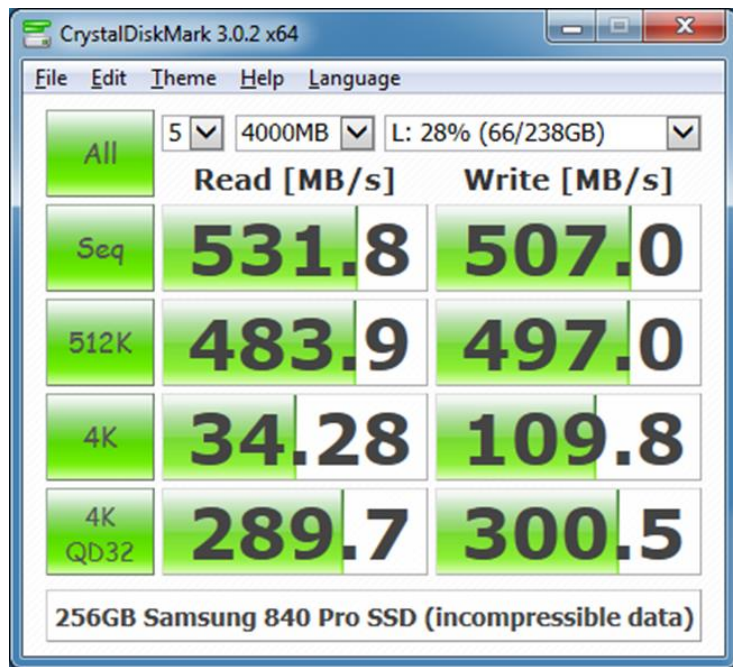
Magnetic Storage vs. Flash-Based Storage

- Magnetic storage performance
 - Sequential: 100-200 MB/sec per disk
 - Random: 100-200 IOPS per disk
- Flash-based storage performance
 - Sequential 3/6/12Gbps: 275/550/1100 MB/sec per drive
 - Sequential PCIe: 1000-6500 MB/sec per card
 - Random: SAS/SATA drives do about 100,00 IOPS
 - PCIe cards can do up to about 1.3 million IOPS
 - Enterprise flash storage trades some raw performance for consistency

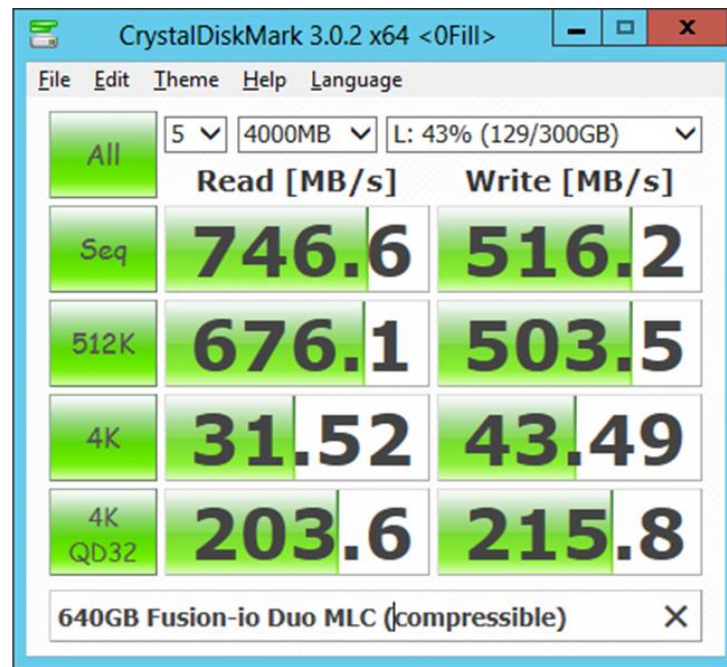
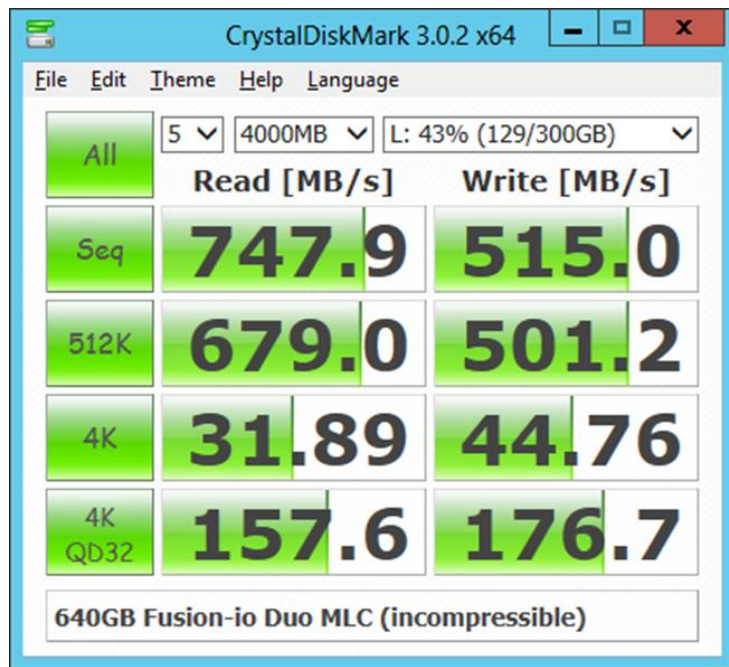
Traditional Magnetic Drive Performance



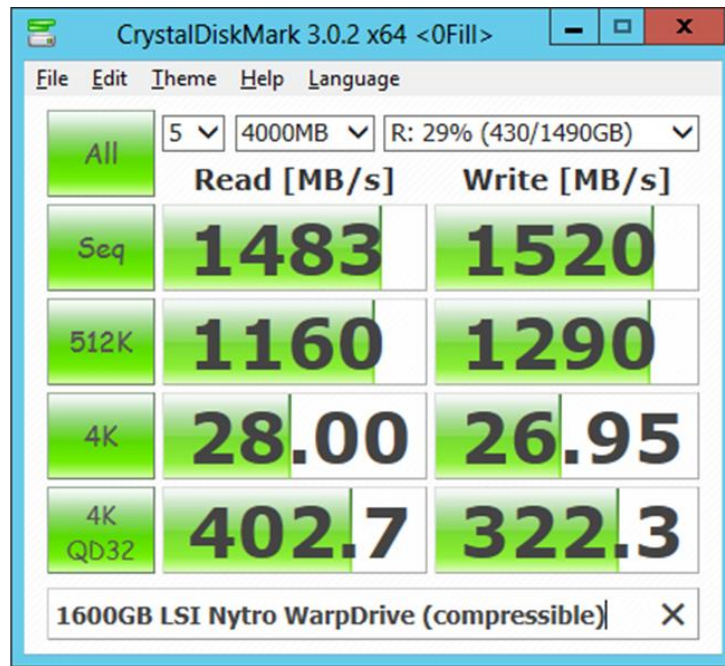
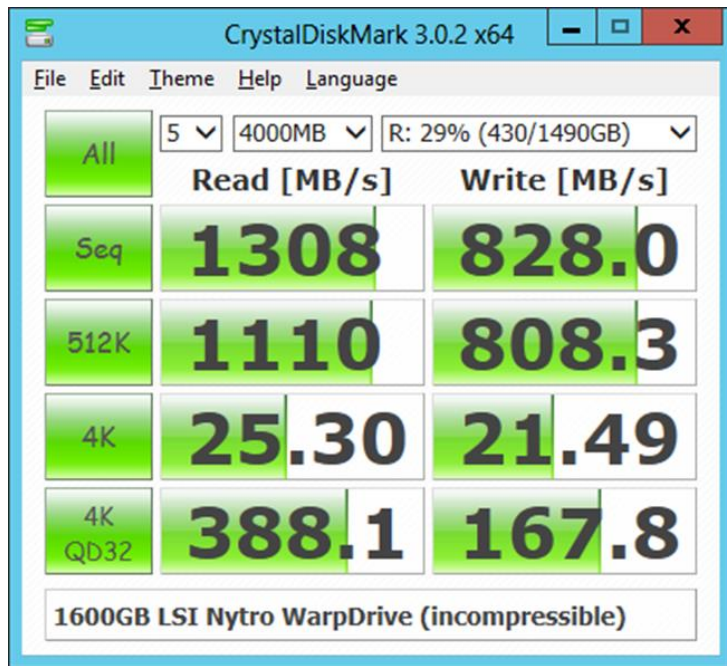
Consumer SSD Drive Performance



Older Enterprise PCIe Drive Performance



Newer Enterprise PCIe Drive Performance



Comparative Sequential Performance

Drive Type	Sequential Reads	Sequential Writes
(2) 15K magnetic SAS in RAID 1	154.6 MB/sec	126.9 MB/sec
(6) 15K magnetic SAS in RAID 10	531.7 MB/sec	414.6 MB/sec
256GB Samsung 840 Pro	531.8 MB/sec	507.0 MB/sec
256GB OCZ Vector	513.1 MB/sec	508.5 MB/sec
640GB Fusion-io MLC (random)	747.9 MB/sec	515.0 MB/sec
640GB Fusion-io MLC (0Fill)	746.6 MB/sec	516.2 MB/sec
1.6TB LSI Nytro WarpDrive (random)	1308.0 MB/sec	828.0 MB/sec
1.6TB LSI Nytro WarpDrive (0Fill)	1483.0 MB/sec	1520.0 MB/sec
2.0TB Intel DC P3700 (random)	1695.8 MB/sec	2021.4 MB/sec

Comparative Random 4K QD32 Performance

Drive Type	Random Reads	Random Writes
(2) 15K magnetic SAS in RAID 1	790 IOPS	1,028 IOPS
(6) 15K magnetic SAS in RAID 10	2,867 IOPS	3,048 IOPS
256GB Samsung 840 Pro	70,727 IOPS	73,371 IOPS
256GB OCZ Vector	74,606 IOPS	71,198 IOPS
640GB Fusion-io MLC (random)	38,454 IOPS	43,115 IOPS
640GB Fusion-io MLC (0Fill)	49,678 IOPS	52,655 IOPS
1.6TB LSI Nytro WarpDrive (random)	94,696 IOPS	40,943 IOPS
1.6TB LSI Nytro WarpDrive (0Fill)	98,259 IOPS	78,641 IOPS
2.0TB Intel DC P3700 (random)	135,759 IOPS	97,262 IOPS

Summary

- Three main I/O performance metrics
 - Latency, IOPS, and sequential throughput
- Different types of SQL Server workloads affect I/O patterns
 - OLTP, DW, OLAP, mixed, database maintenance, etc.
- Different SQL Server file types have different I/O patterns
 - Data files, log files, tempdb files, backup files, etc.
- Make sure to actually test your I/O subsystem!
 - CrystalDiskMark, SQLIO, DiskSpd

References

- Install SQL Server with SMB fileshare as a storage option
 - <http://bit.ly/1qWaLy6>
- Windows Server 2012 R2: Which version of the SMB protocol (SMB 1.0, SMB 2.0, SMB 2.1, SMB 3.0 or SMB 3.02) are you using?
 - <http://bit.ly/18uOEI4>
- Testing Windows Server and the Scale-Out File Server – What should your lab look like?
 - <http://bit.ly/1vC9WMF>