
# Storage Best Practices for Microsoft Server Applications

Dennis Martin
President, Demartek

Principal Research Contributor
Performance Lab Portal
wikibon.org


# Agenda

- Brief Company Overview
- Storage Technology Overview
  - Devices, Interfaces, RAID
- Recommendations and Best Practices for Microsoft Server-based Applications
- Performance Comparisons (solid state and spinning disks)
- References


# Demartek Company Overview

- Industry analysis with on-site test lab
- Lab includes servers, networking and storage infrastructure
  - Fibre Channel:  4 & 8 Gbps
  - Ethernet: 1 & 10 Gbps (with FCoE and iSCSI)
  - Servers: 8 cores, up to 96 GB RAM
  - Virtualization: ESX, Hyper-V, Xen
- Web: www.demartek.com

# Storage Technology Overview

- Devices
  - Solid State and Spinning Disk Drives
- RAID
- Interfaces

- Comment: “overlap”


# Storage Devices

- Solid State Disks (SSD)
  - DRAM
  - NAND-Flash
- Spinning Disk Drives
  - Enterprise
  - Desktop
  - Others


# Solid State Disks

- Memory technology designed to appear as an online storage (disk) device
- Very fast, no moving parts
- Variety of form factors
- Prices dropping
- Some SSDs use DRAM and NAND-Flash together


# SSD: *DRAM*

- IOPS (I/O per second) range from 70K to 5M +
- Latencies measured in microseconds
- Almost always includes battery-backup and/or disk-drive for safety
- Can be used as a cache in front of other storage


# SSD: *NAND-Flash*

- Non-volatile
- IOPS range 10K – 35K reads, writes are slower
- Single device up to 1 TB+
- Variety of interfaces
- Quiet, low-power, low-weight, low-heat
- Two basic types (SLC & MLC)

# NAND-Flash: *SLC*

- Single-Level Cell (SLC)
  - One bit per cell, faster, lower capacity
  - Lower error probability and longer life (100,000+ write cycles)
  - More expensive than MLC
  - More suited to enterprise-class applications


# NAND-Flash: *MLC*

- Multi-Level Cell (MLC)
  - Multiple bits per cell, slower, higher capacity
  - Higher error probability and shorter life (10,000+ write cycles)
  - Less expensive than SLC
  - Generally used for consumer applications


# NAND-Flash Storage Today

- Form factors:
  - Disk drive: 3.5-inch, 2.5 inch, 1.8 inch
    - Interfaces: same choices as spinning disk
  - Card installed in PCI-Express bus
  - Blade-server mezzanine cards
  - Consumer device cards, sticks, etc.
- Current issues:
  - How to measure the usable life of NAND-flash
  - Determining the right mix: SSD & spinning disk drives

# NAND-Flash Storage Futures

- Expect some overlap between enterprise-class and consumer-grade technologies
  - Enterprise feature/function with MLC
- Interfaces: SAS? USB?
- On motherboards
- Other internal and external connection types


# Spinning (magnetic) Disks

- Disk drive technology is well-known
- Market requirements dictate differences in drive types, from enterprise to consumer devices
- Reasonably fast, but slow when compared to CPU and memory
  - IOPS (random I/O) range: 100-1000
- Good pricing with steady price declines and increasing capacities

# Types of Spinning Disk Drives

- Categories
  - Enterprise
  - Desktop
  - Notebook
  - Consumer
- Some overlap between drive categories
- Drive type does not necessarily dictate interface type

# Disk Drives: *Enterprise*

- Rotation speeds: 10K and 15K RPM
- Can tolerate higher vibration in racks
- Designed for 7x24x365 operation
- Moving to 2.5 inch form factor
  - Reduced power, heat, space, weight
- MTBF: 1M+ hours
- Warranty: 5 – 7 years


# Disk Drives: *Desktop*

- Rotation speeds: 5400 and 7200 RPM
- 3.5 inch form factor
- Very large capacities (up to 2 TB)
- Some are only designed for 8x5 operation
- Warranty: 3 – 5 years

# Disk Drives: *Notebook & Consumer*

- Meets the needs of laptop and consumer device market
  - Low power consumption (slower RPM)
  - Light weight
  - Smaller form factors (2.5, 1.8, 1.0 inch)
  - Can tolerate some physical shock
- Not appropriate for server-based applications

# Spinning Disk Drive Comparison

| Device | Enterprise | Desktop | Notebook | Consumer |
|---|---|---|---|---|
| Avg. seek time | 3 – 5 ms | 8 – 11 ms | 10 – 15 ms | 12 – 15 ms |
| Transfer rate (MB/s)* | 70 – 170 | 60 – 120 | 30 – 80 | 6 – 40 |
| RPM (K) | 10, 15 | 5.4, 5.9, 7.2, 10 | 4.2, 5.4, 7.2 | 3.6, 4.2 |
| Capacities | Medium (2.5”) Large (3.5”) | Very large | Medium | Small |
| Command Queuing | TCQ or NCQ | NCQ | NCQ | - |
| Power need | Medium (2.5”) Large (3.5”) | Medium | Low | Very low |
| Warranty | 5 – 7 years | 3 – 5 years | 1 – 5 years | 1 – 3 years |



* Maximum device transfer rate in megabytes per second from drive surface to buffer. Sustained rates are lower. This is not the same as the interface transfer rate.

# Disk Drives: *Recommendations*

- For Microsoft Server Applications only consider enterprise and desktop drives
- Enterprise drives → Performance
  They will run out of capacity before they run out of performance
- Desktop drives → Capacity
  They will run out of performance before they run out of capacity

# RAID (**R**edundant **A**rray of **I**ndependent **D**isks)

- Disks working together to increase reliability or performance or both
  - **RAID 0**: Interleaving or “striping” data across two or more disks
  - **RAID 1**: Disk mirroring – same data written on two different disks *(rebuild possible)*
  - **RAID 5**: Data striping with single parity across multiple disks *(rebuild possible)*
  - **RAID 6**: Data striping with double parity across multiple disks *(rebuild possible)*
- RAID 1 can be combined with other RAID types

# RAID Comparison

| RAID Type | Transactional I/O Performance | Capacity Utilization | Disk Failure and Rebuild Performance |
|---|---|---|---|
| RAID 0 | Good | Best | Poor |
| RAID 1 | Best | Poor | Best |
| RAID 5 | Good | Good | Moderate |
| RAID 6 | Good | Moderate | Good |
| RAID 10 | Best | Poor | Best |

# RAID Recommendations

- Choose RAID to spread data over multiple disks (“spindles”) to get better performance and reliability than using individual disks
- Best overall performance: generally RAID 10
- Best capacity (with recoverability): generally RAID 5

# Storage Interface Comparison

| | SATA | SAS | Fibre Channel | FCoE | Infiniband | USB |
|---|---|---|---|---|---|---|
| **Number of devices** | 1 | 16K | 16M | 16M | 16M | 127 |
| **Maximum distance (meters)** | 1 | 10 | 10K+ | 10, very long | 20, very long | 5 |
| **Cable type** | Copper | Copper | Optical | Copper, Optical | Copper, Optical | Copper, Wireless |
| **Interface device** | On-board, HBA | On-board, HBA | HBA | CNA, 10Gb NIC | HCA | On-board |
| **Transfer speeds (MB/sec)** | 150, 300 | 300, 600 | 100, 200, 400, 800 | 400, 1000 | 1000, 2000, 4000 | 0.15, 1.5, 50, 500* |

MB/sec = Megabytes per second, which is generally calculated as megabits/ second (Mbps) divided by 10 for planning purposes

* *SuperSpeed USB devices expected in 2010*


# Interface Futures

- As interface speeds increase, expect increased usage of fiber-optic cables and connectors for most interfaces
  - At higher Gigabit speeds, copper cables and interconnects become too “noisy” except for short distances
- Expect to see SAN-like types of features for interfaces such as SAS, USB and perhaps others

# Fibre Channel and iSCSI

- Each addresses a different market that has different needs with respect to performance, reliability, scalability and manageability
- Although there are different “plumbing” characteristics between FC and iSCSI, the applications storing data on them can’t tell the difference


# Best Practices

# General Recommendations

- Configure storage for application servers with performance and availability as design criteria
  - Many server applications must satisfy high transaction rates
- Use more disks and faster disks for best performance
  - If you choose “desktop” disk drives, you’re often emphasizing capacity above performance (this choice may also reflect your budget)

# Windows Storage Formatting

- Disk Alignment
- Format Allocation (Cluster) Size
- Stripe Size
- Thin Provisioning Storage
- SSD-aware operating system versions


# Disk Alignment

- Windows Server 2003 or older:
  Align the file system to the disk offset recommended by the storage hardware vendor. If unknown use an offset of 64K.
  - Diskpart command:
    `create partition primary align=64`
- Windows Server 2008 uses default alignment of 1MB


# Format Allocation Size

- Exchange Server
  - Databases: **64K**
  - Logs: can use default size (typically 4K)
- SQL Server: use **64K** for volumes dedicated to SQL Server
  - The SQL Server page size is 8K
  - SQL Server allocates disk from the operating system in units known as “extents” of 8 pages

# Stripe Size

- Since SQL Server accesses disk storage in 64K blocks, the optimum disk array stripe size Microsoft SQL Server volumes is **64K**
- Similar recommendations for Exchange Server


# Thin Provision Storage

- Thin provisioning storage systems use pointers, linked-lists and other similar techniques to minimize the consumption of capacity
- Always use the “quick format” option in Windows
- Defragmentation is not necessary from the operating system

# SSD-aware Operating Systems

- Operating systems need to detect the presence of NAND-flash SSDs
  - Windows 7
  - Windows Server 2008 R2
- No defragmenting
- Trim – notify the underlying device regarding data that is no longer needed

# Exchange Server

- Consider performance before capacity
- Exchange Server is sensitive to disk read and write latencies
  - Exchange Server wants average read latencies < 20 msecs.
- Place Exchange logs on lowest latency disks
- Place databases and logs on separate RAID sets


# SQL Server TempDB

- For best performance, the number of TempDB data files should equal the number of CPU cores in the server
- TempDB is an excellent candidate for SSD


# SQL Server & SharePoint

- Ideally, use separate RAID sets for:
  - TempDB: RAID10 (write-heavy)
  - Transaction logs: RAID10 (write-heavy)
  - Search database: RAID10 (read-write mix)
  - Content databases: RAID10 (read-heavy)


# Performance Comparisons

# Performance Tests

1. Same storage hardware, five different RAID configurations for the databases
   - RAID-5 with 5, 10, & 15 drives
   - RAID-10 with 8 & 16 drives
   - Logs on one RAID-10 set of 8 drives
2. Different storage hardware, same application configuration
   - Two different SSDs
   - SAS and SATA disk drives in various RAID configurations


# Technology Environment

- Servers:
  - Dell PowerEdge 2900, dual Intel Xeon E5345 (2.33 GHz, 8 cores), 48GB & 32GB RAM, Windows Server 2008 x64 (qty. 2)
  - IBM System x3650, dual Intel Xeon E5345 (2.33 GHz, 8 cores), 32GB RAM, Windows Server 2008 x64
  - Intel Server S5000PSL with dual-Xeon E5320 (1.86 GHz, 8 cores), 4GB RAM, Windows Server 2003 x64
- Fibre Channel infrastructure:
  - Brocade 200e, 16-port, 4 Gbps FC switch
  - Brocade 300, 24-port, 8 Gbps FC switch
  - Emulex LPe11002, dual-port, 4 Gbps HBA (in IBM server)
  - Emulex LPe12002, dual-port, 8 Gbps HBA (in Dell servers)
- Storage:
  - Fusion-IO ioDrive, 160GB, SLC NAND-flash, PCI-express 1.1 interface, no cache
  - IBM DS3400 with 48 drives, SAS, 300GB, 15K RPM, 4-port, 4 Gbps FC
  - Intel SRCSASJV, 512MB Cache, supports up to 240 SAS or SATA disk drives
  - Seagate Barracuda 7200.11, SATA, 500GB, 7200 RPM, 32MB cache (qty. 10)
  - Seagate Cheetah 15K.5, SAS, 146GB, 15K RPM, 16MB cache (qty. 10)
  - Texas Memory Systems RAMSAN-400, 128GB, DRAM SSD, 8-port, 4 Gbps FC


# Performance Test Tools

1. Microsoft SQLIOSim
   - Microsoft SQL Server(c) Simulator Stress Test Version 9.00.1399.05
   - Simulates SQL Server I/O workloads
2. Microsoft Exchange Jetstress
   - Microsoft Exchange Server Jetstress Version 08.02.0060.000
   - Simulates Exchange Server 2007 workloads

# Test 1 –SQL Server

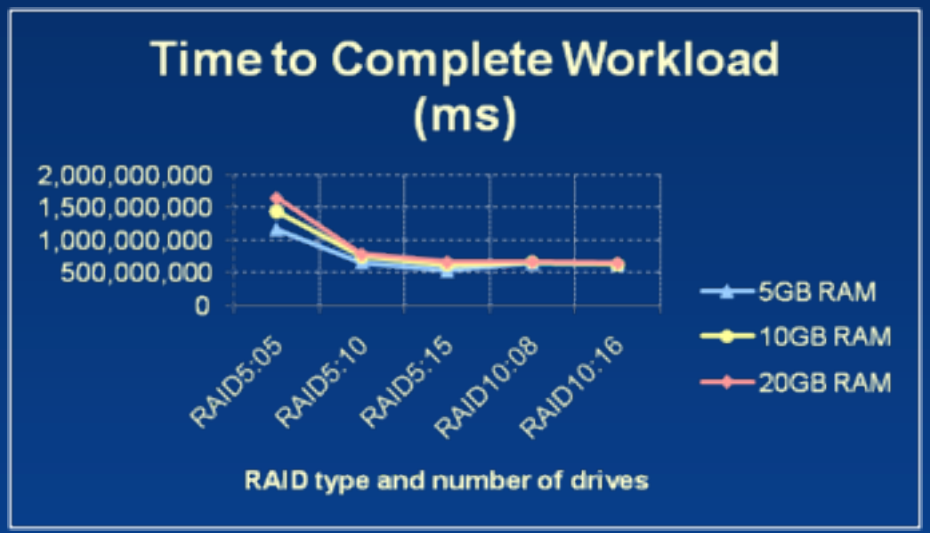
- Microsoft SQLIOSim
  - Different database for each RAID sets
  - Log on same RAID10 set
- Database parameters
  - InitialSize = 25000 MB, MaxSize = 50000 MB, Increment = 100 MB, LogFile = No, Shrinkable = No, Sparse = No
- MaxMemory: 5GB, 10GB, 20GB

# Test 1 – Exchange

- Exchange Server 2007 profile
  - 500 mailboxes
  - Mailbox size = 250MB
  - Exchange IOPS = 0.5 (heavy user)
  - Threads = Auto (2)
  - Storage Groups = 1
  - DB volume size = 1000GB

# Test 1 – SQL Server

(Run on large server with 48GB RAM and external 48-drive array)

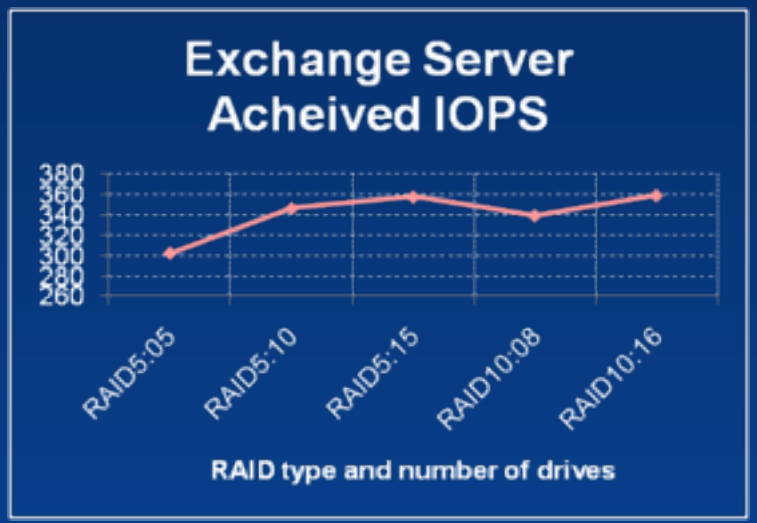


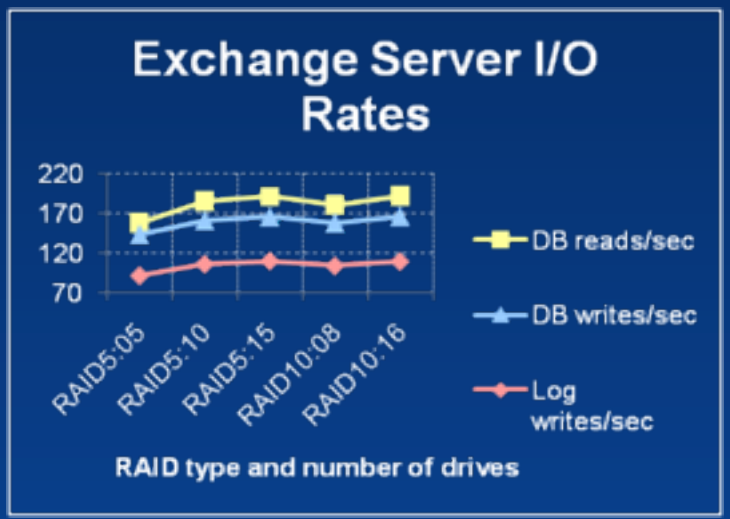

Delayed I/O was reported frequently with the smaller number of drives and smaller memory sizes

| RAID:Qt | 5 GB RAM | 5 GB RAM | 10 GB RAM | 10 GB RAM | 20 GB RAM | 20 GB RAM |
|---|---|---|---|---|---|---|
| RAID5:0 | 24,557 | 1,166.85 | 24,368 | 1,433.27 | 23,014 | 1,641.20 |
| RAID5:1 | 23,577 | 661,576 | 22,311 | 741,603 | 19,434 | 788,324 |
| RAID5:1 | 22,256 | 527,367 | 22,653 | 618,407 | 20,649 | 666,746 |
| RAID10: | 23,636 | 644,863 | 19,275 | 661,547 | 13,434 | 672,392 |
| RAID10: | 22,802 | 625,489 | 17,954 | 624,709 | 11,228 | 653,976 |

# Test 1 – Exchange

(Run on large server with 48GB RAM and external 48-drive array)




| RAID: Qty | Achie ved | DB Disk | DB Disk | Log writes/ | DB Avg Disk | DB Avg Disk | Log Avg |
|---|---|---|---|---|---|---|---|
| RAID5 | 201.0 | 158.40 | 143.42 | 91.206 | 006 | 005 | 001 |
| RAID5 | 346.6 | 185.37 | 161.32 | 105.88 | 005 | 006 | 001 |
| RAID5 | 357.0 | 191.75 | 166.19 | 109.79 | 004 | 006 | 001 |
| RAID1 | 339.4 | 181.15 | 158.31 | 103.61 | 005 | 006 | 001 |
| RAID1 | 359.3 | 192.58 | 166.80 | 109.25 | 004 | 006 | 001 |

# Test 2 – SQL Server

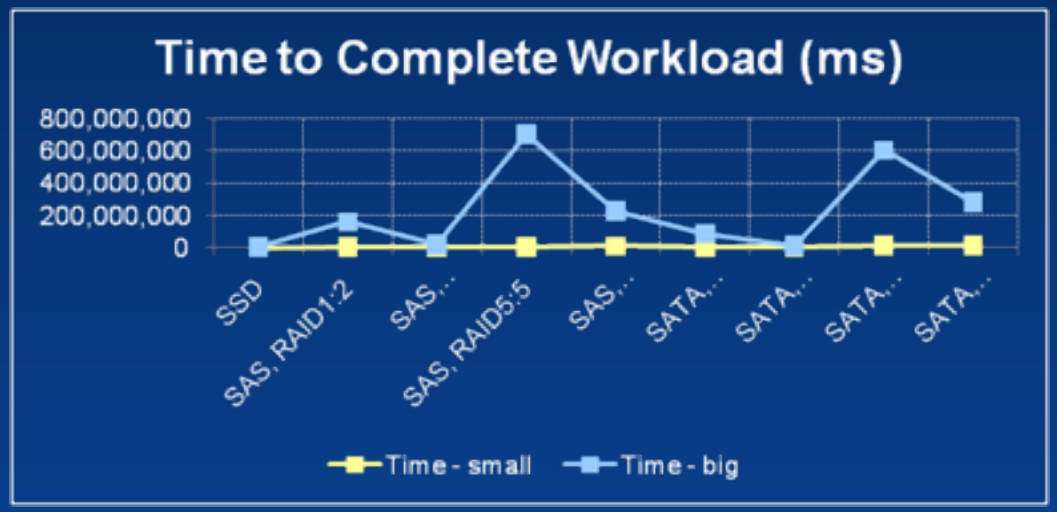
- Microsoft SQLIOSim
  - Log on same RAID1 set
- Two sets of tests with different database sizes
  1. InitialSize = 500 MB, MaxSize = 1000 MB, Increment = 50 MB, LogFile = No, Shrinkable = No, Sparse = No
  2. InitialSize = 5000 MB, MaxSize = 10000 MB, Increment = 500 MB, LogFile = No, Shrinkable = No, Sparse = No

# Test 2 – Exchange

- Exchange Server 2007 profile
  - 500 mailboxes (550 for SATA drives)
  - Mailbox size = 250MB
  - Exchange IOPS
    - SSD (no RAID) drives: 0.5 (heavy user)
    - SAS (RAID-1) drives: 0.5 (heavy user)
    - SATA (RAID-1) drives: 0.3 (light user)
  - Two sets of runs
    - A – Large server with 32GB RAM
    - B – Small server with 4GB RAM

# Test 2 – SQL Server

(Run on small server with 4GB RAM and internal drives)




| | Small | | Big | |
|---|---|---|---|---|
| Device | Throughput | Total Time | Throughput | Total Time |
| SSD | 0 | 87,966 | 0 | 151,256 |
| SAS | 1,412,753 | 25,613 | 150,80 | |
| SAS | 432,640 | 56 | 27,568 | 20,113 |
| SAS | 1,348,205 | 1 | 20,781 | 702,11 |
| SAS | 225,209 | 2 | 20,485 | 228,20 |
| SATA | 1,730,732 | 2 | 21,740 | 87,221 |
| SATA | 921,264 | 1 | 22,614 | 11,981 |
| SATA | 806,947 | 1 | 22,622 | 602,08 |
| SATA | 207,421 | 20 | 20,611 | 281,41 |

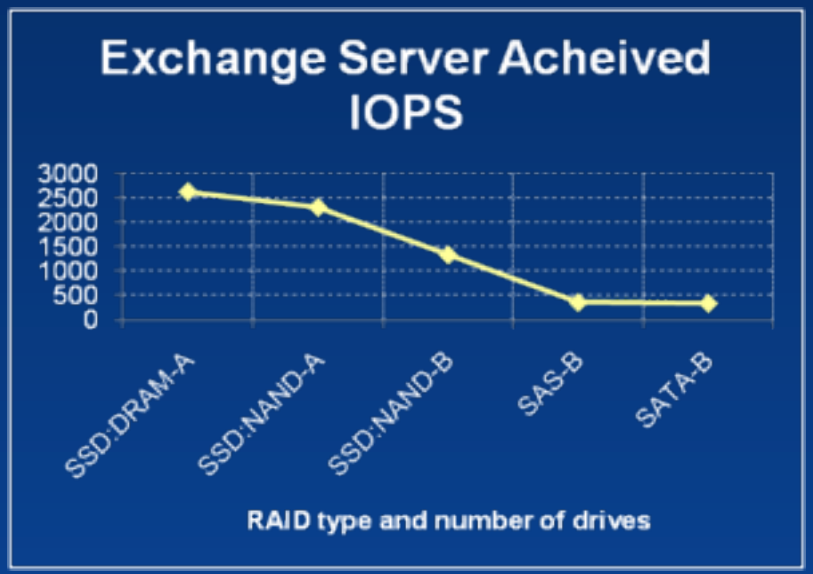
Small = 0.5GB – 1GB database
Big = 5GB – 10GB database

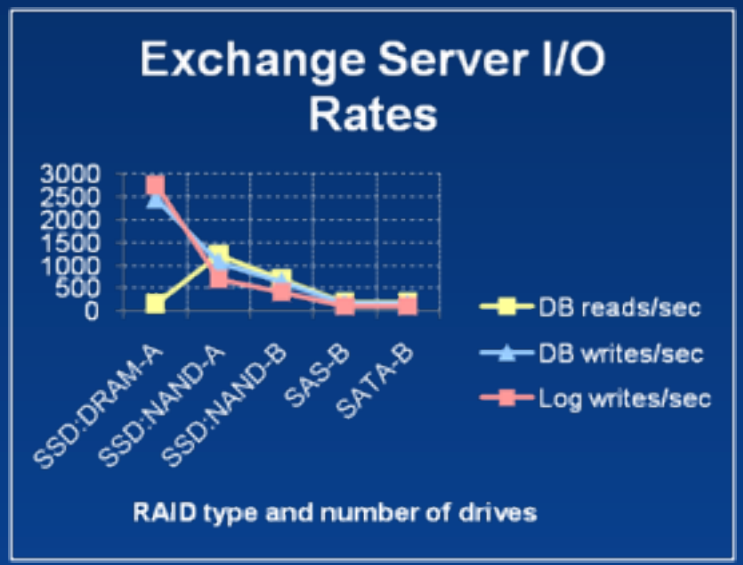
For "Big" databases, SSD and RAID10 (8 drive) configurations had no delayed I/O. All other configurations of "Big" databases had delayed I/O of at least 15 seconds.

SSD = NAND-Flash SLC

# Test 2 – Exchange

(Run on large server with 32GB RAM “A” and small server with 4GB RAM “B”)




| RAID: | Achi | DB | DB | Log | DB | DB | Log |
|---|---|---|---|---|---|---|---|
| SSD:D | 2623 | 166.4 | 2457 | 2761 | 0.000 | 0.000 | 0.000 |
| SSD:N | 2303 | 1236 | 1066 | 699.4 | 0.000 | 0.000 | 0.000 |
| SSD:N | 1327 | 688.0 | 639.7 | 414.9 | 0.000 | 0.000 | 0.000 |
| SAS-B | 357.9 | 191.7 | 166.1 | 109.7 | 0.004 | 0.006 | 0.001 |
| SATA- | 330.4 | 181.1 | 158.3 | 103.6 | 0.005 | 0.006 | 0.001 |

# Futures

- I believe that at the current rate of price decreases, flash SSDs will become the new standard for tier 1 storage within three years
- Research is underway on other types of memory technology that may become good candidates for storage devices
  - Other sessions here at the Flash Memory Summit will discuss these other technologies


# References: SQL Server

- SQL Server 2005 pre-deployment I/O Best Practices http://www.microsoft.com/technet/prodtechnol/sql/bestpractice/pdpliobp.mspx
- SQL Server 2005 I/O Basics http://www.microsoft.com/technet/prodtechnol/sql/2005/iobasics.mspx
- Database Engine I/O Requirements: http://support.microsoft.com/kb/967576
- Tempdb requirements: http://support.microsoft.com/kb/917047/en-us


# References: Exchange

- Exchange Server 2007 Storage Guidelines
  http://technet.microsoft.com/en-us/library/bb124518.aspx


# References: SharePoint

- SharePoint Server 2007 Best Practices
  http://technet.microsoft.com/en-us/office/sharepointserver/bb736746.aspx


# Contact Information

Dennis Martin, President
Demartek
(303) 940-7575
dennis@demartek.com

www.linkedin.com/in/dennismartin
http://twitter.com/demartek

blog: www.wikibon.org