COMPUTERWORLD SNIA

SNW

Driving Innovation Through the Information Infrastructure

SPRING 2011

Deep Dive on Solid State Storage

The Technologies & Architectures

Dennis Martin

President

Demartek

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)
  - Servers: 8+ cores, very large RAM
  - Virtualization: ESX, Hyper-V, Xen
- We prefer to run real-world applications to test servers and storage solutions
  - Currently testing various SSD and FCoE technologies
  - We create our own data sets for application workloads
- Web: www.demartek.com

Solid-State Storage Overview

- Uses memory technology as the storage media and appears as a disk drive to the O.S.
- Very fast, no moving parts
- Variety of form factors
- Prices dropping
- Some SSDs use DRAM and NAND-Flash together
- Capacities doubling almost yearly

Acronyms & Buzzwords

- SSD: Solid-State Drive (or Disk)
- SSS: Solid-State Storage
- SLC: Single-Level Cell
- MLC: Multi-Level Cell
- P-E Cycle: Program-Erase Cycle
- EFD: Enterprise Flash Drive
- SCM: Storage Class Memory

NAND-Flash SSD

- IOPS
  - 10K – 250K reads per device
    - Enterprise HDD: 100-200 IOPS
    - Desktop HDD: < 100 IOPS
  - Writes are generally slower than reads
- Capacities
  - Individual devices
    - Drive form factor: up to 1TB
    - PCIe card: up to 1.2TB
  - Arrays: Up to 250TB (“all-SSD” arrays)

NAND-Flash: What Is It?

- A specific type of EEPROM
  - EEPROM: Electrically Erasable Programmable Read-Only Memory
  - The underlying technology is a floating-gate transistor that holds a charge
- Bits are erased and programmed in blocks
  - Process is known as the Program-Erase (P-E) cycle
  - Flash blocks are typically 4KB, some larger

NAND Flash Technologies

- Single-Level Cell (SLC) – One bit per cell
- Multi-Level Cell (MLC) – Two or more bits per cell
  - Triple Level Cell (TLC) – Three bits per cell

	SLC	MLC-2	MLC-3	MLC-4
Bits per cell	1	2	3	4
Performance	Fastest	⟷	⟷	Slowest
Endurance	Longest	⟷	⟷	Shortest
Capacity	Smallest	⟷	⟷	Largest
Error Prob.	Lowest	⟷	⟷	Highest
Price per GB	Highest	⟷	⟷	Lowest
Applications	Enterprise	Mostly Consumer	Consumer	Consumer

The first announcements of MLC-3 and MLC-4 were made in 2009.

NAND Flash: Endurance & Price

- Endurance
  - SLC typically 10-20 times better than MLC-2
  - SLC typical life of 100,000 write cycles
    - Newer “enterprise SLC” may have 3x write cycles
  - MLC-2 is much better than MLC-3 or MLC-4
  - MLC typical life 10,000 or fewer cycles
    - Newer “enterprise MLC” may have 3x write cycles
- Price
  - SLC typically greater than 2x the price of MLC-2 for the same capacity

NAND Flash: General Trends

- Process sizes are shrinking
  - History: 90, 72, 50 nm
  - 2009: 34, 32 nm
  - 2010-2011: mid-20's nm
- Page sizes, block sizes, and Error Correction Code (ECC) requirements are increasing

NAND Flash: General Trends

- Data retention, endurance, and performance are decreasing as bits per cell increase
  - For consumer applications, endurance becomes less important as density and capacity increase
- Power consumption increases somewhat as bits per cell increase beyond 2 bits per cell
- Newer NAND flash controllers bring some SLC features to MLC flash

Power and Cooling

Device type	RPM	Form factor	Interface	Watts Typical	Watts Idle
Spinning disk	15K	3.5”	FC/SAS	13 – 19	8 – 14
Spinning disk	15K	2.5”	SAS	8 – 14	5 – 7
Spinning disk	10K	3.5”	FC/SCSI	11 – 18	6 – 13
Spinning disk	10K	2.5”	SAS	8 – 14	3 – 6
Spinning disk	7.2K	3.5”	SAS/SATA	7 – 13	3 – 9
Spinning disk	7.2K/5.4K	2.5”	SATA	1 – 4	0.7 – 1
SSD: SLC-flash	-	*	SAS/SATA	1 – 8	0.05 – 4
SSD: MLC-flash	-	*	SAS/SATA	0.1 – 2	0.05 – 0.5

Typically in datacenters, every watt of power consumed by computing equipment requires another watt of power to cool it.

* SSDs are available in 3.5”, 2.5” and 1.8” HDD form factors and other form factors

Flash in Enterprise Products

- Disk array vendors
  - Primary storage: SSDs in standard HDD slots
  - Cache: SSD technology used as cache
- Appliance vendors – “Accelerators”
- Server vendors
  - Add flash on a PCI-Express bus card
  - Add flash directly onto the motherboard
  - Blade server mezzanine cards
- Is enterprise flash storage or memory?

Vendor Product Trends

- Automated data movement
  - Applies to primary storage
  - Moves hot data to SSD tier
  - Scheduled by minutes, hours, days, etc.
  - LUN level and beginning to see sub-LUN level automated data movement
- SSDs together in cache and primary storage
- External disk array controllers and internal RAID adapters are adapting to SSD speeds

O.S. Behavior with Flash

- Operating systems need to behave differently with flash SSDs
  - Trim – notify the underlying device regarding data that is no longer needed
    - Trim is currently available for SATA interfaces only. The SAS committee has added UNMAP to the SAS/SCSI spec.
  - Windows 7 and Windows Server 2008 R2
    - Defragmenting is off by default for flash SSDs
  - RHEL 6 with EXT4 only, but Trim is not enabled by default
- Utilities (Intel RapidStorage 9.6+, etc.)

SSD: Cache

- Caching controller identifies any frequently accessed data (“hot data”)
- Caching controller automatically moves hot data to SSD media
- Multiple applications can benefit from the SSD cache simultaneously
- Performance improves over time, as cache is populated with data
- Overall HDD I/O load is reduced: fewer I/Os

SSD: Primary Storage

- User decides what data to place on SSD
- User decides when to place data on SSD
- User moves specific data to SSD
- SSD benefits only the applications that use the data placed on the SSD
- Performance improves instantly
- Automation software can help select and move data to SSD

SSD Performance Test: Web Server Workload

- Must maintain consistent response times
- Must handle sudden increases in traffic
- Must be cost-effective

Web Server Response Time

- Jakob Nielsen's Alertbox, June 21, 2010
  http://www.useit.com/alertbox/response-times.html
- Response-Time Limits
  - **0.1 seconds:** gives the feeling of instantaneous response
  - **1 second**: user's flow of thought is seamless
  - **1-10 seconds:** users feel at the mercy of the computer and wish it was faster
  - **10+ seconds:** users start thinking about other things

Web Server Setup

- Windows Server 2008 R2 with IIS 7.5
  - Default data compression disabled
  - 8dot3 short names removed and disabled
- 40GB of web server content
- 1.48 million files
  - 80000 HTML text pages
  - 1.4 million graphic images (JPEG and PNG)
- Represents a web hosting server with many sites and many pages

Two Sets of Tests

1. Comparing small configuration of desktop-class 7200 RPM SATA interface disk drives to SLC SSDs in caching configuration

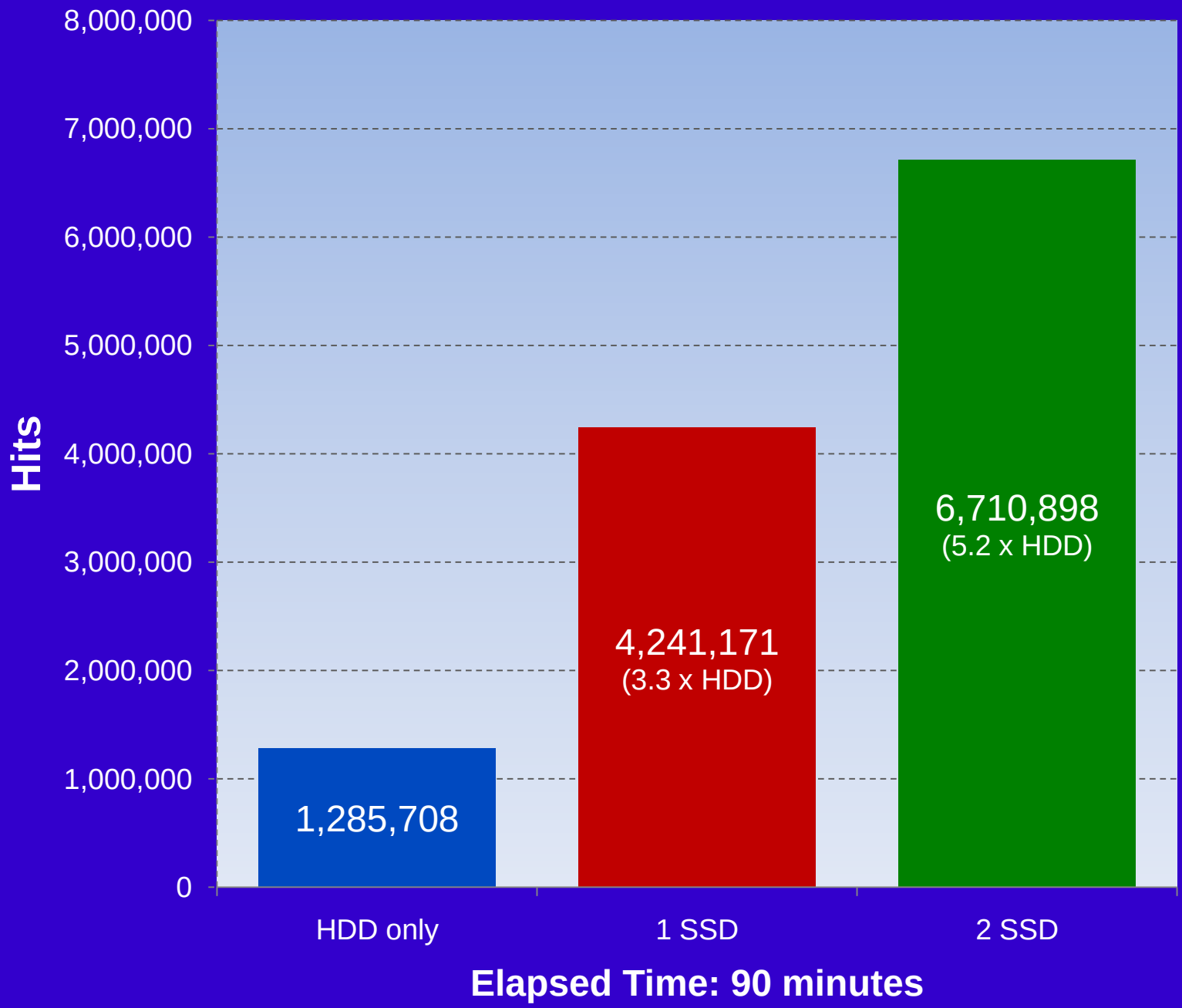
2. Comparing large configuration of enterprise-class 15K RPM SAS interface disk drives to PCIe SSD accelerator card in primary storage configuration

- Each test ran for 90 minutes

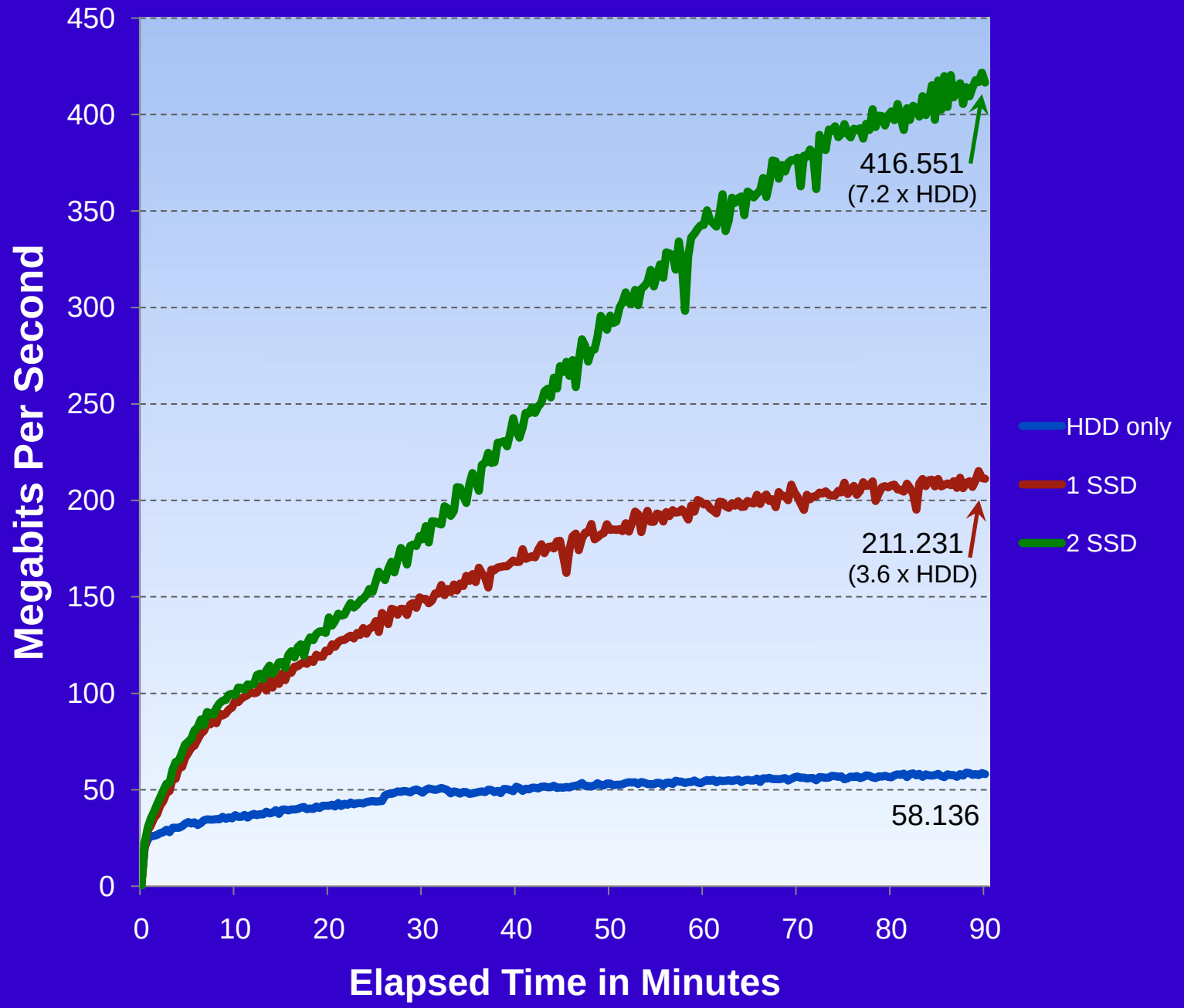
Test 1 – Caching
O.S. and Web content

- Configuration 1
  - 6 disk drives: 500GB SATA, 7200 RPM, 3.5-inch, RAID10
- Configuration 2
  - 6 disk drives: 500GB SATA, 7200 RPM, 3.5-inch, RAID10
  - 2 SSDs (cache): 32GB, SLC, 2.5-inch, SATA interface
- Network: 1GbE
  - Teamed NICs

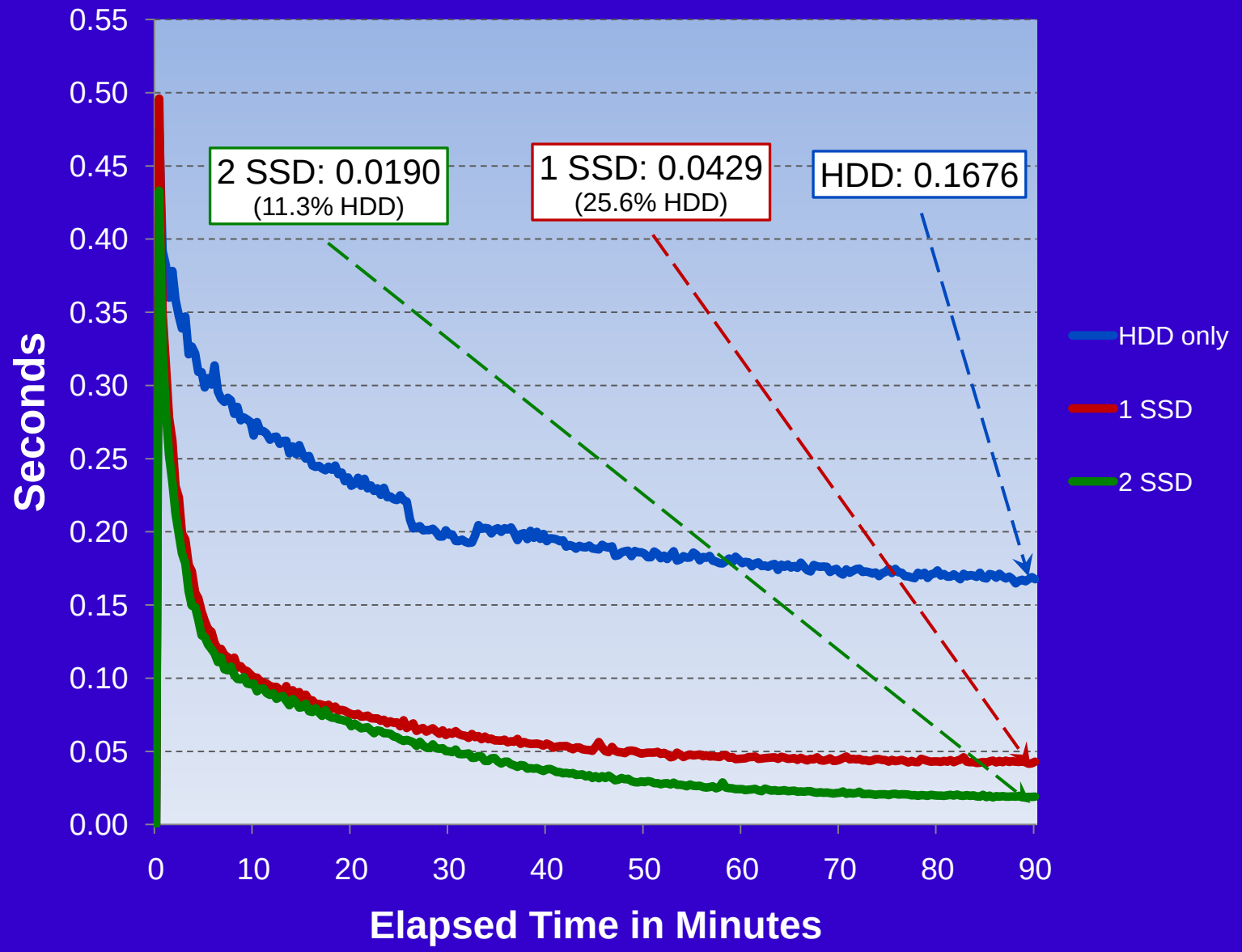
Total Hits – Caching Configuration



Throughput – Caching Configuration



Average Page Response Time
Caching Configuration
(Lower is better)

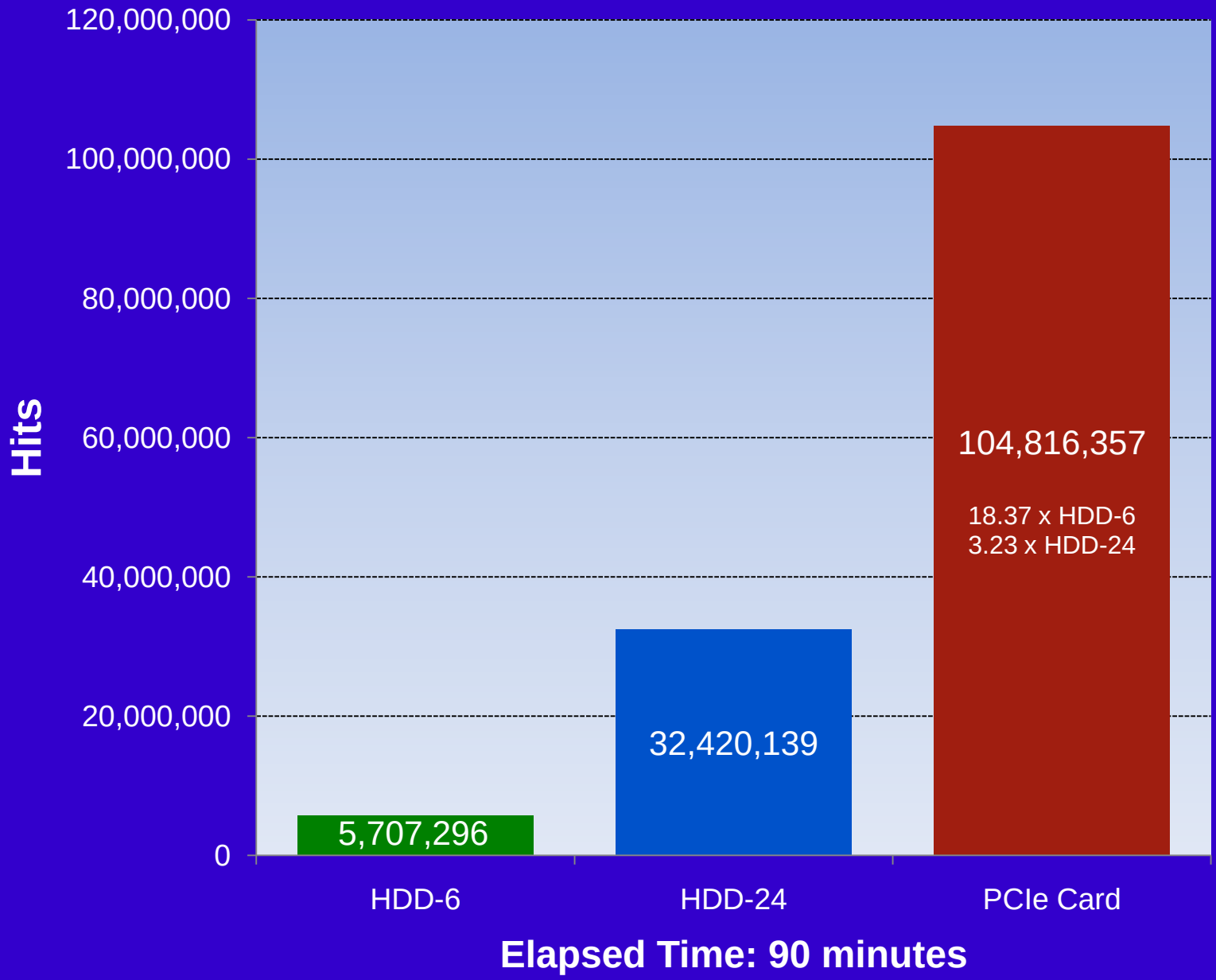


Test 2 – Primary Storage

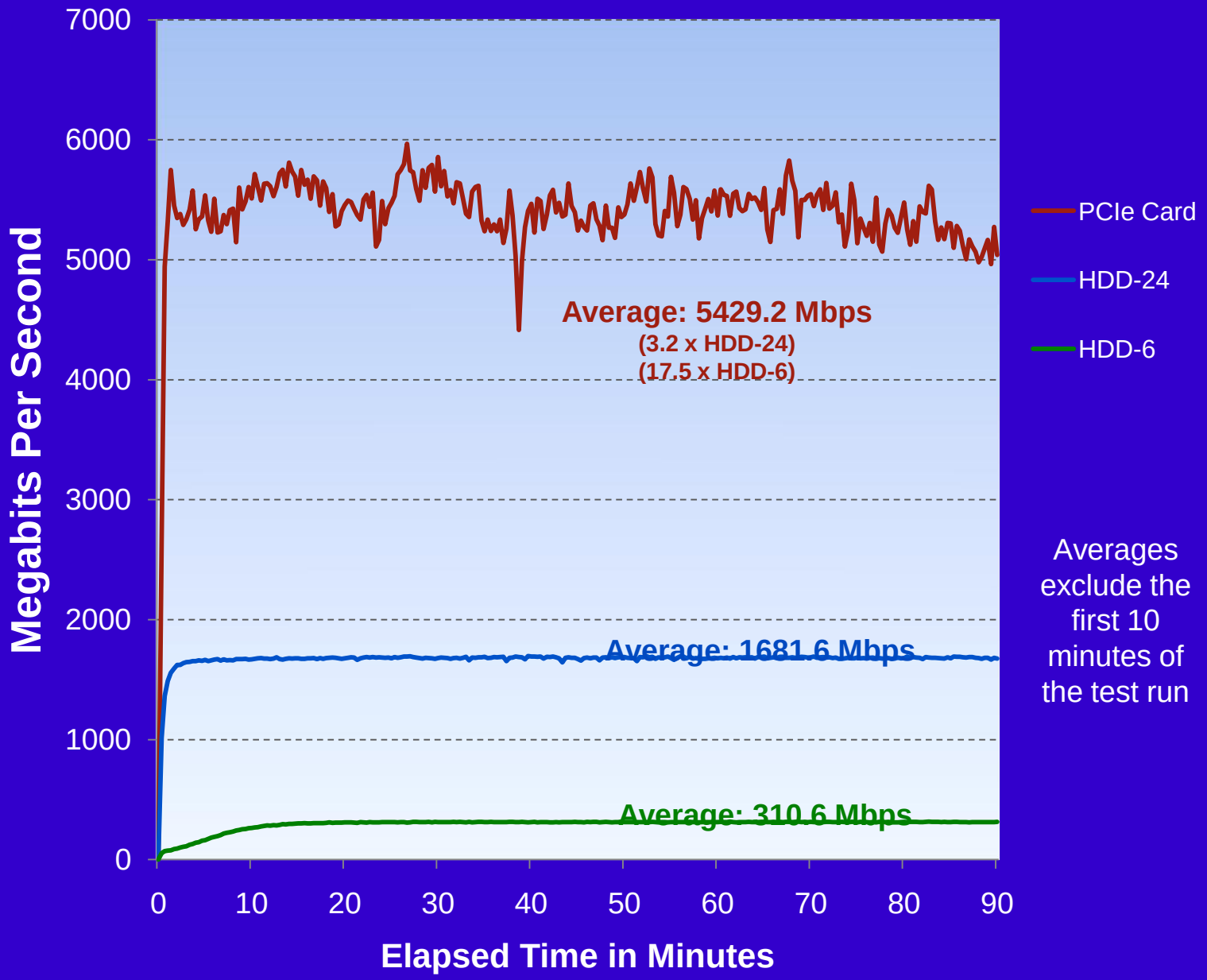
Web content only

- Configuration 1
  - 6 disk drives: 73GB 6Gbps SAS, 15K RPM, 2.5-inch, RAID10
- Configuration 2
  - 24 disk drives: 73GB 6Gbps SAS, 15K RPM, 2.5-inch, RAID 10
- Configuration 3
  - 1 PCIe SSD: 300GB SLC Flash
- Network: 10GbE
  - SSDs and 10GbE go well together

Total Hits – Primary Storage Configuration

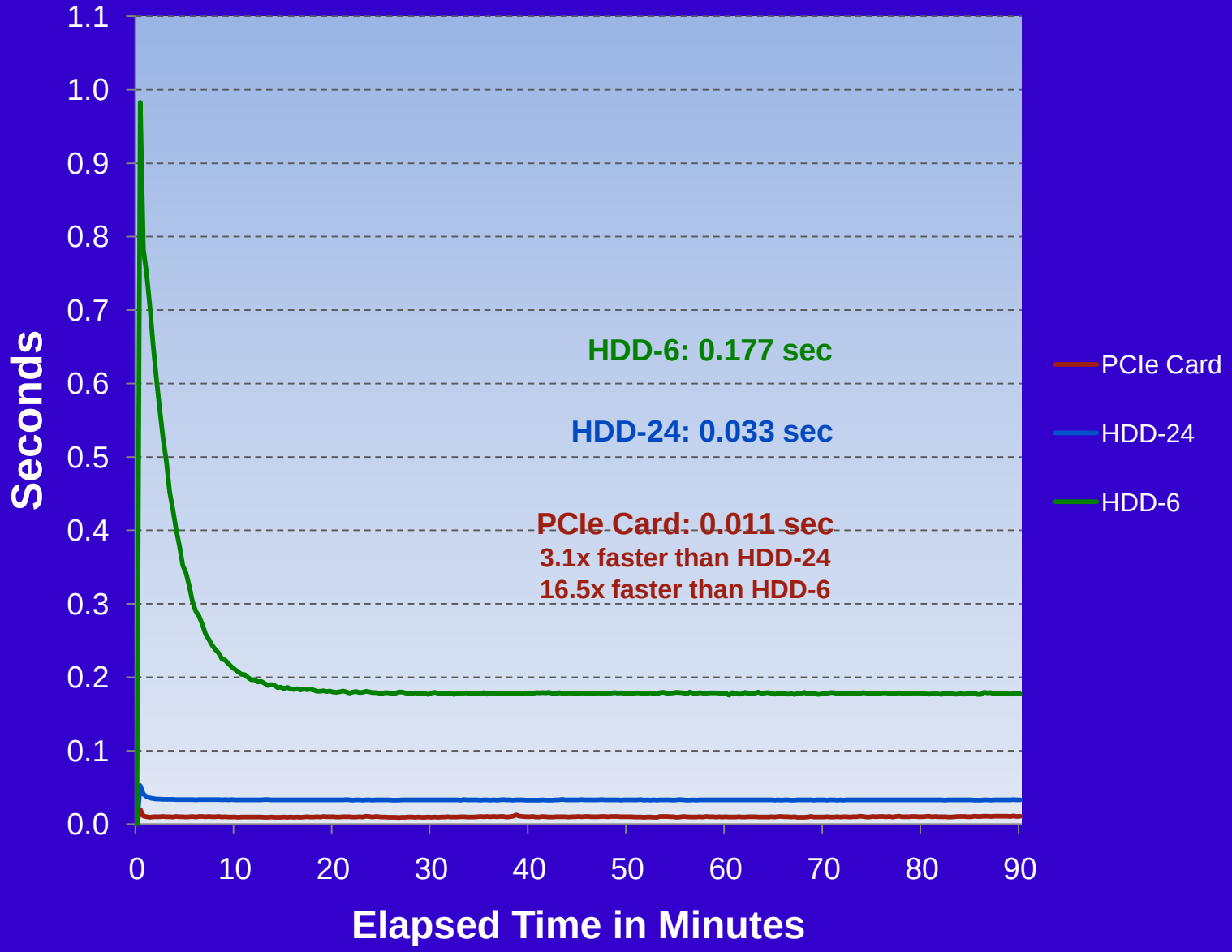


Throughput – Primary Storage Configuration

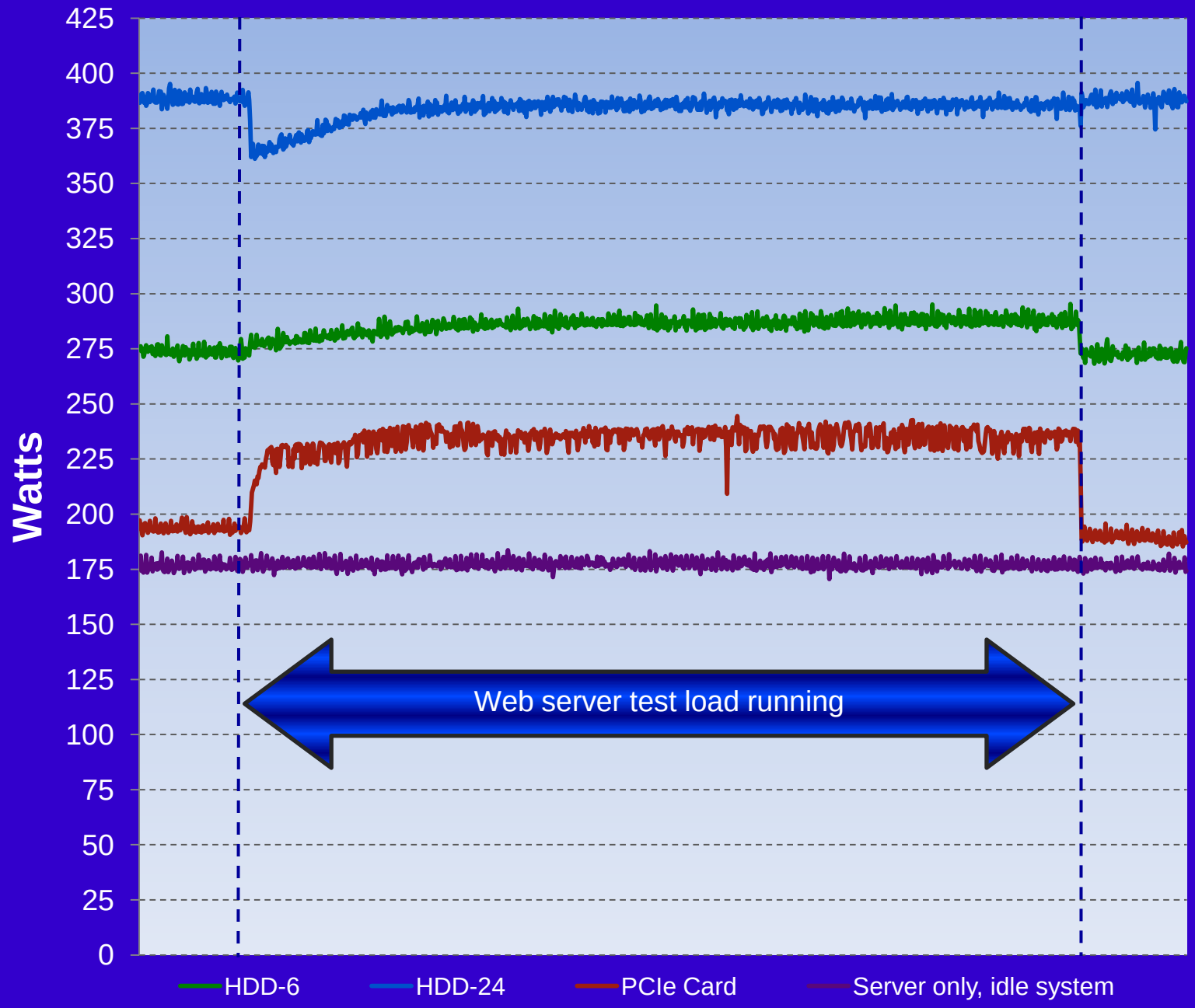


Averages exclude the first 10 minutes of the test run

Average Page Response Time
Primary Storage Configuration
(Lower is better)



Web Server Power Consumption



Performance Comments

- SSD technology can move the bottleneck to unexpected places
  - The 1GbE network was the bottleneck during the initial PCIe card test, requiring us to go to the 10GbE network to get the full performance of the PCIe card
- SSD technology can drive up CPU utilization
  - Considerably more work can get done with SSD technology, which can significantly increase CPU utilization

Future

- Emerging technologies, especially in the flash controllers, will enable MLC flash to become suitable for the enterprise
- Opinion: I believe that at the current rate of price decreases and capacity increases, SSDs (probably flash) will become the new standard for tier-1 storage by 2012.

Ongoing Research

- Other types of memory technology that may become good candidates for storage devices (within 3-5 years)
  - PCM: Phase Change Memory (PC-RAM)
  - Solid Electrolyte
  - MRAM: Magnetic RAM (Racetrack)
  - FeRAM: Ferroelectric RAM
  - RRAM: Resistive RAM (Memristor)

Demartek SSD Resources

- Demartek SSD Zone
  - http://www.demartek.com/SSD.html
- Look for my article ***Making the Case for Solid-State Storage*** in June online edition of Storage Magazine
  - http://searchstorage.techtarget.com
- Demartek Storage Interface Comparison
  - http://www.demartek.com/Demartek_Interface_Comparison.html
  - Or search for “storage interface comparison”

Free Monthly Newsletter

- Demartek publishes a free monthly newsletter highlighting recent reports, articles and commentary. Look for the newsletter sign-up at www.demartek.com.

Contact Information

(303) 940-7575
www.demartek.com
http://twitter.com/Demartek
YouTube: www.youtube.com/Demartek
Skype: Demartek

Dennis Martin, President
dennis@demartek.com
www.linkedin.com/in/dennismartin