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, iSCSI)
  - Servers: 8 cores, up to 96GB 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 solutions
- Web: www.demartek.com

Agenda

- Solid State Storage technology overview (DRAM and NAND Flash)
- Performance vs. cost
- Power & cooling
- Plug-in flash (using disk drive interfaces)
- How vendors are adding flash to their systems
- New architectures, integrations and trends
- Demartek lab results

Solid-State Storage Overview

- Uses memory 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

New 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

DRAM SSD

- Same type of memory that is in servers
- Volatile: needs battery or disk backup
- Highest IOPS: 70K – 5M+
- Latencies in microseconds
- Can be used as a cache in front of other storage

NAND-Flash SSD

- Non-volatile
- Quiet, low-power, low-weight, low-heat
- Types: SLC & MLC
- Variety of form factors
  - Disk drive
  - PCIe card
  - Motherboard module

NAND-Flash SSD

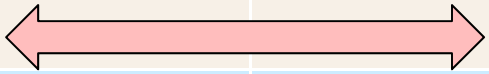
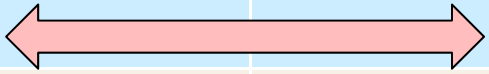
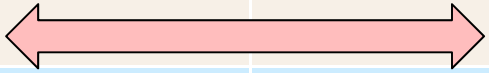
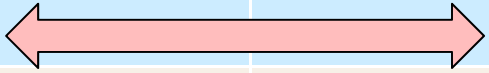
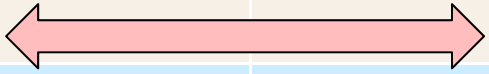
- IOPS
  - 10K – 250K reads per device
  - Writes can be slower than reads
- Capacities
  - Individual devices
    - Drive form factor: 300GB
    - PCIe card: 1.2TB
  - Arrays: Up to 105TB (“all-SSD” arrays)

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

NAND Flash Technologies

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

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
  - MLC-2 is much better than MLC-3 or MLC-4
  - MLC typical life 10,000 or fewer cycles
  - Recent announcements of “Enterprise MLC”
- Price
  - SLC typically greater than 2x the price of MLC-2

NAND Flash: General Trends

- Process sizes are shrinking
  - History: 90, 72, 50 nm
  - 2009: 34, 32 nm
  - Future: Expect mid-20s nm in 2010-2011
- 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

NAND Flash: Controllers

- 2009: NAND Flash controllers bring SLC features to MLC Flash
- 2010-2011: Expect more enterprise MLC solutions

NAND Flash: Memory Wear

- Bad Block management
  - Mitigate for NAND errors
- Wear Leveling
  - Arrange for an even distribution of erase counts across the flash media
- Additional Features
  - Extra, un-advertised capacity
  - ECC (Error Correction Code)

Performance vs. Cost

- Several vendors offer cost comparisons of their storage arrays with only hard disk drives (HDDs) vs. their arrays with a mix of HDDs and SSDs
- Mixed-drive configurations can be less expensive than single-drive configurations

Performance vs. Cost

- High-end array
  - Traditional configuration (55 TB usable):
    - 244 x 300GB, 15K FC Disks
  - New configuration (55 TB usable):
    - 8 x 73GB Flash
    - 136 x 300GB 15K FC Disks
    - 32 x 1TB SATA Disks

60% More IOPS	28% Fewer drives
21% Less power & cooling	17% Lower cost

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	-	*	SATA	1 – 3	0.05 – 0.1
SSD: MLC-flash	-	*	SATA	0.1 – 0.5	0.05 – 0.1

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

Plug-in Flash

- 3.5, 2.5 & 1.8 inch disk drive form factor
  - Interfaces available today: SATA, SAS, PATA/IDE, FC, U320 SCSI, Narrow SCSI
  - Trends: drive vendors are moving to 6Gb SAS, SATA (6Gb & 3Gb), and some USB 3.0
- PCI-Express bus cards
- Mounted on server motherboards
  - JEDEC MO-258A

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.
  - Used at LUN level today; beginning to see sub-LUN level automated data movement
- SSDs together in cache and primary storage
- SSD-only arrays for file and block storage

Moving the Bottleneck

- External disk array controllers and internal RAID adapters must adapt
  - New class of high-speed devices
  - Higher speed interfaces
- Higher speed interfaces
  - PCI-Express bus (version 1.x and 2.0)
  - 6-Gbps serial: SAS (2009) and SATA-3 (2010)
  - 5-Gbps USB 3.0 (2010)
  - 10-Gbps FCoE and iSCSI
  - 16-Gbps FC in 2011?

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, but Trim is not enabled by default
- Utilities (Intel RapidStorage 9.6+, etc.)

Optimal I/O Workloads

- Database indexes and temporary tables
- Any application data that is stored on “short-stroked” spinning disks, especially if you’re using 10% of drive capacity
- Any application where the entire dataset can be stored on solid-state storage

Demartek Lab Recent Results

- PCI-Express bus:
  - Jetstress IOPS up to 40,000+ with multiple PCI-Express SSD cards in a server
- 6Gb/sec SAS RAID controller:
  - IOmeter random read IOPS more than 100,000 with four SSDs in RAID0 stripe
  - IOmeter random write IOPS more than 20,000 with four SSDs in RAID0 stripe

SSD Performance Comments

- Enterprise applications only need small amount of SSD relative to total HDD capacity for significant performance gains
- Demartek tests with caching solutions show huge gains (5x – 8x) with only one or two SSDs in one disk enclosure

Future

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

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 2010 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

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

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


COMPUTERWORLD
SNIA
SNW
OCTOBER
11-14
2010
The Gaylord Texan, Dallas, Texas