The New Era of Solid State Drives: Technical Opportunities and Challenges

October 20, 2008

Daniel Donggi Lee, Ph.D
Flash Solution, Memory Division
Samsung Electronics Co., Ltd.

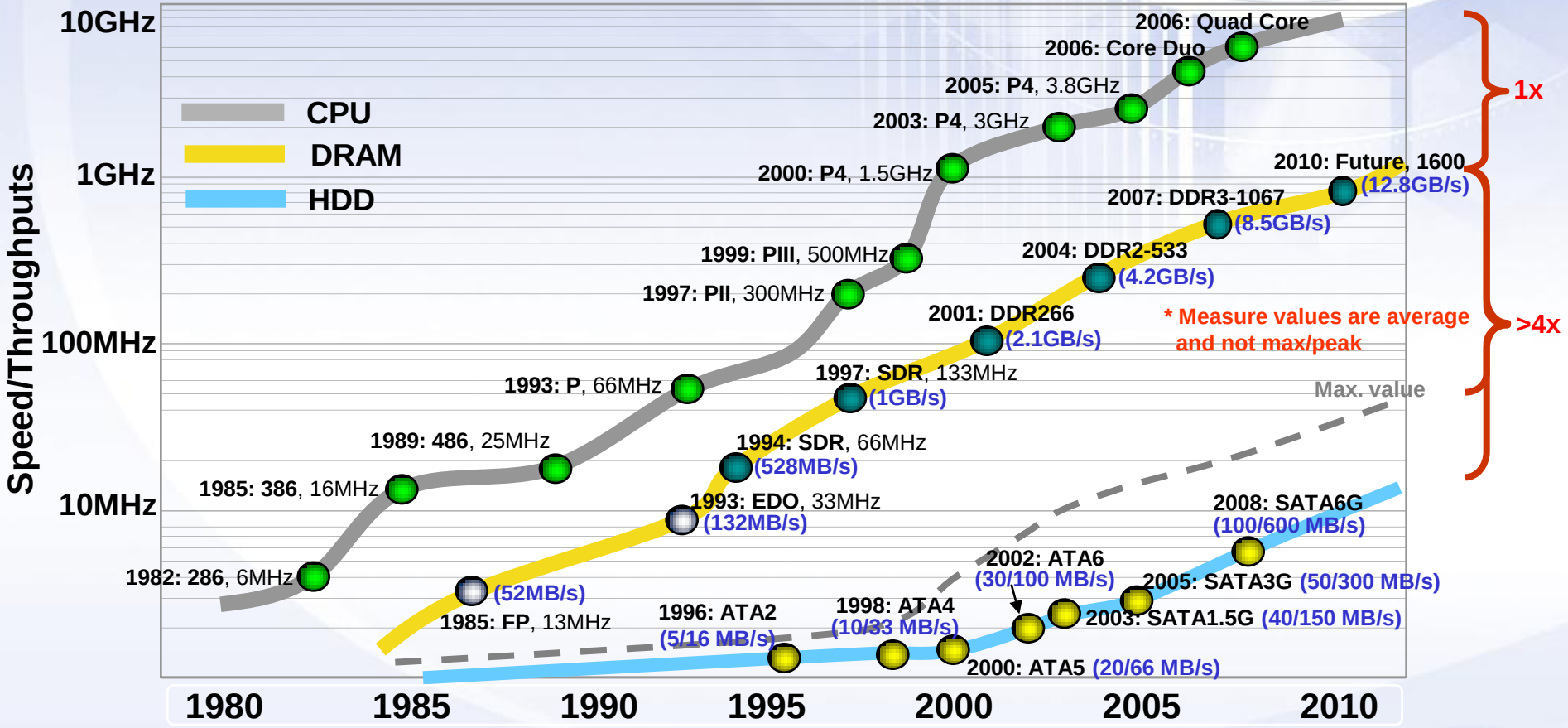
NVRAMOS 2008

SAMSUNG

Contents

A New Hurdle in Computing Platform

- Need to bridge performance gap in computing platform
  - CPU cache bridges CPU and system memory (DRAM) gap
  - Need to bridge the gap between system memory and HDD



Source: Samsung

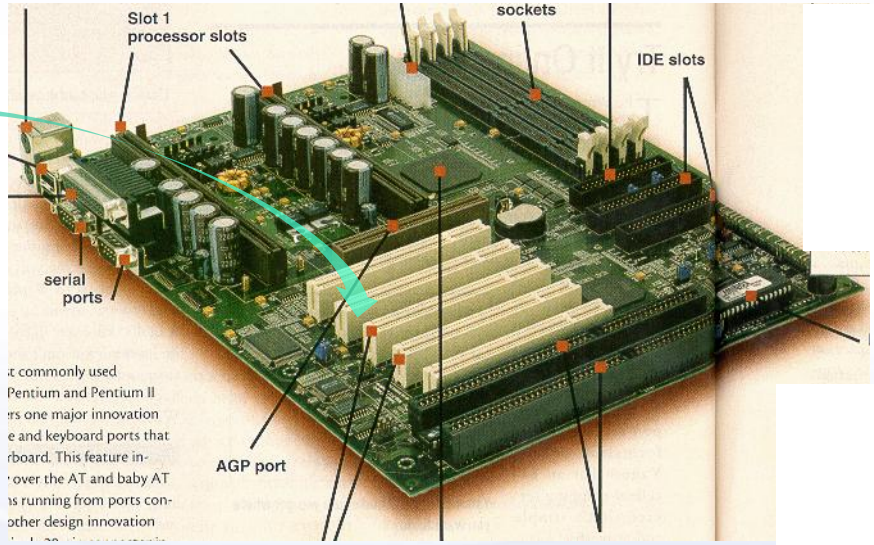
SAMSUNG

Existing Approaches - I

❑ NVC (Non-Volatile Cache)

- **Flash memory is added as a storage cache**
  - **Pros**
    - **No HDD SKU proliferation for PC OEMs**
  - **Cons**
    - **Data separation problem: Inconsistency may happen if either HDD or flash memory is detached from the system**
    - **System synchronization required in every power on cycle**

FUSION-IO AMD The future is fusion denali intel MARVELL



Source: www.fusionio.com

Existing Approaches - II

❑ SSD (Solid State Drive)

- Data storage device that uses sold-state memory
- Utilizing existing storage device interfaces
  - **Pros**
    - No additional system variation for computing platform
    - Fast TAT time for production launch
  - **Cons**

    - Cost, System optimization, variety of SKUs (over 83 vendors)

SAMSUNG intel MTRON STEC INDILINX SanDisk TOSHIBA Leading Innovation

Texas Memory System – RamSan-440

4500MB/s sustained throughputs
600k random IOPS
Fiber Channel I/F, DRAM SSD

Source: www.superssd.com

Violin Memory Inc – Violin 1010

Scalable Memory Architecture (VXM)
84 VIMMs (Violin Intelligent Memory Modules)
1M random IOPS
PCIe x4/x8 I/F, DRAM/Flash SSD

Source: www.violin-memory.com

Why NAND Flash is Good Solutions ?

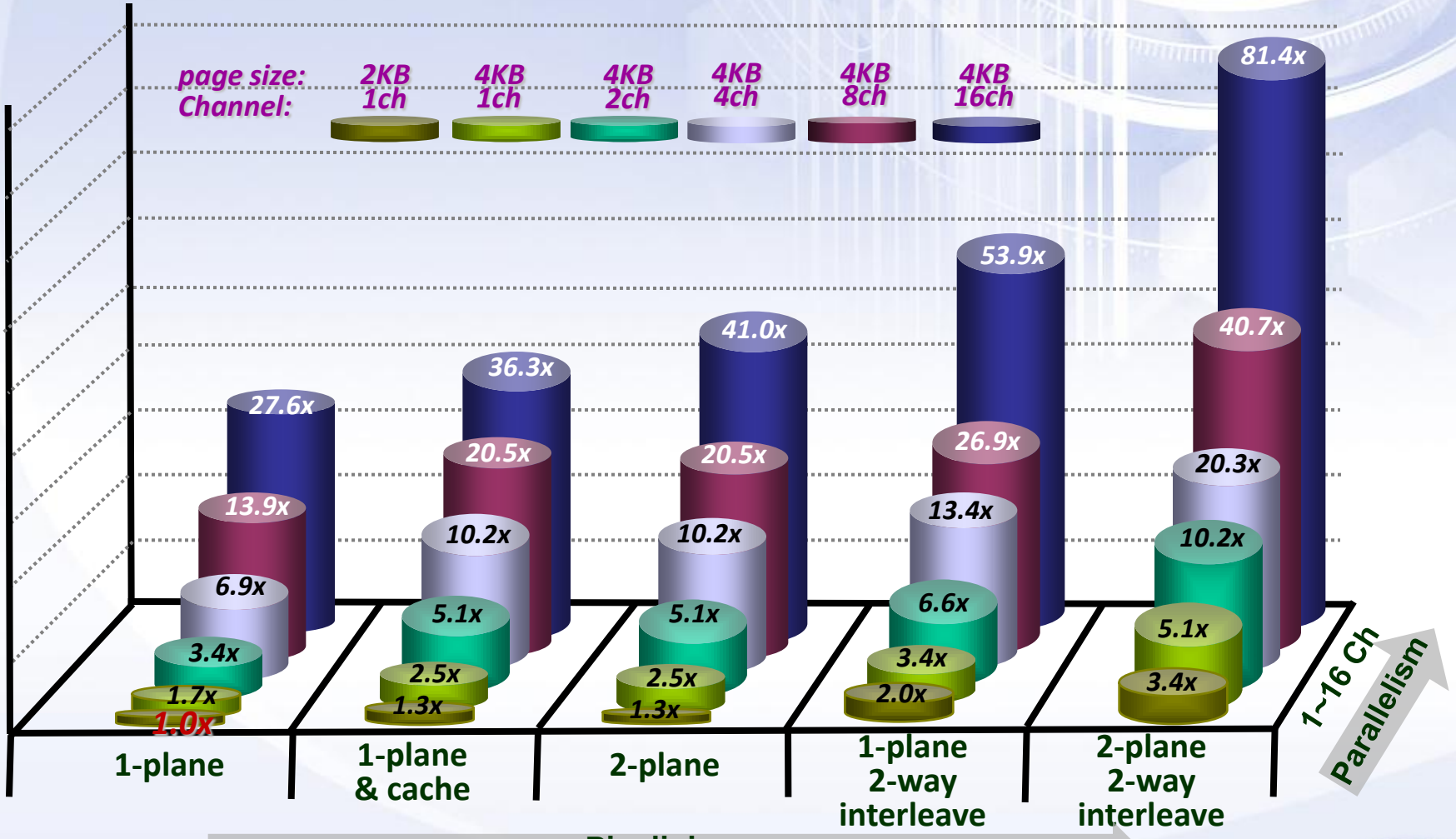
- **Flash complements HDD for Fast Seek and Low Power**
  - **Consistent performance at sequential and random access**
  - **Low power and instant standby and activation**

SAMSUNG NAND Flash		2.5” 5400rpm HDD
R: >34MB/s, W: >17MB/s	Sequential Access	R/W: 22 ~ 48MB/s
R: 34MB/s, W: 17MB/s	Random Access	R/W: 5 ~ 10MB/s
4/8/16Channels, N Chip *	Multi-Interleave	Not Available
54mW	Active Power	2.5W
90uW	Standby Power	2.5W

*Using Interleave, Flash can support the best read and write performance

NAND Flash like Toy of Lego™

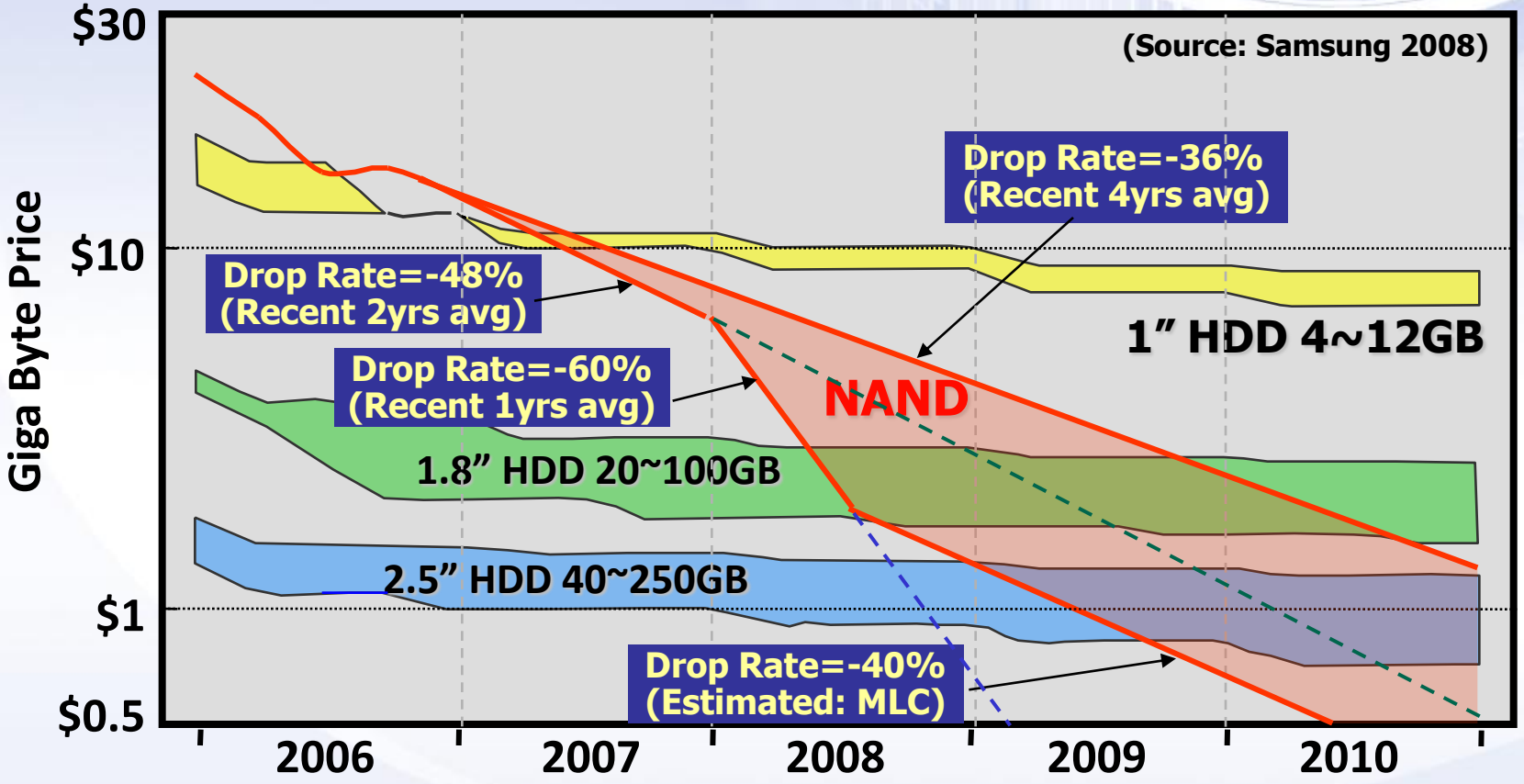
- 17MB/s with a single 50nm 8Gb SLC in one plane mode
- Performance = F {multi-plane, multi-way/channel…}



(Source: 2007 Samsung)

No More NAND Price is Limit !

- Corp. Notebook: 30~40GB/Home User: 48GB (eWeek.com '06 Aug.)
  - OS (XP: 3~4GB, Vista: 10GB) + App. (6~7GB) + User data (~20GB)
- 2010: 1.8" HDD 43Mpcs, 2B$ forecasted



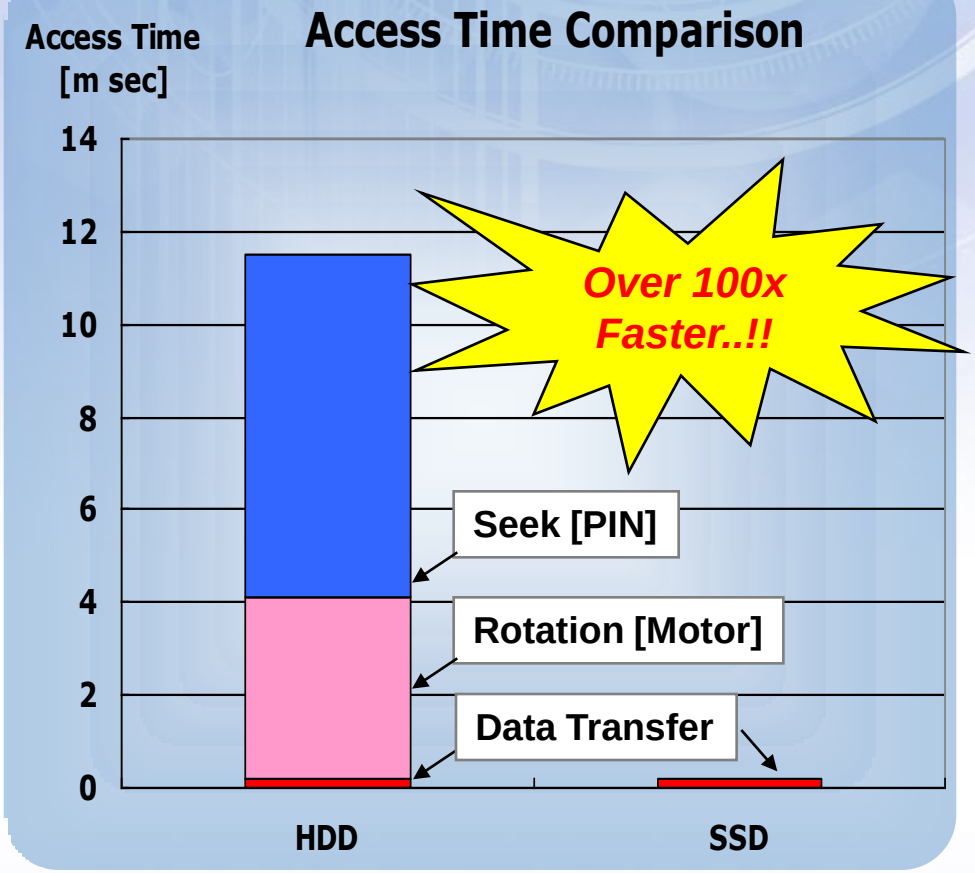
Contents

SSD Overview

❑ Overall Comparison


2.5" SATAII SSD (RBA MLC based)		2.5" SATAII HDD (7200rpm based)
Solid NAND Flash Based	Mechanism	Magnetic Rotating Platters
~256GB	Density	80GB
79g	Weight	365g
Read : 250MB/s Write : 160MB/s	Performance	Read : 59MB/s Write : 60MB/s
0.8W	Active Power Consumption	3.86W
20G (10~2000Hz)	Operating Vibration	0.5G (22~350Hz)
1500G/0.5ms	Shock Resistance	300G/2.0ms
MTBF > 1M hrs.	Endurance	MTBF < 700k hrs.

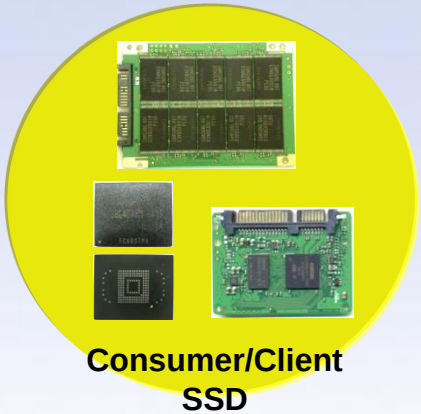
❑ Access Time Comparison



- SSD is superior to HDD considering all aspects..!!
- Especially, SSD has ultra fast access time thanks to no rotation platters..!!

SSD Lineup

❑ Product Categories



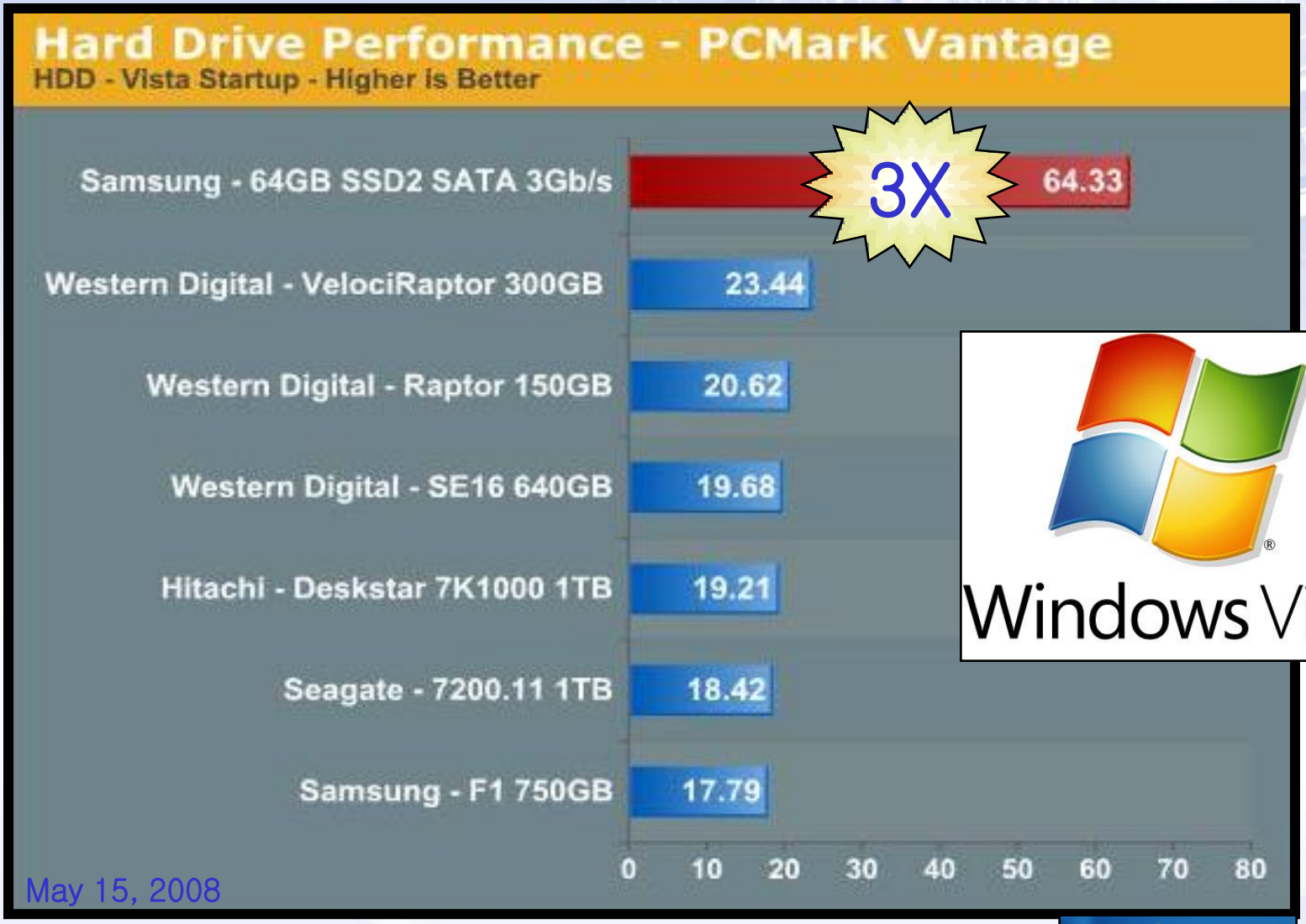
Consumer/Client SSD

Enterprise/Server SSD

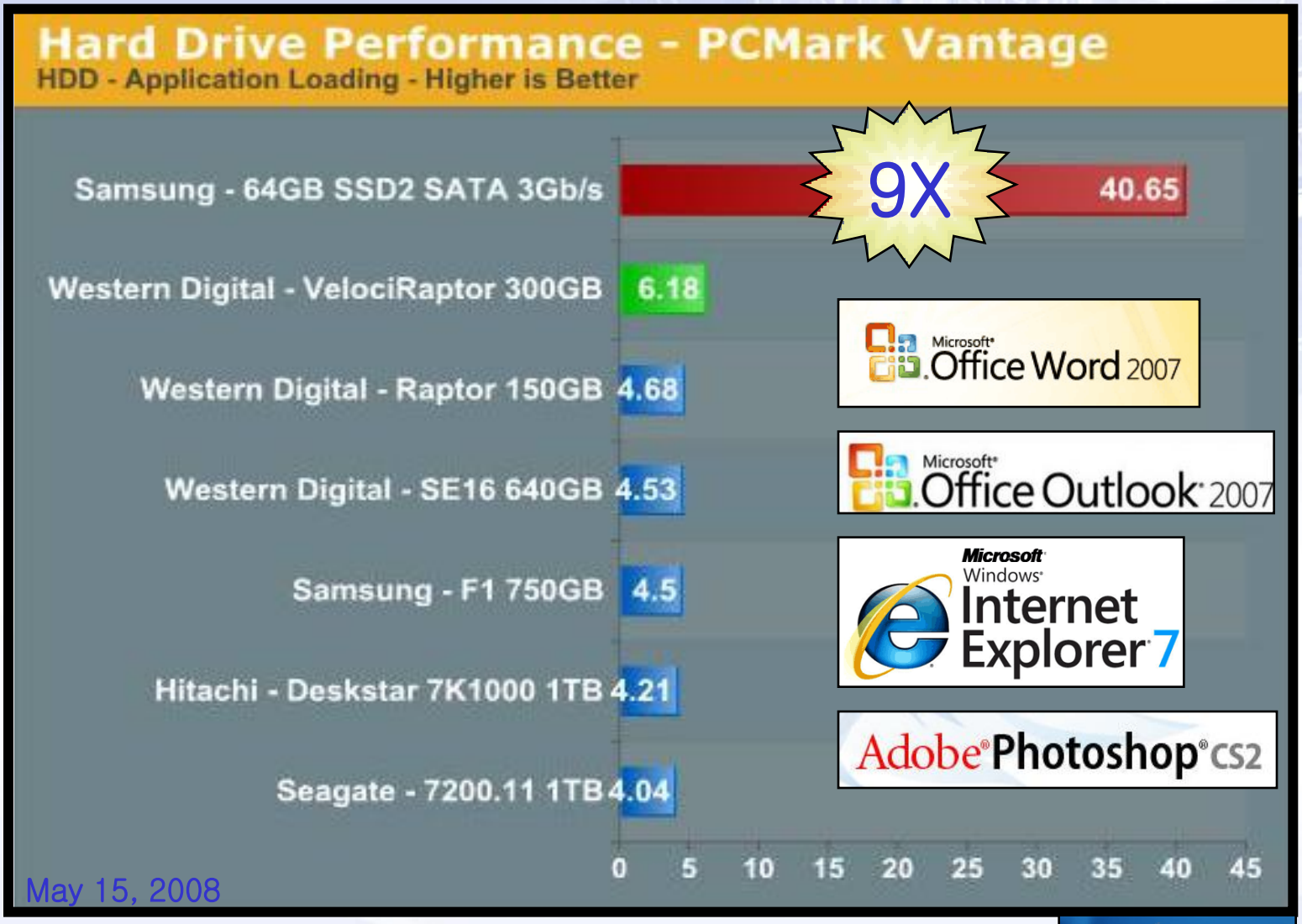
Military/Industrial SSD

Major Target	Consumer/Laptop Storage	Enterprise/Server Storage	Military/ Industrial Storage
Optimized For	Lowest $$/GB	Performance, Power	Reliability, Durability, Rugged, Security
Applications	Mixed	OLTP, Office Productivity, Data Mining	Data Recording, Mission Media
Trend	Moving to MLC	Higher MTBF	Encryption

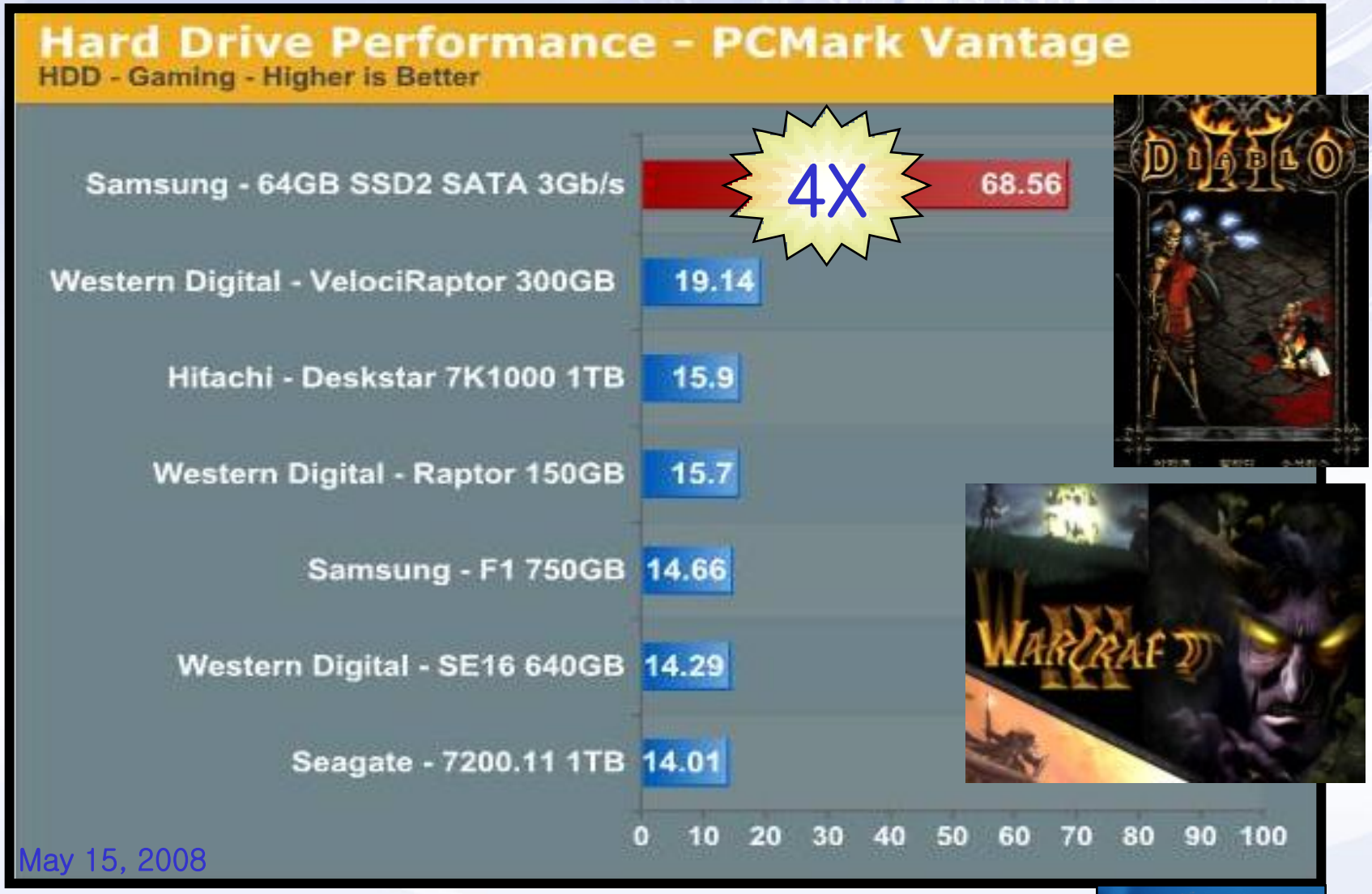
Fast Startup



Fast User Experiences

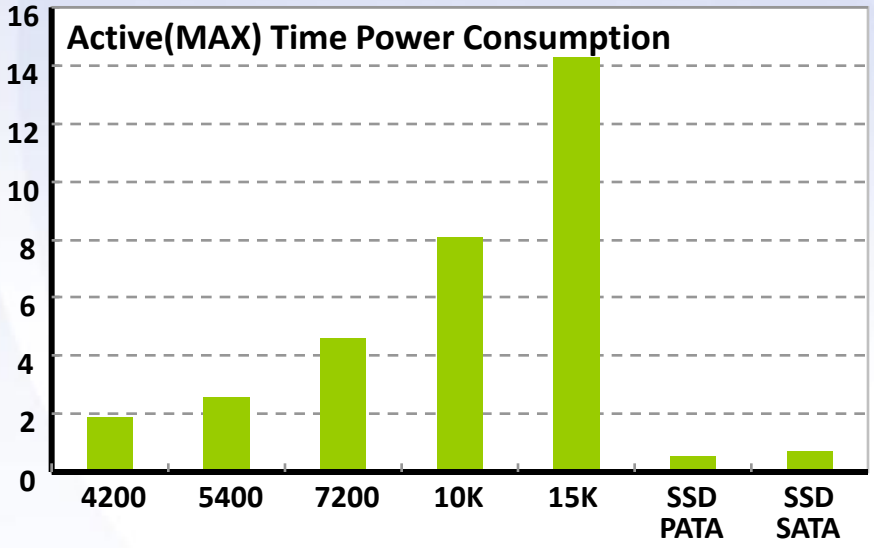


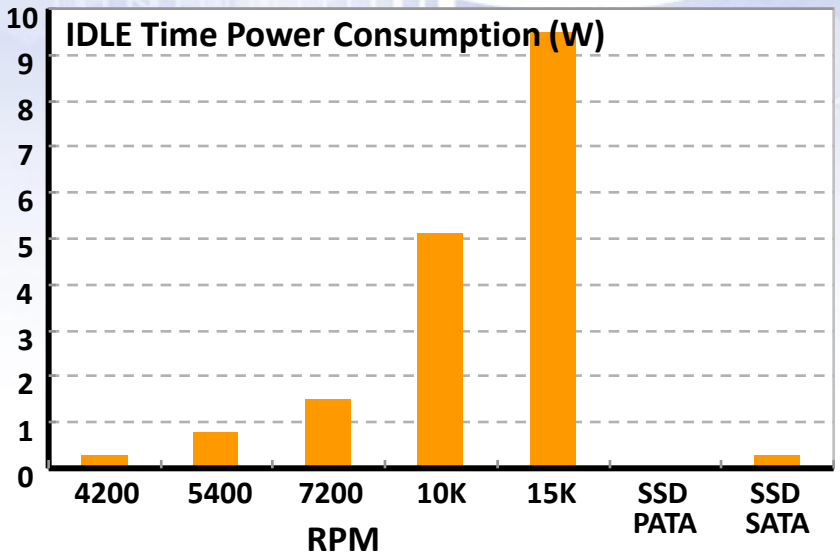
Fast Gaming Experiences



Power Savings

- **SSD has 9X to 14X power savings compared HDD**
  - **15K RPM 3.5” HDD, operating - 14W, idle – 9W**
  - **SSD, under 1W at both operating & idle**



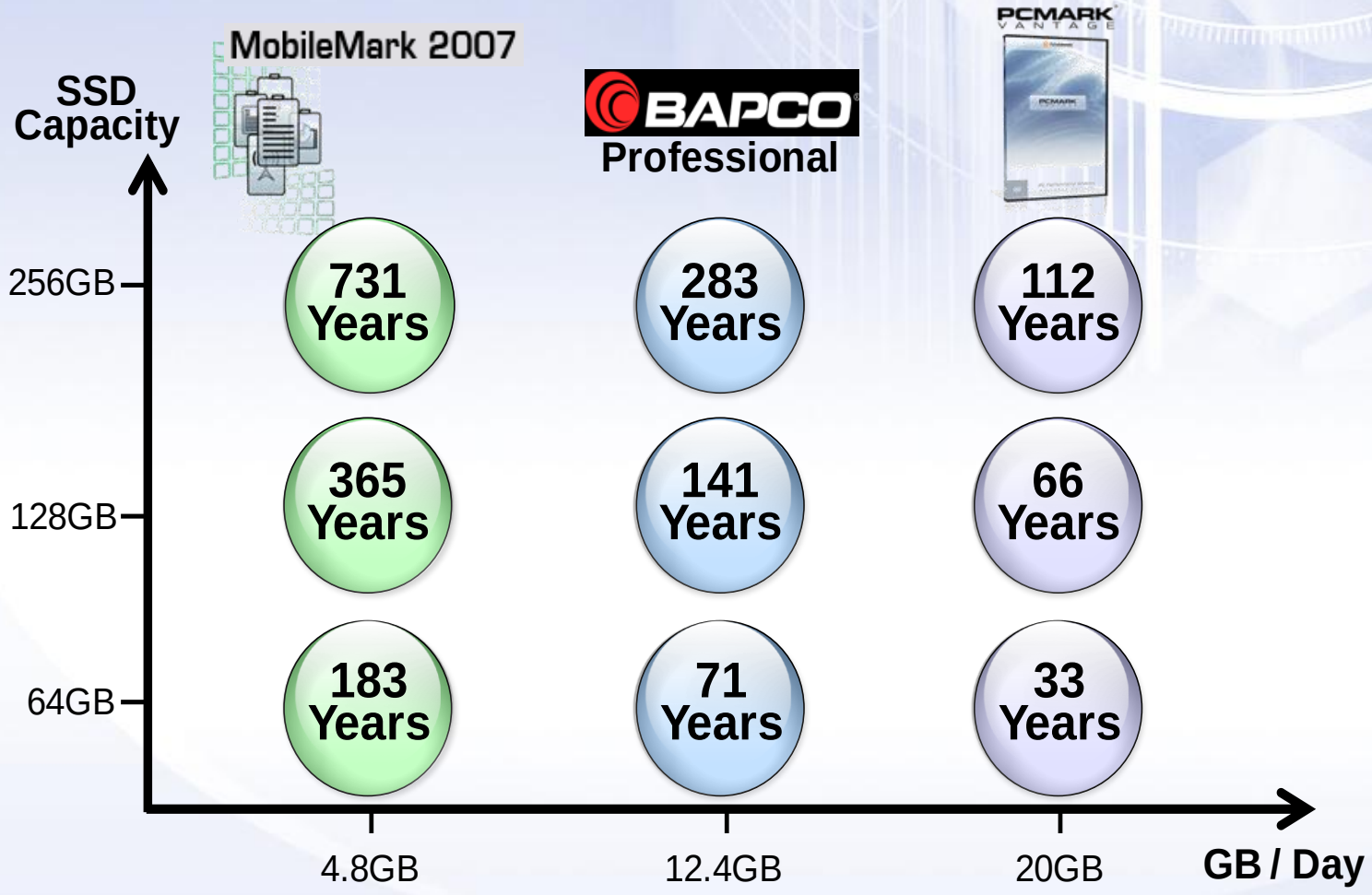


* HDD DATA back ground

RPM	Drive
4200	Hitachi GST Travelstar 4k120
5400	Toshiba MK1032GSX
7200	Hitachi GST Travelstar 7k100
10K	Seagate Savvio 10K.1
15K	Seagate Cheetah 15K.4

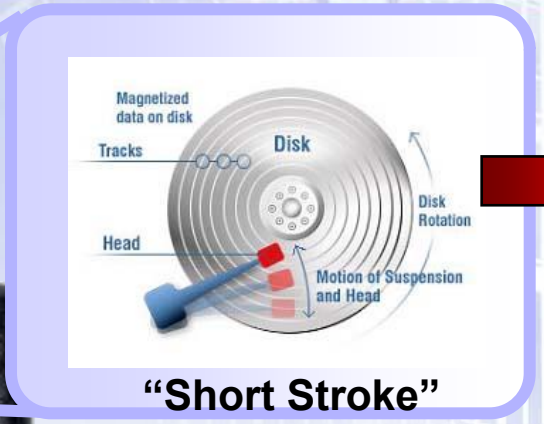
More Lifetime

❑ MLC SSD Estimated Notebook Lifetime



Server Opportunity

- SSD's Performance/Power overcomes HDD's

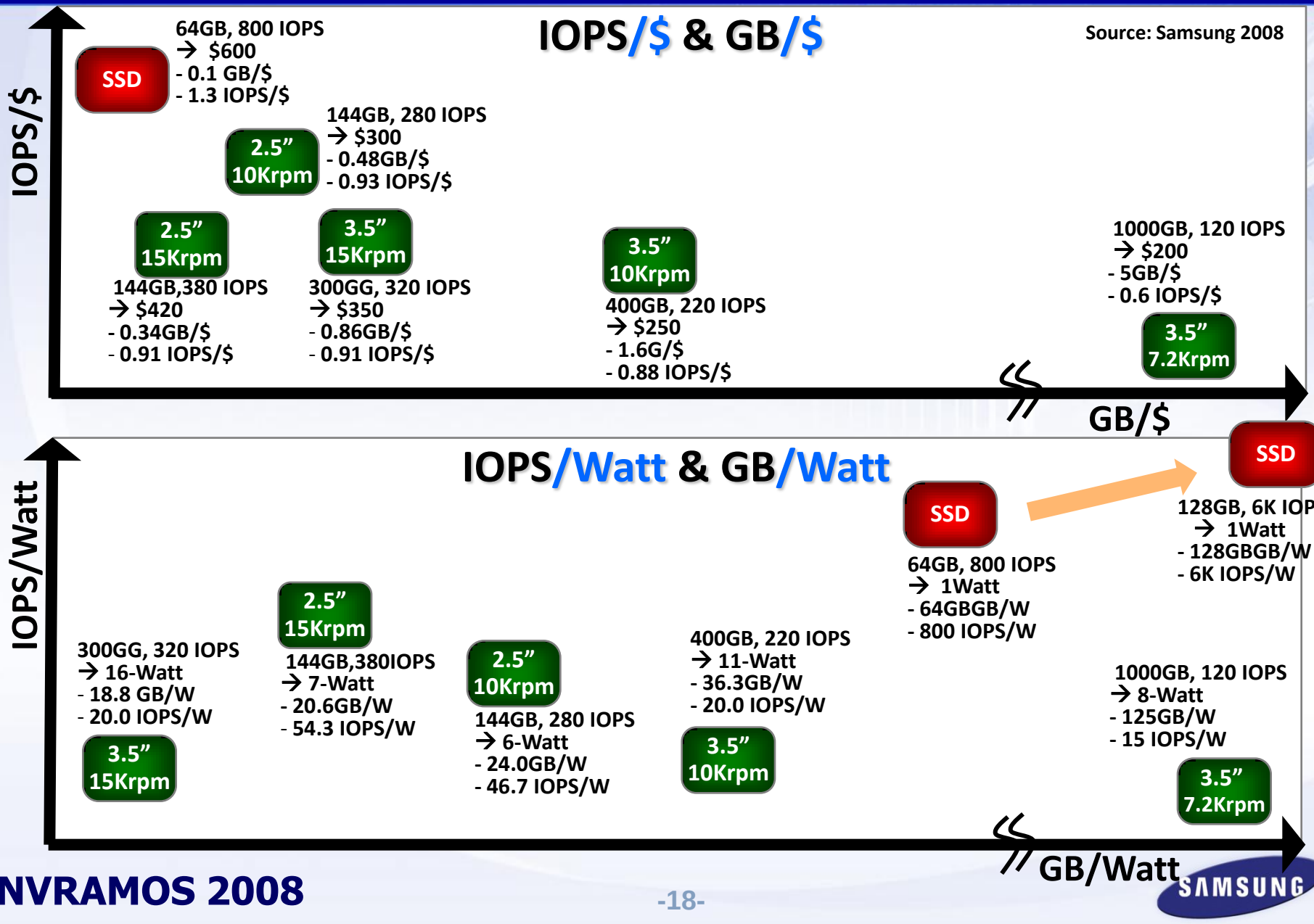


	15K HDD 3.5"	15K HDD 2.5"	SSD 2.5"
Capacity	300GB	144GB	64GB
Read	93 MB/s	108 MB/s	100 MB/s
IOPS	320	380	800
Price Est.	~ $350	~ $420	~ $600
Avg. Watts	16 W	7W	1W
GB / $	0.86GB (1.2x)	0.34GB (2.9x)	1.0GB
IOPS / W	20 (40X)	54 (15X)	800

*SSD: IOPS: random read @ 512B, Data (SEC Marketing)

*HDD: Price and Data from storagereview.com; IOPS: average read speed

Better IOPS/$, IOPS/W in Enterprise



Contents

SSD Related Technical Issues – Big Pictures

* Tens or hundreds of MBytes

Operating System

File System

Memory Manager
(ex. VM, Buffer)

Storage System
Architecture
(ex. Hierarchy, RAID)

Storage System

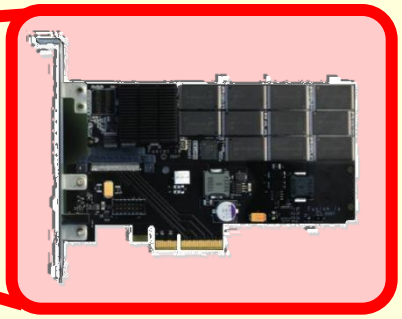
Internal Parallelism
(RAID+Buffer*Inside)

Host Interface

CPU

NANDs

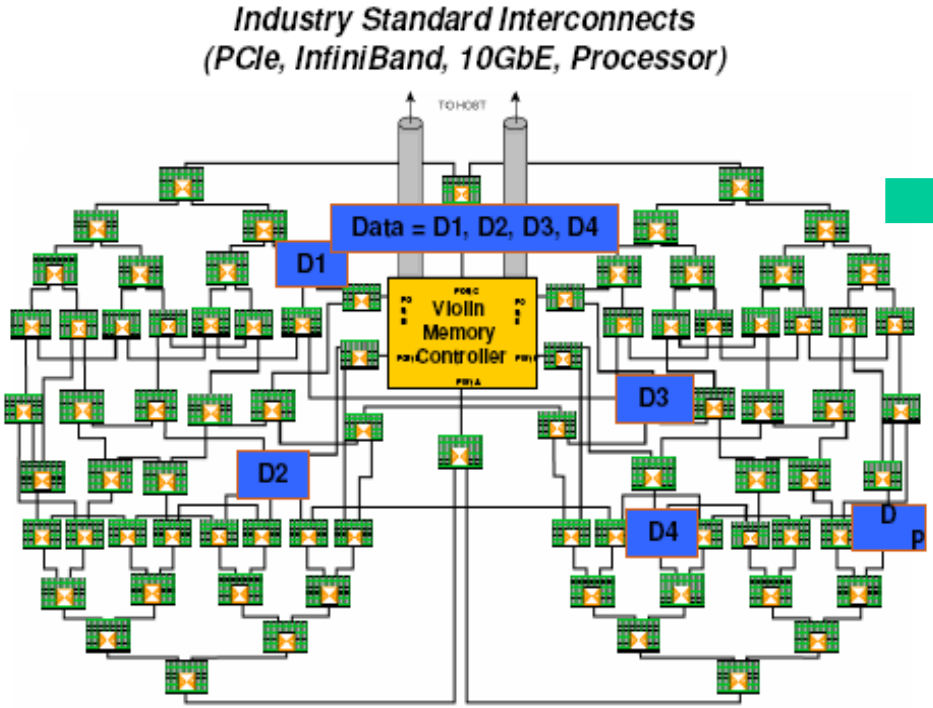
DRAMs



2-layered Parallelism
(RAID on RAID)

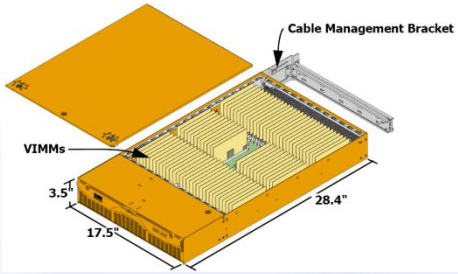
Storage System Architecture

- Violin-Memory Inc. (www.violin-memory.com)
- System Architecture
  - VIMM + VXM architecture (SSD unit ➔ system level)
  - Multi-level Flash RAID Hierarchy
    - Tightly integrated RAID/flash controller
  - Asynchronous IO over multi-VIMMs



Violin Switched Memory (VXM)

4TB Appliance



Source: www.violin-memory.com

Operating System

❑ OS has been optimized for HDD over decades

- Especially, file system and memory management
- Ex. delayed write, sequentially-optimized allocation, read-ahead, swap operation

❑ HDD optimization = SSD optimization?

Common

- Favors large I/O
- Favors delayed write
- Segmented FS layout

Different

- Page/block boundary
- Large read-ahead is a waste
- No seek time, but write perf. is dependent on position & fragmentation

❑ CPU utilization vs Storage Utilization

- OS kernel's efficiency related to resource balancing
- Resource scheduling optimization issue

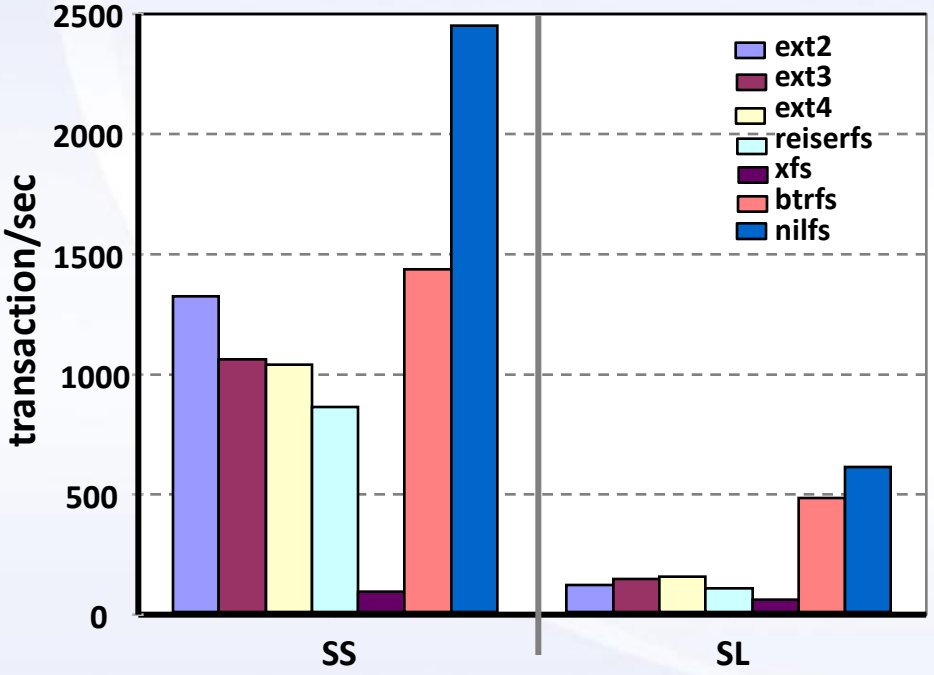
❑ Not only OS, but also Applications

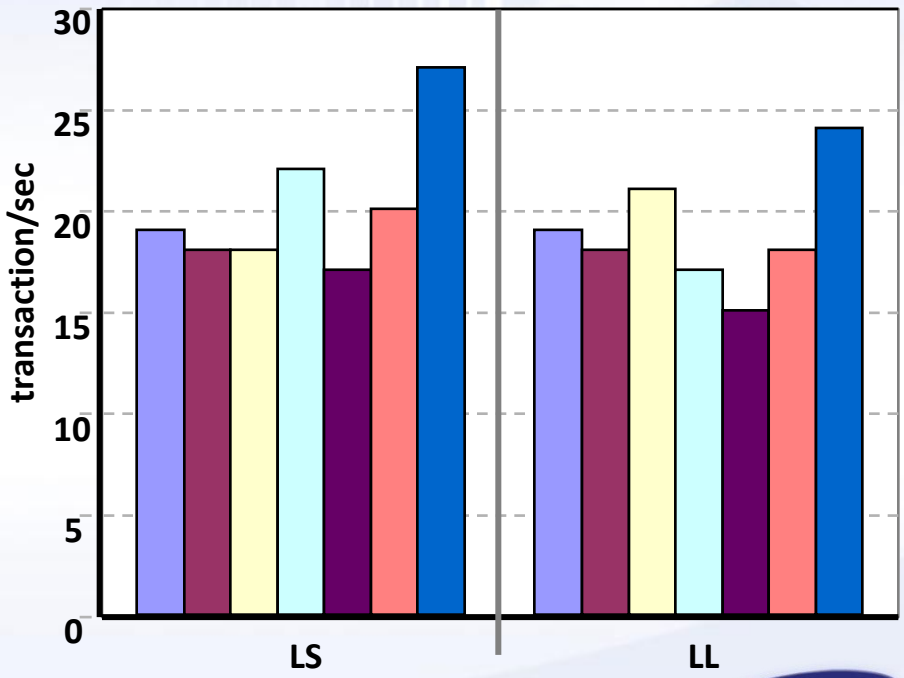
- Application should utilize the internal I/O parallelism of SSD (like CPU parallelism)
- App-level I/O parallelism: multi-tasking I/O, asynchronous I/O

File System

- **Linux FS Benchmark**
  - **Running postmark on various Linux FS**
- **Log-structured and COW FS performs better than traditional FS on SSD**

Workload	File size	# of file (work-set)
SS	9-15K	10,000
SL	9-15K	100,000
LS	0.1-3M	1,000
LL	0.1-3M	4,250





Internal SSD Architecture

❑ Storage Buffer Management

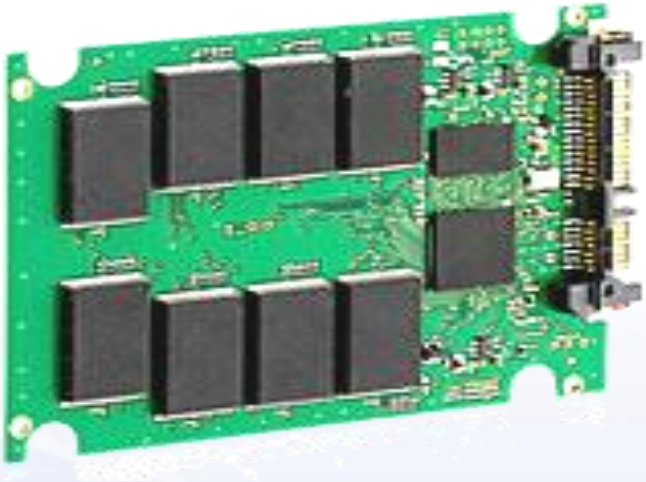
- SSD has larger buffer than HDD (similar to RAID buffer)
- Efficient buffer coordination between host and storage device will be helpful

❑ How to coordinate with existing RAID system?

- Multi-layered hierarchy of RAID system with (duplicated) buffers
- Translated host I/O pattern into SSD-friendly pattern (ex. Pre-fetch, deferred write)

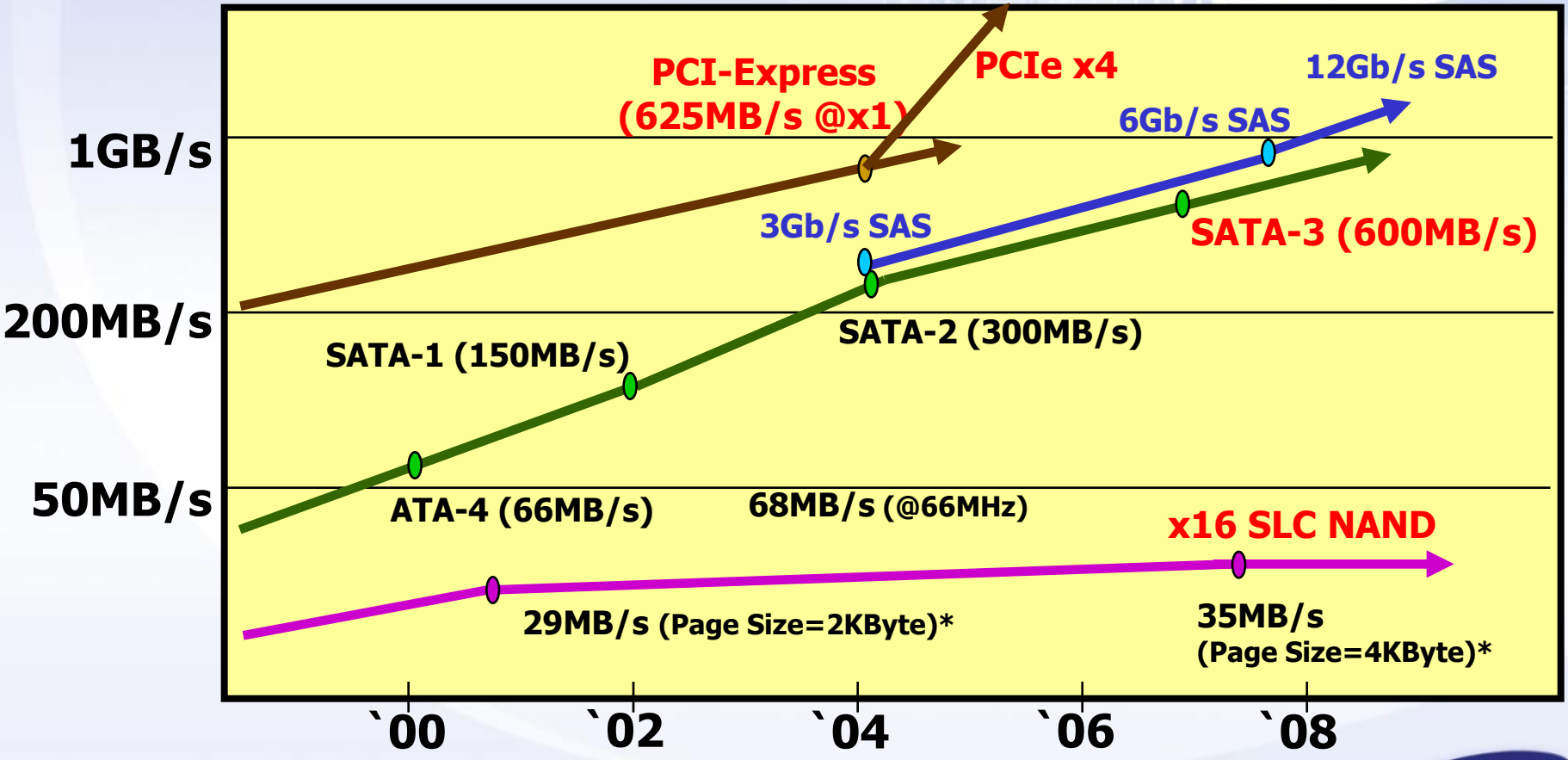
❑ How to utilize internal parallelism?

- Interface for multiple I/O command – T10 TCQ (common practice : 32 ~ 256)
- Cf. number of flash chips in SSD – 64 (Gen3)



Host Interface

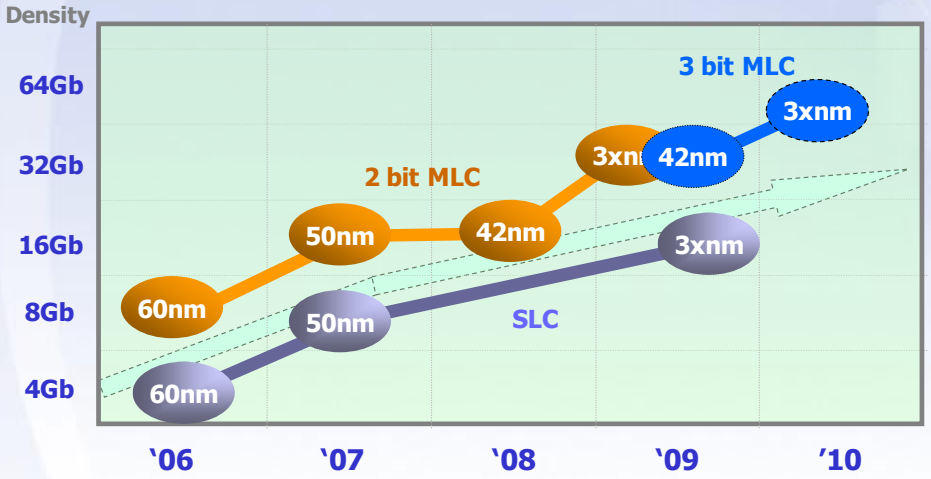
- NAND Flash is no more performance bottleneck
  - Host interface would be performance limit
  - However, complexity in protocol and SI/PI increase

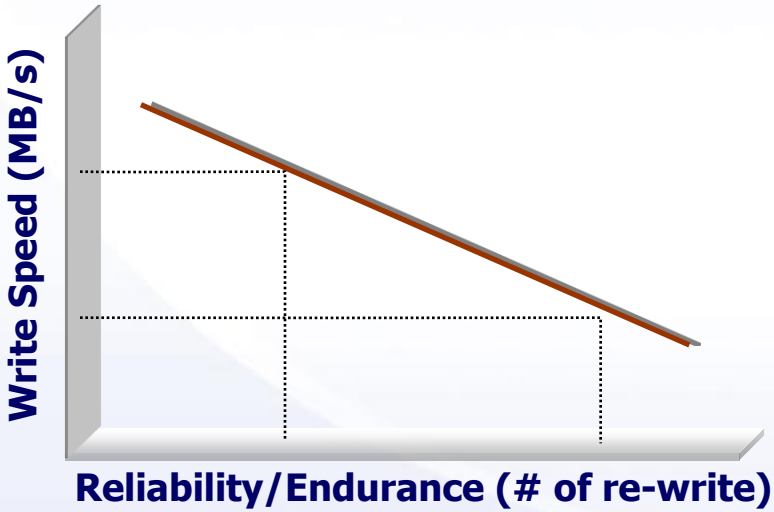


Note *) tRC=50ns, tR=20us

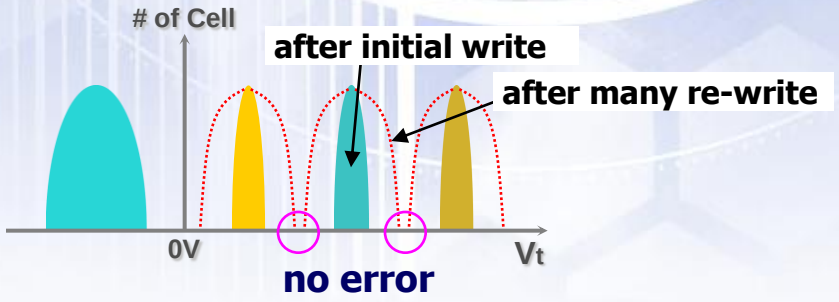
NAND Internals

❑ Endurance and Write speed is reverse-proportional

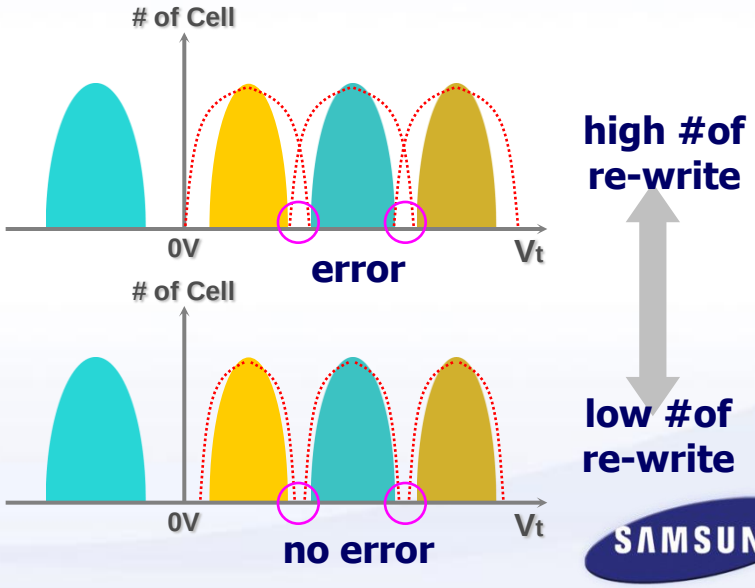




- High reliability (high #of re-write), Slow write



- High speed write, low #of re-write



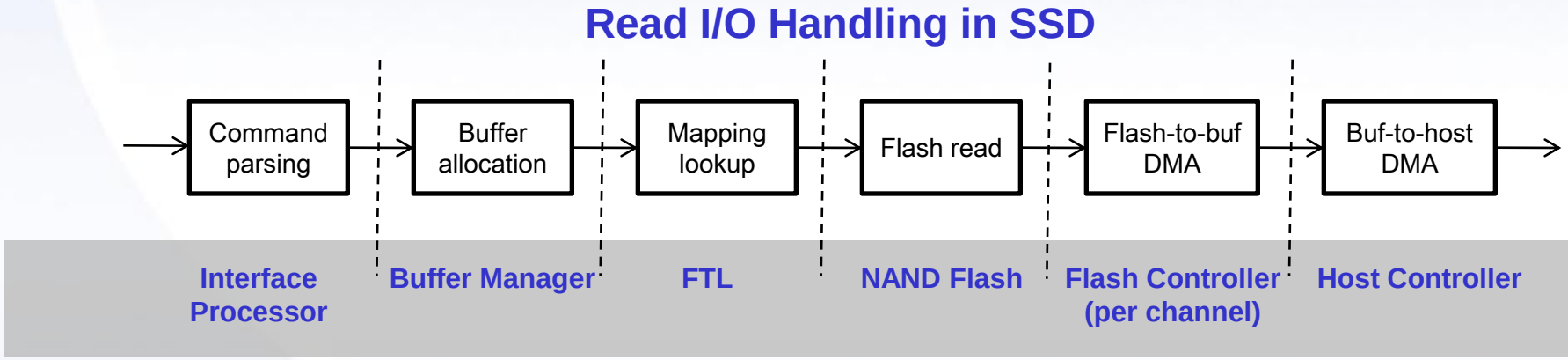
SSD Resource Management

❑ How I/O is processed in SSD?

- Internal processing is a sequence of pipelined stages
- I/O stage is parallelized into multiple hardware resources (ex. Chip, channel)

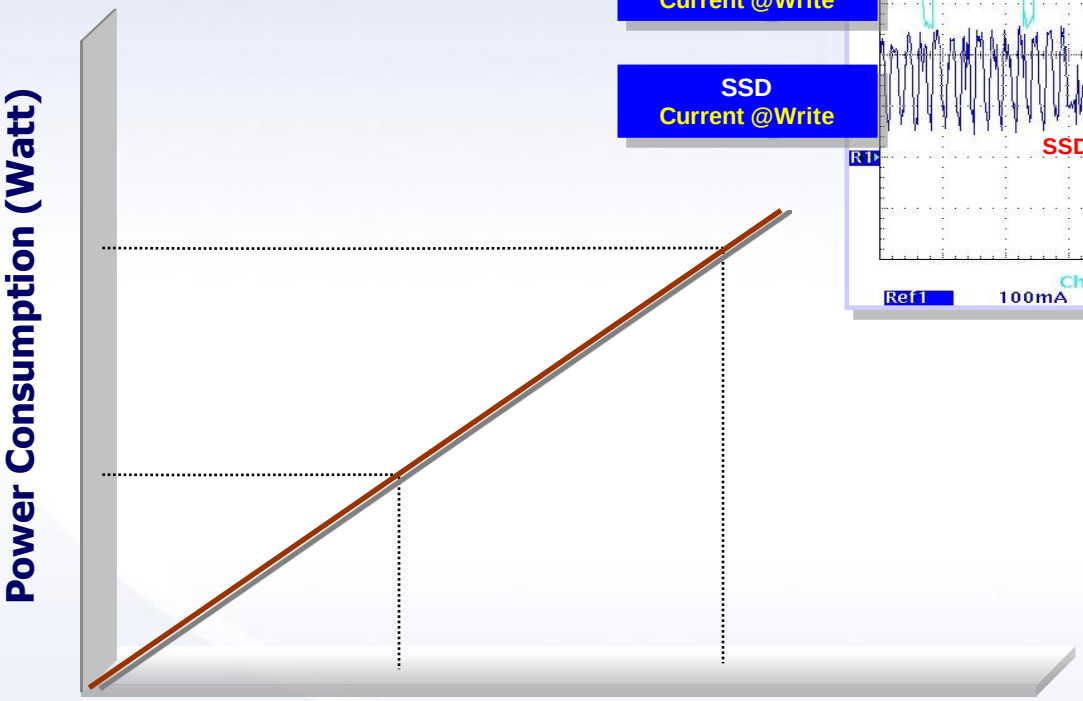
❑ Resource scheduling of SSD

- How to speed up the entire I/O processing throughput?
- Similar to super-scalar pipeline problem (without inter-request dependency)



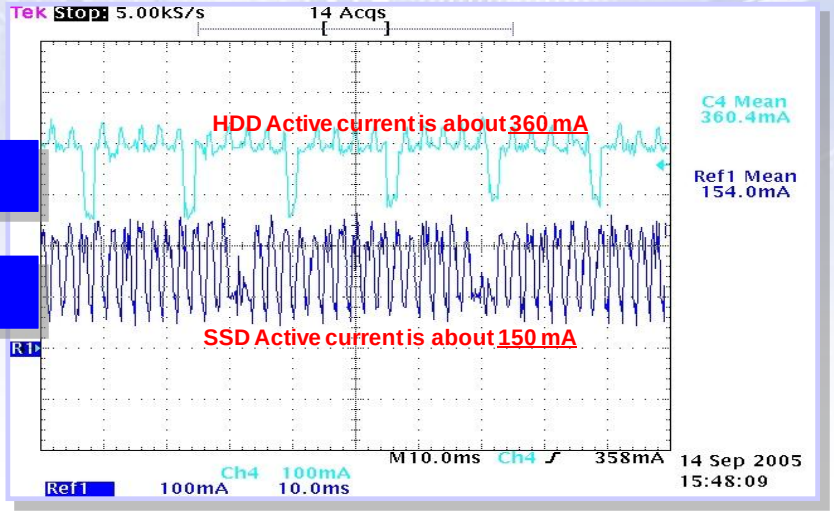
Low Power Management

- **Power consumption directly proportional to Storage Parallelism**
  - **Low power resource Managements**



1.8" HDD Current @Write

SSD Current @Write



Related Works (Research-side)

Flash Translation Layer (FTL)

- A space-efficient flash translation layer for CompactFlash Systems (IEEE Transactions on Consumer Electronics, 2002)
- A superblock-based flash translation layer for NAND flash memory (EMSOFT, 2006)
- A Log Buffer Based Flash Translation Layer Using Fully-Associative Sector Translation (ACM Transaction on Embedded Computing System, 2007)
- A Re-Configurable FTL(Flash Translation Layer) Architecture for NAND Flash based Applications (RSP, 2007)

Flash-optimized Buffer Management

- CFLRU: a replacement algorithm for flash memory (CASES, 2006)
- FAB: flash-aware buffer management policy for portable media players (IEEE Transactions on Consumer Electronics, 2006)
- BPLRU: A Buffer Management Scheme for Improving Random Writes in Flash Storage (USENIX Conference on File and Storage Technologies, 2008)

Flash-optimized File System

- A New Type of NAND Flash-based File System: Design and Implementation (WiCOM, 2006)
- An Efficient NAND Flash File System for Flash Memory Storage (IEEE Transactions on Computers, 2006)
- Embedded NAND Flash file System for Mobile Multimedia Device (IEEE Transactions on Computer, 2008)

Related Works (Industry-side)

Btrfs	
Developer	Oracle Corporation
Full name	Btrfs
Introduced	June 12, 2007 (Linux)
Structures	
Directory contents	btree
File allocation	extents
Limits	
Max file size	16 EiB
Max number of files	2^{64}
Max filename length	255 bytes
Max volume size	16 EiB
Allowed characters in filenames	All bytes except NUL and '/'
Features	
Attributes	POSIX
File system permissions	POSIX, ACL
Transparent encryption	No
Supported operating systems	Linux

ORACLE

ZFS	
Developer	Sun Microsystems
Full name	ZFS
Introduced	November 2005 (OpenSolaris)
Structures	
Directory contents	Extensible hash table
Limits	
Max file size	16 EiB
Max number of files	2^{48}
Max filename length	255 bytes
Max volume size	16 EiB
Features	
Forks	Yes (called Extended Attributes)
Attributes	POSIX
File system permissions	POSIX, ACL
Transparent compression	Yes
Transparent encryption	Yes (currently beta)[1]
Supported operating systems	Sun Solaris, Apple Mac OS X Server 10.5, FreeBSD, Linux via FUSE

Sun microsystems

WAFL	
Developer	NetApp
Full name	Write Anywhere File Layout
Limits	
Max file size	16TB (limited by containing aggregate size)
Max volume size	16TB (limited by containing aggregate size)
Allowed characters in filenames	selectable (UTF-8 default)
Features	
Dates recorded	atime, ctime, mtime
File system permissions	UNIX permissions and ACLs
Transparent compression	No
Transparent encryption	No (possible with 3rd party appliances like Decru DataFort)
Single Instance Storage	Yes (*FAS Dedup*: periodic offline scans, block based; *VTL Dedup*: online byte-range based)

NetApp

Source: Wikipedia

SAMSUNG

Research Hints

❑ Operating System

- File system - alignment, allocation policy, design (ex. COW)
- Block layer - SSD-optimized I/O scheduler
- Volume manager: alignment, allocation
- Virtual memory: read-ahead

❑ Storage System

- On-storage buffer management
- RAID-like optimization algorithm
- Exchange hint info with host

❑ Lessons from flash FS

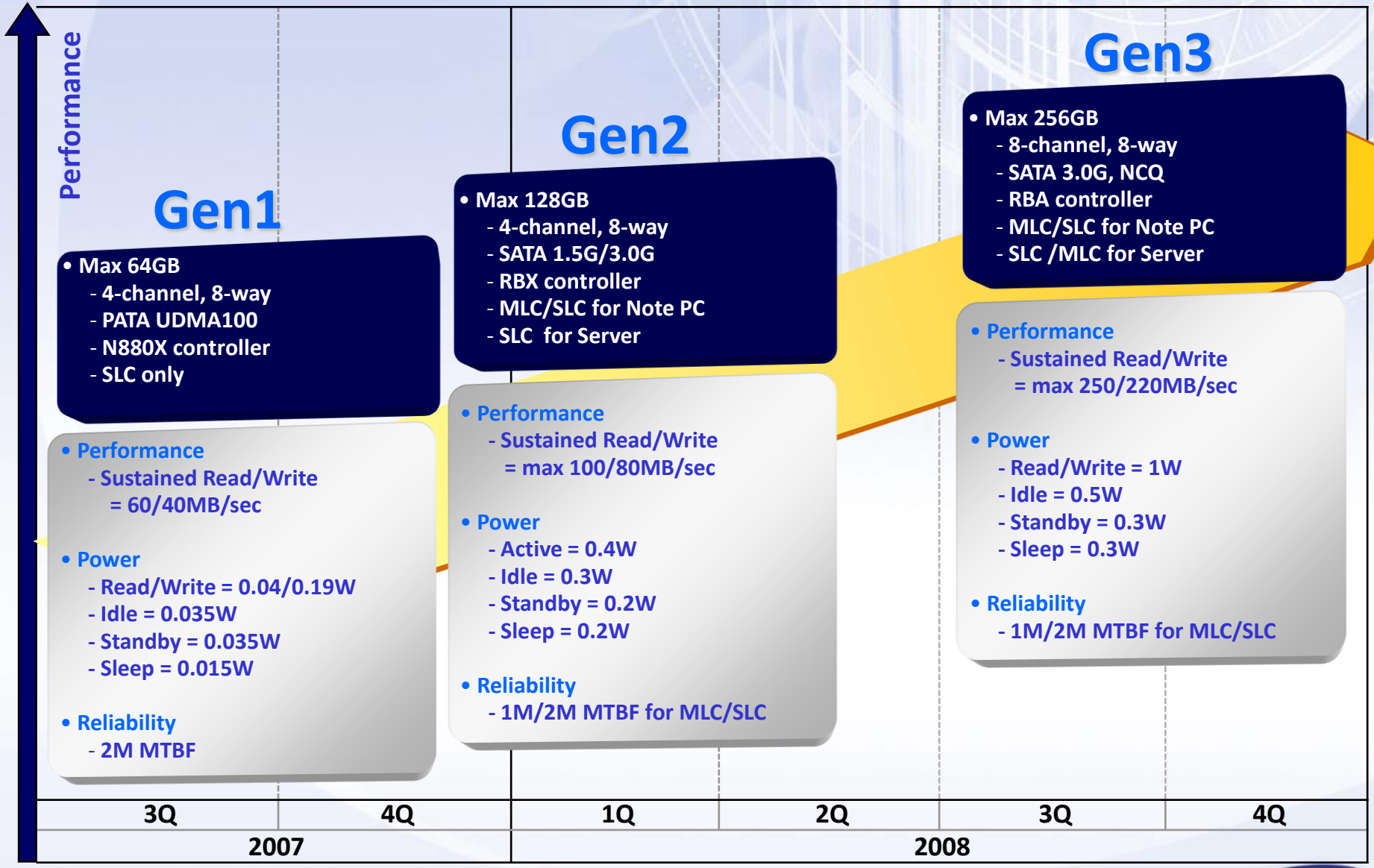
- Sequential writing at multiple logging points
- Wandering tree
  - Trace-off between sequentiality vs. amount of write
  - Cf. space map (Sun ZFS)
- Need to optimize garbage collection overhead
  - Either FS itself or FTL in SSD

Research Issue Summary

Technology	Revisiting existing technology	New technology for flash storage
System Architecture for Flash Storage	Host RAID System Storage management	Violin, TMS
Operating System	N/A	T13: Trim command
File System	Log-structured FS COW FS	N/A
Buffer Management	N/A	CFLRU, FAB, BPLRU
RAID	RAID Buffer management	N/A
Low Power Management	Dynamic power management	N/A

Contents

Samsung SSD Roadmap



The 1st SSD Only NotePC - Lenovo X300™

- First "Mainstream" SSD Only Notebook
- 64GB Solid State Drive
  - No HDD Option or "Opt Out"
- Represents an Irreversible Industry Trend…

Samsung SSD Press Release - Client

Client MLC

SAMSUNG Develops World's Fastest and Largest Capacity 2.5-inch, MLC-based (256GB) SSD with SATA II Interface

Taipei, Taiwan on May 26, 2008

Taipei, Taiwan - May 26, 2008 : Samsung Electronics, the world leader in advanced memory technology, announced today that it has developed the world's fastest, 2.5-inch, 256 Gigabyte (GB) multi-level cell (MLC) based solid state drive (SSD) using a SATA II interface at the fifth annual Samsung Mobile Solution Forum held at the Westin Taipei Hotel. Samsung's new 256GB SSD is also the thinnest drive with the largest capacity to be offered with a SATA II interface.

Half Slim

Samsung Introduces High-performance, Low-density SATA II SSDs for Low-priced PC Market

Seoul, Korea on Aug 27, 2008

Seoul, Korea - Aug. 27, 2008 - Samsung Electronics Co., Ltd., the world leader in advanced semiconductor technology, announced today that it has begun sampling low-density, higher-performance solid state drives (SSDs) that are only 30 percent of the size of 2.5-inch SSDs and highly cost-efficient to manufacture. With the introduction of these smaller, low-capacity SSDs, Samsung now offers an attractive replacement for existing hard drives used in low-cost PCs. Available in densities of 8GB, 16GB and 32GB, the new multi-level-cell SSDs will be mass produced beginning next month.

Samsung SSD Press Release - Server

Blade Launch

Samsung Solid State Drive Selected for New HP ProLiant Virtualization Blade Server

By: HP News Desk
Oct. 10, 2008 01:45 PM

news desk

Samsung Electronics Co., Ltd., the world leader in advanced semiconductor technology, announced today that its 32 gigabyte (GB) and 64 GB solid state drives (SSDs) have been selected, after extensive testing, for use in the HP ProLiant BL495c virtualization blade server. This represents the first time that an SSD has been qualified for use in a server optimized for virtualization. The server is the world's first server blade designed specifically to host virtual machines, and is designed for use in virtualized environments that require significant memory, data storage and network connections to optimize server performance.

High Endurance SLC

SAMSUNG Collaborates with Sun Microsystems to Develop New Ultra-Endurance Flash Memory for SSD Products in Server Applications

Seoul, Korea on Jul 17, 2008

Seoul, Korea - July 17, 2008 : Samsung Electronics Co., Ltd., the world leader in advanced semiconductor technology, announced today that it has collaborated with Sun Microsystems to develop a single-level-cell NAND flash memory device for use in solid state drives that offers much higher endurance levels than any other flash memory device on the market today.

World Recognized Samsung SSD

❑ Why Samsung ?

The Samsung SSD has been voted one of the ten most brilliant and bold ideas that are set to change the world-recognized for pushing the envelope of technology and for design and engineering innovation.

Tom's Hardware 2008 Best of Tom's Hardware

CES Innovations products 2008

CeBIT Highlight products 2007

Popular Mechanics Breakthrough Awards 2007

THANK YOU