Become a SSD expert in minutes!

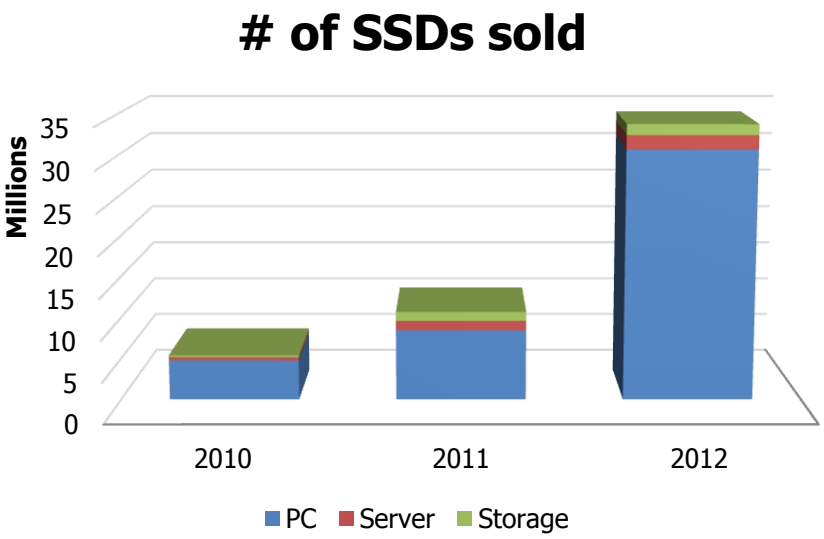
Ryan Smith

ryan.smith@ssi.samsung.com

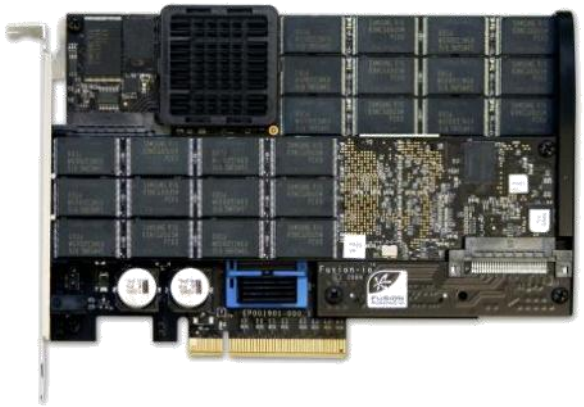
408-205-8889

What is a SSD?

- SSD = Solid State Drive
- RAM-based introduced in 1970's
- Flash-based version in 1990's
- Today, it typically uses NAND Flash
- 2012 is a big year for SSDs
- Don't complicate it.. it's just a really fast drive!

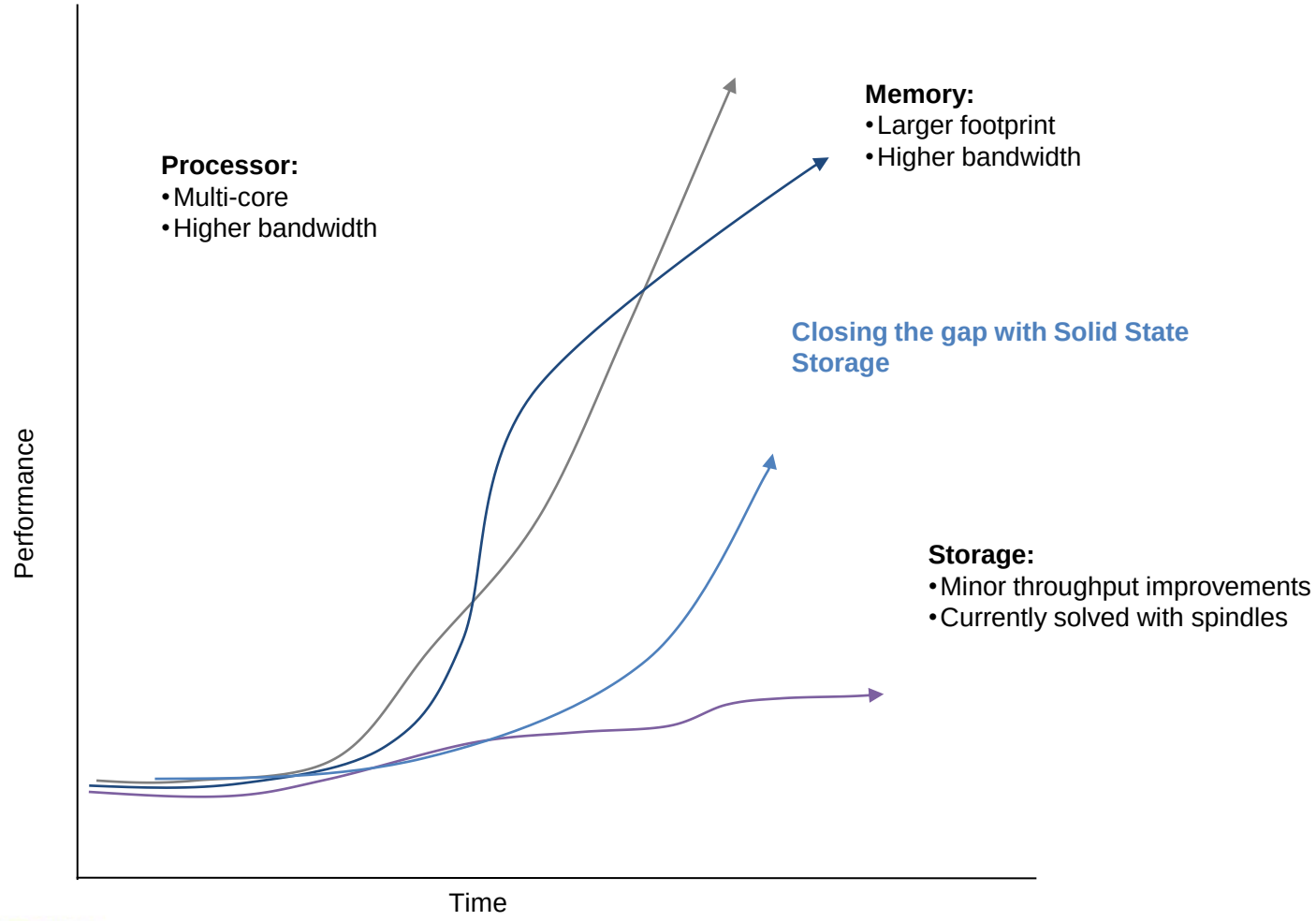


Source : Samsung



Why an SSD?

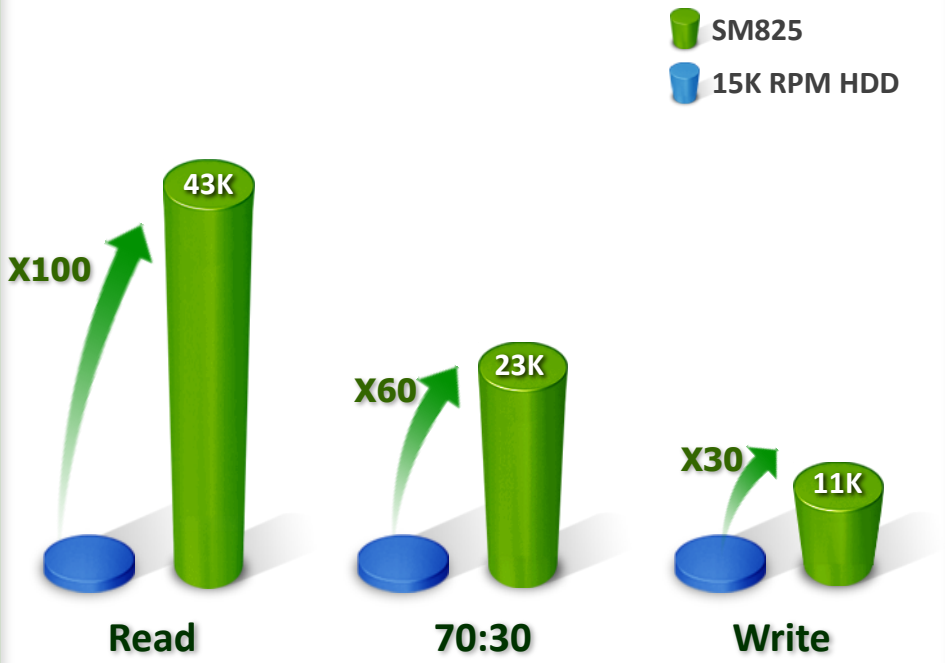
- **Three things that dictate the speed of your PC/Server:**
  - **CPU, DRAM, and HDD**
- **Everything is speeding up.. Except the HDD**



Why an SSD?

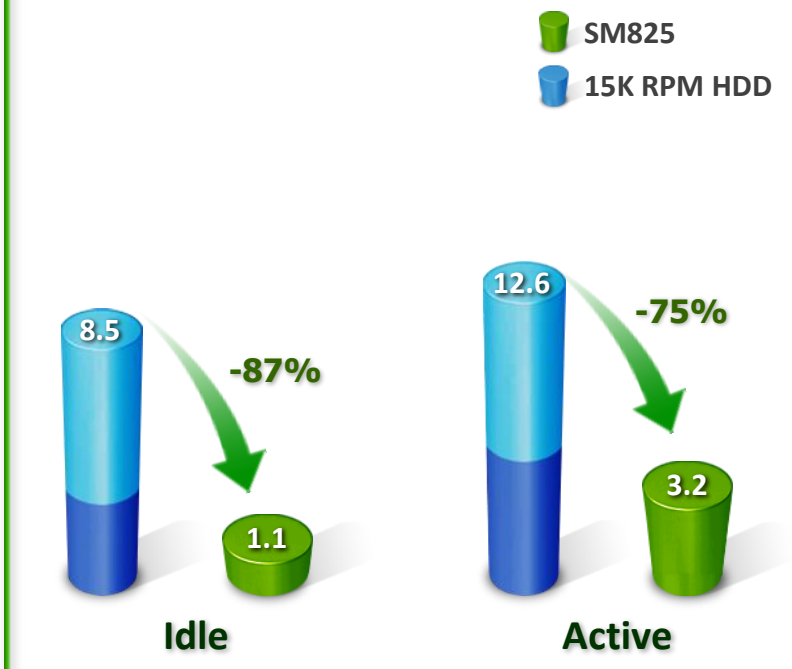
- Lower response times (latency)
- Higher IOPS and Throughput
- Lower Power
- No RVI Issues, More reliable

Random Performance (IOPS)



Test Environment : Intel SR2600UR Server / IOMeter2008

Power Consumption (Watt)



Test Environment : Intel SR2600UR Server / IOMeter2008 / 4KB RND R70:W30

Source : Samsung

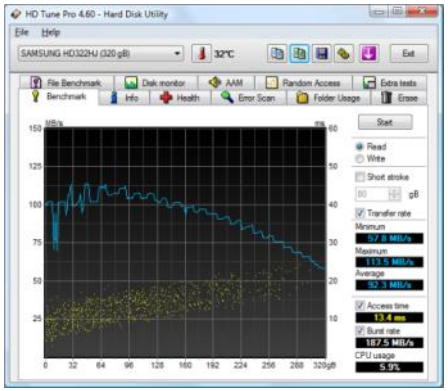
So what's there to know about an SSD?

SSD Key Characteristics

SSD Components
NAND Characteristics
P/E Cycles
WAF
TBW
SMART
Host Interface
Sustained vs. Peak Performance
Benchmarking

SSD Influencers

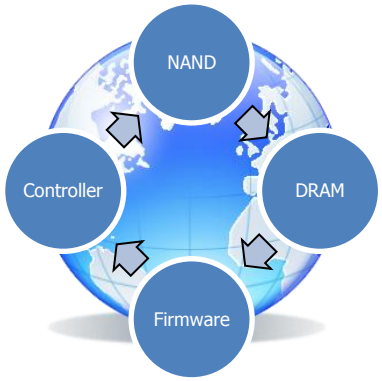
TRIM
Over-provisioning
Changing Workload

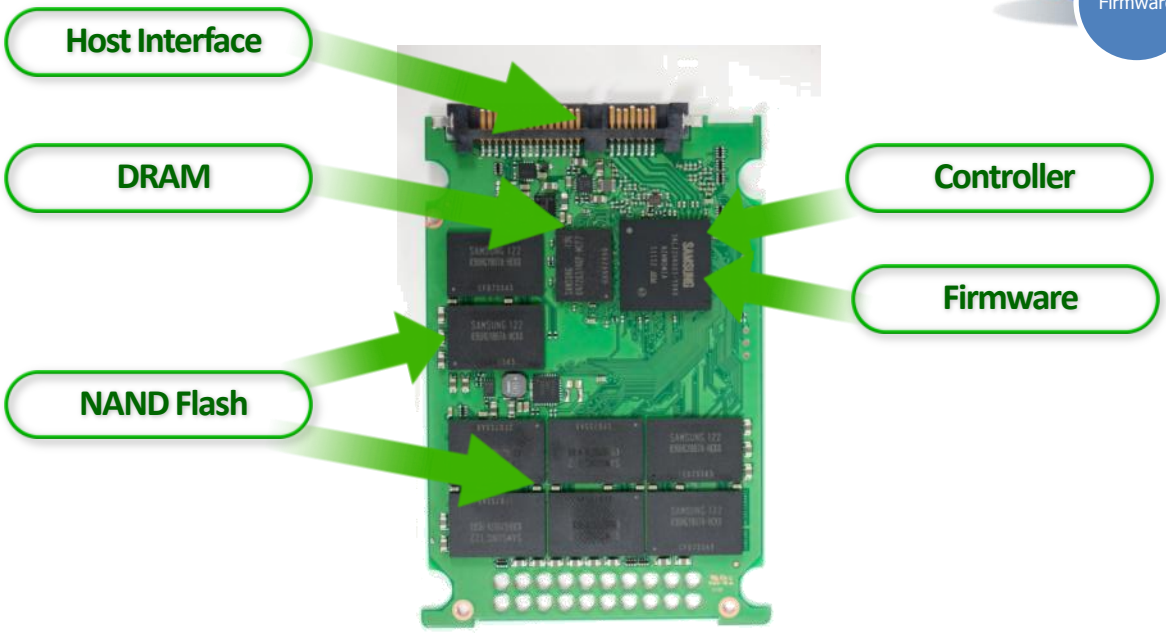


SSD Key Characteristics

SSD Components

- **Host/NAND Controller**
- **Firmware**
- **NAND Flash**
- **DRAM**
- **Capacitors (optional)**





All components work closely together

SSD Image Source : Anandtech

NAND Characteristics

- Types of NAND
  - TLC
  - MLC
  - E-MLC
  - SLC

TLC — 500-1K P/E Cycles, 1 year retention

MLC — 3-5K P/E Cycles, 1 year retention

E-MLC — 10-30K P/E Cycles, 3 month retention

SLC — 90-100K P/E Cycles, 3 mo – 1 yr retention

PC → Enterprise

- **Geometry / Lithography**
  - 4xnm, 3xnm, 2xnm
  - Smaller = Less Cost

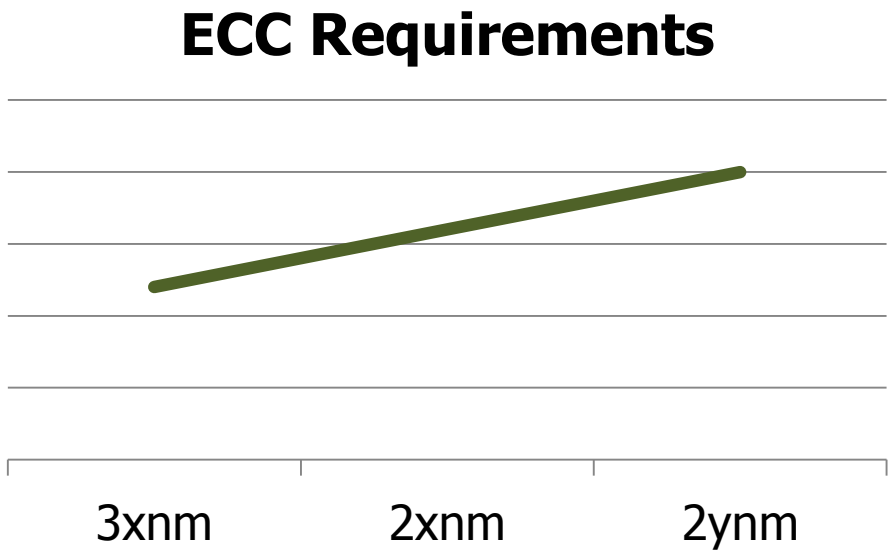
- **NAND Hierarchy**
  - Pages: Smallest unit that can be read/written (e.g., 8KB)
  - Erase block: Groups of pages (e.g., 64 pages @ 8KB = 512KB)

P/E Cycles

Program / Erase Cycles

The # of times a given NAND cell can be programmed & erased

- As geometries shrink, error correction must get better
- It's like a car warranty!
  - 3 years or 50,000 miles
  - 3 years or 3,000 P/E Cycles
- Not a useful characteristic by itself



Write Amplification Factor (WAF)

Write Amplification Factor

Bytes written to NAND versus bytes written from PC/Server

$$WAF = \frac{\text{Bytes written to NAND}}{\text{Bytes written from Host}}$$

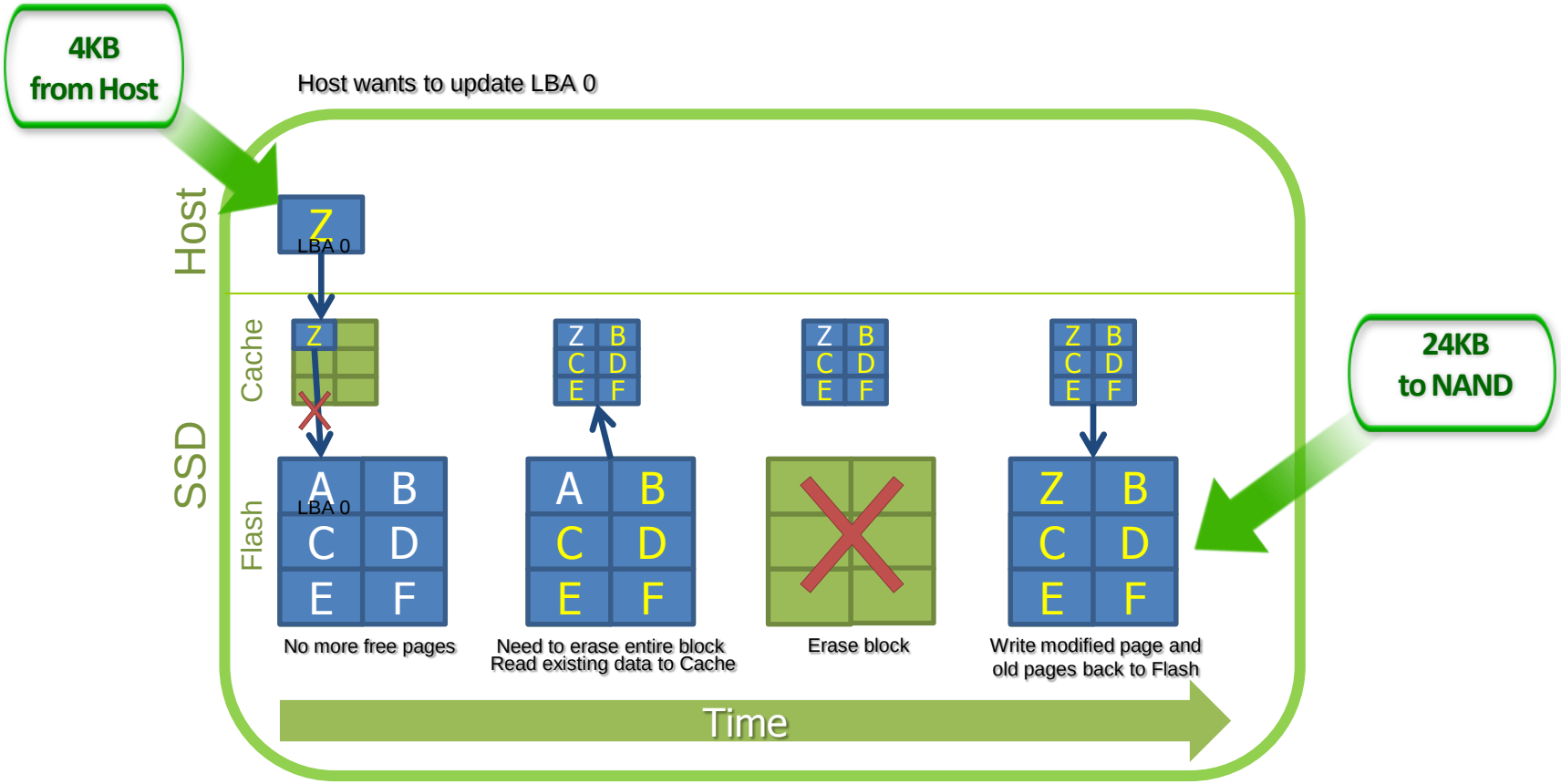
- **WAF 1 means 1MB from host writes 1MB to NAND**
- **WAF 5 means 1MB from host writes 5MB to NAND**

- **Factors that can affect WAF:**

Controller	Flash Translation Layer (FTL)
	Wear Leveling
	Over-provisioning
	Garbage Collection
Host Application	Write Profile (Ran vs. Seq)
	Free user space / TRIM

Write Amplification (WAF) Example

- **Below example illustrates WAF of 6**



TBW

TeraBytes Written

# of terabytes you can write to the drive over it's useful life

$$TBW = \frac{(\text{Capacity GB}/1000) \times \text{PE Cycles}}{\text{WAF}}$$

Examples:

((128GB / 1000) * 3000) / 5 = 76.8 TBW
((128GB / 1000) * 3000) / 2.5 = 153.6 TBW
((256GB / 1000) * 3000) / 5 = 153.6 TBW
((128GB / 1000) * 30000) / 5 = 768 TBW

SMART

- **Look at health and various statistics**
- **Allows for predictable maintenance windows**
- **Calculate WAF, TBW**
- **Host GB written = [ID241] / (2/1024/1024)**
- **NAND GB written = [ID177] * Capacity GB**
- **WAF = NAND GB / Host GB**
- **Expected Life (yrs) = Warranty PE * ([ID9]/24/365) / [ID177]**

ID	Attribute Name
5	Reallocated Sector Count
9	Power-on Hours
12	Power-on Count
177	Wear Leveling Count
179	Used Reserved Block Count
180	Unused Reserved Block Count
181	Program Fail Count
182	Erase Fail Count
187	Uncorrectable Error Count
195	ECC Error Count
199	CRC Error Count
241	Total LBA Written

Host Interface

- This is how you communicate to the SSD
- So many choices..
  - SATA
  - SAS
  - PCIe (NVMe, SCSIe, SATAe, Proprietary)
- Which is right for you?

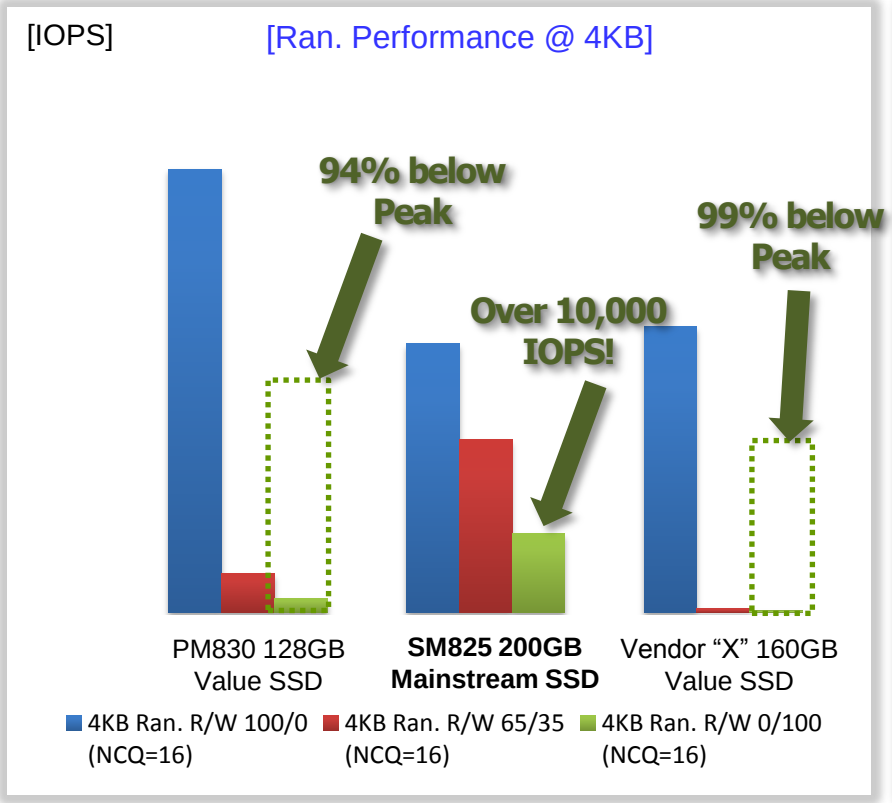
PC	Server	External Storage
SATA PCIe	**SATA** SAS PCIe	SATA + SAS bridge **SAS**

Sustained vs. Peak Performance

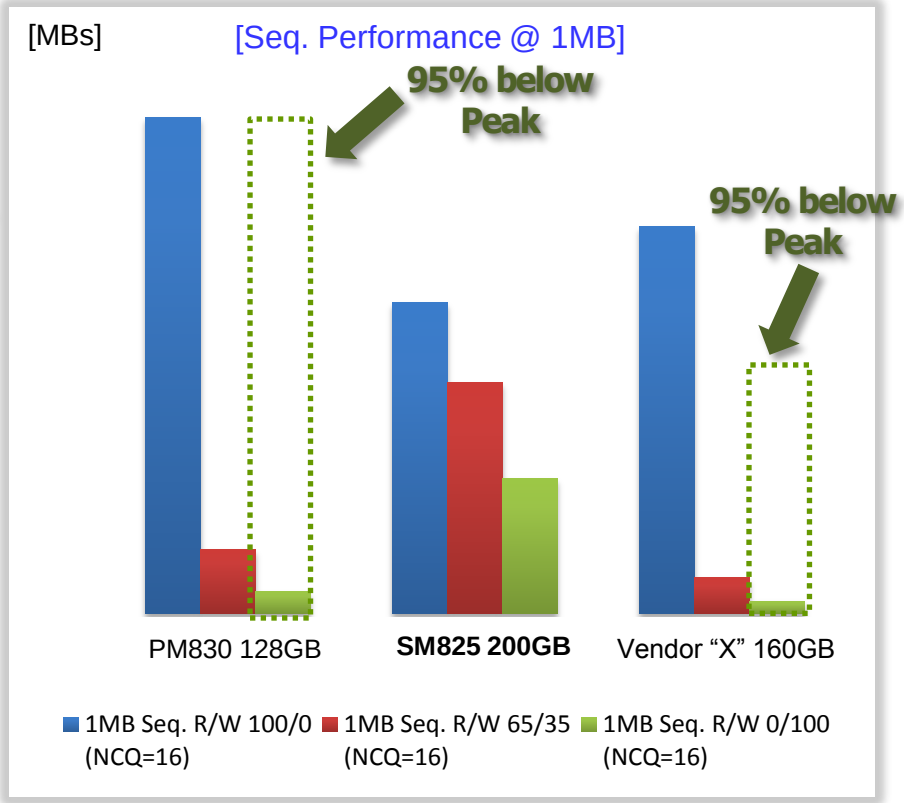
- There can be significant differences in sustained vs. peak
- Run enterprise benchmark (e.g., SNIA RTP 2.0)
- Or even better, run your own workload (or simulated)

There is a BIG difference between "Value" and "Mainstream/Enterprise" SSDs when you have any degree of writes in your workload

Samsung PM830 vs Vendor "X"
11x Sustained Random Writes



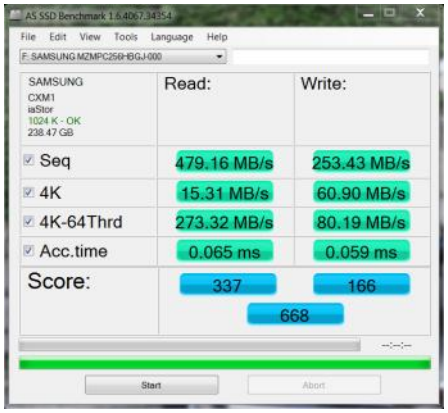
Samsung PM830 vs Vendor "X"
2x Sustained Sequential Writes

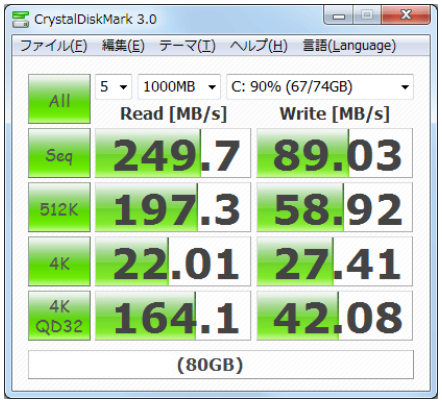


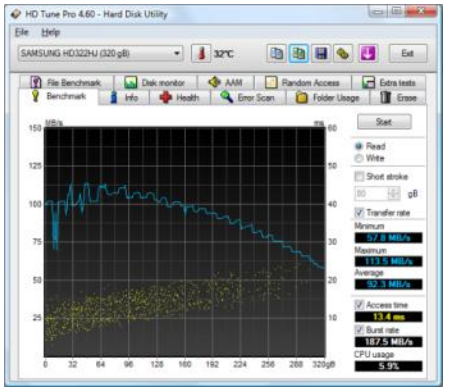
Source : Samsung / SNIA RTP2.0 Benchmark

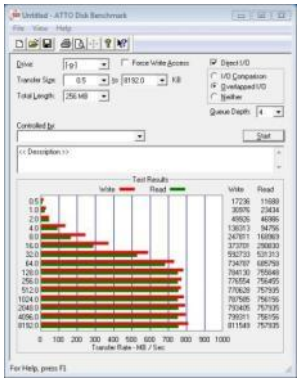
Benchmarking

- **Synthetic or actual workload & take measurements**





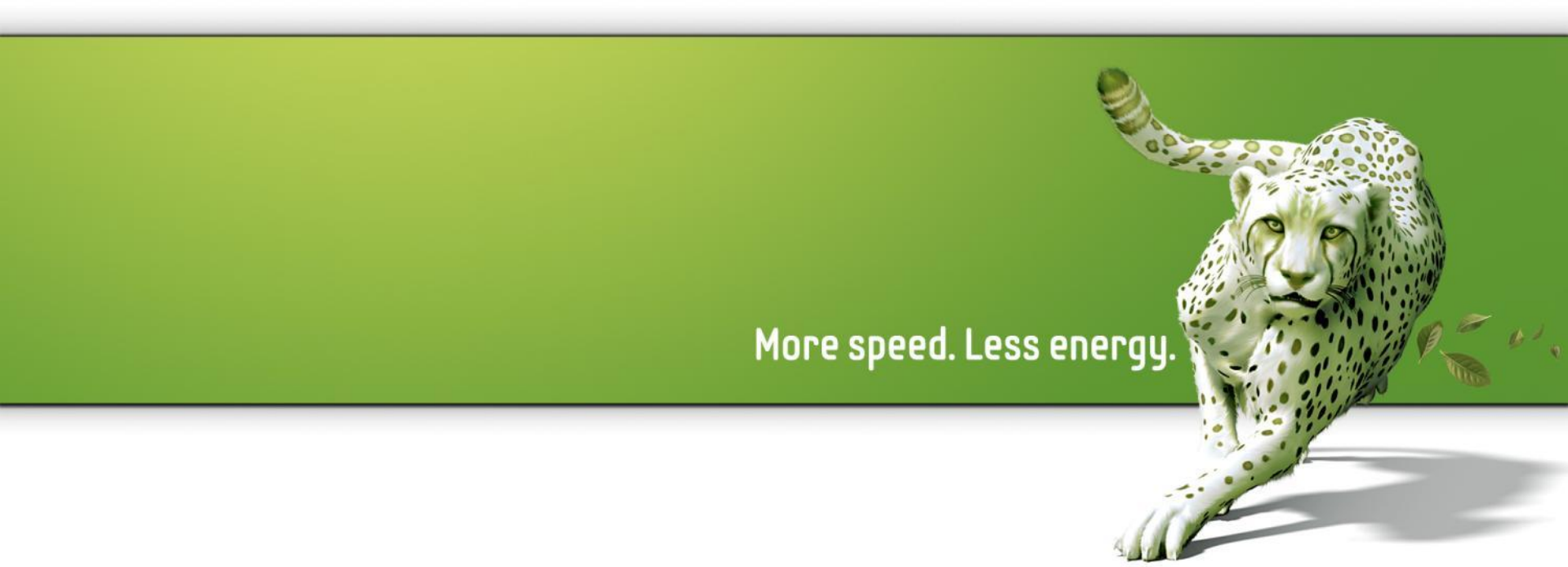




Benchmark	URL
SNIA RTP 2.0	http://www.snia.org/tech_activities/standards/curr_standards/pts
Iometer	http://sourceforge.net/projects/iometer/
ATTO Disk	http://www.attotech.com/products/product.php?sku=Disk_Benchmark
CrystalDiskMark	http://crystalmark.info/software/CrystalDiskMark/index-e.html
HD Tune Pro	http://www.hdtune.com/
AS SSD (SSD)	http://alex-is.de/PHP/fusion/downloads.php?download_id=9
Anvil (SSD)	http://thessdreview.com/latest-buzz/anvil-storage-utilities-releases-new-storage-and-ssd-benchmark/
Scripts	Have multiple "dd" running with best guess workload, capturing timing/speeds
Real Workload	Capture trace during real workload and playback (ioapps, blktrace/btereplay)

SSD Reviewers

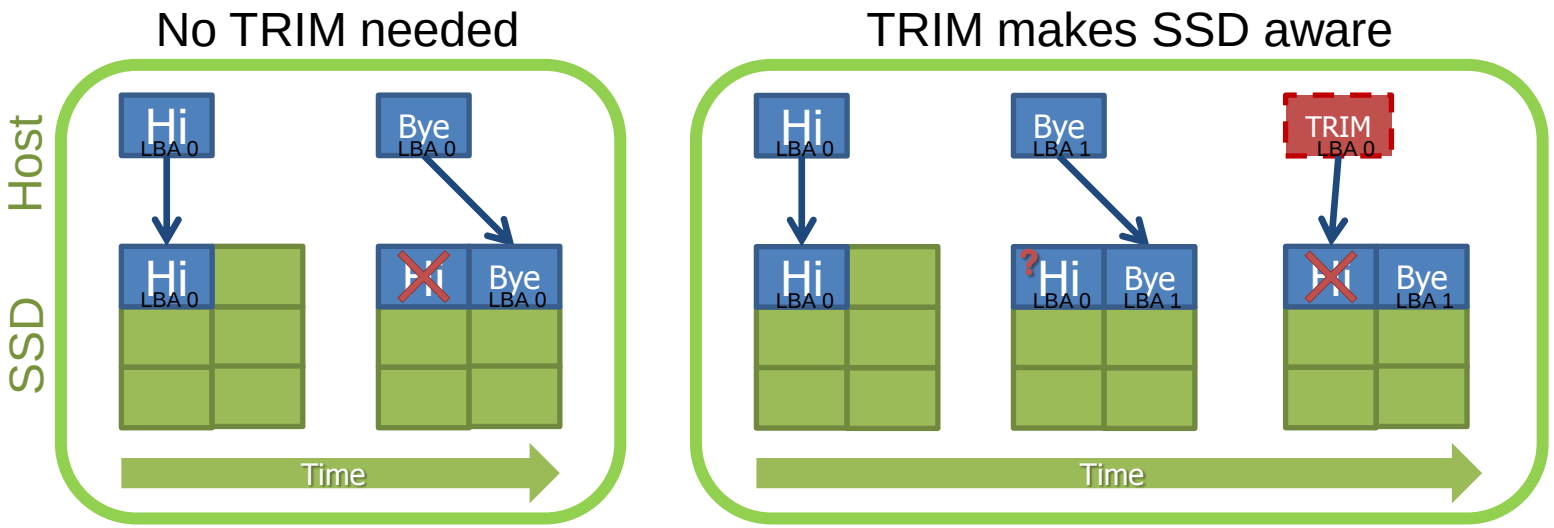
- **Good SSD Review sites available..**



SSD Influencers

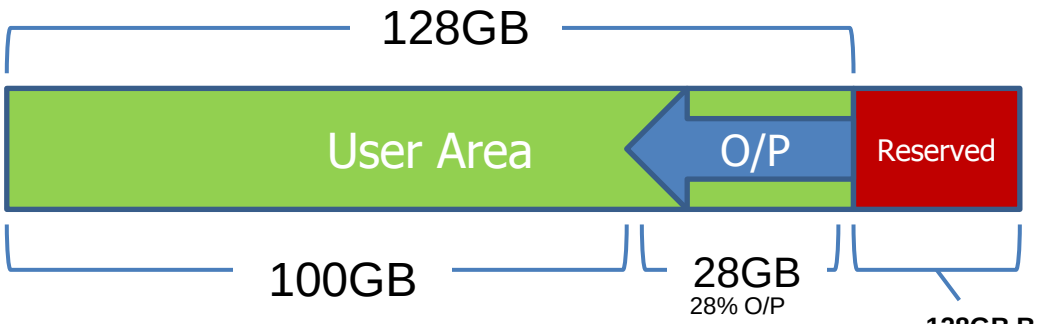
TRIM

- **Helps the SSD know which blocks aren't used**
- **Widely supported standard: Windows, Mac OS X, Linux, hdparm**
- **Better sustained performance and extends TBW**
- **Without TRIM, SSD only knows block isn't used once the same LBA is written to**



Over-Provisioning

- **Helps a few things:**
  - **Improves Write Performance**
  - **Reduces WAF, Increases TBW**



128GB Base-2 to Base-10 conversion:
137,438,953,472 to 128,000,000,0000 (6.9%)

Sample 128GB SSD	120GB		100GB
Over-Provisioning	7%		28%
Random Read (8K) IOPS	80K		80K
Random Write (8K) IOPS	1,800	3.5x	6,300
Sequential Read (64K) MB/s	500		500
Sequential Write (64K) MB/s	400		400
4KB Random WAF	5	-73%	1.35
4KB Random TBW	15	3x	45

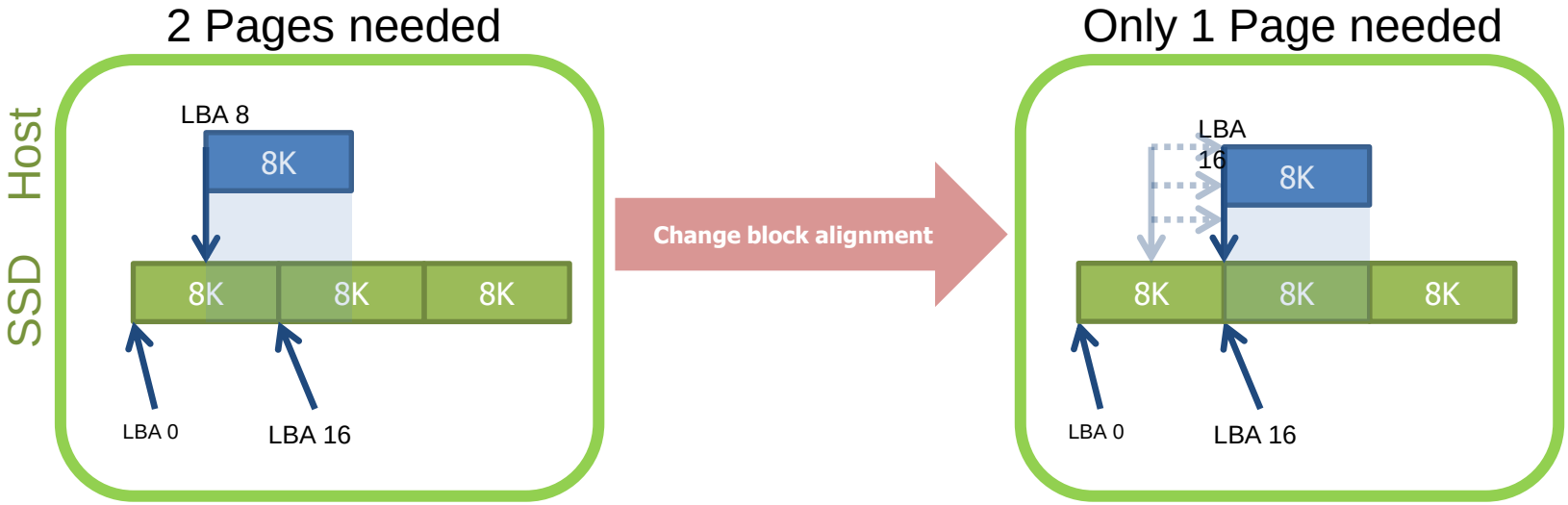
These performance numbers are fictitious but do represent the actual benefits seen during tests

Change Write Workload

- **Write sequentially instead of random to reduce WAF**
  - **If you have control of the I/O to the disk, this will pay off**

	Random		Sequential
MLC 512GB SSD	60 TBW	20x	1250 TBW

- **Align your writes with the page boundaries (e.g., 8KB)**

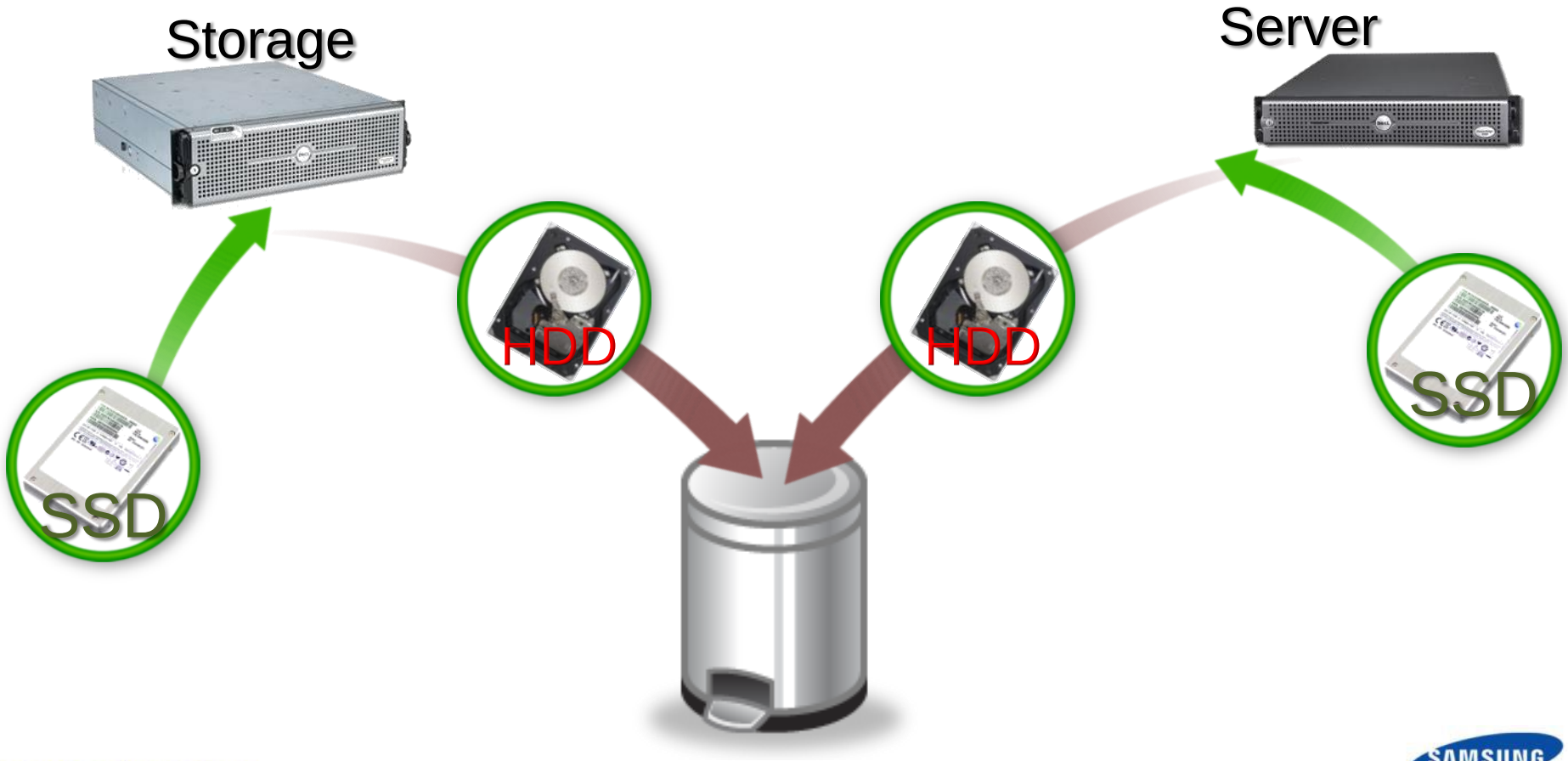


- **If alignment is too hard to implement, just increase your IO size**

Applications of SSDs

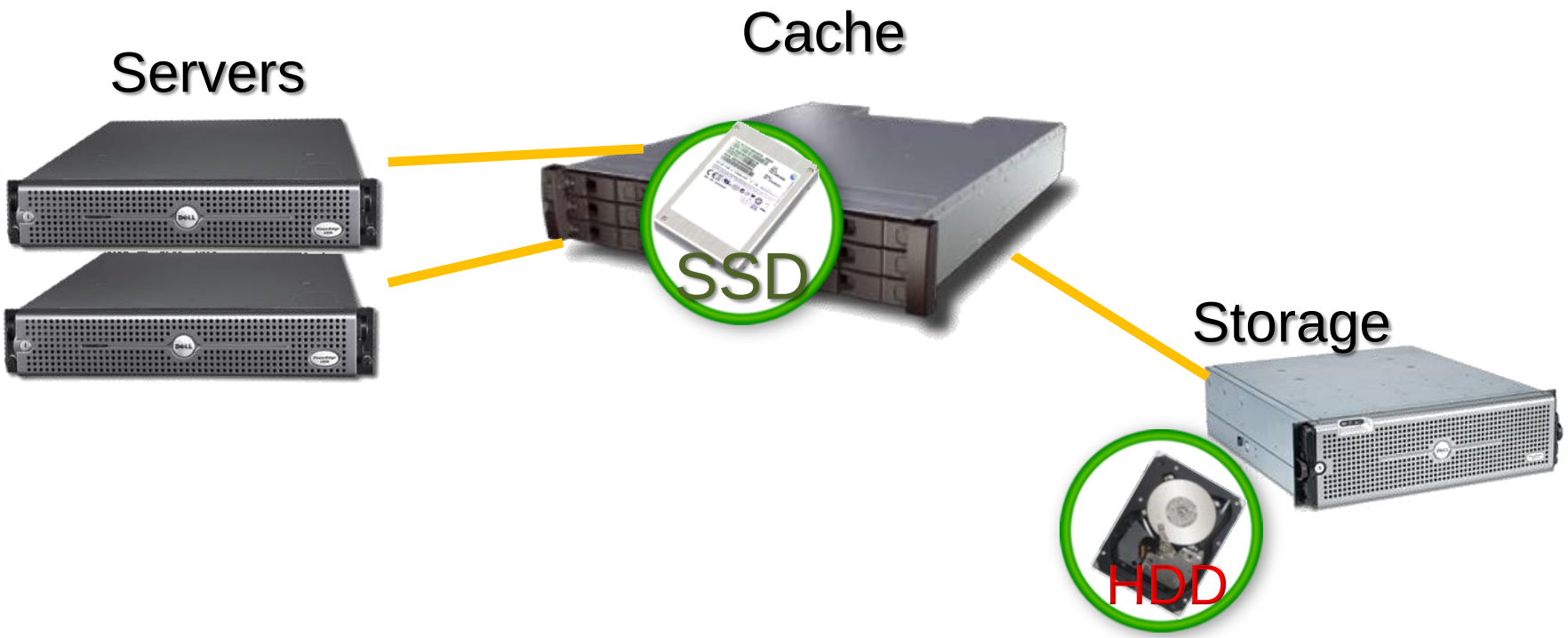
HDD Replacement

- **Replace boot drive or main storage**
- **Fastest and easiest way to experience SSDs**



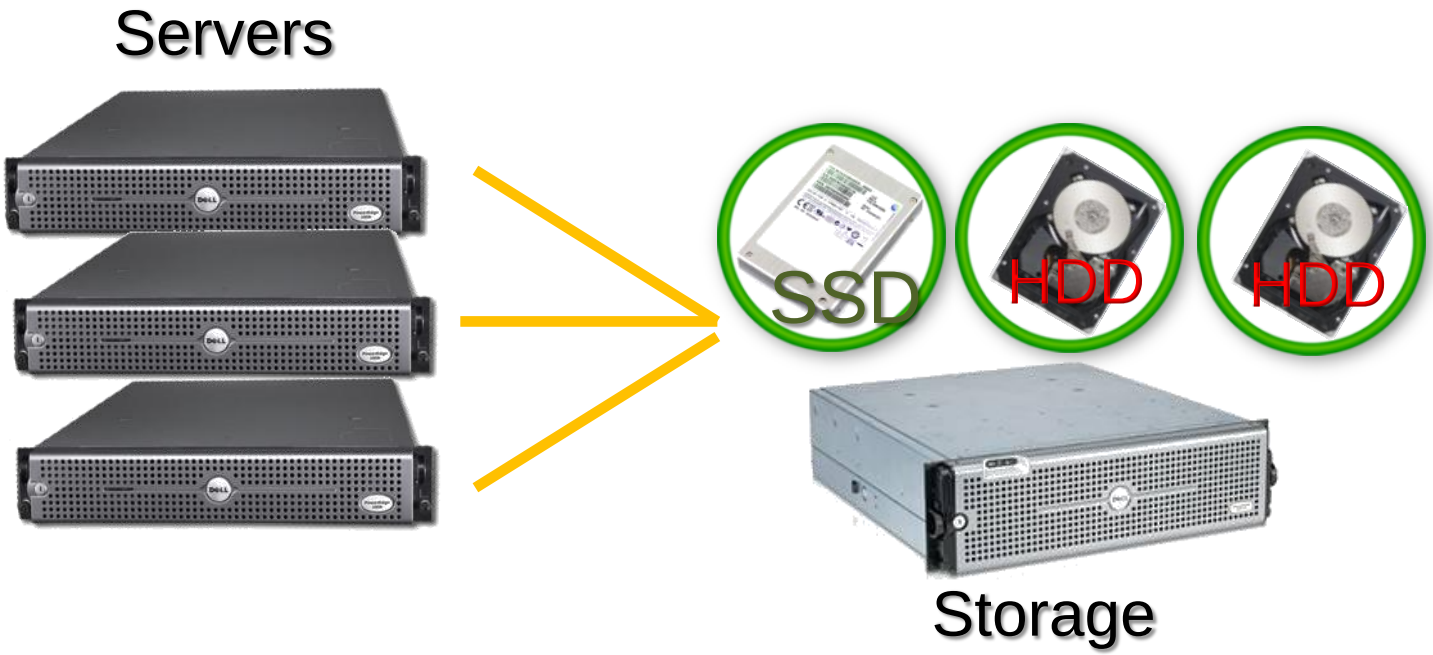
Caching Appliance

- **Read and/or Write Cache**
- **Sits between servers and storage, typically in a SAN**
- **Used to speed up legacy or slower storage**



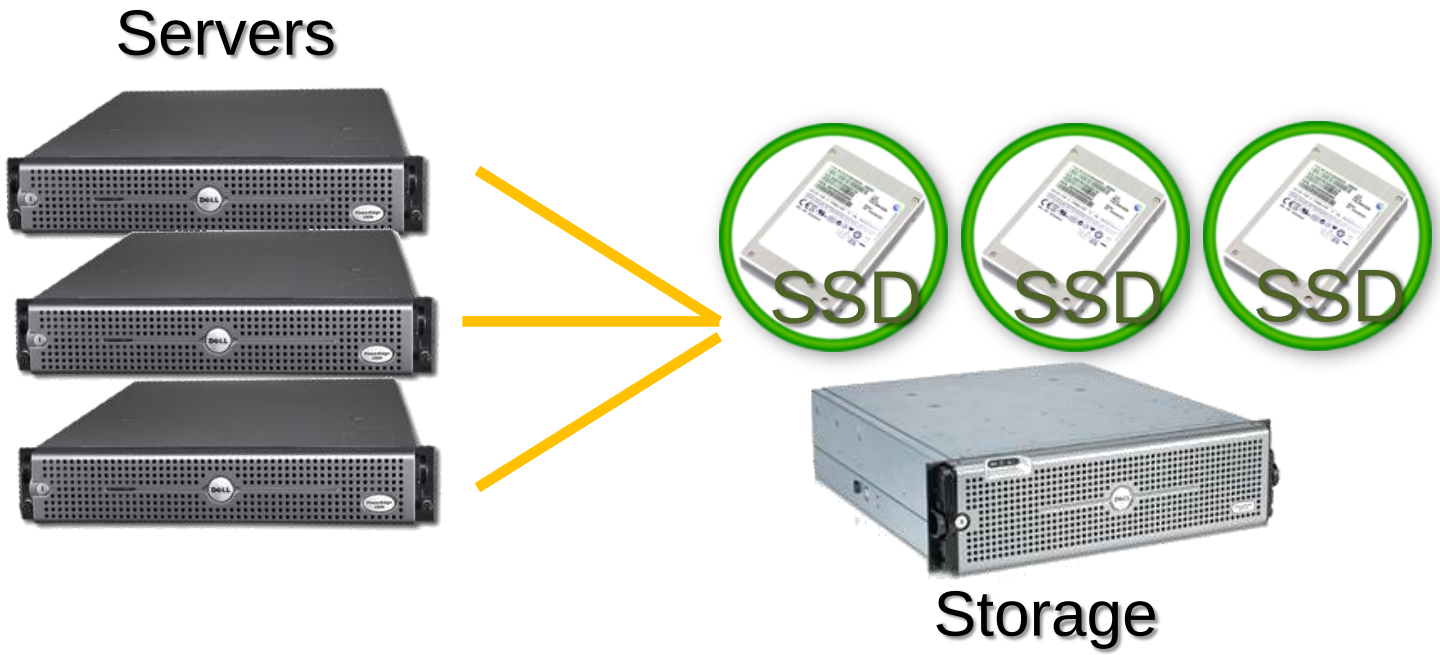
Tiered Storage

- **An external storage device (NAS, SAN)**
- **Only puts "hot" or "critical" data on SSD**
- **Most of the storage is still on HDD**



All Flash Storage

- External storage based on 100% SSD/Flash
- Typically uses MLC and de-duplication/compression to achieve better pricing
- Designers of these systems are Flash experts



Thank You!

Introducing...
SAMSUNG
Green Memory
web site

www.samsung.com/greenmemory

New Green Memory App Available

- Applications

Green Memory Apps for Android OS	Green Memory Apps for iOS	Green Memory Mobile Homepage
QR code for Galaxy Tab & Galaxy S	QR code for iPad & iPhone	QR code for Mobile Homepage

SAMSUNG