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!

# of SSDs sold

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

Processor:
- Multi-core
- Higher bandwidth

Memory:
- Larger footprint
- Higher bandwidth

Storage:
- Minor throughput improvements
- Currently solved with spindles

Closing the gap with Solid State Storage
Why an SSD?

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

**Random Performance (IOPS)**

- **Read**: X100, 43K
- **70:30**: X60, 23K
- **Write**: X30, 11K

**Power Consumption (Watt)**

- **Idle**: 8.5, -87%
- **Active**: 12.6, -75%

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

Source: Samsung

samsung.com/greenmemory
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

samsung.com/greenmemory
SSD Key Characteristics
SSD Components

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

All components work closely together

SSD Image Source: Anandtech
samsung.com/greenmemory
NAND Characteristics

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

- 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

<table>
<thead>
<tr>
<th>ECC Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

3,000

samsung.com/greenmemory
Write Amplification Factor (WAF)

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:
  - Flash Translation Layer (FTL)
  - Wear Leveling
  - Over-provisioning
  - Garbage Collection
  - Write Profile (Ran vs. Seq)
  - Free user space / TRIM
Write Amplification (WAF) Example

Below example illustrates WAF of 6

- Host wants to update LBA 0
- No more free pages
- Need to erase entire block
- Read existing data to Cache
- Erase block
- Write modified page and old pages back to Flash

4KB from Host

Host

Cache

SSD

Flash

24KB to NAND

Time

samsung.com/greenmemory
# TeraBytes Written

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

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

Examples:

\[
\begin{align*}
((128\text{GB} / 1000) \times 3000) / 5 &= 76.8 \text{ TBW} \\
((128\text{GB} / 1000) \times 3000) / 2.5 &= 153.6 \text{ TBW} \\
((256\text{GB} / 1000) \times 3000) / 5 &= 153.6 \text{ TBW} \\
((128\text{GB} / 1000) \times 30000) / 5 &= 768 \text{ TBW}
\end{align*}
\]
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]

<table>
<thead>
<tr>
<th>ID</th>
<th>Attribute Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Reallocated Sector Count</td>
</tr>
<tr>
<td>9</td>
<td>Power-on Hours</td>
</tr>
<tr>
<td>12</td>
<td>Power-on Count</td>
</tr>
<tr>
<td>177</td>
<td>Wear Leveling Count</td>
</tr>
<tr>
<td>179</td>
<td>Used Reserved Block Count</td>
</tr>
<tr>
<td>180</td>
<td>Unused Reserved Block Count</td>
</tr>
<tr>
<td>181</td>
<td>Program Fail Count</td>
</tr>
<tr>
<td>182</td>
<td>Erase Fail Count</td>
</tr>
<tr>
<td>187</td>
<td>Uncorrectable Error Count</td>
</tr>
<tr>
<td>195</td>
<td>ECC Error Count</td>
</tr>
<tr>
<td>199</td>
<td>CRC Error Count</td>
</tr>
<tr>
<td>241</td>
<td>Total LBA Written</td>
</tr>
</tbody>
</table>
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?

<table>
<thead>
<tr>
<th>PC</th>
<th>Server</th>
<th>External Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATA</td>
<td>SATA</td>
<td>SATA + SAS bridge</td>
</tr>
<tr>
<td>PCIe</td>
<td>SAS, PCIe</td>
<td>SAS</td>
</tr>
</tbody>
</table>

samsung.com/greenmemory
• There can be significant differences in sustained vs. peak performance.
• 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.
Benchmarking

- Synthetic or actual workload & take measurements

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNIA RTP 2.0</td>
<td><a href="http://www.snia.org/tech_activities/standards/curr_standards/pts">http://www.snia.org/tech_activities/standards/curr_standards/pts</a></td>
</tr>
<tr>
<td>Iometer</td>
<td><a href="http://sourceforge.net/projects/io-meter/">http://sourceforge.net/projects/io-meter/</a></td>
</tr>
<tr>
<td>CrystalDiskMark</td>
<td><a href="http://crystalmark.info/software/CrystalDiskMark/index-e.html">http://crystalmark.info/software/CrystalDiskMark/index-e.html</a></td>
</tr>
<tr>
<td>HD Tune Pro</td>
<td><a href="http://www.hdtune.com/">http://www.hdtune.com/</a></td>
</tr>
<tr>
<td>AS SSD (SSD)</td>
<td><a href="http://alex-is.de/PHP/fusion/downloads.php?download_id=9">http://alex-is.de/PHP/fusion/downloads.php?download_id=9</a></td>
</tr>
<tr>
<td>Scripts</td>
<td>Have multiple “dd” running with best guess workload, capturing timing/speeds</td>
</tr>
<tr>
<td>Real Workload</td>
<td>Capture trace during real workload and playback (ioapps, blktrace/btereplay)</td>
</tr>
</tbody>
</table>
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

No TRIM needed

TRIM makes SSD aware
Over-Provisioning

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

Sample 128GB SSD

<table>
<thead>
<tr>
<th></th>
<th>120GB</th>
<th>100GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over-Provisioning</td>
<td>7%</td>
<td>28%</td>
</tr>
<tr>
<td>Random Read (8K) IOPS</td>
<td>80K</td>
<td>80K</td>
</tr>
<tr>
<td><strong>Random Write (8K) IOPS</strong></td>
<td><strong>1,800</strong></td>
<td><strong>6,300</strong></td>
</tr>
<tr>
<td>Sequential Read (64K) MB/s</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Sequential Write (64K) MB/s</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>4KB Random WAF</td>
<td>5</td>
<td>1.35</td>
</tr>
<tr>
<td>4KB Random TBW</td>
<td>15</td>
<td>45</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Sequential</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLC 512GB SSD</td>
<td>60 TBW</td>
<td>1250 TBW</td>
</tr>
</tbody>
</table>

- 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
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