NAND Flash Solid State Storage
Performance and Capability -- an In-depth Look

lance l. smith, Fusion-io
The material contained in this tutorial is copyrighted by the SNIA.

Member companies and individual members may use this material in presentations and literature under the following conditions:
- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA Education Committee.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
NAND Flash Solid State Storage Performance and Capability

"This presentation provides an in-depth examination of the fundamental theoretical performance, capabilities, and limitations of NAND Flash-based Solid State Storage (SSS). The tutorial will explore the raw performance capabilities of NAND Flash, and limitations to performance imposed by mitigation of reliability issues, interfaces, protocols, and technology types. Best practices for system integration of SSS will be discussed. Performance achievements will be reviewed for various products and applications."
Moore’s continues to beat Newton’s Law

- Mechanical Drives have hit their limits
  - Platter stability degrades at higher speeds
  - Short-stroking reduces capacity for seek time
  - Capacity is limited by smaller form factors

- Solid State Storage continues to evolve
  - Greatest bit density (bits per cubic volume)
  - Random IOPS are 250 times greater
  - MLC increases capacity and lowers costs
  - Advanced error correction improves reliability
  - Performance and Capacity are intertwined
Moore’s continues to beat Newton’s Law

• Mechanical Drives have hit their limits
  • Platter stability degrades at higher speeds
  • Short-stroking reduces capacity for seek time
  • Capacity is limited by smaller form factors

• Solid State Storage continues to evolve
  • Greatest bit density (bits per cubic volume)
  • Random IOPS are 250 times greater
  • MLC increases capacity and lowers costs
  • Advanced error correction improves reliability
  • Performance and Capacity are intertwined
Data Integrity + Performance

There can be no data integrity trade-off for performance
Media Reliability / Availability

❖ The GOOD
   ❖ No moving parts
   ❖ Post infant mortality (catastrophic) device failures are rare
   ❖ Predictable wear out

❖ The BAD
   ❖ Relatively high bit error rate, which increases with wear
   ❖ Higher density and MLC increases bit error rate
   ❖ Program and Read Disturbs

❖ The UGLY
   ❖ Partial Page Programming
   ❖ Data retention is poor at high temperature and wear
   ❖ Infant mortality is high (large number of parts…)

NAND Flash Solid State Storage Performance and Capability
© 2009 Storage Networking Industry Association. All Rights Reserved.
Controller Reliability Management

- Wear leveling & Spare Capacity
- Read & Program Disturb control
- Data & Index Protection
  - ECC Correction
  - Internal RAID
  - Data Integrity Field (DIF)
- Management

Poor Media + Great Controller = Great SSS Solution
Data Integrity versus Performance

Performance

99% 99.9% 99.99% 99.999% ...

Data Integrity

GOOD BAD

NAND Flash Solid State Storage Performance and Capability
© 2009 Storage Networking Industry Association. All Rights Reserved.

Tuesday, August 18, 2009
Performance is about ROI

**Lower OpEx**
- Less HW Maintenance
- Less SW Maintenance
- Greater Uptime
- Less Power/Cooling
- Fewer Diverse Skills

**Lower CapEx**
- Fewer CPUs
- Less RAM
- Less Network Gear
- Fewer SW Licenses
- Less Space

**HIGHER Productivity**
Media Performance

❖ The GOOD

❖ Performance is excellent (wrt HDDs)
❖ High performance per power (IOPS/Watt)
❖ Low pin count: shared command / data bus \(\rightarrow\) good balance

❖ The BAD

❖ Not really a random access device
  > Block oriented
  > R/W access speed imbalance
❖ Slow effective write (erase/transfer/program) latency
❖ Performance changes with wear

❖ The UGLY

❖ Some controllers do read/erase/modify/write
❖ Others use inefficient garbage collection
Performance Drivers – SSS Design

- Number of NAND Flash Chips (Die)
- Number of Channels (Real / Pipelined)
- Interconnect
- Data Protection (internal/external RAID; DIF; ECC…)
- SLC / MLC Flash Type
- Effective Block Size (LBA; Sector)
- Write Amplification Efficiency
- Garbage Collection (GC) Efficiency
- Bandwidth Throttling
- Buffer Capacity & Mgmt
Simplified Theoretical Analysis

- **Bandwidth Only (Not IOPS)**
  - Large Transfers (Data length = Integer times die count)
  - Infinite Buffer
  - Reads/Writes queued for maximum bandwidth
  - No system latency

- **Read/Write Ratio %’s fixed**
  - 100/0, 75/25, 50/50, 25/75, 0/100
  - Steady State, 100% Efficient GC (EB erase/EB written = 1)

- **Maximum Total BW for SATA-II and PCI-e X4**
  - No overhead considered
## Bandwidth Depends on Die Count

<table>
<thead>
<tr>
<th></th>
<th>SLC</th>
<th>MLC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transfer Rate (MB/s)</strong></td>
<td>tRC &amp; tWC</td>
<td>400</td>
</tr>
<tr>
<td><strong>Page Program (us)</strong></td>
<td>tProgram</td>
<td>200</td>
</tr>
<tr>
<td><strong>EB Erase (us)</strong></td>
<td>tErase</td>
<td>3000</td>
</tr>
<tr>
<td><strong>Load Page (us)</strong></td>
<td>tR (tRead)</td>
<td>25</td>
</tr>
<tr>
<td><strong>Capacity per die</strong></td>
<td>0.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

---

**Theoretical BW (MB/s) v Number of Die (SLC, MLC)**

- SLC BW (0/100, 50/50, 75/25) MB/s
- MLC BW (50/50) MB/s
- MLC BW (25/75) MB/s
- MLC BW (0/100) MB/s

Presumes 8 die per bus & 4 CS per bus
Read / write performance imbalance closed with additional banks
Greater R/W imbalance in MLC requires more banks
## Features directly affecting performance measurements

<table>
<thead>
<tr>
<th></th>
<th>SATA (A)</th>
<th>SATA (B)</th>
<th>PCI (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capacity (GB)</strong></td>
<td>32</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td><strong>Bus/Link</strong></td>
<td>SATA-II (3 Gb/s)</td>
<td>SATA-II (3 Gb/s)</td>
<td>PCI-E X4 1.1</td>
</tr>
<tr>
<td><strong>Memory Type</strong></td>
<td>SLC</td>
<td>SLC</td>
<td>SLC</td>
</tr>
<tr>
<td><strong>Adjustable Reserve Capacity</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>SSS Internal RAID</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>-- Running during test</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>K-IOPS (RMS)</strong></td>
<td>8</td>
<td>27</td>
<td>88</td>
</tr>
<tr>
<td><strong>K-IOPS (RMS) / WATT</strong></td>
<td>3</td>
<td>?</td>
<td>7</td>
</tr>
<tr>
<td><strong>Bandwidth (RMS, MB/s)</strong></td>
<td>56</td>
<td>208</td>
<td>743</td>
</tr>
<tr>
<td><strong>ECC correction</strong></td>
<td>7 bits in 512B</td>
<td>4 bits in ?</td>
<td>11 bits in 240B</td>
</tr>
</tbody>
</table>
Measured vs Theoretical Bandwidth

Note: Theoretical Max BW with 24 channels (4 die per bus, 4 CS per bus) is identical to the PCI-C, 24 channel shown in these charts.

Capacity Multiplier:
- SATA-B: 1
- PCI-C: 2
Access Process (Physics Ignored)

➢ Read Access
  ➢ Address Chip / EB / Page
  ➢ Load Page into Register
  ➢ Transfer Data From Register 1-byte per cycle

➢ Write Access
  ➢ Address Chip / EB
  ➢ Erase EB

➢ …some time later…
  ➢ Address Chip / EB / Page
  ➢ Transfer Data To Register 1-byte per cycle
  ➢ Program Register to Page

Typical NAND Flash Die:
  • 2000 Erase Blocks (EB)
  • 64 Pages per EB
  • 4000 Bytes per Page
  • 500 MByte Total Capacity
### Example 1: Read/Erase/Modify/Write

#### Time = t1

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>b c</td>
</tr>
<tr>
<td>1</td>
<td>j k l</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>q r</td>
</tr>
</tbody>
</table>

#### Time = t2

Starting State

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>b c</td>
</tr>
<tr>
<td>1</td>
<td>j k l</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>q r</td>
</tr>
</tbody>
</table>

Write Buffer & W,X,Y

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>b c W X</td>
</tr>
<tr>
<td>1</td>
<td>j Y k l</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>q r</td>
</tr>
</tbody>
</table>

#### Time = t3

Write Buffer & Z,A,B’,C’,R’

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>B' C' w x</td>
</tr>
<tr>
<td>1</td>
<td>j Y k l</td>
</tr>
<tr>
<td>2</td>
<td>m Z A</td>
</tr>
<tr>
<td>3</td>
<td>q R'</td>
</tr>
</tbody>
</table>

Buffer holds data while EB-1 Erased
Example 2: Read/Modify/Write

<table>
<thead>
<tr>
<th>Time = t1</th>
<th>Time = t2</th>
<th>Time = t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting State</td>
<td>Data to Buffer (not shown)</td>
<td>Data to Buffer (not shown)</td>
</tr>
<tr>
<td></td>
<td>Erase EB-1 (not shown)</td>
<td>Erase EB-1 (not shown)</td>
</tr>
<tr>
<td></td>
<td>Write Buffer &amp; W,X,Y to EB-1</td>
<td>Write Z,A &amp; Replace b,c,r with B’,C’,R’ &amp; Write EB-1</td>
</tr>
</tbody>
</table>

Implicit wear leveling; EB-1 → EB-2 → EB-3
Presumes that destination EB-2 & EB-3 erased prior to transfer of data → higher performance (than previous “Read/Erase/Modify/Write” example)
Example 3: Garbage Collection

<table>
<thead>
<tr>
<th>Time = t1</th>
<th>Time = t2</th>
<th>Time = t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Garbage Collect EB-1</td>
<td>EB-1 GC’d to EB-2</td>
<td>EB-1 erase</td>
</tr>
<tr>
<td>W,X,Y added</td>
<td>b,c,r replaced by B’,C’,R’</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
<th>Page</th>
<th>Erase Block 1</th>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>b</td>
<td>0</td>
<td>b</td>
<td>0</td>
<td>b</td>
</tr>
<tr>
<td>1</td>
<td>j</td>
<td>1</td>
<td>j</td>
<td>1</td>
<td>j</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
<td>2</td>
<td>m</td>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>--</td>
<td>3</td>
<td>--</td>
<td>3</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 2</th>
<th>Page</th>
<th>Erase Block 2</th>
<th>Page</th>
<th>Erase Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>--</td>
<td>0</td>
<td>W</td>
<td>0</td>
<td>W</td>
</tr>
<tr>
<td>1</td>
<td>--</td>
<td>1</td>
<td>Y</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>--</td>
<td>2</td>
<td>m</td>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>--</td>
<td>3</td>
<td>--</td>
<td>3</td>
<td>--</td>
</tr>
</tbody>
</table>

NAND Flash Solid State Storage Performance and Capability
© 2009 Storage Networking Industry Association. All Rights Reserved.
In this example,

- **COPIED DATA:** \{b, c, j, k, l, m, q, r\} 8 blocks
- **NEW DATA:** \{W, X, Y, B', C', Z, A, R'\} 8 blocks

- 50% (8 of 16) writes are user initiated
- 50% (8 of 16) writes are internal movement (overhead)

**Important:**

- 50% of EB-1 was “invalid data”
- What if only 10% had been “invalid data?”
- GC efficiency is dependent upon % of reserve capacity
Tower of Hanoi

Want to do this in fewer moves? Add more pegs!
GC: Pathological Write Conditions

- If a high percentage of total storage capacity utilized
  AND
- A High percentage of data has no correlation-in-time
  AND
- Continuous writing (no recovery time for GC)
  THEN...

Efficiency of GC greatly diminished
Pathological Write Condition

Graph showing the performance of various storage capacities over time. The graph includes different lines representing user capacity formatted out of total capacity, with capacities ranging from 28 GiB to 70 GiB.

User Capacity Formatted of Total
- 30 GiB of 80G PCI-C
- 40 GiB of 80G PCI-C
- 60 GiB of 80G PCI-C
- 70 GiB of 80G PCI-C
- 74 GiB of 80G PCI-C
- 28 GiB of 30G SATA-B

Y-axis: MB/s
X-axis: Seconds

NAND Flash Solid State Storage Performance and Capability
© 2009 Storage Networking Industry Association. All Rights Reserved.

Tuesday, August 18, 2009
Performance vs R/W Ratio

Read/Write Collisions → Drop in Mixed Performance
Scalability versus R/W Ratio

R/W Ratio and Number of Devices in Parallel
Performance vs Block Size (75/25)

75/25 R/W IOPS

Block Size

75/25 R/W Bandwidth (MB/s)

Block Size

NAND Flash Solid State Storage Performance and Capability
© 2009 Storage Networking Industry Association. All Rights Reserved.
SATA-A Scalability vs R/W vs Block Size
SATA-B Scalability vs R/W vs Block Size
System Level Considerations

- Data / Index Protection (RAID and DIF)
- Scalability
- Compare system- or data-center-level
  - Not device
- Best case: test on real application
  - Not benchmark
  - Plan to do tuning to reach top perf. / objectives
  - Applications may have contra-indicated optimizations
    - Keeping data in close physical proximity (short stroking)
    - Caching algorithms
Questions to Ask: Things to Know

- Bandwidth / IOPS at
  - Block size(s) you need
  - R/W ratio you use
  - Steady State / Burst
  - Reserve capacity used
  - Data’s temporal relationship
  - Scalability
  - RAIDing
  - BOL / EOL

- Design impacts on data integrity; life; failures & perf.
  - ECC robustness
  - Write amplification / GC efficiency
  - Internal RAID
  - Bandwidth throttling
  - Partial Page Programming

- Test Conditions
  - Workload
  - Temporal Relationships
  - User capacity / reserve capacity
Please send any questions or comments on this presentation to SNIA: tracksolidstate@snia.org

Many thanks to the following individuals for their contributions to this tutorial.
- SNIA Education Committee

Jonathan Thatcher
Khaled Amer
Phil Mills
Rob Peglar
Marius Tudor