NAND Flash Solid State Storage
Reliability and Data Integrity -- an In-depth Look

jonathan thatcher, 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.
Abstract

NAND Flash Solid State Storage Reliability and Data Integrity

Abstract: “This tutorial provides an in-depth examination of NAND Flash as a storage medium and control techniques used to ensure enterprise-grade Solid State Storage (SSS) reliability and data integrity. Several aspects of the topic are discussed, including reliability characteristics unique to the medium (e.g. wear out, shelf life, infant mortality, failure modes), and implications for architecture, implementation, system integration, and deployment of NAND Flash-based SSS. These topics will be reviewed with a focus on device failure rates, data integrity, and data availability. Finally, recommendations are made regarding best practices, and information that users should obtain from potential vendors to properly assess NAND Flash SSS device reliability and data integrity.”
The GOOD:
- No moving parts
- Catastrophic device failures are rare (post infant mortality)

The BAD:
- Relatively high bit error rate, increasing with wear
  - MLC wear rate (higher capacity density) worse than SLC
  - Higher density NAND Flash will increase bit error rate
- Program and Read Disturbs

The UGLY:
- Partial Page Programming
- Data retention is poor at high temperature
- Infant mortality is high (large number of parts….)
Controller Reliability Management

- Wear leveling & Spare Capacity (e.g. Spare Blocks)
- Read & Program Disturb Controls
- Data & Index Protection
  - ECC Correction
  - Internal RAID
  - Data Integrity Field (DIF)
- Management

Poor Media + Great Controller → Great SSS Solution

Note: Multipage Programming Should Not Be Done
Is the following example about Performance or Reliability?
Company: Wine.com

- On-line internet Retail

- Representative Markets:
  - Transaction Processing
  - Data Mining

- Problem (4Q07)
  - Systems not capable → Contract out data mining
  - Performance hitting 100% → Loss of revenue
  - Cost to meet growth: prohibitive
Wine.com

Before (3Q2008)

- Front End
  - Internet
  - Back End
  - Payment Processing

After (4Q2008)

- Internet
- Application level replication for redundancy
- 33% CapEx avoidance
- 50% OpEx reduction
- 33% Footprint reduction
- 12x improvement on write
  - Latency down from 4 to <1ms
- 14x improvement on read
  - Latency down from 12 to <1ms

“Enough capacity to cover 24 months of growth”

Geoffrey Smalling - CTO
### RAS Improvements

<table>
<thead>
<tr>
<th>Change</th>
<th>Before</th>
<th>After</th>
<th>Reliability Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td># Servers</td>
<td>8</td>
<td>4</td>
<td>Substantial</td>
</tr>
<tr>
<td># Software Licenses</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td># Ethernet Routers</td>
<td>2</td>
<td>0</td>
<td>Substantial</td>
</tr>
<tr>
<td># NAS Appliances</td>
<td>2</td>
<td>0</td>
<td>Substantial</td>
</tr>
<tr>
<td># HDDs</td>
<td>28</td>
<td>0</td>
<td>Substantial</td>
</tr>
<tr>
<td># Ethernet Switches</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td># SSS Devices</td>
<td>0</td>
<td>6</td>
<td>(Minimal)</td>
</tr>
</tbody>
</table>

And some think SSS is only about performance

Think Again!
Performance

↓

Reliability
## Product Specifications

### 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 -- Running during test</strong></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></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>?</td>
<td>11 bits in 240B</td>
</tr>
</tbody>
</table>
IF

product **performance** is **10X**,

can the **life** be **0.1X**?

Can **infant mortality** be **10X**?
With repeated P/E Cycles, electrons get “stuck” in the oxide, preventing a reliable “read” of a logical 0 and a logical 1.
Memory Cell Wear Out

Higher BER

< Data Integrity

< Life

NAND Flash Solid State Storage Reliability and Data Integrity
© 2009 Storage Networking Industry Association. All Rights Reserved.
Raw Erase Error Rate (SLC, MLC)
Read Disturb Raw Error Rate (MLC)
Retirement v Temperature (MLC)

Error Rate Decreases with Increase in Temperature
But,
Data Retention Decreases with Temperature
Erase Cycle Effect on Read Disturbs

Read Disturbs Bit Errors
Vendor A, MLC

Bit Error Rate (BER)

0 200000 400000 600000 800000 1e+06

Read Cycles

0.0001

0.00001

1e-08

1e-07

1e-06

1e-05

Read BER after 0 Cycles
Read BER after 2K Cycles
Read BER after 4K Cycles
Read BER after 6K Cycles
The Need For Robust ECC

The reason we have SSS is:

Thumb drives
Digital cameras
iPOD
Etc.

Drive the price down and capacity up.

These applications do not need:

Robust data integrity
Large number of write cycles
ECC Extends the PEB Life (SLC)
ECC Extends the PEB Life (MLC)

ECC Comparison: Manufacturer Recommendation vs Aggressive
Probability of PEB Retirement
Vendor A, MLC

Longer Life
MFG Recommendation

Robust ECC / Aggressive Retirement

Vendor B -- MLC: Probability of PEB Retirement with Temperature
Manufacturer ECC Recommendation: 1 bit Run, 4 bits total, 512 Bytes

3X Longer Life
Probability of Exceeding ECC (MLC)

- Longer Life
- Greater Data Integrity

![Graph showing the probability of exceeding ECC correction for different ECC levels and erase-program cycles.](image)

NAND Flash Solid State Storage Reliability and Data Integrity
© 2009 Storage Networking Industry Association. All Rights Reserved.
Data Integrity V. Extended Life

Extended Life

BAD

GOOD

99%  99.9%  99.99%  99.999%  …

Data Integrity

NAND Flash Solid State Storage Reliability and Data Integrity
© 2009 Storage Networking Industry Association. All Rights Reserved.
## Life (“TBW : Terabytes Written”)

<table>
<thead>
<tr>
<th></th>
<th>Life Expectancy</th>
<th>Average Rate</th>
<th>Used per Day</th>
<th>Consumed per Year</th>
<th>Life Expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consumer Vehicle</strong></td>
<td>200K miles</td>
<td>45 MPH</td>
<td>1.5 Hrs</td>
<td>25K Miles</td>
<td>8.1 years</td>
</tr>
<tr>
<td><strong>Commercial Vehicle</strong></td>
<td>5,000K miles</td>
<td>55 MPH</td>
<td>11 Hrs</td>
<td>221K Miles</td>
<td>22.6 years</td>
</tr>
<tr>
<td><strong>Consumer (MLC)</strong></td>
<td>32,000 TBW</td>
<td>100 MB/s</td>
<td>7 Hrs</td>
<td>920 TB</td>
<td>34.8 years</td>
</tr>
<tr>
<td><strong>Enterprise (SLC)</strong></td>
<td>160,000 TBW</td>
<td>250 MB/s</td>
<td>24 Hrs</td>
<td>7,884 TB</td>
<td>20.3 years</td>
</tr>
</tbody>
</table>

*Your Mileage May Vary*
## Life Calculation

<table>
<thead>
<tr>
<th>“Enterprise”</th>
<th>&quot;Consumer&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000,000</td>
<td>100,000</td>
</tr>
<tr>
<td>Program-Erase Cycles</td>
<td>Capacity (B)</td>
</tr>
<tr>
<td>160,000,000,000</td>
<td>320,000,000,000</td>
</tr>
<tr>
<td>160,000,000,000,000,000</td>
<td>Total Bytes Written</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>Hours used per day</td>
<td>Seconds used per day</td>
</tr>
<tr>
<td>86,400</td>
<td>25,200</td>
</tr>
<tr>
<td>Write Rate (B/s)</td>
<td></td>
</tr>
<tr>
<td>250,000,000</td>
<td>100,000,000</td>
</tr>
<tr>
<td>21,600,000,000,000</td>
<td>Bytes Written per day</td>
</tr>
<tr>
<td>7,884,000,000,000,000</td>
<td>Bytes Written per yr (B/yr)</td>
</tr>
<tr>
<td>20.3</td>
<td>34.8</td>
</tr>
<tr>
<td>Life in years</td>
<td></td>
</tr>
</tbody>
</table>

NAND Flash Solid State Storage Reliability and Data Integrity  
© 2009 Storage Networking Industry Association. All Rights Reserved.
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**

<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</td>
</tr>
<tr>
<td>2</td>
<td>m</td>
</tr>
<tr>
<td>3</td>
<td>q r</td>
</tr>
</tbody>
</table>

**Time = t1**

**Starting State**

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>W X</td>
</tr>
<tr>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>m Z A</td>
</tr>
<tr>
<td>3</td>
<td>q R'</td>
</tr>
</tbody>
</table>

**Time = t2**

**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</td>
</tr>
<tr>
<td>2</td>
<td>m Z A</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</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.
Assumptions
- Simplified to show erase blocks with 4 pages, each page having 4 data blocks
- Invalid (erased or replaced) data is indicated by “—”
- Old data is indicated by lower case letters
- New data is indicated by CAPs; Replacement data is indicated by “prime” (e.g. c \(\rightarrow\) C’)

Detail T = t1 to T = t2 transition
- Data is read from EB-1
- EB-1 is erased
- New data \{W, X, Y\} modifies previous invalid data
- Data is written back to EB-1

Detail T = t2 to T = t3 transition
- Data is read from EB-1 into data buffer
- EB-1 is erased
- New data \{B’, C’, Z, A, R’\} modifies previous data in data buffer
- Data is written back to EB-1

Note: backup material for those reviewing or looking at presentation without audio/video
**Example 2: Read/Modify/Write**

<table>
<thead>
<tr>
<th>Time = ( t_1 )</th>
<th>Time = ( t_2 )</th>
<th>Time = ( t_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Starting State</strong></td>
<td><strong>Data to Buffer (not shown)</strong></td>
<td><strong>Data to Buffer (not shown)</strong></td>
</tr>
<tr>
<td><strong>Erase Block 1</strong></td>
<td><strong>Erase EB-1 (not shown)</strong></td>
<td><strong>Erase EB-1 (not shown)</strong></td>
</tr>
<tr>
<td><strong>Erase Block 2</strong></td>
<td><strong>Write Buffer &amp; W,X,Y to EB-1</strong></td>
<td><strong>Write Z,A &amp; Replace b,c,r with B’,C’,R’ &amp; Write EB-1</strong></td>
</tr>
<tr>
<td><strong>Erase Block 3</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page</th>
<th>Erase Block 1</th>
<th>Page</th>
<th>Erase Block 2</th>
<th>Page</th>
<th>Erase Block 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>b  c  --  --</td>
<td>0</td>
<td>b  c  W  X</td>
<td>0</td>
<td>B’  C’  w  x</td>
</tr>
<tr>
<td>1</td>
<td>j  --  k  l</td>
<td>1</td>
<td>j  Y  k  l</td>
<td>1</td>
<td>j  y  k  l</td>
</tr>
<tr>
<td>2</td>
<td>m  --  --  --</td>
<td>2</td>
<td>m</td>
<td>2</td>
<td>m  Z  A</td>
</tr>
<tr>
<td>3</td>
<td>--  q  r</td>
<td>3</td>
<td>q  r</td>
<td>3</td>
<td>q  R’</td>
</tr>
</tbody>
</table>

Implicit wear leveling; EB-1 \( \rightarrow \) EB-2 \( \rightarrow \) EB-3
Presumes that destination EB-2 & EB-3 erased prior to transfer of data \( \rightarrow \) higher performance (than previous “Read/Erase/Modify/Write” example)
“Write Amplification Impact”

In this example,

- Data written t1 to t2: 16 blocks
  - NEW DATA {W, X, Y} 3 blocks; Copied Data {b, c, j, k, l, m, q, r} 8 blocks
  - Null Data: 5 blocks
- Data written t2 to t3: 16 blocks
  - NEW DATA {B’, C’, Z, A, R’} 5 blocks; Copied Data {w, x, j, y, k, l, m, q} 8 blocks
  - Null Data: 3 blocks
- (2) EB erasures
  - 25% (8 of 32) writes are user initiated
  - 75% (24 of 32) writes are internal data movement (overhead)

Important:

- Amount of valid or invalid data in EB-1 is irrelevant to performance impact
- “Write Amplification” is workload (access pattern) dependent (e.g., what if the write of R’ above was not coincident with B’ & C’)

© 2009 Storage Networking Industry Association. All Rights Reserved.
SATA-B: IOPS vs Transfer Size

Possible evidence of “write amplification”
### Example 3: Garbage Collection

<table>
<thead>
<tr>
<th>Time = t1</th>
<th>Start Garbage Collect EB-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>b c -- --</td>
</tr>
<tr>
<td>Page 1</td>
<td>j -- k l</td>
</tr>
<tr>
<td>Page 2</td>
<td>m -- -- --</td>
</tr>
<tr>
<td>Page 3</td>
<td>-- -- q r</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time = t2</th>
<th>EB-1 GC’d to EB-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>b c -- --</td>
</tr>
<tr>
<td>Page 1</td>
<td>j -- k l</td>
</tr>
<tr>
<td>Page 2</td>
<td>m -- -- --</td>
</tr>
<tr>
<td>Page 3</td>
<td>-- -- q r</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time = t3</th>
<th>EB-1 erase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page 0</td>
<td>W b c X</td>
</tr>
<tr>
<td>Page 1</td>
<td>Y j k l</td>
</tr>
<tr>
<td>Page 2</td>
<td>m q r</td>
</tr>
<tr>
<td>Page 3</td>
<td>C’ Z A R’</td>
</tr>
</tbody>
</table>
Assumptions
- New data blocks and data blocks being garbage collected are interleaved

Details
- At Time = t0, erase block 1 (EB-1) is identified for Garbage Collection (GC)
- At Time = t1, good data is moved from EB-1 to EB-2 (it is implicit that an index is updated accordingly); New data W, X, and Y are added while the GC is taking place. EB-1 is then ready to be erased
- At Time = t2, EB-2 is erased; Data for b, c, & r have been updated with B’, C’ & R’; b, c & r are indicated as “Invalid.”

Note: backup material for those reviewing or looking at presentation without audio/video
GC Performance Impact

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
GC: Pathological Write Conditions

- IF high percentage of total storage capacity utilized
  
  AND

- 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

User Capacity Formatted of Total

- 30GiB of 80G PCI-C
- 40GiB of 80G PCI-C
- 60GiB of 80G PCI-C
- 70GiB of 80G PCI-C
- 74GiB of 80G PCI-C
- 28GiB of 30G SATA-B

MB/s vs. Seconds

NAND Flash Solid State Storage Reliability and Data Integrity
© 2009 Storage Networking Industry Association. All Rights Reserved.
Your Mileage Will Vary
SSS Solutions have these features now

MANY DO NOT

Do your Homework
Choose Wisely
Caveat Emptor
SSS Data Protection Mechanisms

- Robust ECC
  - Increased need for MLC
  - Increased need as NAND Flash Capacity Increases
- Index Protection (ECC, DIF, Other)
  - Eliminates Index Induced Silent Data Corruption
- Internal RAID
  - Overcomes High Infant Mortality
  - Continued Protection Over Life
SSS Data Protection Mechanisms

- Power Loss
  - Storage of metadata / indexes
  - Completion of acknowledged writes
- Data Scrubbing
- Read / Program Disturb Management
- Data Retention Management
- Thermal Management
- Partial Page Programming Avoidance
SSS Life Extension Mechanisms

- Robust ECC
- RAID (Internal)
- Wear Leveling – Reduce Hot Spots
- Wear-out Prediction – Planned Maintenance
- Write Amplification Avoidance
- Garbage Collector Efficiency
Reliability – External Conditions

- System Configuration
  - External RAID
  - Write Caching
  - DIF Support

- Application Characteristics
  - Read/Write Ratios
  - Temporal Randomness of Access
  - Steady State v Bursting

- Configuration / Tuning
  - Reserve Capacity Setting (% of used capacity)

- Device Management
  - Device Retirement (device wear affects error probabilities)
Contra-indicated?

- **Bandwidth Throttling**
  - May improve device life
  - May not translate to the system / data center or more global reliability gains.

- **Partial Page Programming**
  - May decrease write amplification
  - **Severely** reduces data integrity

- **SSS tuned for performance at a block size you don’t use**
  - Write amplification at block size you do use
  - Increased wear
  - Decreased data integrity
  - Reduced life
Global v Local Optimization

- Compare at data center / system level
  - Performance
  - Cost (CapEx; OpEx)
  - Power
  - Reliability, Availability

- Tune for Your Needs!
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

Khaled Amer
Rob Peglar
Marius Tudor