A Study of SSD Reliability in Large Scale Enterprise Storage Deployments

Stathis Maneas, Kaveh Mahdaviani, Tim Emami, Bianca Schroeder

USENIX FAST ‘20
Reliability of SSD-based enterprise storage systems

• What we know:
  - Four field studies (distributed data center storage systems).
Reliability of SSD-based enterprise storage systems

• What we know:
  ▪ Four field studies (distributed data center storage systems).

• We focus on enterprise storage systems:
  ▪ Different drives, workloads, and reliability mechanisms.
  ▪ High-end drives, reliability is ensured through RAID, etc.
Reliability of SSD-based enterprise storage systems

• What we know:
  ▪ Four field studies (distributed data center storage systems).

• We focus on enterprise storage systems:
  ▪ Different drives, workloads, and reliability mechanisms.
  ▪ High-end drives, reliability is ensured through RAID, etc.

• Factors that have not been studied before:
  ▪ 3D-TLC NAND.
  ▪ Large Capacity Drives (e.g., 8TB and 15TB).
  ▪ Firmware Versions.
  ▪ RAID Groups.
Systems Description

• 1.4 million SSDs.
• 2.5 years of data.
• SLC, cMLC, eMLC, 3D-TLC drives.
• 3 manufacturers.
• 18 drive models:
  ▪ 12 different capacities.
• Varying age, usage, and system configurations.
Replacement Types

- Issues can be reported by a drive, the storage layer, the file system, etc.

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL1</td>
<td>Predictive Failures</td>
</tr>
<tr>
<td></td>
<td>Threshold Exceeded</td>
</tr>
<tr>
<td></td>
<td>Recommended Failures</td>
</tr>
<tr>
<td>SL2</td>
<td>Aborted Commands</td>
</tr>
<tr>
<td></td>
<td>Disk Ownership I/O Errors</td>
</tr>
<tr>
<td></td>
<td>Command Timeouts</td>
</tr>
<tr>
<td>SL3</td>
<td>Lost Writes</td>
</tr>
<tr>
<td>SL4</td>
<td>SCSI Errors</td>
</tr>
<tr>
<td></td>
<td>Unresponsive Drive</td>
</tr>
</tbody>
</table>
Issues can be reported by a drive, the storage layer, the file system, etc.

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL1</td>
<td>Predictive Failures</td>
<td>12.78</td>
</tr>
<tr>
<td></td>
<td>Threshold Exceeded</td>
<td>12.73</td>
</tr>
<tr>
<td></td>
<td>Recommended Failures</td>
<td>8.93</td>
</tr>
<tr>
<td>SL2</td>
<td>Aborted Commands</td>
<td>13.56</td>
</tr>
<tr>
<td></td>
<td>Disk Ownership I/O Errors</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>Command Timeouts</td>
<td>1.81</td>
</tr>
<tr>
<td>SL3</td>
<td>Lost Writes</td>
<td>13.54</td>
</tr>
<tr>
<td>SL4</td>
<td>SCSI Errors</td>
<td>32.78</td>
</tr>
<tr>
<td></td>
<td>Unresponsive Drive</td>
<td>0.60</td>
</tr>
</tbody>
</table>
### Replacement Types

- Issues can be reported by a drive, the storage layer, the file system, etc.

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL1</td>
<td>Predictive Failures</td>
<td>12.78</td>
</tr>
<tr>
<td></td>
<td>Threshold Exceeded</td>
<td>12.73</td>
</tr>
<tr>
<td></td>
<td>Recommended Failures</td>
<td>8.93</td>
</tr>
<tr>
<td>SL2</td>
<td>Aborted Commands</td>
<td>13.56</td>
</tr>
<tr>
<td></td>
<td>Disk Ownership I/O Errors</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>Command Timeouts</td>
<td>1.81</td>
</tr>
<tr>
<td>SL3</td>
<td>Lost Writes</td>
<td>13.54</td>
</tr>
<tr>
<td>SL4</td>
<td>SCSI Errors</td>
<td>32.78</td>
</tr>
<tr>
<td></td>
<td>Unresponsive Drive</td>
<td>0.60</td>
</tr>
</tbody>
</table>

- **SCSI Errors** dominate!
- One third of drive replacements are merely preventative based on predictions (Category SL1)!
- SSDs rarely become completely unresponsive!
How frequently are SSDs replaced?

• *Annual Replacement Rate (ARR)*:

\[
ARR = \frac{\#\text{Failed Devices}}{\#\text{Device years}}
\]
The average ARR across the entire population is 0.22%, but rates vary widely (0.07 - 1.2%).
Drive Replacements

- Annual Replacement Rate (ARR):

  The average ARR across the entire population is 0.22%, but rates vary widely (0.07 - 1.2%)!
Drive Replacements

- **Annual Replacement Rate (ARR):**

  - The average ARR across the entire population is 0.22%, but rates vary widely (0.07 - 1.2%).

ARR for Google’s Data Centers (1-2.5%)
Drive Replacements

- **Annual Replacement Rate (ARR):**

  - The average ARR across the entire population is 0.22%, but rates vary widely (0.07 - 1.2%)!
Drive Replacements

• **Annual Replacement Rate (ARR):**

\[ ARR = \frac{\#Failed Devices}{\#Device \ years} \]

• **Which factors impact flash reliability?**
  - Flash Type (SLC, cMLC, eMLC, 3D-TLC).
  - Lithography.
  - Usage and Age.
  - Firmware Version.
  - Other factors (see the paper).
Flash Type

• **Common expectation**: Lower failure rates for SLC ($$$) versus cMLC/eMLC and 3D-TLC.
Flash Type

- **Common expectation:** Lower failure rates for SLC ( $$$ ) versus cMLC/eMLC and 3D-TLC.
• **Common expectation:** Lower failure rates for SLC ($$\$$) versus cMLC/eMLC and 3D-TLC.

- *SLC* drives not necessarily better than *MLC* drives.
- *eMLC* drives not necessarily better than *cMLC* drives.
• **Common expectation:** Lower failure rates for **SLC ($$\$$)** versus **cMLC/eMLC** and **3D-TLC**.

- **SLC** drives not necessarily better than **MLC** drives.
- **eMLC** drives not necessarily better than **cMLC** drives.
- **3D-TLC** drives have the highest replacement rates.
• Compare models with the same flash type.

• **Common expectation**: Higher failures rates for higher densities.
• Compare models with the same flash type.

• **Common expectation:** Higher failures rates for higher densities.

- **eMLC:** models with higher densities (1xnm) have higher replacement rates.
- **3D-TLC:** models with lower densities (V2) have higher replacement rates (the trend is reversed)!
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Percentage of P/E cycles limit used so far.
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Percentage of P/E cycles limit used so far.

- **eMLC**: The effect of infant mortality is evident!
- **3D-TLC**: The differences are not pronounced, other effects at play (capacity, age).
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Drive’s age (time deployed in production), as an indicator of wear-out.
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Drive’s age (time deployed in production), as an indicator of wear-out.

Source: https://www.nrel.gov/
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Drive’s age (time deployed in production), as an indicator of wear-out.

Source: [https://www.nrel.gov/](https://www.nrel.gov/)
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Drive’s age (time deployed in production), as an indicator of wear-out.

Source: https://www.nrel.gov/
Age

- Usage affects the reliability of SSDs, due to wear-out of their cells.
- Drive’s age (time deployed in production), as an indicator of wear-out.

- Infant mortality is significant (12–15 months)!

Source: [https://www.nrel.gov/](https://www.nrel.gov/)
• Usage affects the reliability of SSDs, due to wear-out of their cells.
• Drive’s age (time deployed in production), as an indicator of wear-out.

- Infant mortality is significant (12–15 months)!
- It takes a long time to stabilize (1.5–2 years)!
Firmware Version

• Compare individual firmware versions within the same model:
  ▪ Most SSDs (70%) have the same firmware version in our observation window.
• Consider SSDs which have seen little usage (< 1%).
• Compare individual firmware versions within the same model:
  ▪ Most SSDs (70%) have the same firmware version in our observation window.
• Consider SSDs which have seen little usage (< 1%).
• Compare individual firmware versions within the same model:
  ▪ Most SSDs (70%) have the same firmware version in our observation window.
• Consider SSDs which have seen little usage (< 1%).
• Compare individual firmware versions within the same model:
  ▪ Most SSDs (70%) have the same firmware version in our observation window.
• Consider SSDs which have seen little usage (< 1%).

- A drive’s firmware version has a tremendous impact on reliability (by a factor of 3-10X)!
- Firmware updates must be made as easy as possible for customers!
• How frequently do double failures occur?
  ▪ 2% of RAID groups see > 1 failure in our observation window.
Failure correlations within a RAID group

- How frequently do double failures occur?
  - 2% of RAID groups see > 1 failure in our observation window.

- How quickly after the first does the second failure happen?
Failure correlations within a RAID group

- How frequently do double failures occur?
  - 2% of RAID groups see > 1 failure in our observation window.

- How quickly after the first does the second failure happen?

- 46% of successive failures occur on the same day!
• How frequently do double failures occur?
  ▪ 2% of RAID groups see > 1 failure in our observation window.

• How quickly after the first does the second failure happen?

  46% of successive failures occur on the same day!
  Probability of 2nd failure within a week: 2.54%!
Failure correlations within a RAID group

• How frequently do double failures occur?
  ▪ 2% of RAID groups see > 1 failure in our observation window.

• How quickly after the first does the second failure happen?
  ▪ 46% of successive failures occur on the same day!
  ▪ Probability of 2nd failure within a week: 2.54%!
  ▪ The chance of a follow-up failure does not show a direct relationship with RAID group size!

• How are they related to RAID group size?
Conclusion – Final Remarks

• Many aspects different from expectations:
  ▪ A long period of infant mortality!
  ▪ Higher densities not always experience higher replacement rates.
  ▪ SLC not generally more reliable than MLC.

• Firmware versions can have a significant impact on replacements:
  ▪ Make firmware updates as easy and painless as possible!

• Temporally correlated failures within the same RAID group:
  ▪ No evidence that follow-up failures are correlated with RAID group size.
  ▪ Single-parity RAID configurations, data loss analysis, etc.

• Several other metrics and factors that were not presented:
  ▪ Capacity, Bad Blocks, Spare Blocks consumed, etc.
  ▪ Statistical tests.