Lessons and Actions: What We Learned from 10K SSD-Related Storage Failures

Erci Xu, Mai Zheng, Feng Qin, Yikang Xu, Jiesheng Wu
SSD-Based Storage System Powers

*The Life Essentials*
Concerns of SSD Reliability

• Wear out
• Limited Program/Erase Cycles
• New failure modes
• Program/Erase Error
• Metadata corruption
• Sensitive to environment
• NAND in heated air
Flash Reliability in Production: The Expected and the Unexpected

Abstract
As solid state drives based on flash technology are becoming a staple for persistent data storage in datacenters, it is important to understand their reliability characteristics. While there is a large body of work based on experiments with individual flash chips, controlled lab environments under synthetic workloads, there is a lack of information on their behavior in the field. This paper provides large-scale field study covering many millions of drive-days, the different drive models, different flash technologies (MLC, MLC2, SLC) over 6 years of production use in Google’s data centers. We study a wide range of reliability characteristics and come to a number of unexpected conclusions. For example, raw bit error rate (BER) grows at a much slower rate with respect to the operational data rate compared to and more than, in some cases, the FRU insertion rate. Unlike the bathtub curve, failures are not predicted by the number of reads or write, but rather by the number of bytes. We use this technique to better understand failures in the field.

Previous Large Scale SSD Studies

• E.g.: Failure rate curve
• not bathtub

• FTL impact
• Thermal Throttling

• Uncorrectable errors
Our Study

• Focus on RASR failures
  • Reported As SSD-Related

• Lessons and actions from three perspective:
  • Software Design
  • Hardware Architecture
  • System Administration
Outline

- INTRODUCTION
- SYSTEM ARCHITECTURE & DATASET
- RASR FAILURES OVERVIEW
- LESSONS AND ACTIONS
- CONCLUSIONS & FUTURE WORK
Alibaba Cloud Infrastructure

Cloud Service
- Block Service Software Stack
- NoSQL Service Software Stack
- Big Data Service Software Stack

Cluster

Rack
- Node Single Setup
  - 1 SSD
  - Sys. SSD
- Node Multiple Setup
  - 12-18 SSD
  - Sys. SSD
- Node Hybrid Setup
  - 12-36 HDD
  - 2 SSD
  - Sys. SSD

RASR Failures
SSD Fleet in Our Study

• Near half million SSDs from 3 vendors spanning over 3 years deployment

<table>
<thead>
<tr>
<th>Model</th>
<th>Capacity</th>
<th>Lithography</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>480GB</td>
<td>20nm</td>
<td>2-3 yrs</td>
</tr>
<tr>
<td>M2</td>
<td>800GB</td>
<td>20nm</td>
<td>2-3 yrs</td>
</tr>
<tr>
<td>M3</td>
<td>480GB</td>
<td>16nm</td>
<td>1-2 yrs</td>
</tr>
<tr>
<td>M4</td>
<td>480GB</td>
<td>20nm</td>
<td>2-3 yrs</td>
</tr>
<tr>
<td>M5</td>
<td>480GB</td>
<td>20nm</td>
<td>1-2 yrs</td>
</tr>
</tbody>
</table>

different SSD models

different SSD usages

<table>
<thead>
<tr>
<th>Service</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Service</td>
<td>Journaling</td>
</tr>
<tr>
<td></td>
<td>Persistence</td>
</tr>
<tr>
<td>NoSQL</td>
<td>Journaling</td>
</tr>
<tr>
<td></td>
<td>Persistence</td>
</tr>
<tr>
<td>Big Data</td>
<td>Temporary</td>
</tr>
</tbody>
</table>
RASR Failures Overview

• We have collected around 130K failure tickets over 3 years
• Around 6% of them is RASR Failures. Around 10K events.
RASR Failures Overview (cont.)

- 5 Symptoms of RASR Failures
L&A for Hardware Architects

• One DC in our deployment has higher-than-usual Media Error affected rate.
  Under same drive model
  Under same cloud service
Passive Heating in Hardware Architecture

Passive Heating: Heating on *idle* SSDs by neighboring active SSDs
Passive Heating Impacts

Can heat up *idle* SSDs by 28 Celsius Degrees
Passive Heating Solutions

Routine Scanning (~4 hrs)
- Software Based
- Close Temperature Monitoring

FTL Support
- Efficient Monitoring/Correcting
- Firmware Modification
L&A for Software Developers

• Certain cloud services may cause unbalanced usage of SSDs

<table>
<thead>
<tr>
<th>Service</th>
<th>Average Value Per Hour</th>
<th>Coefficient of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>7.69GB</td>
<td>35.5%</td>
</tr>
<tr>
<td>Big Data</td>
<td>1.57GB</td>
<td>1.8%</td>
</tr>
<tr>
<td>NoSQL</td>
<td>6.10GB</td>
<td>3.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service</th>
<th>Host Read</th>
<th>Host Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>7.69GB</td>
<td>6.56GB</td>
</tr>
<tr>
<td>Big Data</td>
<td>1.57GB</td>
<td>1.22GB</td>
</tr>
<tr>
<td>NoSQL</td>
<td>6.10GB</td>
<td>5.28GB</td>
</tr>
</tbody>
</table>

Block storage service has much higher CV which indicates the usage among SSD is not balanced.
Service Imbalance

- Histogram of usage with a step of 0.5GB/hr.
- The majority of SSDs under both NoSQL and Big Data Analytics services have similar values.
- The SSDs under the block storage service shows diverse values.
The updated chunk always write back to the same SSD.
The updated chunk is re-allocated to a *new SSD*.
L&A for System Admins: Part I

- 5 Symptoms of RASR Failures

![Bar chart showing the percentage of RASR failures for different symptoms]

- Drive Unfound
- Media Error
- Buffer IO Error
- File System Unmoutable
- Node Unbootable
UCRC errors indicate bad cables

SSDs with heavy UCRC errors are 2.7X more likely to lead to “Drive Unfound” failures
L&A for System Admins: Part II

• How to quickly identify root cause of failures?

<table>
<thead>
<tr>
<th>Fix</th>
<th>Percentage</th>
<th>Root Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rebooting</td>
<td>11.9%</td>
<td>Transient</td>
</tr>
<tr>
<td>Mount Options Check</td>
<td>0.4%</td>
<td>Human Mistake</td>
</tr>
<tr>
<td>FSCK</td>
<td>16.5%</td>
<td>Undetermined</td>
</tr>
<tr>
<td>Data Check</td>
<td>6.0%</td>
<td>Undetermined</td>
</tr>
<tr>
<td>Slot Check</td>
<td>20.1%</td>
<td>Human Mistake</td>
</tr>
<tr>
<td>Replacing Cable</td>
<td>13.9%</td>
<td>Faulty Cable</td>
</tr>
<tr>
<td>Replacing SSD</td>
<td>31.2%</td>
<td>Failed Device</td>
</tr>
</tbody>
</table>
L&A for System Admins: Part II

• Over 20% of SSD-related OS-level error events are caused by incorrect manual operations
  • “Wrong Slot” is a dominant case: an SSD is plugged into an incorrect slot.
Our Solution

• OIOP: One Interface One Purpose
  • Different SSD interfaces: M.2/U.2 besides SATA
  • E.g., in a hybrid setup with multiple SSDs, the system drive uses the M.2 interface, while storage SSDs still use the SATA interface

https://www.avadirect.com/blog/m-2-vs-u-2-vs-sata-express/
Conclusions & Future Work

• A systematic view of RASR failures in three perspectives
  • Hardware Architecture
    • Suboptimal intra-node and inter-node stacking can lead to passive heating
    • Two possible solutions for passive heating
  • Software Design
    • 15-20% of SSDs are overly used under block storage service
    • Mitigated by shared log structure
  • System Administration
    • Leveraging UCRC Errors for failure root diagnosis
    • OIOP for Wrong Slot Failure

• Next steps
  • Predicting device errors or system failures
  • Related Researches on NVMe SSDs.
Thank You!

Q&A

Erci Xu  
Ohio State University

Mai Zheng  
Iowa State University

Feng Qin  
Ohio State University

Yikang Xu  
Aliyun  
Alibaba

Jiesheng Wu  
Aliyun  
Alibaba