Understanding SSD Reliability in Large-Scale Cloud Systems

Erci Xu
Ohio State University

Mai Zheng
Iowa State University

Feng Qin
Ohio State University

Yikang Xu
Aliyun

Jiesheng Wu
Aliyun

Flash-Memory-Based Solid-Stata Drives is popular in today’s DC

- Enterprises and Individuals are increasingly use SSD as storage media

- 2018 Q1: 45.46 million units

- Estimated Enterprise SSDs: 30 million units in 2018, up from 20 million units in 2016

Estimate of Shipments of hard and solid state disk (HDD/SSD) drives worldwide
SSD Reliability Concerns

- Wear out issue
  - Limited Life Cycles

- Complex failure modes
  - Program/Erase Error

- Sensitive to environment
  - NAND in heated environment
Previous Large Scale SSD Studies

• Reveal important aspects of flash-based drives in the field
  • Life Curve
    • Not Bathtub
  • Uncorrectable Errors are popular
    • Program/Erase Error
  • NAND in heated environment

SSD Reliability in the Cloud Systems are more than the device
Outline of Our Talk

• 1. Architecture Overview

• 2. Human Errors

• 3. Service Imbalance

• 4. Transmission Errors

• 5. Conclusions & Future Works
Architecture Overview

- **Service**
  - Block Storage
  - NoSQL Table Storage
  - Big Data Analytics

- **Cluster Level (Distributed File System)**
  - Chunk Master Logs
  - Chunk Server Logs

- **Node Level**
  - Operating System Logs
  - System Monitoring Logs

- **Device Level (SSD)**
  - Self-Monitoring, Analysis, and Reporting Technology (SMART)
SSD Fleet in Our Research

Targeting different cloud products to achieve general findings

Targeting representative services with different functionalities and avoiding biased results from different workloads
Datasets in Our Research: Errors and Failures

<table>
<thead>
<tr>
<th>Level</th>
<th>Event</th>
<th>Definition</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>Buffer IO Error</td>
<td>A failed read/write from file system to SSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Media Error</td>
<td>Software detected actual data corruption</td>
<td></td>
</tr>
<tr>
<td></td>
<td>File System Unmountable</td>
<td>Unable to load the file system on a SSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drive Missing</td>
<td>OS unable to find a plugged SSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong Slot</td>
<td>SSD has been plugged to the Wrong SATA slot</td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>Host Read</td>
<td>Total amount of LBA read from the SSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Host Write</td>
<td>Total amount of LBA write from the SSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Program Error</td>
<td>Total # of errors in NAND write operations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raw Bit Error Rate</td>
<td>Total bits corrupted divided by total bits read</td>
<td>Daily</td>
</tr>
<tr>
<td></td>
<td>End-to-End Error</td>
<td>Total # of parity check failures between interfaces</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uncorrectable Error</td>
<td>Total # of data corruption beyond ECC’s ability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UDMA CRC Error</td>
<td>Total # of CRC check failures during Ultra-DMA(UDMA)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: **Events Collected in the Target Storage System.** *Freq.*: Frequency, event logs can be updated daily (“Daily”) or upon new events (“Event”):

**Including Node level failure to build correlation and understand impact**
Outline of Our Talk

1. Architecture Overview
2. Human Errors
3. Service Imbalance
4. Transmission Errors
5. Conclusions & Future Works
Human Errors

• Observation: Over 20% of OS-level error events are caused by incorrect manual operations

• Cause: Wrong Slot is a dominant cause. A SSD is plugged into an incorrect slot.
Wrong Slot

• Root cause: the device is mapped to the SATA slot in 1-to-1 fashion

• Possible solutions: adding an indirect translation layer between file system mounting points and hardware slots

• Drawbacks:
  1. Global scale changes
  2. Kernel Level Modifications
Wrong Slot: Our Solution

• OIOP: One Interface One Purpose

• New interfaces, such as M.2 and U.2

• We switch the mapping from mounting-to-SATA to mounting-to-interface

• Example: In the hybrid setup, the system drive is plugged to the M.2 interface while storage SSDs still use the SATA interface
Outline of Our Talk

• 1. Architecture Overview

• 2. Human Errors

• 3. Service Imbalance

• 4. Transmission Errors

• 5. Conclusions & Future Works
Service Imbalance

• Observation: Certain cloud services may cause unbalanced usage of SSDs

<table>
<thead>
<tr>
<th></th>
<th>Host Read</th>
<th>Host Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Value/Hour</td>
<td>Block</td>
<td>7.69GB</td>
</tr>
<tr>
<td></td>
<td>BigData</td>
<td>1.57GB</td>
</tr>
<tr>
<td></td>
<td>NoSQL</td>
<td>6.10GB</td>
</tr>
<tr>
<td>CV Block</td>
<td></td>
<td>35.5%</td>
</tr>
<tr>
<td>CV BigData</td>
<td></td>
<td>1.8%</td>
</tr>
<tr>
<td>CV NoSQL</td>
<td></td>
<td>3.2%</td>
</tr>
</tbody>
</table>

Table 5: Host Read and Host Write Comparison between Services. CV: Coefficient of Variance, the ratio of standard deviation to mean.

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

- Each dot in the line equals the cumulative count of SSDs that have hourly host read (or write) amount falls into a range along the X axis, with a step of 0.5GB/hr and starting from 0.5.
- The majority of SSDs under both NoSQL and Big Data Analytics services have similar values (i.e., one major spike in the corresponding curve).
- On the other hand, the SSDs under the block storage service shows diverse values (i.e., two spikes far apart) as marked in the figure. The distribution of host write is similar.
Service Imbalance: Further Breakdown

- Pangu 1.0 Data Layout for Block storage
  - In-place updates

- User can generate imbalance of workload to SSDs

- 15%-20% of SSDs have higher workload
Outline of Our Talk

1. Architecture Overview

2. Human Errors

3. Service Imbalance

4. Transmission Errors

5. Conclusions & Future Works
Transmission Error occurs when data fails to pass the CRC checking after SSD-to-Host transmission and would trigger an automatic retry.
Transmission Error: CRC Errors by SSD Model

Observation: CRC errors are concentrated around 6% of all drives
Transmission Error: CRC Errors with Age

- Observation: errors are stably generated
- The biased distribution and stable generation indicate there can be non-transient factors behind the generation of CRC errors
Transmission Errors: CRC Errors vs. Other Errors

- **Results:** SPCC results with device level events
- **Observation:** UCRC errors not correlated with device-level errors. UCRC is an independent type of errors
Transmission Errors: CRC Errors vs. Failures

• Results: Failure rate results with OS-level error events
• Observation: UCRC error can lead to a higher rate of OS-level error events
Outline of Our Talk

• 1. Architecture Overview

• 2. Human Errors

• 3. Service Imbalance

• 4. Transmission Errors

• 5. Conclusions & Future Works
Field Study Summary

• SSD failure is more than just data errors

• Human Error
  • Issue: Wrong Slot Plugging
  • Root Cause: 1-to-1 fashion mapping

• Service Imbalance
  • Issue: 15-20% of SSDs have higher workload
  • Root cause: in-place update

• Impact of Transmission Errors
  • Issue: UCRC error is not necessarily benign in the long run
  • Root cause: UCRC error is a useful indicator for faulty interconnection
Thank You!

Q&A

Understanding SSD Reliability in Large-Scale Cloud Systems

Erci Xu
Ohio State University

Mai Zheng
Iowa State University

Feng Qin
Ohio State University

Yikang Xu
Aliyun

Jiesheng Wu
Aliyun Alibaba