What Can We Learn from Four Years of Data Center Hardware Failures?

Guosai Wang, Lifei Zhang, Wei Xu
Motivation: Evolving Failure Model

• Failures in data centers are common and costly
  - Violate service level agreement (SLA) and cause loss of revenue

• Understand failures: reduce TCO

• Today’s data centers are different
  - 😊 Better failure detection systems, experienced operators
  - 😢 Adoption of less-reliable, commodity or custom ordered hardware, more heterogeneous hardware and workload
  - **Result:** more complex failure model

• **Goal:** comprehensive analysis of hardware failures in modern large-scale IDCs
We Re-study Hardware Failures in IDCs

Our work:

- **Large scale**: hundreds of thousands of servers with 290,000 failure operation tickets
- **Long-term**: 2012-2016
- **Multi-dimensional**: components, time, space, product lines, operators’ response, etc.
- Reconfirm or extend previous findings + Observe new patterns
Interesting Findings Overview

**Common beliefs**

- Failures are uniformly randomly distributed over time/space
- Failures happen independently
- HW unreliability shapes the software fault tolerance design

**Our findings**

- HW failures are not uniformly random
  - at different time scales
  - sometimes at different locations
- Correlated HW failures are common in IDCs
- It is also the other way around: software fault tolerance indulges operators to care less about HW dependability
Failure Management Architecture

- Hardware Monitor System (HMS)
- HMS Agent
- Servers
- Failure Records
- Operators
- Failure Pool
- Operators/Programs
- Log
- Failure Operation Tickets
- Repair Process
Failure Management Architecture

- **HMS agents** detect failures on servers
Failure Management Architecture

- **HMS agents** detect failures on servers
- **HMS** collects failure records, and store them in a failure pool
Failure Management Architecture

- **HMS agents** detect failures on servers
- **HMS** collects failure records, and stores them in a failure pool
- **Operators/programs** generate a FOT for each failure record
Dataset: 290,000+ FOTs

- The failure operation tickets (FOTs) contain many fields

**Diagram Notes**

- **id**, **hostname**, **host idc**, **error device**, **error type**, **error time**, **error position**, **op time**, **error detail**, etc.
Multi-dimensional Analysis on the Dataset

- We study the failures on different dimensions based on different fields of FOTs

![Diagram of multi-dimensional analysis](image)

- **id**, **hostname**, **host idc**, **error device**, **error type**, **error time**, **error position**, **op time**, **error detail**, etc.
Multi-dimensional Analysis on the Dataset

- We study the failures on different dimensions based on different fields of FOTs

**Time:** *error time*

**Product lines:** *hostname*

**Space:** *hostname, host idc*

**Operators’ response:** *error time, op time*

**Components:** *error device*

**id, hostname, host idc, error device, error type, error time, error position, op time, error detail, etc.*
## Failure Percentage Breakdown by Component

<table>
<thead>
<tr>
<th>Device</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard Disk Drive</td>
<td>81.84%</td>
</tr>
<tr>
<td>Miscellaneous*</td>
<td>10.20%</td>
</tr>
<tr>
<td>Memory</td>
<td>3.06%</td>
</tr>
<tr>
<td>Power</td>
<td>1.74%</td>
</tr>
<tr>
<td>RAID card</td>
<td>1.23%</td>
</tr>
<tr>
<td>Flash card</td>
<td>0.67%</td>
</tr>
<tr>
<td>Motherboard</td>
<td>0.57%</td>
</tr>
<tr>
<td>SSD</td>
<td>0.31%</td>
</tr>
<tr>
<td>Fan</td>
<td>0.19%</td>
</tr>
<tr>
<td>HDD backboard</td>
<td>0.14%</td>
</tr>
<tr>
<td>CPU</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

*"Miscellaneous" are manually submitted or uncategorized failures*
• About half of HDD failures are related to *SMART values* or *prediction error count*

### Failure Type Breakdown of HDD

- SMARTFail
- PredictErr
- RaidPdPreErr
- RaidPdFailed
- Missing
- NotReady
- MediumErr
- RaidPdMediaErr
- BadSector
- PendingLBA
- TooMany
- DStatus
- Others

---

**Some HDD SMART value exceeds the threshold**

**SMART = Self Monitoring Analysis and Reporting Technique**

**The prediction error count exceeds the threshold**
Failure Types for Hard Disk Drive

• About half of HDD failures are related to **SMART values** or **prediction error count**

Failure Type Breakdown of HDD:

- SMARTFail
- PredictErr
- RaidPdPreErr
- RaidPdFailed
- Missing
- NotReady
- MediumErr
- RaidPdMediaErr
- BadSector
- PendingLBA
- TooMany
- DStatus
- Others

Some HDD SMART value exceeds the threshold

**SMART** = **Self Monitoring Analysis and Reporting Technique**

The prediction error count exceeds the threshold
Outline

• Dataset overview

➢ Temporal distribution of the failures

• Spatial distribution of the failures

• Correlated failures

• Operators’ response to failures

• Lessons Learned
FR is **NOT** Uniformly Random over Days of the Week

**Hypothesis 1.** The average number of component failures is uniformly random over different days of the week.

- A chi-square test can reject the hypothesis at 0.01 significance level for all component classes.
FR is **NOT** Uniformly Random over Hours of the Day

**Hypothesis 2.** The average number of component failures is uniformly random during each hour of the day.
FR is **NOT** Uniformly Random over Hours of the Day

- **Possible Reasons**
  - High workload results in more failures
  - Human factors
  - Components fail in large batches
FR is **NOT** Uniformly Random over Hours of the Day

- **Possible Reasons**
  - High workload results in more failures
  - Human factors
  - Components fail in large batches
FR is **NOT** Uniformly Random over Hours of the Day

- **Possible Reasons**
  - High workload results in more failures
  - Human factors
  - Components fail in large batches
FR is NOT Uniformly Random over Hours of the Day

- Possible Reasons
  - High workload results in more failures
  - Human factors
  - Components fail in large batches
FR of each Component Changes During its Life Cycle

• Different component classes exhibit different FR patterns.
FR of each Component Changes During its Life Cycle

• Infant mortalities:
FR of each Component Changes During its Life Cycle

• Wear out
Outline

• Dataset overview
• Temporal distribution of the failures
  ➢ Spatial distribution of the failures
• Correlated failures
• Operators’ response to failures
• Lessons Learned
Physical Locations Might Affect the FR Distribution

- **Hypothesis 3.** The failure rate on each rack position is independent of the rack position.

- In general, at 0.05 significance level:
  - can not reject the hypothesis in 40% of the data centers
  - can reject it in the other 60%
FR Can be Affected by the Cooling Design

- FRs are higher at rack position 22 and 35

• Possible reasons
  - Design of IDC cooling and physical structure of the racks
Outline

• Dataset overview
• Temporal distribution of the failures
• Spatial distribution of the failures
  ➢ Correlated failures
• Operators’ response to failures
• Lessons Learned
Correlated Failures are Common

• Correlated failures: *batch failures, correlated component failures, repeating synchronous failures*

• Fact: 200+ HDD failures on each of 22.5% of the days

• Case study
  - Nov. 16th and 17th, 2015
  - 5,000+ servers, or 32% of all the servers of the product line, reporting hard drive *SMARTFail* failures
  - 99% of these failures were detected between 21:00 on the 16th and 3:00 on the 17th.
  - Operators replaced about 1,600, decommissioned the remaining 4000+ out-of-warranty drives
  - Failure reason not clear yet
Causes of Correlated Failures

All the following have happened before 😱
- Environmental factors (e.g., humidity)
- Firmware bugs
- Single point of failure (e.g., power module failures)
- Human operator mistakes
- ...
Outline

• Dataset overview
• Temporal distribution of the failures
• Spatial distribution of the failures
• Correlated failures

Operators’ response to failures

• Lessons Learned
Operators’ Response to Failures

• Response time: $RT = op\_time - err\_time$
**RT is Very High in General**

- RT for *D_fixing*: Avg. 42.2 days, median 6.1 days
- 10% of the FOTs: RT > 140 days

- Is it because operators busy dealing with large number of failures?
  - No!
RT in Different Product Lines Varies

- Observation 1: Variation of RT in different product lines is large
- Observation 2: Operators respond to large number of failure more quickly

Who cares?

The REAL problems 😞
OPs are Less Motivated to Respond to HW Failures

Possible reasons

• Software redundancy design
  - Delayed Responding, process failures in batches

• Many hardware failures are no longer urgent
  - E.g., SMART failures may not be fatal

• Repair operation can be costly
  - E.g., Task migration
Outline

• Dataset overview
• Temporal distribution of the failures
• Spatial distribution of the failures
• Correlated failures
• Operators’ response to failures

Lessons Learned
Lessons Learned I

• Much old wisdom still holds.
  - More correlated failures $\Rightarrow$ software design challenge
  - Automatic hardware failure detection & handling: 😊
  - Data center design: avoid “bat spot”
Lessons Learned II

• Strike the right balance among software stack complexity, hardware dependability, and operation cost.

• Data center dependability needs joint optimization effort that crosses layers.
Lessons Learned III

• *Stateful* failure handling system
  - Data mining tool: discover correlation among failures
  - Provide operators with extra information
Thank you! Q&A

Outline

• Dataset overview
• Temporal distribution of the failures
• Spatial distribution of the failures
• Correlated failures
• Operators’ response to failures
• Lessons Learned
Hypothesis 4. *Time between failures (TBF) of all components follows an exponential distribution.*

Hypothesis 5. *TBF of each individual component class follows an exponential distribution.*

Large proportion of small values
Failure Operation Ticket (FOT)

• Categories of FOTs

<table>
<thead>
<tr>
<th>Failure trace</th>
<th>Handling decision</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{\text{fixing}}$</td>
<td>Issue a repair order (RO)</td>
<td>70.3%</td>
</tr>
<tr>
<td>$D_{\text{error}}$</td>
<td>Not repair and set to decommission</td>
<td>28.0%</td>
</tr>
<tr>
<td>$D_{\text{falsealarm}}$</td>
<td>Mark as a false alarm</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

• Fields:
  
id, host id, hostname, host idc, error device, error type, error time, error position, error detail
FR of Misc. Failures During the Lifecycle

• Most manual detection and debugging efforts happen only at deployment time
• Less cost to repair (not much tasks to migrate)
RT for Each Component Class

- Median RTs for SSD and mist. failures are the shortest (hours)
- Median RTs for HDD, fans, and memory are the longest (7-18 days)
- Standard deviation of the RT for HDD: 30.2 days
Self-Monitoring, Analysis and Reporting Technology

• Fields: raw value, worst, threshold, status
• SMART attribute examples (failure related)
  • Reallocated Sectors Count
  • End-to-End error
  • Uncorrectable Sector Count
  • Reported Uncorrectable Errors
  • Current Pending Sector Count
  • Command Timeout
  • ...

### Examples of Failure Types

<table>
<thead>
<tr>
<th>Failure type</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMARTFail</td>
<td>Some HDD SMART value exceeds the predefined threshold.</td>
</tr>
<tr>
<td>RaidPdPreErr</td>
<td>The prediction error count exceeds the predefined threshold.</td>
</tr>
<tr>
<td>Missing</td>
<td>Some device file could not be detected.</td>
</tr>
<tr>
<td>NotReady</td>
<td>Some device file could not be accessed.</td>
</tr>
<tr>
<td>PendingLBA</td>
<td>Failures are detected on the sectors that are not accessed.</td>
</tr>
<tr>
<td>TooMany</td>
<td>Large number of failed sectors are detected on the HDD.</td>
</tr>
<tr>
<td>DStatus</td>
<td>IO requests are not handled by the HDD and are in D status.</td>
</tr>
<tr>
<td>BBTFail</td>
<td>The bad block table (BBT) could not be accessed.</td>
</tr>
<tr>
<td>HighMaxBbRate</td>
<td>The max bad block rate exceeds the predefined threshold.</td>
</tr>
<tr>
<td>RaidVdNoBBU</td>
<td>Abnormal cache setting due to BBU (Battery Backup Unit) is detected, which</td>
</tr>
<tr>
<td></td>
<td>degrades the performance.</td>
</tr>
<tr>
<td>CacheErr</td>
<td></td>
</tr>
<tr>
<td>DIMMCE</td>
<td>Large number of correctable errors are detected.</td>
</tr>
<tr>
<td>DIMMUE</td>
<td>Uncorrectable errors are detected on the memory.</td>
</tr>
</tbody>
</table>
Repeating Failures

• Over 85% of the fixed components never repeat the same failure
• Repair can fail
• 2% of servers that ever failed contribute more than 99% of all failures
Batch Failure Frequency for Each Component

- **r_N**: a normalized counter of how many days during the D days, in which more than N failures happen on the same day.
- Normalized by the total time length D.

<table>
<thead>
<tr>
<th>Device</th>
<th>r_{100}(%)</th>
<th>r_{200}(%)</th>
<th>r_{500}(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDD</td>
<td>55.4</td>
<td>22.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>3.7</td>
<td>1.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Power</td>
<td>0.7</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>Memory</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>RAID card</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Flash card</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>Fan</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Motherboard</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SSD</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPU</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>