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

Guosai Wang, Lifei Zhang, Wei Xu

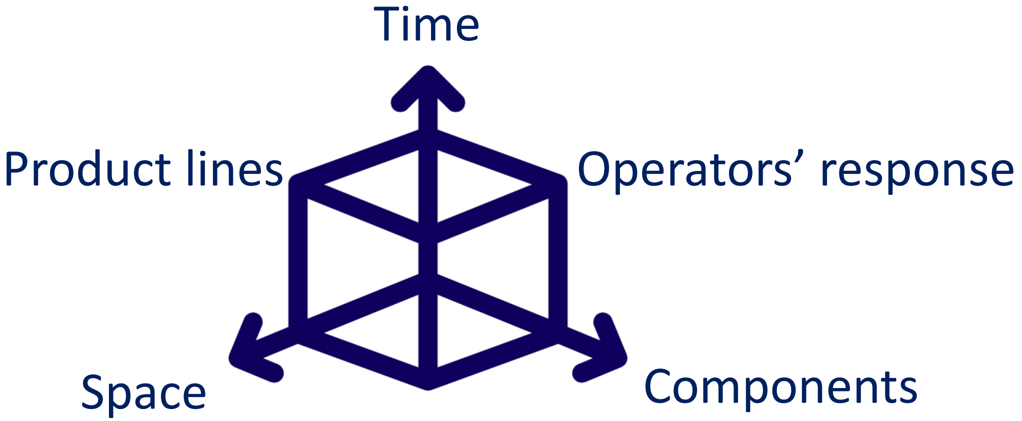
Motivation: Evolving Failure Model

- 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

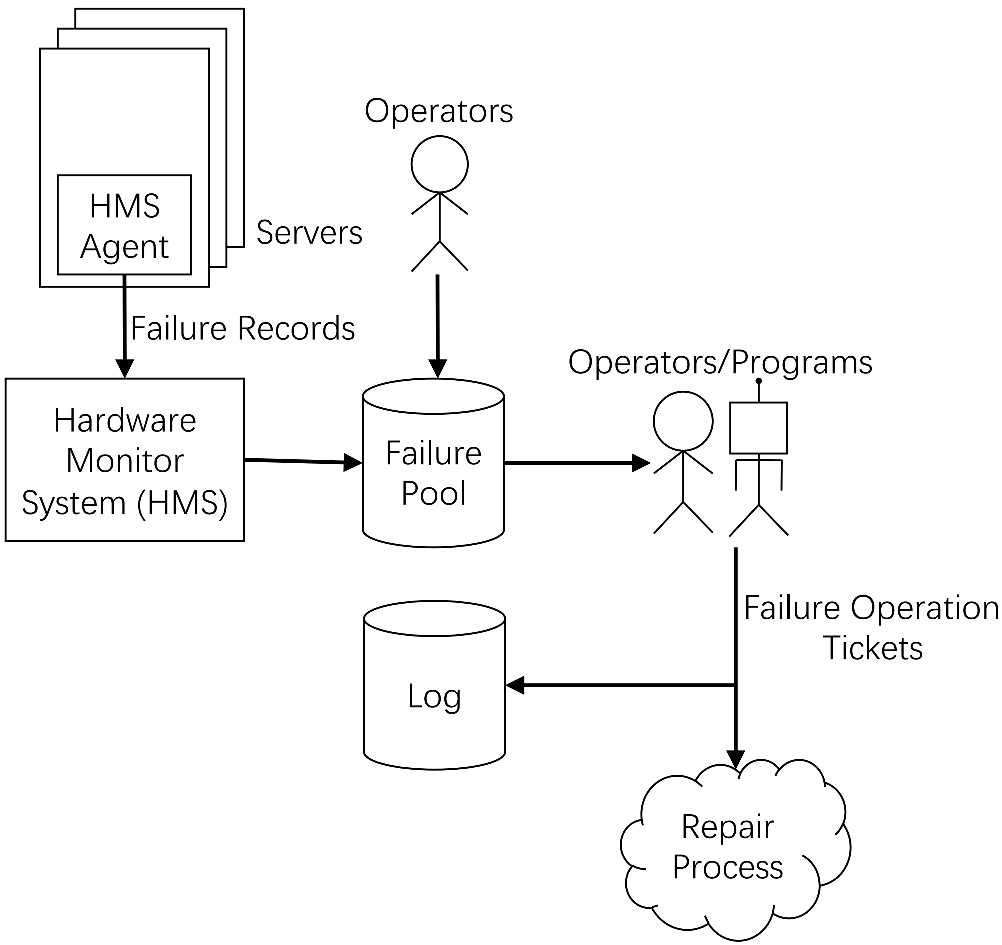
Common beliefs

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

Our findings

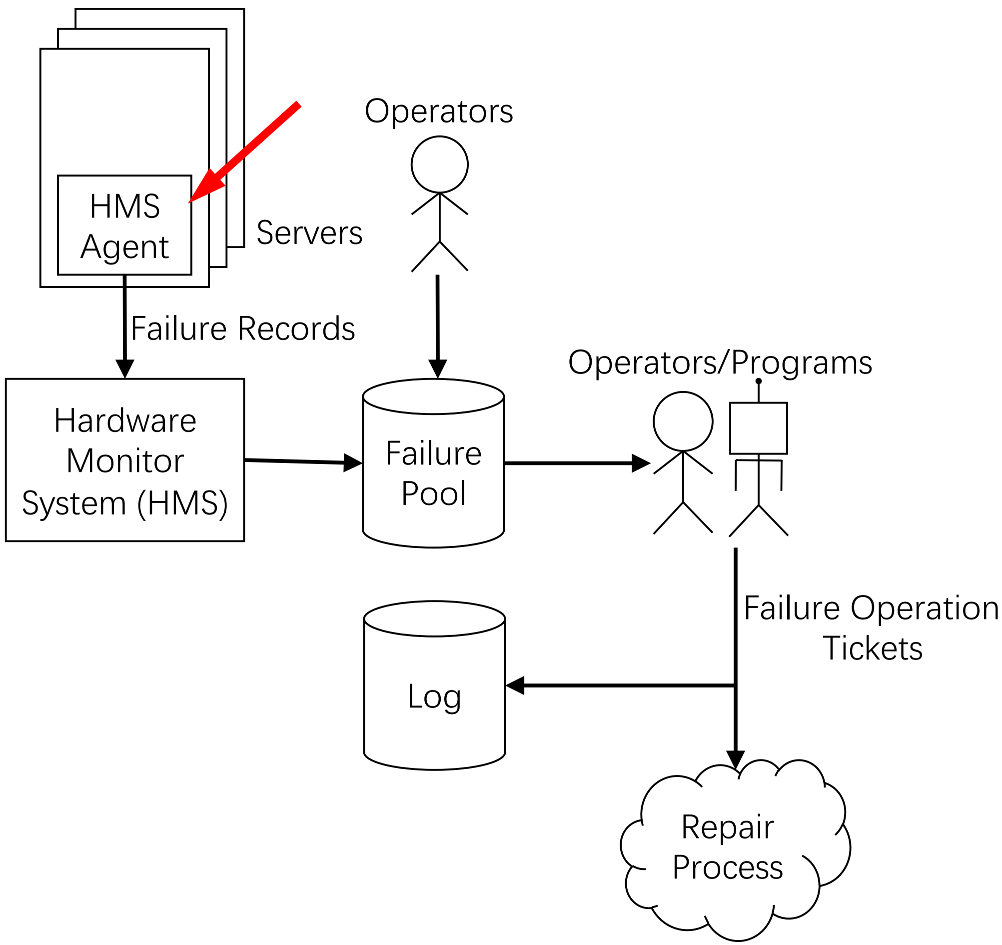
- HW failures are not uniformly random
  - at different time scales
  - sometimes at different locations
- Correlated HW failures are common in IDCs
- It is also the other way around: software fault tolerance indulges operators to care less about HW dependability

Failure Management Architecture



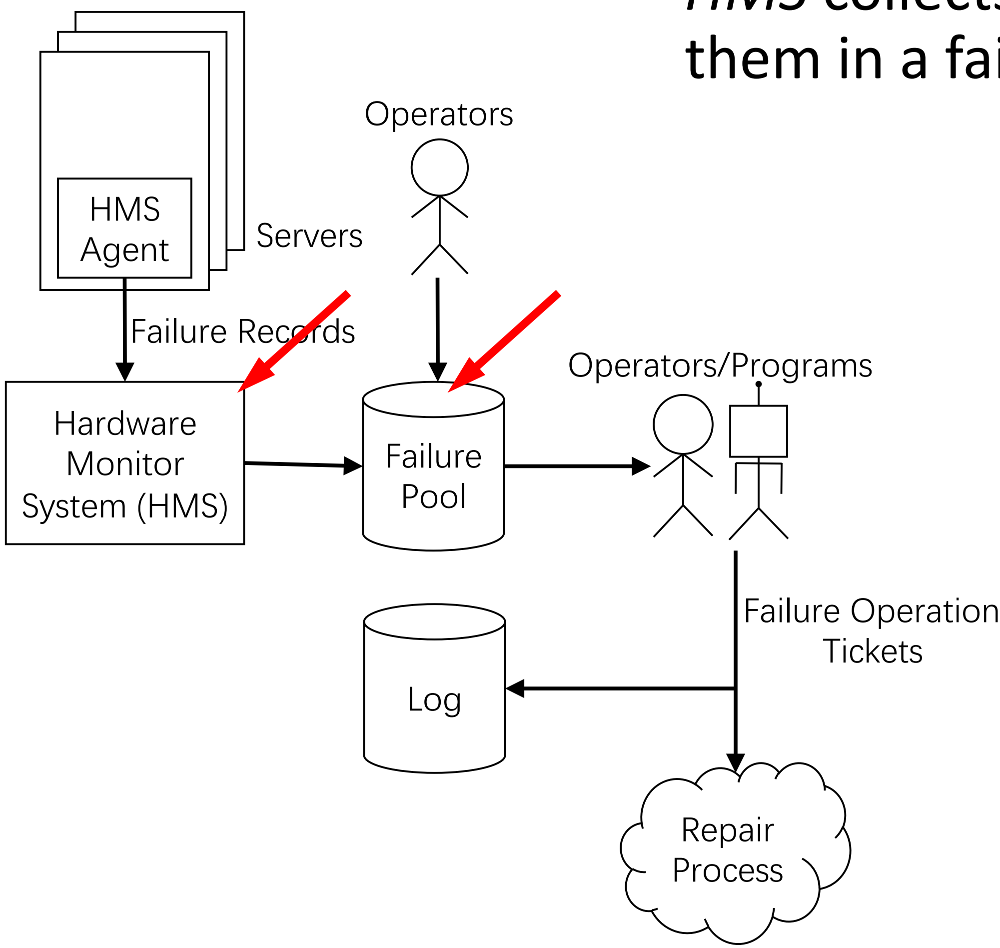
Failure Management Architecture

- *HMS agents* detect failures on servers

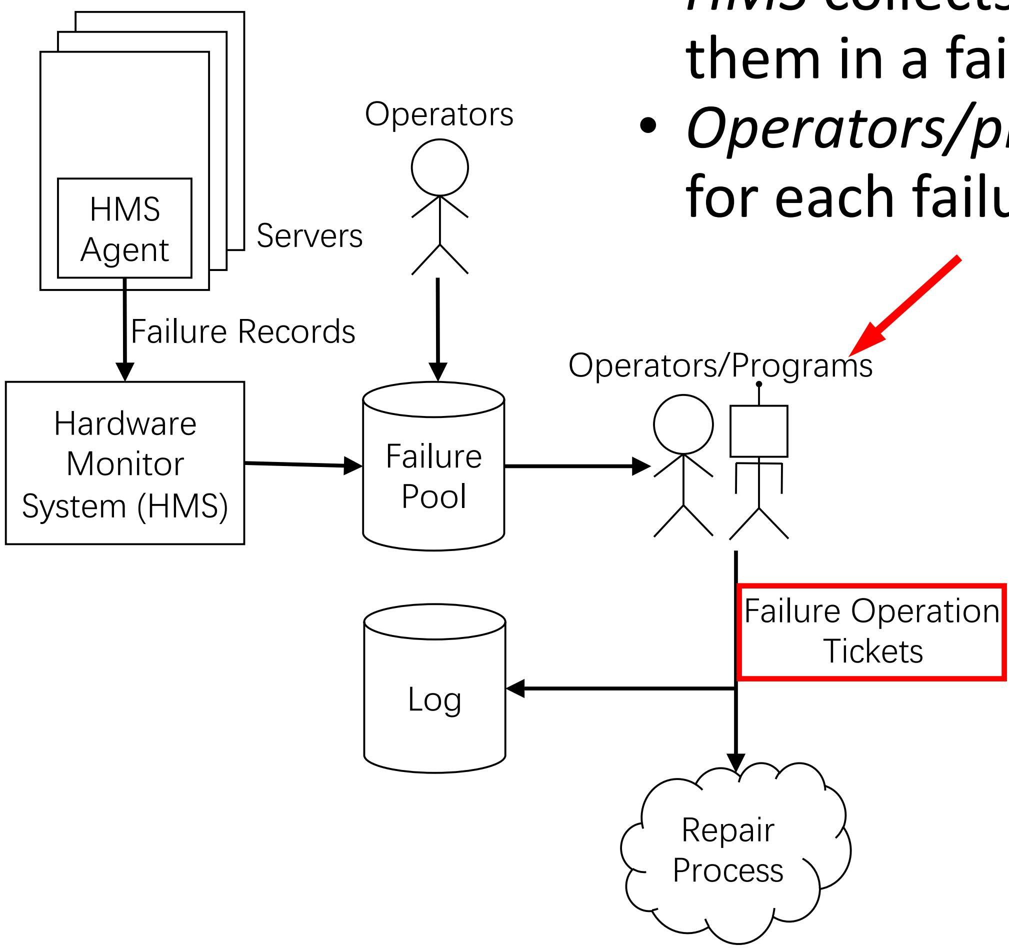


Failure Management Architecture

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



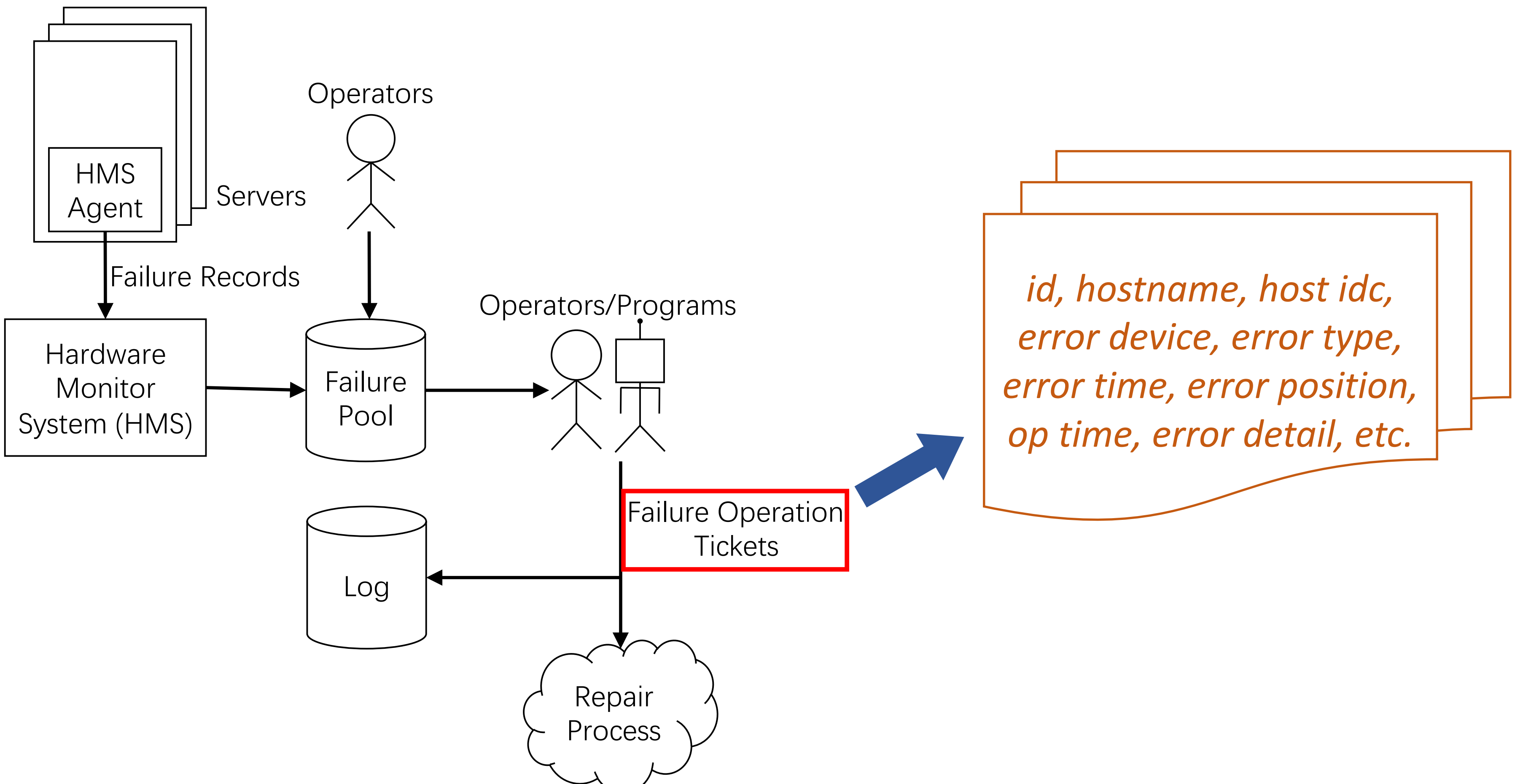
Failure Management Architecture



- *HMS agents* detect failures on servers
- *HMS* collects failure records, and store them in a failure pool
- *Operators/programs* generate a FOT for each failure record

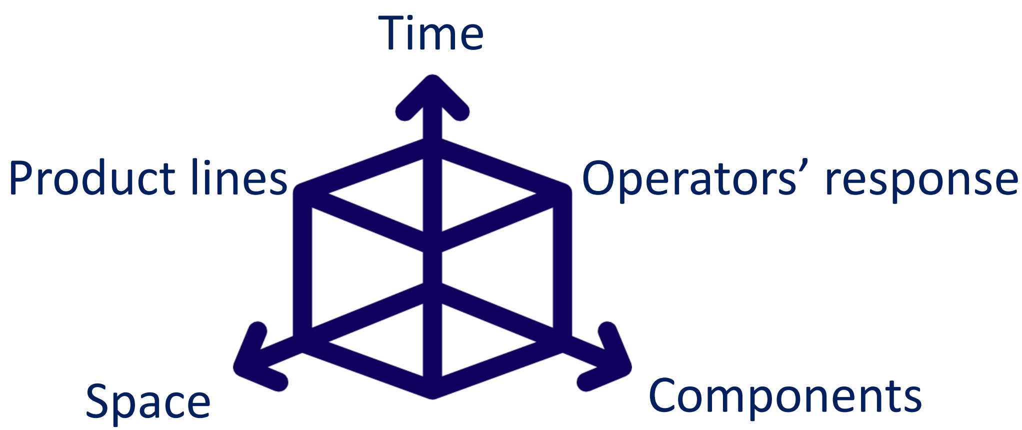
Dataset: 290,000+ FOTs

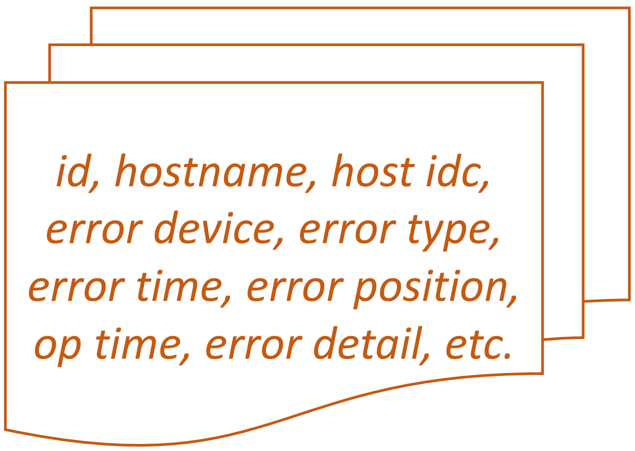
- The failure operation tickets (FOTs) contain many fields



Multi-dimensional Analysis on the Dataset

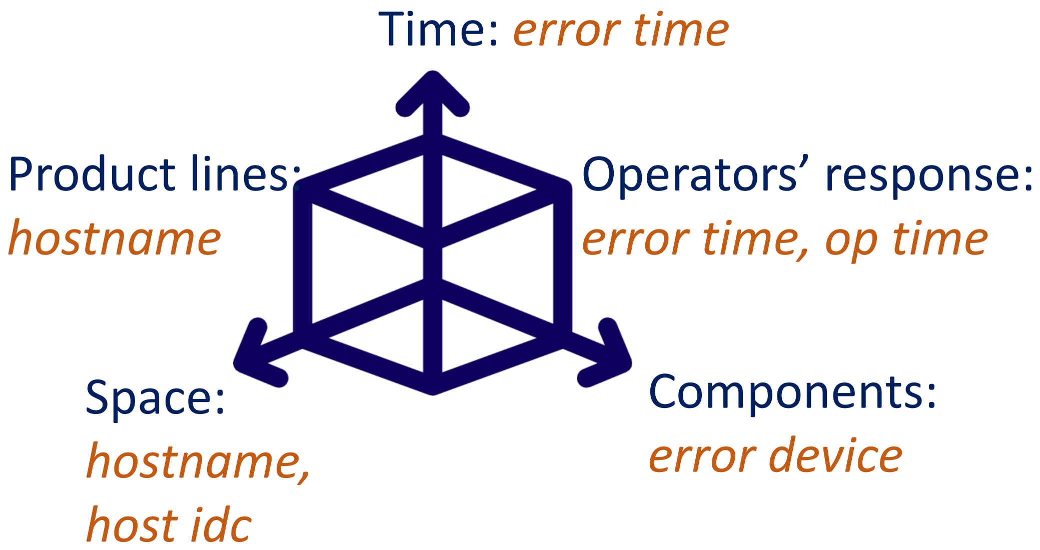
- We study the failures on different dimensions based on different fields of FOTs

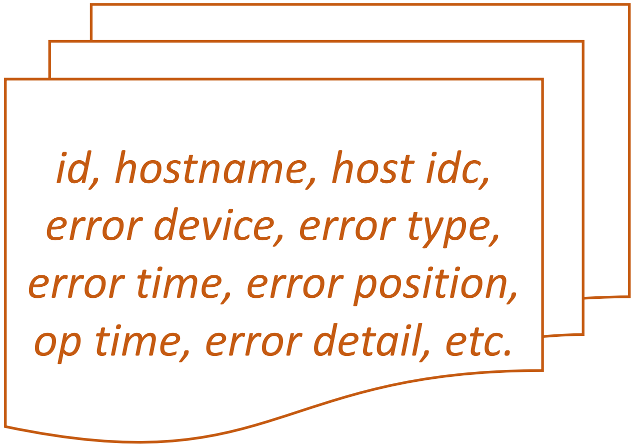




Multi-dimensional Analysis on the Dataset

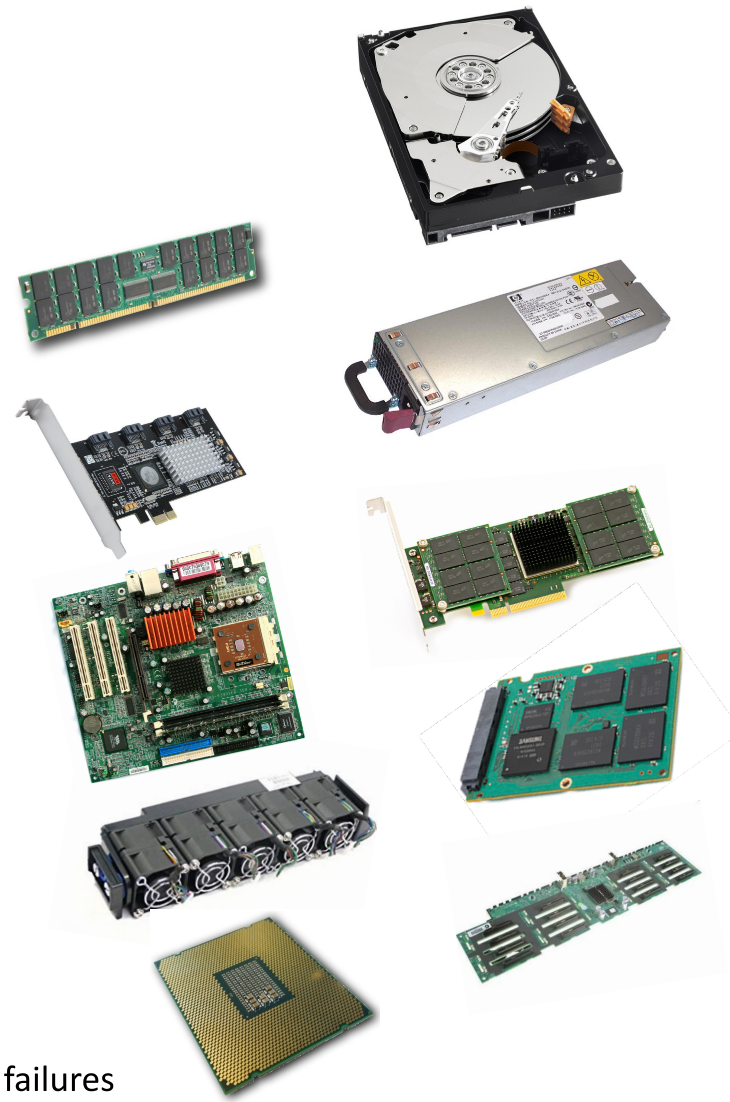
- We study the failures on different dimensions based on different fields of FOTs





Failure Percentage Breakdown by Component

Device	Proportion
Hard Disk Drive	81.84%
Miscellaneous*	10.20%
Memory	3.06%
Power	1.74%
RAID card	1.23%
Flash card	0.67%
Motherboard	0.57%
SSD	0.31%
Fan	0.19%
HDD backboard	0.14%
CPU	0.04%

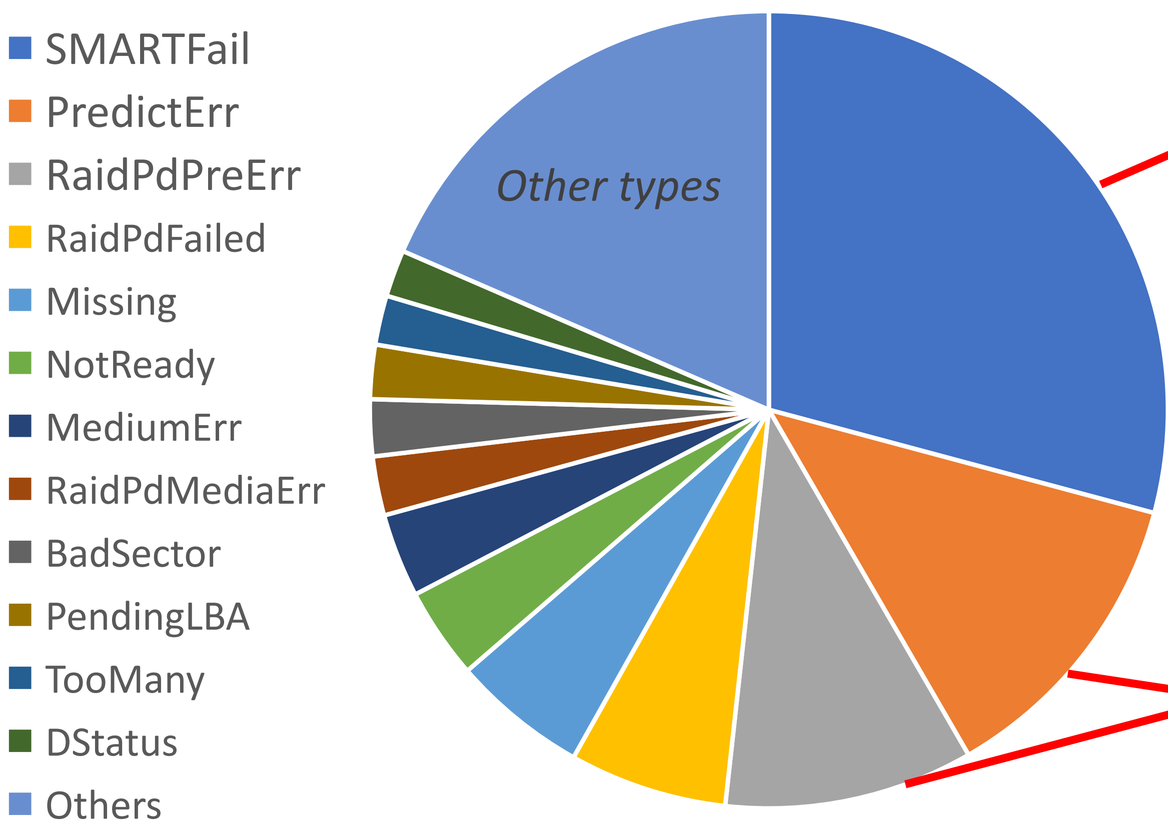


*"*Miscellaneous*" are manually submitted or uncategorized failures

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



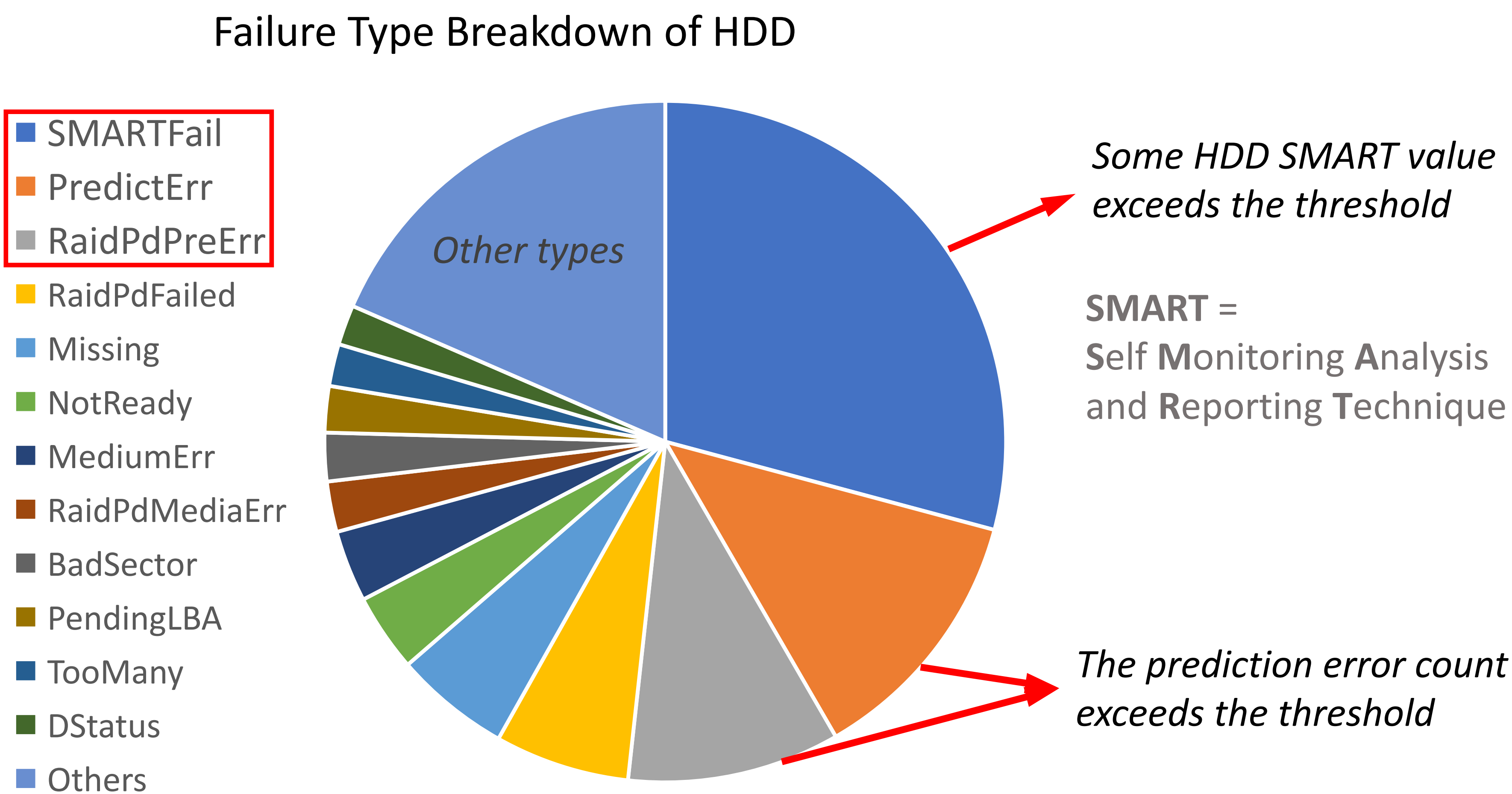
Some HDD SMART value exceeds the threshold

SMART =
Self **M**onitoring **A**nalysis and **R**eporting **T**echnique

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*



SMART =
Self **M**onitoring **A**nalysis and **R**eporting **T**echnique

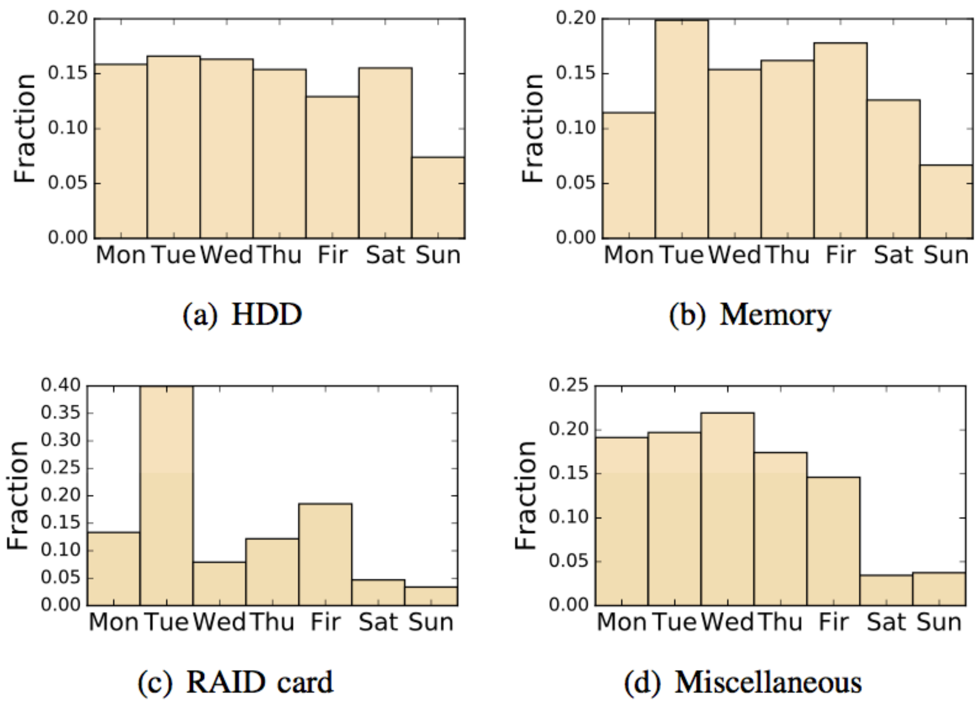
Outline

- 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

REJECTED

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



(a) HDD

(b) Memory

(c) RAID card

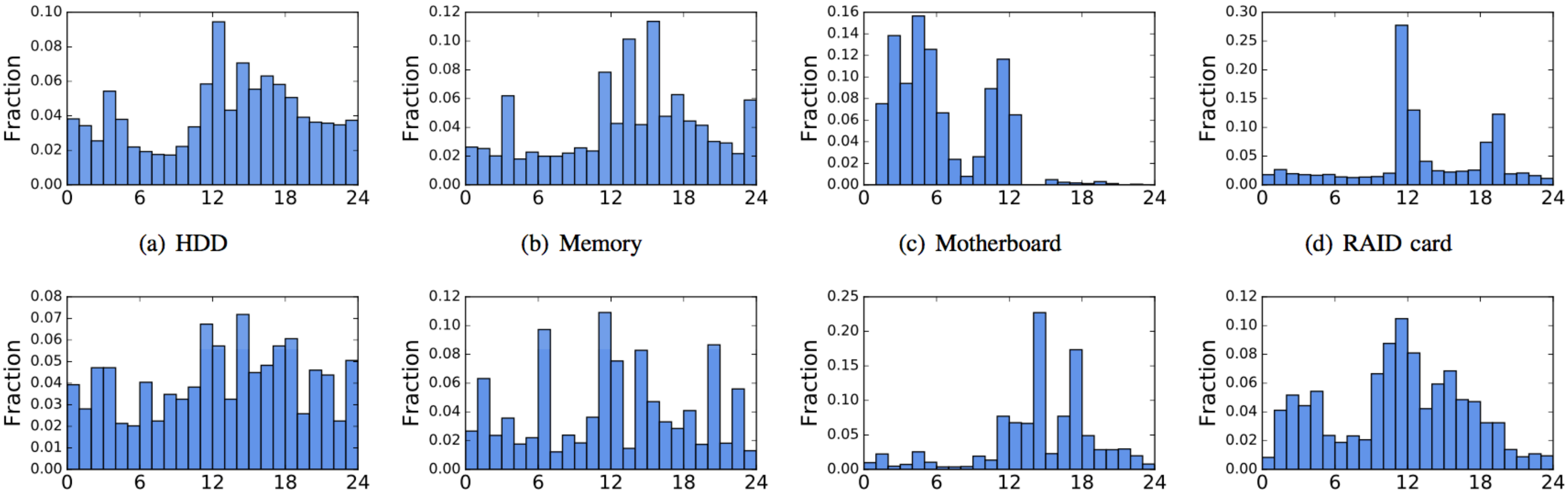
(d) Miscellaneous

- 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

REJECTED

Hypothesis 2. *The average number of component failures is uniformly random during each hour of the day.*

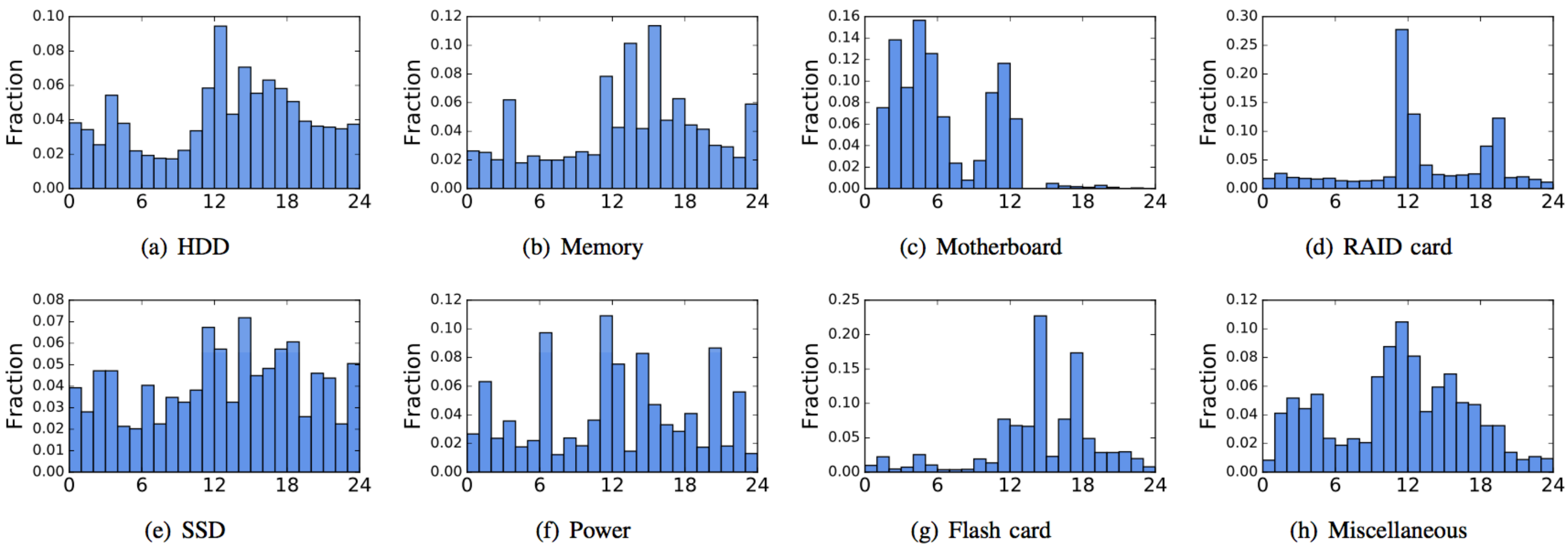


(a) HDD (b) Memory (c) Motherboard (d) RAID card

(e) SSD (f) Power (g) Flash card (h) Miscellaneous

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

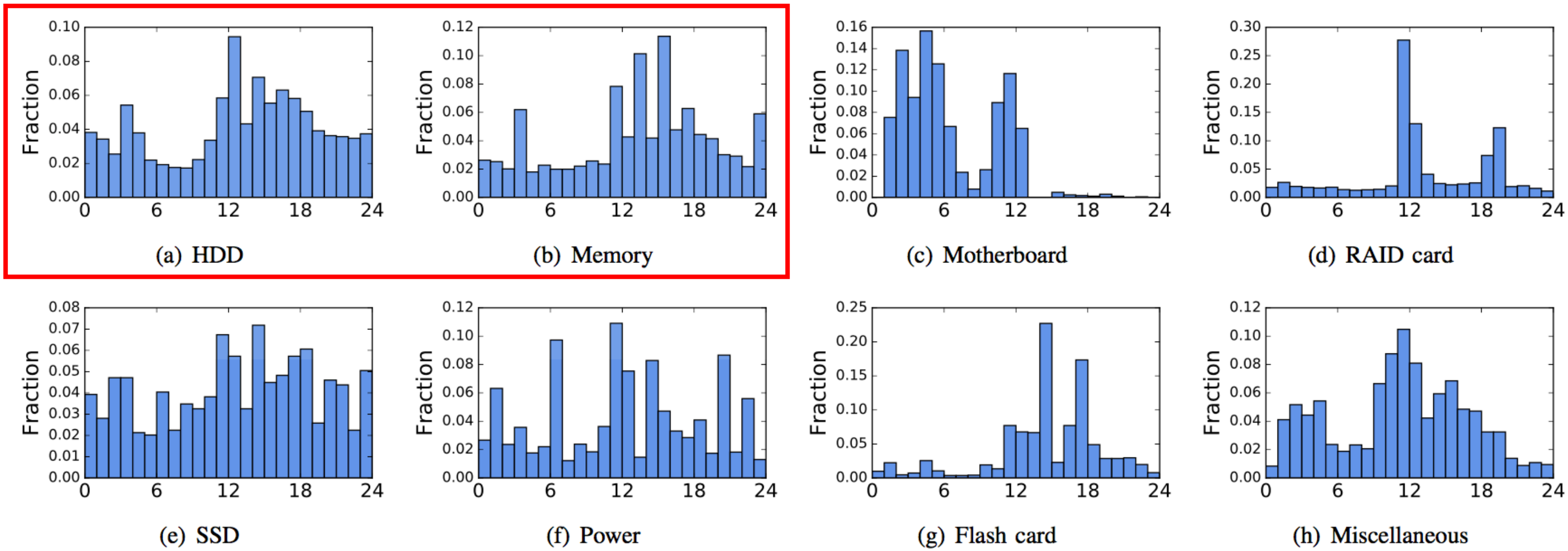


(a) HDD (b) Memory (c) Motherboard (d) RAID card

(e) SSD (f) Power (g) Flash card (h) Miscellaneous

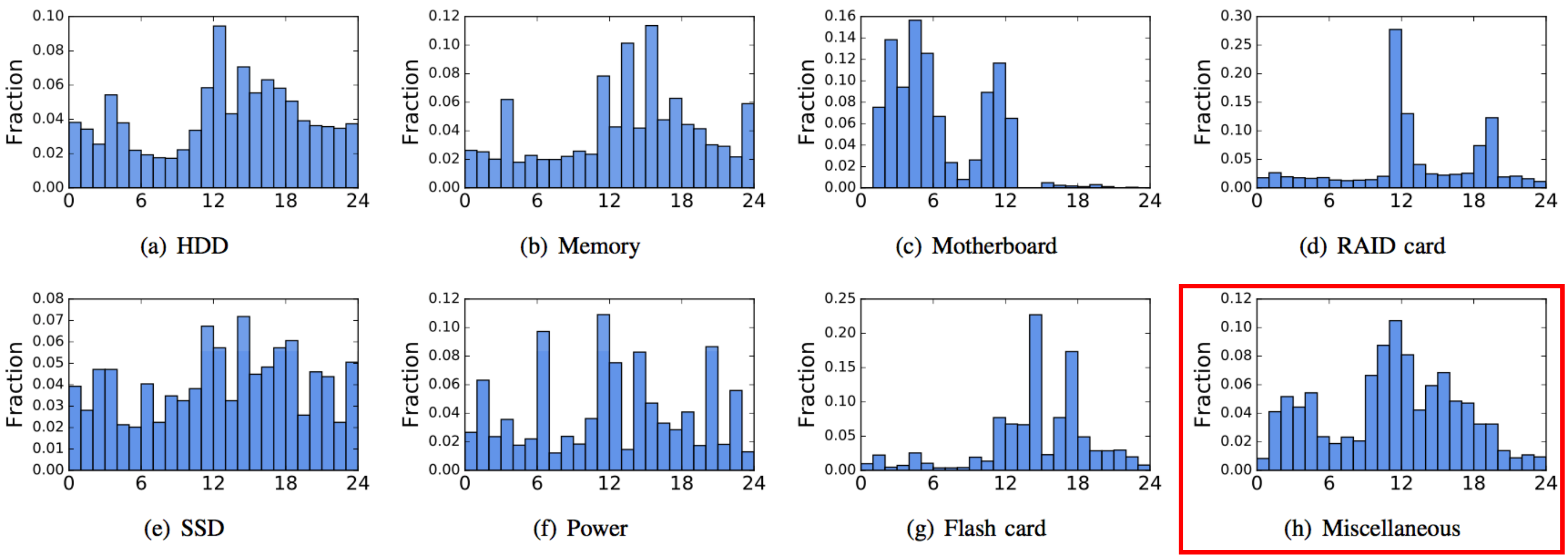
FR is NOT Uniformly Random over Hours of the Day

- 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

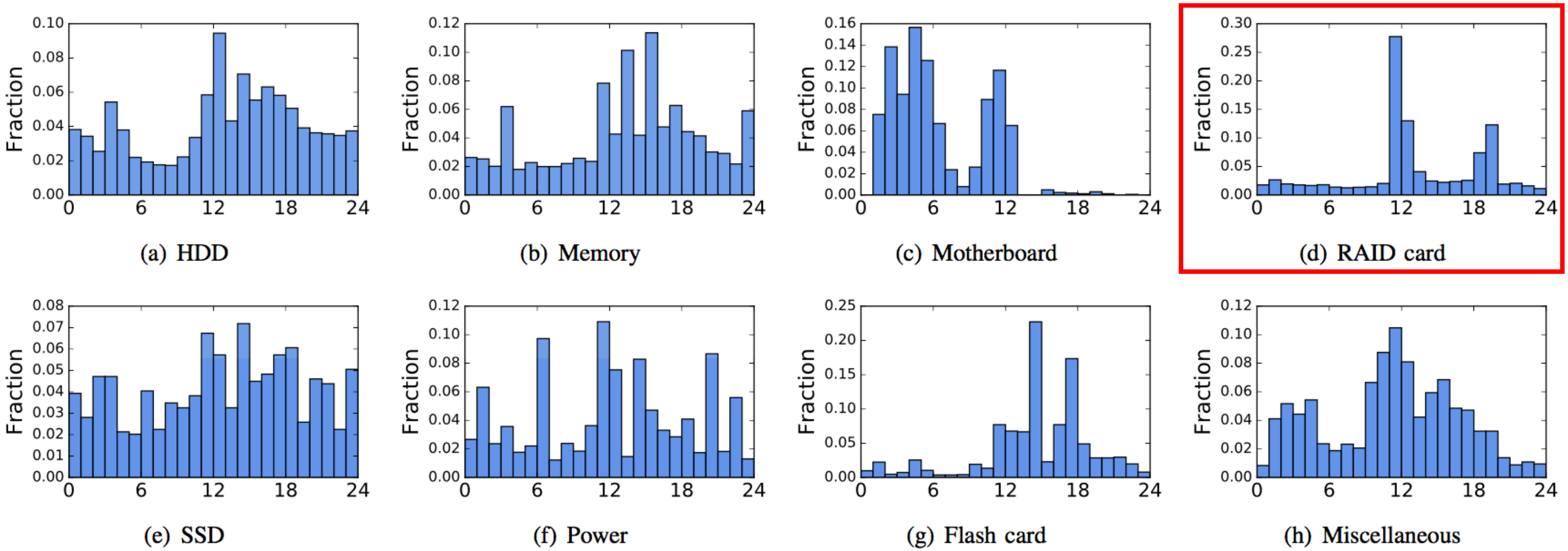


(a) HDD (b) Memory (c) Motherboard (d) RAID card

(e) SSD (f) Power (g) Flash card (h) Miscellaneous

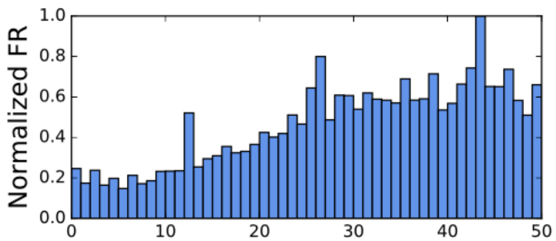
FR is NOT Uniformly Random over Hours of the Day

- Possible Reasons
  - High workload results in more failures
  - Human factors
  - Components fail in large batches

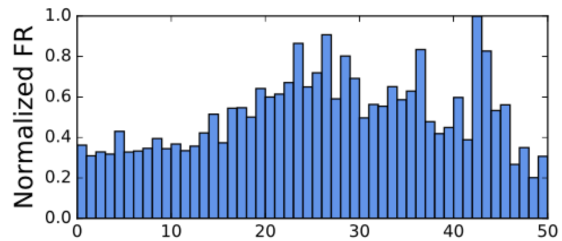


FR of each Component Changes During its Life Cycle

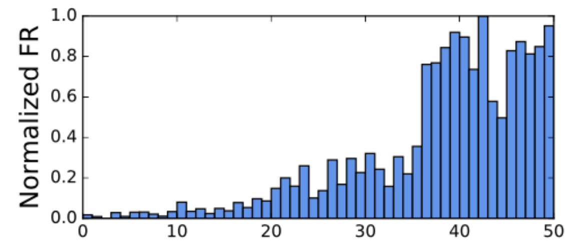
- Different component classes exhibit different FR patterns.



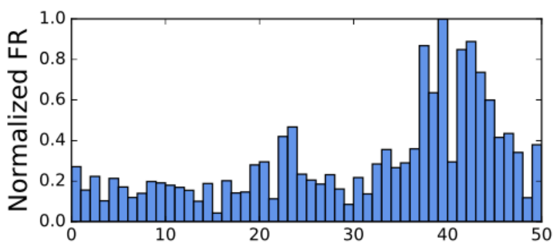
(a) HDD



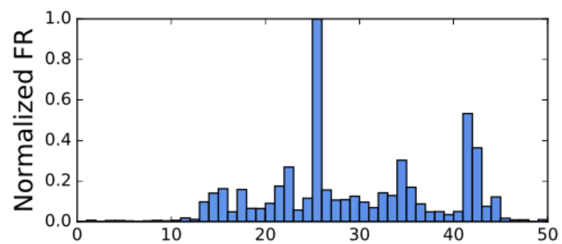
(b) Memory



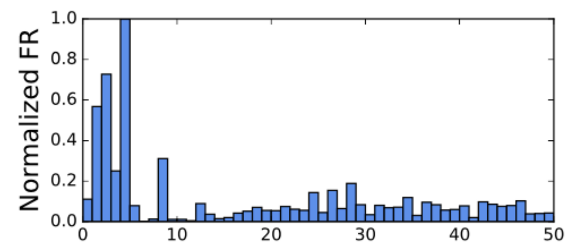
(c) Motherboard



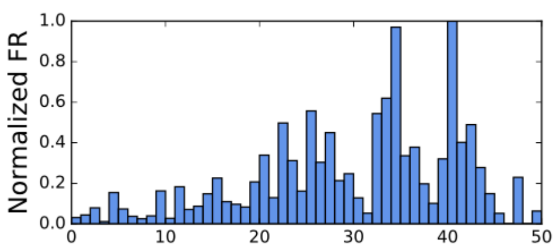
(d) SSD



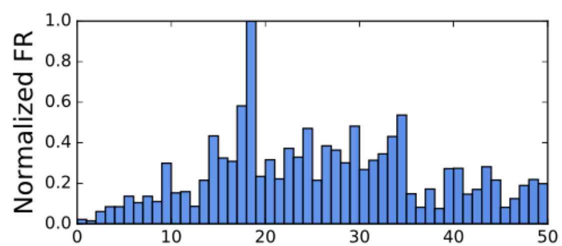
(e) Flash card



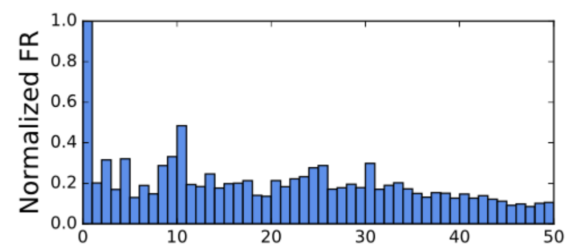
(f) Raid Card



(g) Fan



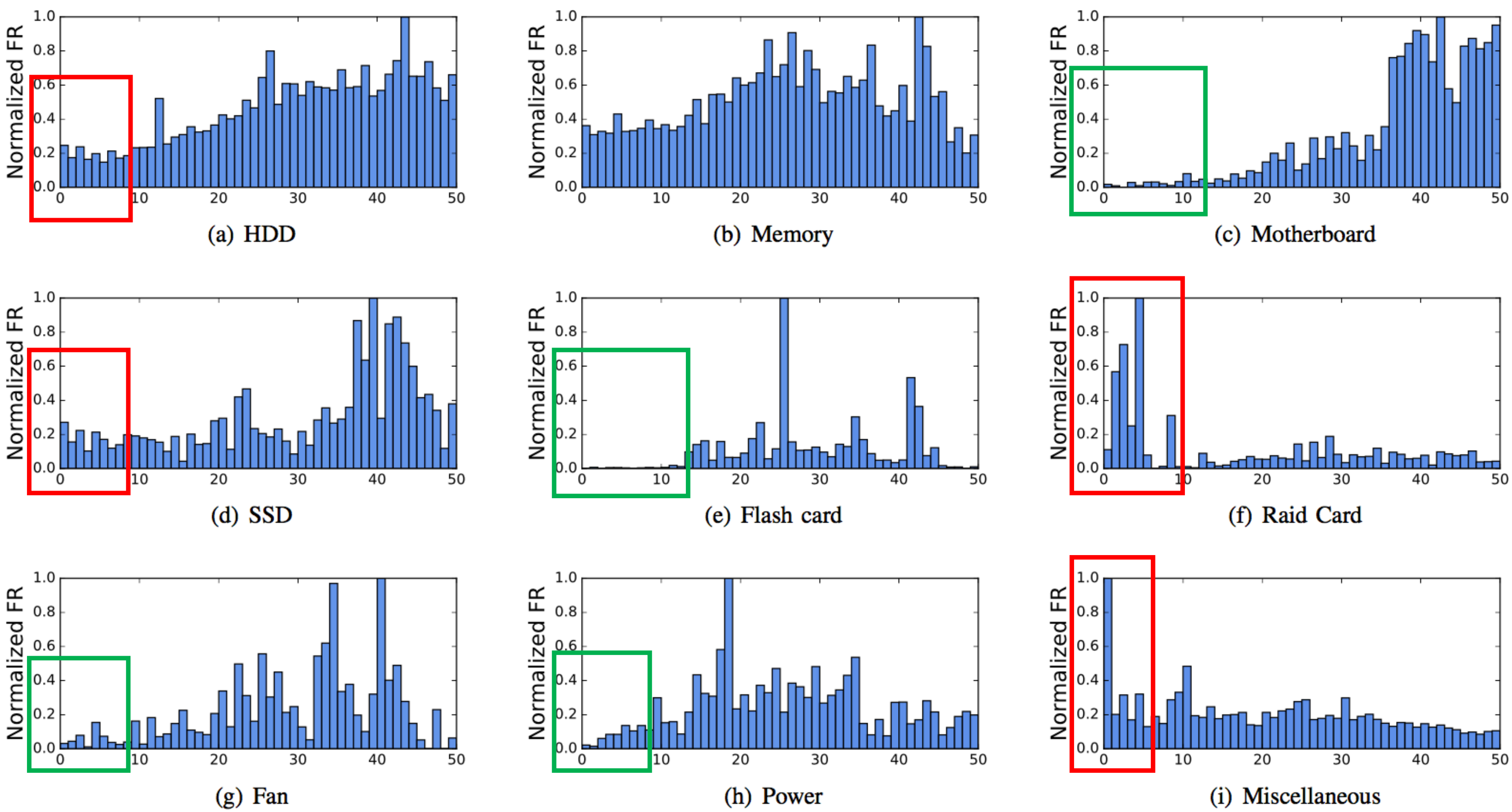
(h) Power



(i) Miscellaneous

FR of each Component Changes During its Life Cycle

- Infant mortalities:



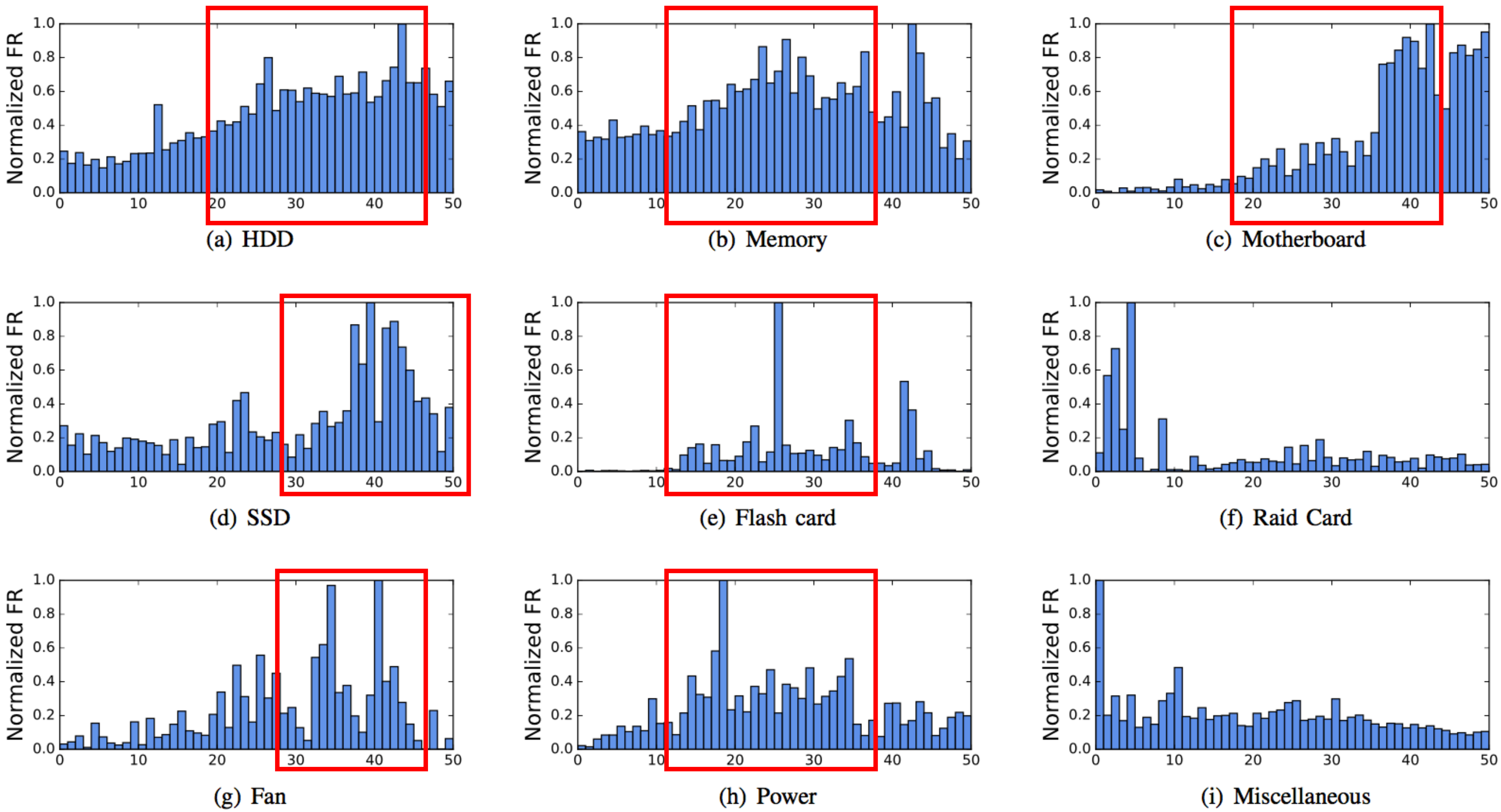
(a) HDD (b) Memory (c) Motherboard

(d) SSD (e) Flash card (f) Raid Card

(g) Fan (h) Power (i) Miscellaneous

FR of each Component Changes During its Life Cycle

- Wear out

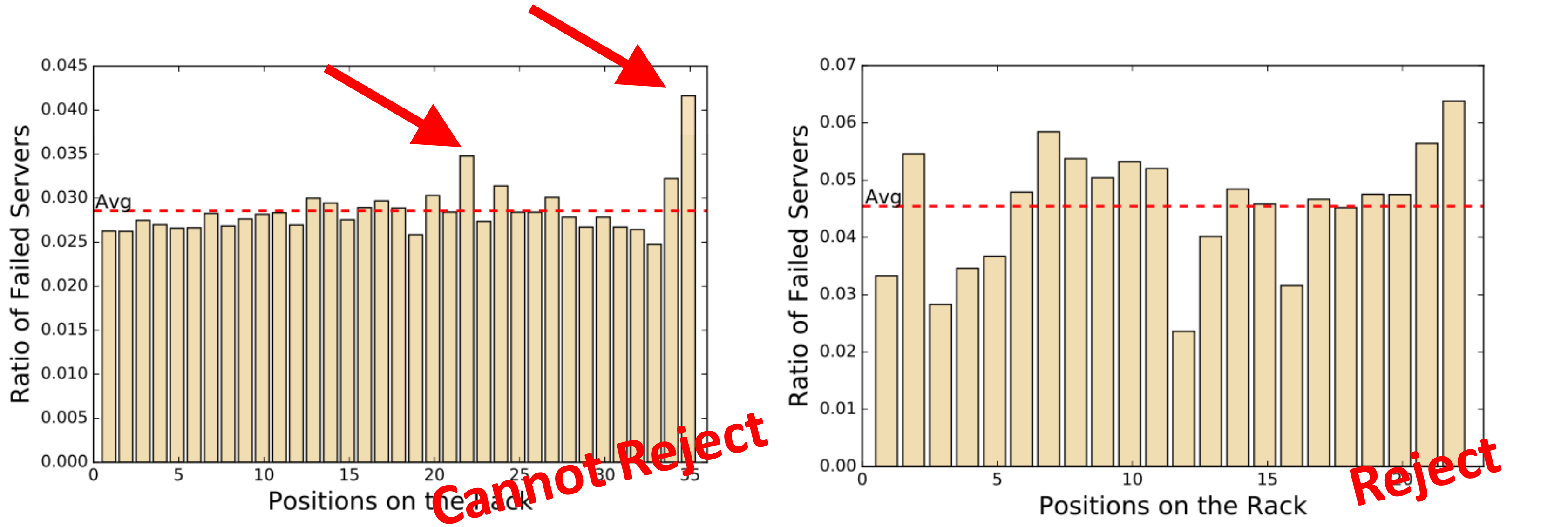


Outline

- 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.*



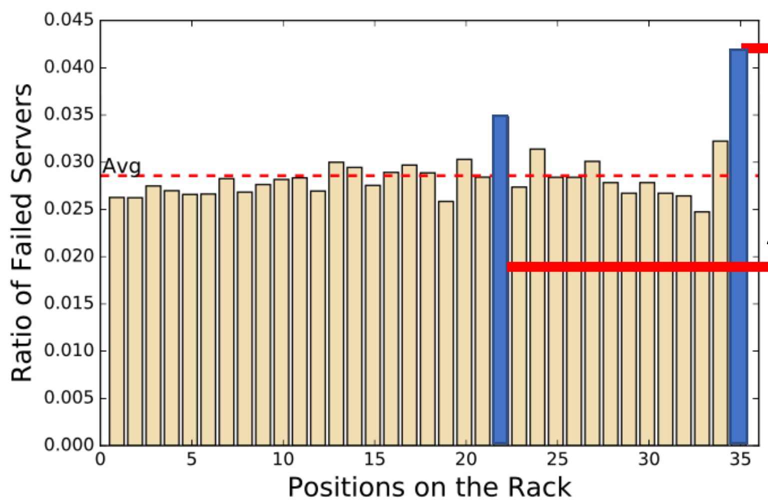
(a) Data center A

(b) Data center B

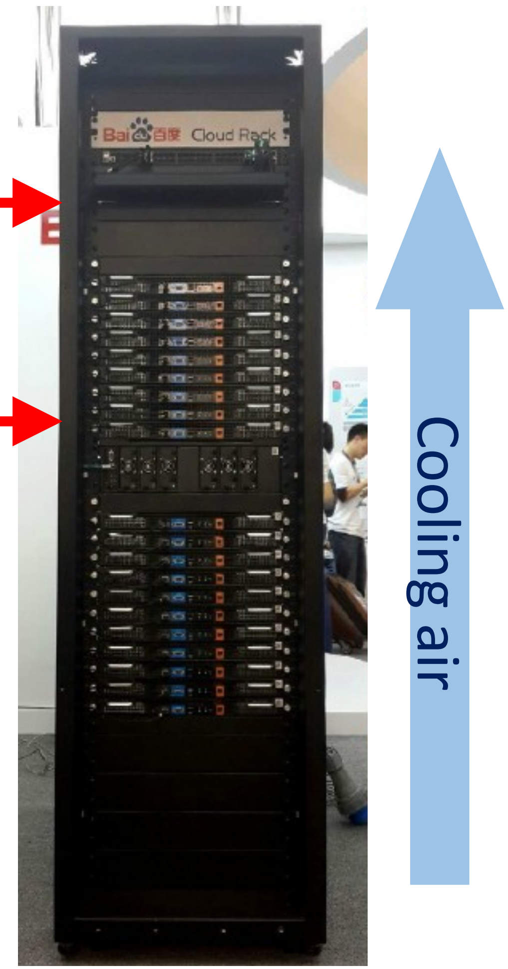
- In general, at 0.05 significance level:
  - can not reject the hypothesis in 40% of the data centers
  - can reject it in the other 60%

FR Can be Affected by the Cooling Design

- FRs are higher at rack position 22 and 35



(a) Data center A



A typical Scorpion rack

- 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

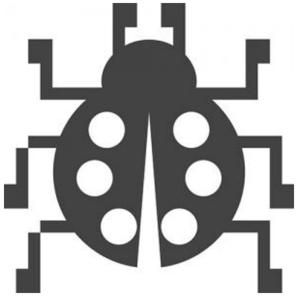
Correlated Failures are Common

- 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
- …

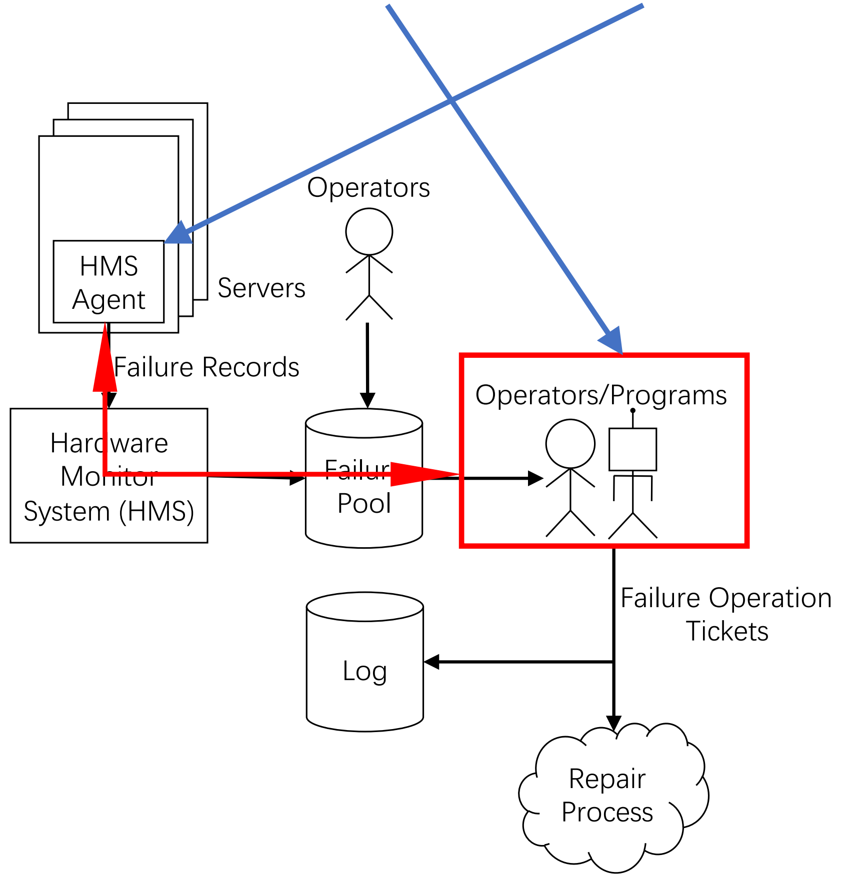


Outline

- Dataset overview
- Temporal distribution of the failures
- Spatial distribution of the failures
- Correlated failures
- ➢ Operators’ response to failures
- Lessons Learned

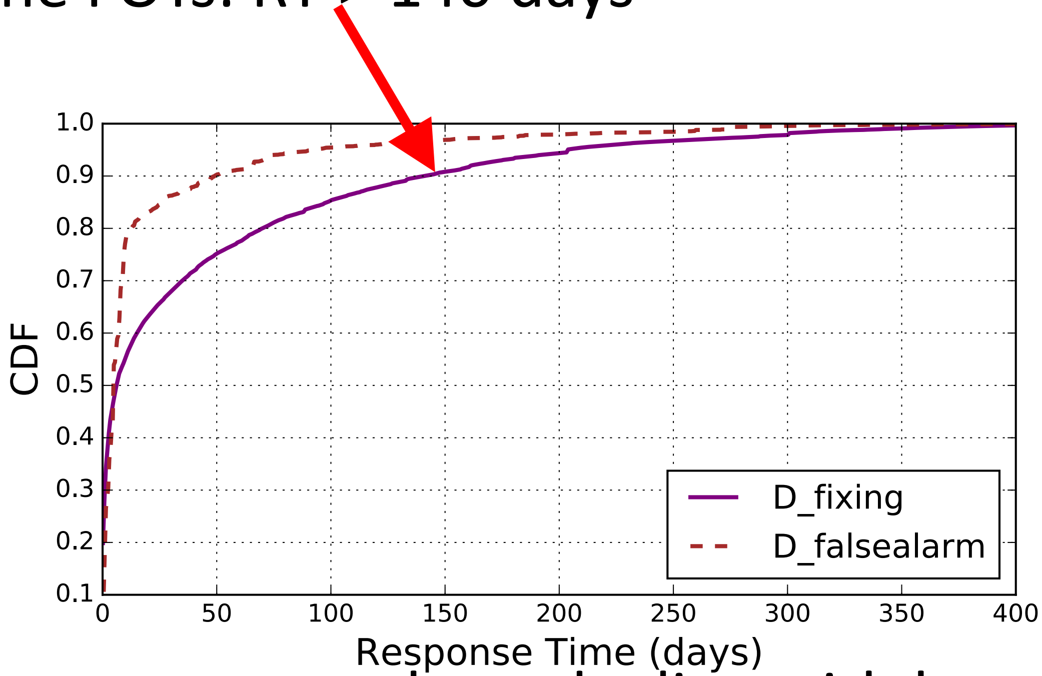
Operators' Response to Failures

- Response time: *RT = op_time – err_time*



RT is Very High in General

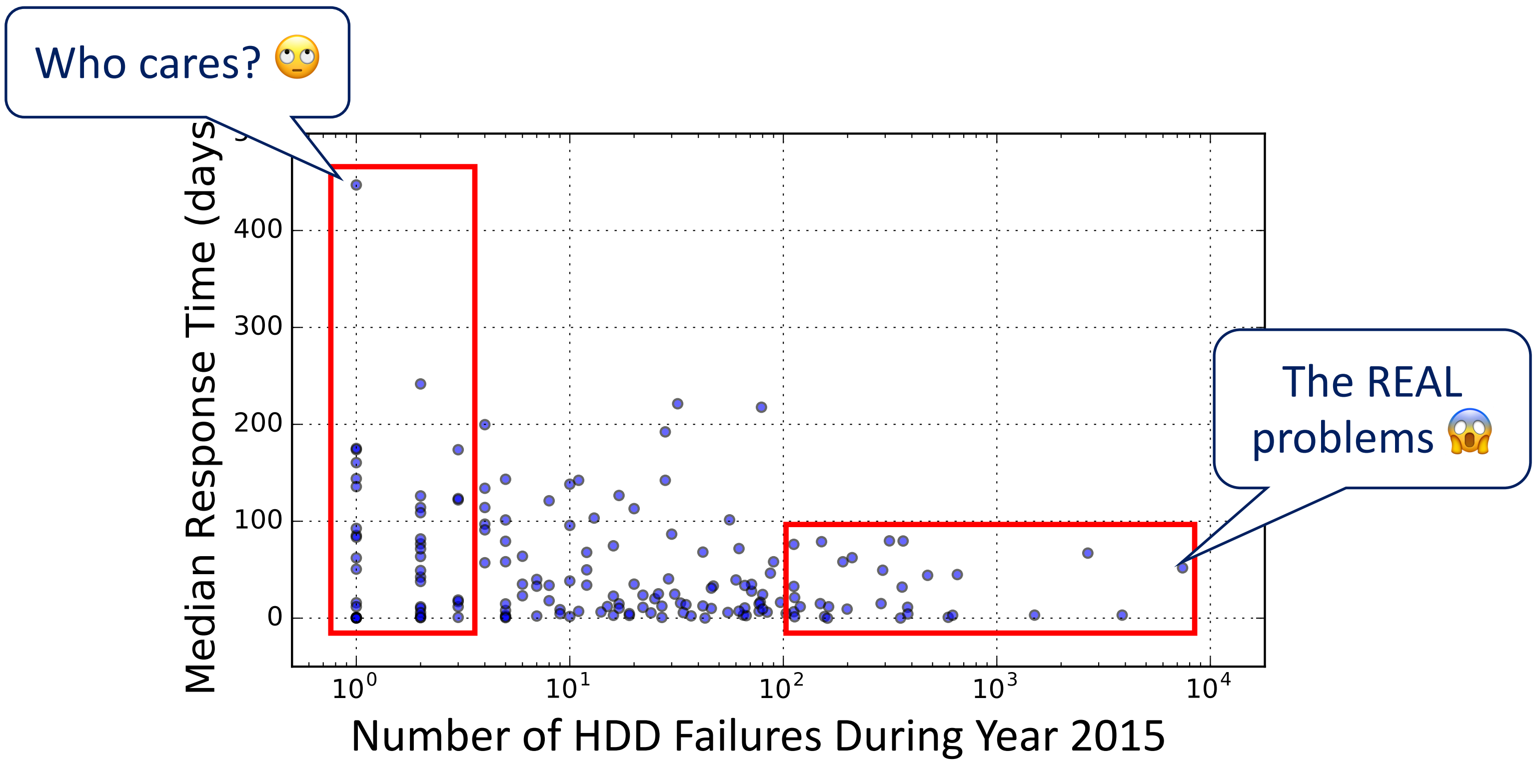
- RT for *D_fixing*: Avg. 42.2 days, median 6.1 days
- 10% of the FOTs: RT > 140 days



- Is it because operators busy dealing with large number of failures?
- No!

RT in Different Product Lines Varies

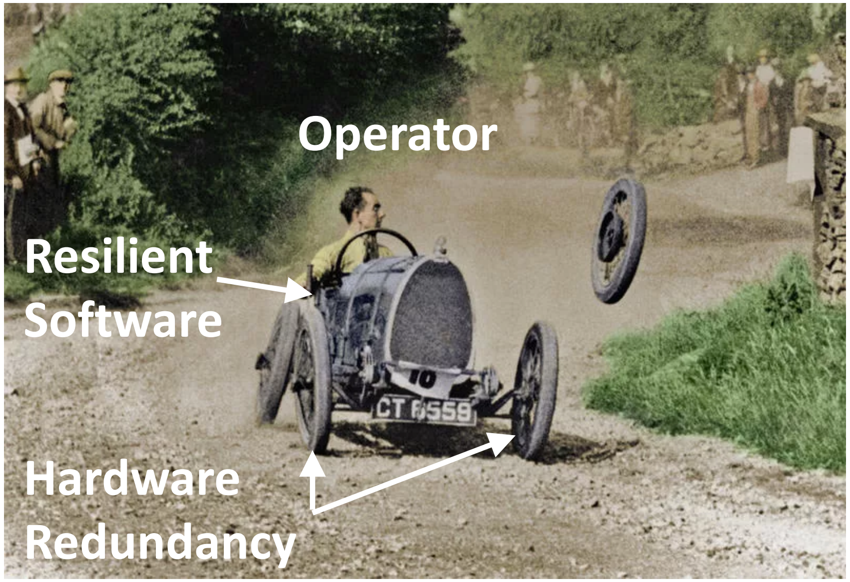
- Observation 1: Variation of *RT* in different product lines is large
- Observation 2: Operators respond to large number of failure more quickly



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

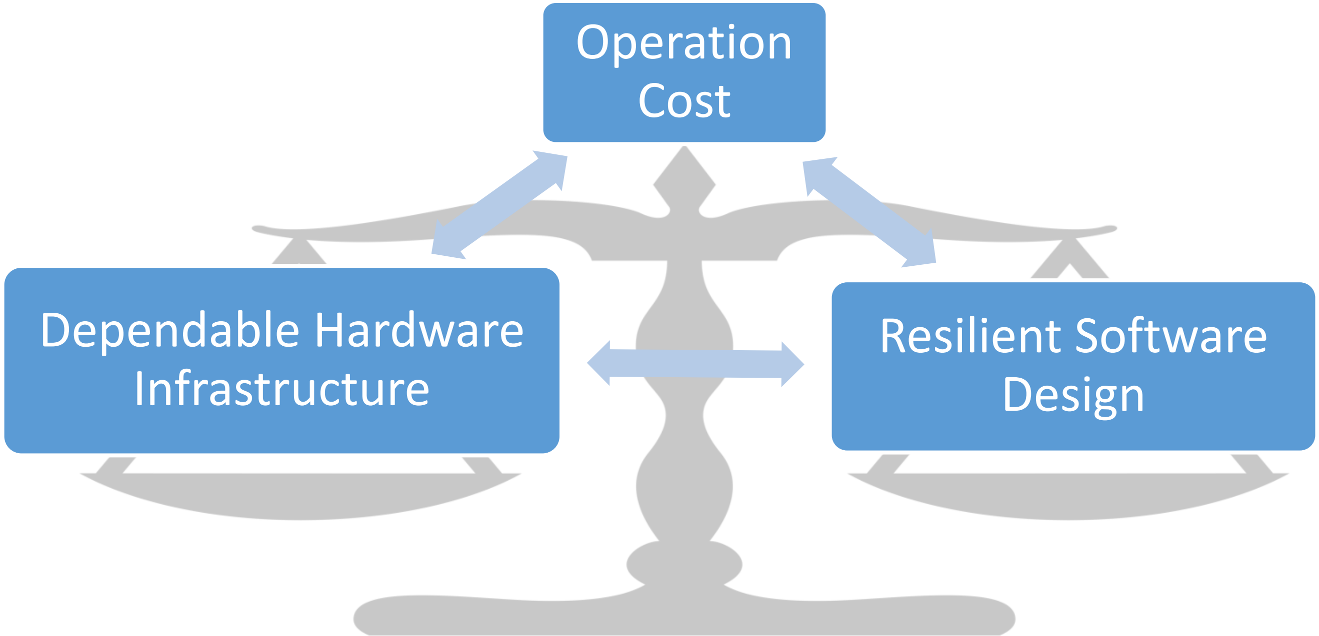
- Dataset overview
- Temporal distribution of the failures
- Spatial distribution of the failures
- Correlated failures
- Operators’ response to failures
- ➢ Lessons Learned

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”

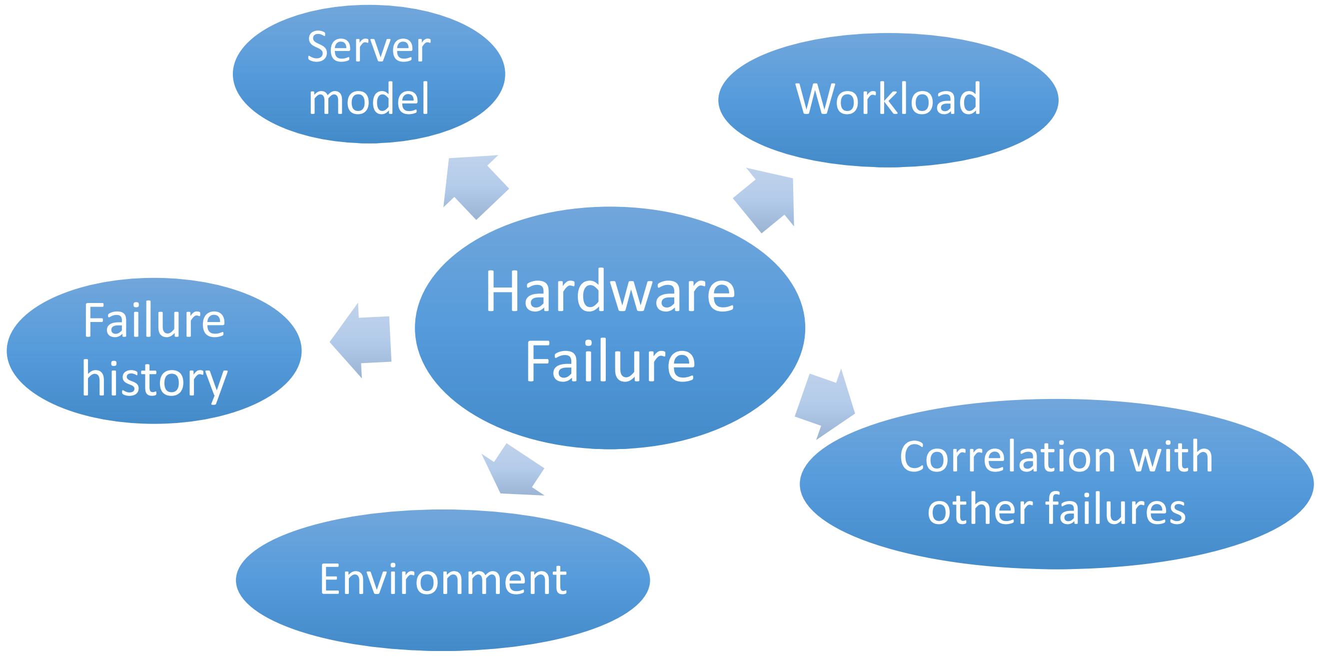
Lessons Learned II

- 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

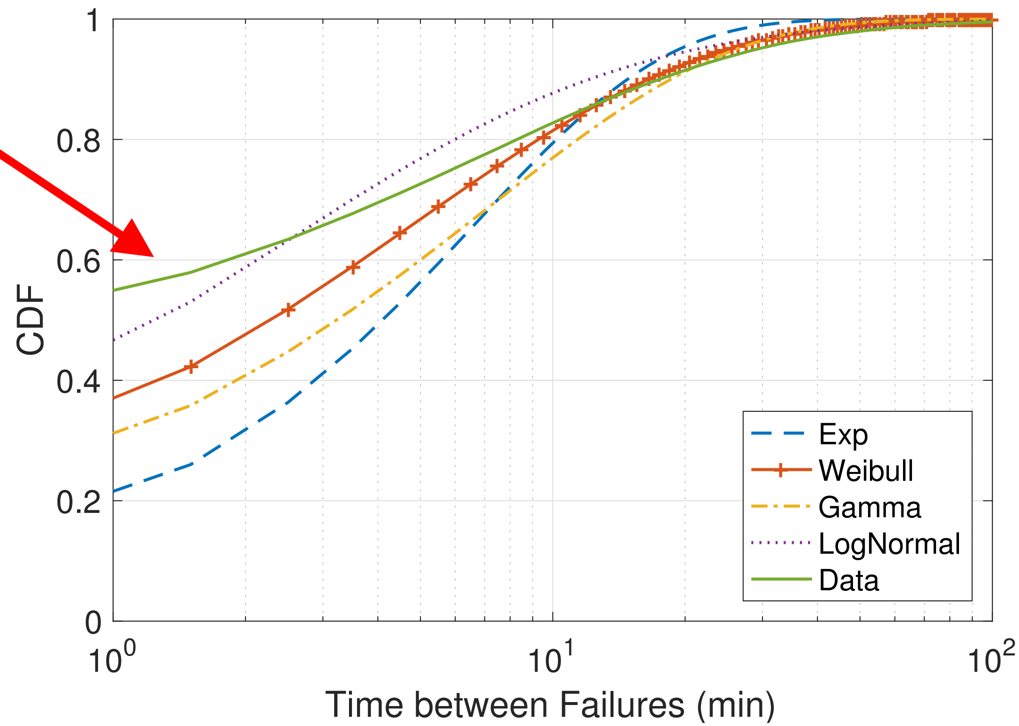
- Dataset overview
- Temporal distribution of the failures
- Spatial distribution of the failures
- Correlated failures
- Operators’ response to failures
- Lessons Learned

TBF Cannot be Well Fitted by Well-known Distributions

REJECTED **Hypothesis 4.** *Time between failures (TBF) of all components follows an exponential distribution.*

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

Large proportion of small values



Failure Operation Ticket (FOT)

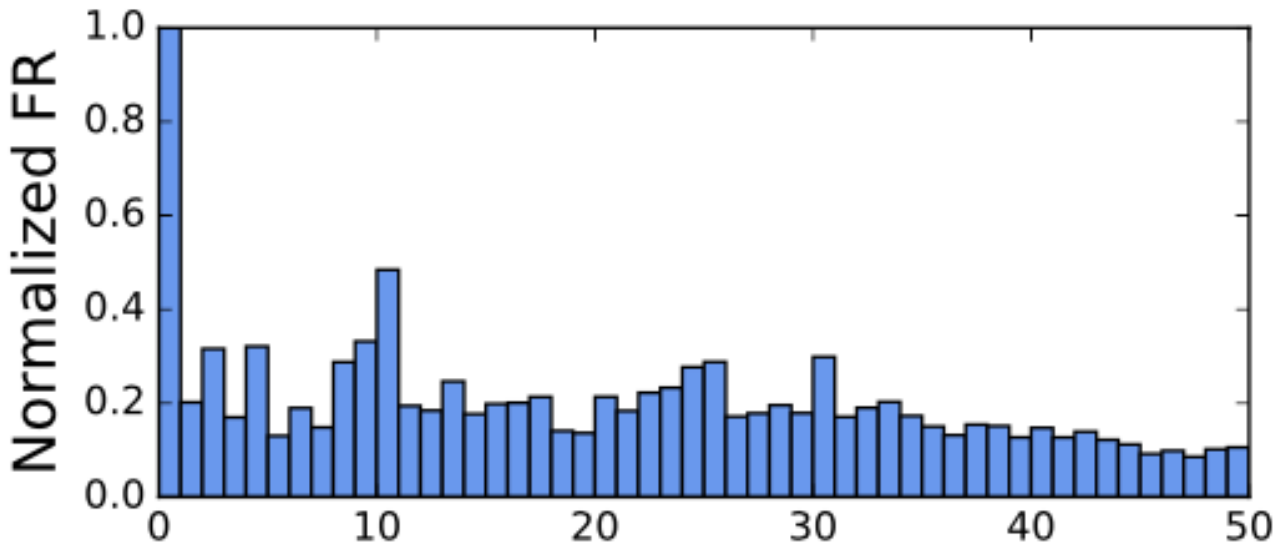
- Categories of FOTs

Failure trace	Handling decision	Percentage
D_fixing	Issue a repair order (RO)	70.3%
D_error	Not repair and set to decommission	28.0%
D_falsealarm	Mark as a false alarm	1.7%

- Fields:
  id, host id, hostname, host idc, error device, error type, error time, error position, error detail

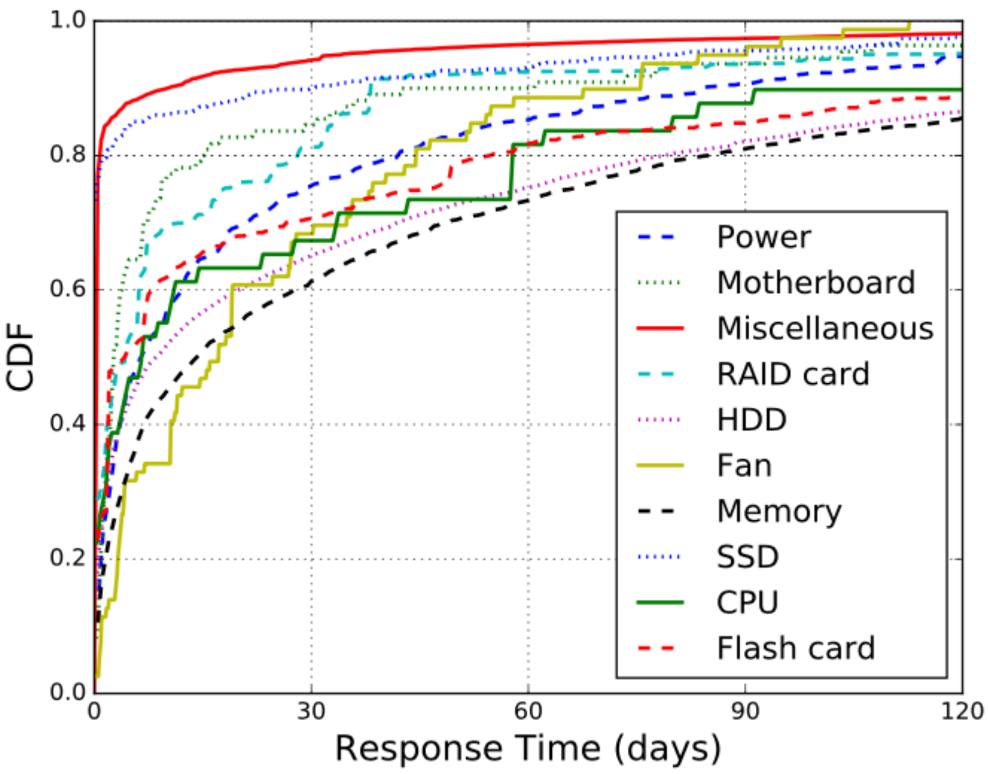
FR of Misc. Failures During the Lifecycle

- 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

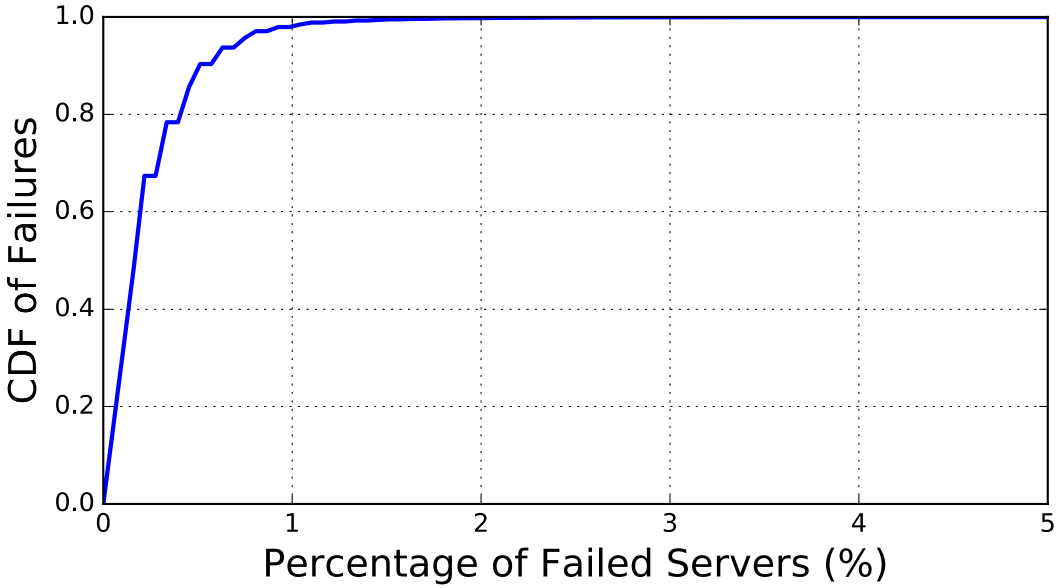
- Fields: raw value, worst, threshold, status
- SMART attribute examples (failure related)
  - Reallocated Sectors Count
  - End-to-End error
  - Uncorrectable Sector Count
  - Reported Uncorrectable Errors
  - Current Pending Sector Count
  - Command Timeout
  - ...

Examples of Failure Types

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

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.

Device	$\mathbf{r_{100}}$(%)	$\mathbf{r_{200}}$(%)	$\mathbf{r_{500}}$(%)
HDD	55.4	22.5	2.5
Miscellaneous	3.7	1.3	0.1
Power	0.7	0.4	0
Memory	0.4	0.4	0.1
RAID card	0.4	0.2	0.1
Flash card	0.1	0.1	0
Fan	0.1	0	0
Motherboard	0	0	0
SSD	0	0	0
CPU	0	0	0