2022 USENIX Annual Technical Conference

USENIX ATC '22

NVMe SSD Failures in the Field: the Fail-Stop and the Fail-Slow

Ruiming Lu, Erci Xu, Yiming Zhang, Zhaosheng Zhu, Mengtian Wang, Zongpeng Zhu, Guangtao Xue, Minglu Li, Jiesheng Wu

Hardware Failures

- **The Achilles' Heel of Modern Data Centers**
  - **Storage (SSD & HDD)**
  - NIC
  - CPU
  - Memory

Source:
Data centers at Alicloud.

Source: https://community.fs.com/blog/different-types-of-server-rack-used-in-data-center.html

Large-Scale SSD Reliability Studies

- **Focus on SAS/SATA SSD**
  - Failure Rate Curve
  - FTL Impact
  - Correlated Failures

A Large-Scale Study of Flash Memory Failures in the Field

Justin Meza, Carnegie Mellon University, meza@cmu.edu
Qiang Wu, Facebook, Inc., qwu@fb.com
Sanjeev Kumar, Facebook, Inc., skumar@fb.com
Onur Mutlu, Carnegie Mellon University, onur@cmu.edu

A Study of SSD Reliability in Large Scale Enterprise Storage Deployments

Stathis Maneas, *University of Toronto*
Kaveh Mahdaviani, *University of Toronto*
Tim Emami, *NetApp*
Bianca Schroeder, *University of Toronto*

An In-Depth Study of Correlated Failures in Production SSD-Based Data Centers

Shujie Han[1], Patrick P. C. Lee[1], Fan Xu[2], Yi Liu[2], Cheng He[2], and Jiongzhou Liu[2]
[1]The Chinese University of Hong Kong [2]Alibaba Group

Abstract

Flash-based solid-state drives (SSDs) are increasingly adopted as the mainstream storage media in modern data centers. However, little is known about how SSD failures in the field are correlated, both spatially and temporally. We argue that characterizing correlated failures of SSDs is critical,

To elaborate, the following questions on correlated failures remain unexplored: (i) How far are SSD failures spaced apart across different scopes in large-scale data centers? (ii) How likely does an SSD fail after another failure occurs in the same scope? (iii) How long is the time interval between two consecutive SSD failures in the same scope? (iv) Do SSD

Changes in NVMe SSD

- **External**
  - NVMe Interface
- **Internal**
  - Favor 3D NAND
  - RAIN (Redundant Array of Independent NAND)
  - LDPC (Low-Density Parity-Check)

Reliability of NVMe SSD ?

USENIX ATC '22 Our Study

- **Comparative Fail-Stop Study**
  - **NVMe** SSD vs. **SAS/SATA** SSD
- **Quantitative Fail-Slow Study**
  - Severity
  - Impact Factors
  - Transition to Fail-Stop

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE
ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL
ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

NVMe SSD Failures in the Field: the Fail-Stop and the Fail-Slow

Ruiming Lu[1*], Erci Xu[2*], Yiming Zhang[3†], Zhaosheng Zhu[4], Mengtian Wang[4], Zongpeng Zhu[4], Guangtao Xue[1†], Minglu Li[1,5], and Jiesheng Wu[4]

[1]Shanghai Jiao Tong University, [2]PDL, [3]Xiamen University, [4]Alibaba Inc., and [5]Zhejiang Normal University

Abstract

NVMe SSD has become a staple in modern datacenters thanks to its high throughput and ultra-low latency. Despite its popularity, the reliability of NVMe SSD under mass deployment remains unknown. In this paper, we collect logs from over one million NVMe SSDs deployed at Alibaba, and conduct extensive analysis. From the study, we identify a series of major reliability changes in NVMe SSD. On the good side, NVMe SSD becomes more resilient to early failures and variances of access patterns. On the bad side, NVMe SSD becomes more vulnerable to complicated correlated failures. More importantly, we discover that the ultra-low latency nature makes NVMe SSD much more likely to be impacted by fail-slow failures.

1 Introduction

NVMe SSD is now the new favorite of modern data centers. With a performance specification of up to 6GB/s bandwidth and microsecond-level latency, NVMe SSD serves as a strong performance upgrade to its SATA-based peers [8, 18, 29–31].

Apart from the performance, the reliability of any hardware under mass deployment is of great concern [3, 5–7, 10, 14, 38, 40, 42, 45]. While there is a spate of work covering the failure characteristics of SATA SSDs in the field [34–36, 41, 47], their findings may not be conclusive for NVMe SSD.

First, with a low-latency interface, NVMe SSD can be especially *prone* to fail-slow failure (aka. gray failure [17, 21, 25, 26, 48]). In a nutshell, the NVMe SSD fail-slow failure causes a drive to exhibit abnormal performance slowdown (e.g., high latency under normal traffic). Unlike SATA SSD, where fail-slow failure may be masked by the relatively high latency (>100μs), NVMe SSD can be easily impacted due to its ultra-low latency nature ($\sim$ 10μs) [23, 27, 28].

Moreover, the NVMe SSD is not just the SATA SSD with an interface upgrade. Instead, the internal architecture of NVMe SSD has gone through considerable changes. An outstanding example is the wide adoption of 3D-TLC NAND in NVMe SSD for larger capacity. Compared to MLC, the denser bits per cell (i.e., TLC) shows lower reliability and the vertical stacking (i.e., 3D flash) can exhibit disparate behaviors or even opposite patterns (e.g., lower error rate under higher temperatures [32]). Also, the vendors have integrated a series of techniques to improve the overall reliability in NVMe SSD, such as Redundant Array of Independent NAND (RAIN) or Low-Density Parity-Check code (LDPC) [43, 50]. Unfortunately, with no large-scale NVMe SSD fail-stop study available at the moment, the influences of recent advancements remain unknown.

In this paper, we study the fail-stop and fail-slow failures of NVMe SSDs deployed at Alibaba. Specifically, we collect and analyze device logs (i.e., SMART [11]), runtime logs (i.e., `iostat`), and failure tickets from over one million NVMe SSDs[1]. Throughout the study, we set our analysis into the context of previous studies to help various parties of interest get a clear picture of NVMe SSD reliability, including the improving and deteriorating failure patterns of fail-stop failures and the characteristics regarding the fail-slow failures.

We start our study by plotting and analyzing the baseline statistics (§3) of the NVMe SSDs, including the drive characteristics (e.g., manufacturer and model), usage characteristics (e.g., power-on time), and health metrics (e.g., annual replacement rate). Then, we comb through the dataset against different impact factors such as age and write amplification (§4). Finally, we lay a special focus on the fail-slow failures (§5), where we rigorously identify the fail-slow drives and perform extensive analysis. Altogether, we obtain 10 major findings and we list the highlights as follows:

- Infant mortality (failures occurring soon after deployment), a concerning failure trend in SATA SSD [35], is not outstanding in NVMe SSD. For nearly all of our models, the failure rate in the first three months is equivalent to or even less than that from later periods.
- High Write Amplification Factor (WAF), unlike SATA SSD [36], is no longer closely correlated with failures. Interestingly, NVMe SSD with low WAF (WAF$\leq$1) exhibits 2.19$\times$ higher ARR than high-WAF ones.
- Co-located (i.e., intra-node/rack) NVMe SSD failure becomes more temporally correlated. For example, compared to SATA SSD, NVMe SSD correlated failure increases up

*Equal contribution.
†Corresponding authors.

[1]We release our dataset at https://tianchi.aliyun.com/dataset/dataDetail?dataId=128972.

Outline

~~INTRODUCTION~~

DATASET

FAIL-STOP

FAIL-SLOW

SUMMARY & TAKE-AWAY POINTS

USENIX ATC '22

Our Dataset

- **1.8+ million enterprise-level NVMe SSDs at Alibaba:**
  - MLC, 3D-TLC, and QLC drives.
  - 3 manufacturers.
  - 11 drive models:
    - 12 different capacities (370GB-4000GB).
  - Varying drive age and usage.
  - Diverse services:
    - Block storage, object storage, big data, buffering, log, streaming, and query.

USENIX ATC '22

Our Dataset

- **1.8+ million NVMe SSDs**
- **Data source**
  - SMART logs (by devices)
  - Failure tickets (by monitoring daemons)
  - Performance logs (iostat)

Our Dataset

- **1.8+ million NVMe SSDs**
- **Data source**
  - SMART logs (by devices)
  - Failure tickets (by monitoring daemons)
  - Performance logs (iostat)
- **Dataset released**
  - https://tianchi.aliyun.com/dataset/dataDetail?dataId=128972

Outline

~~INTRODUCTION~~

~~DATASET~~

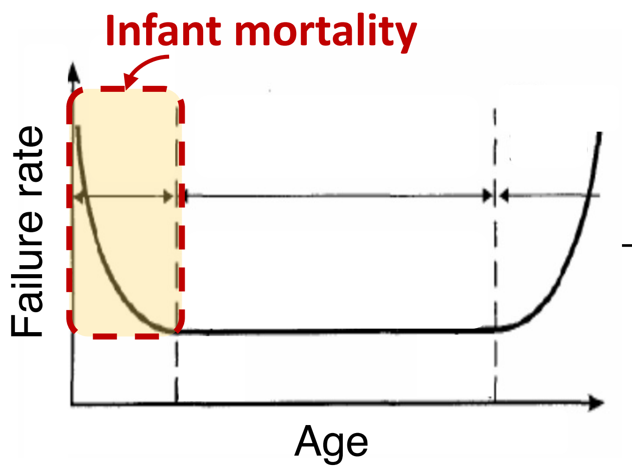
FAIL-STOP

FAIL-SLOW

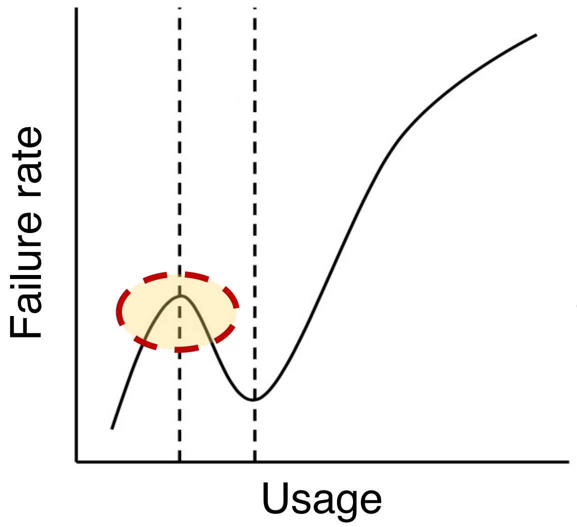
SUMMARY & TAKE-AWAY POINTS

Infant Mortality

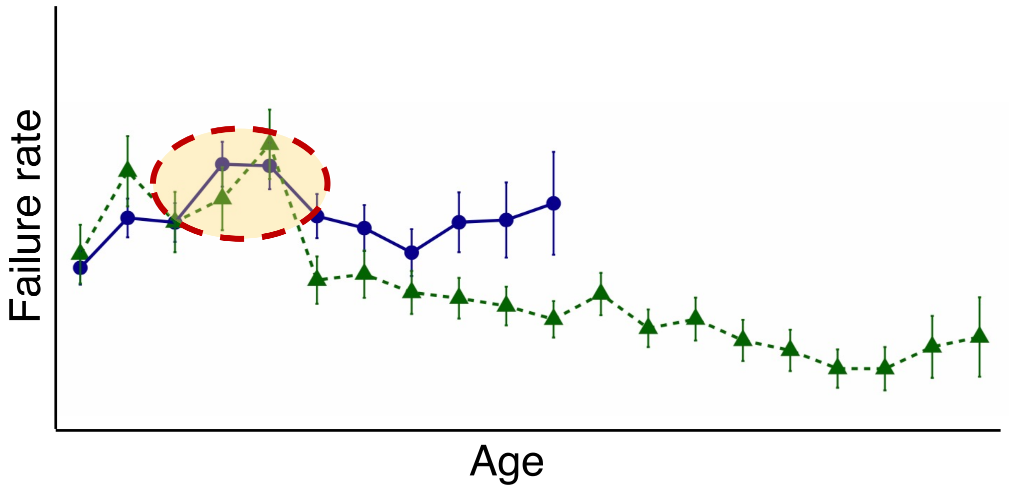
- **How does the storage device failure rate vary with age/usage?**



(a) HDD[1]



(b) SSD[2]



(c) SAS/SATA SSD[3]

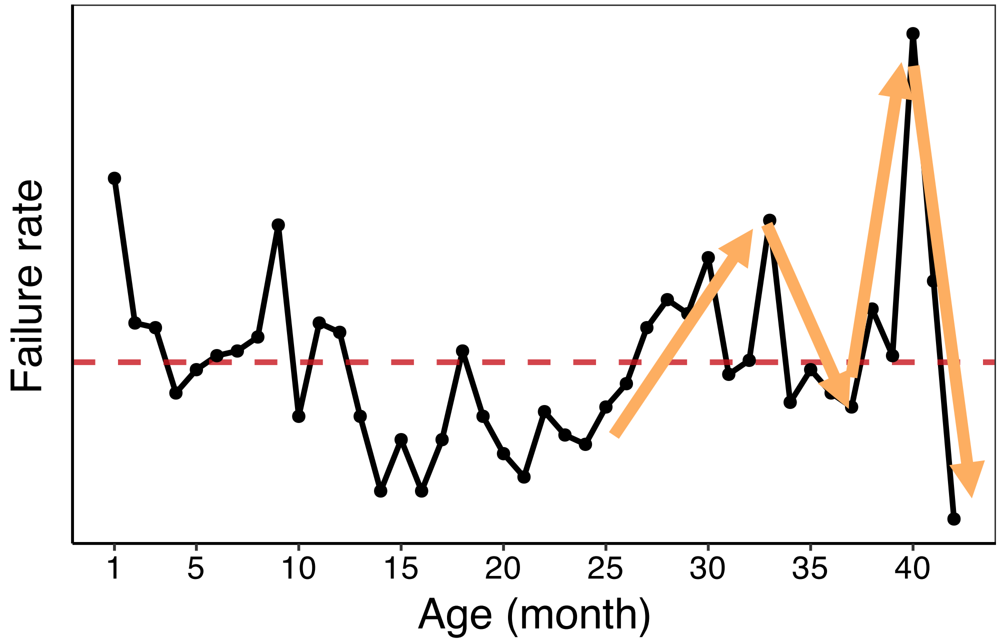
2007 — Facebook, 2015 — NetApp, 2020

[1] Schroder et al. *Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?*
[2] Meza et al. *A large-scale study of flash memory failures in the field.*
[3] Maneas et al. *A study of SSD reliability in large scale enterprise storage deployments.*

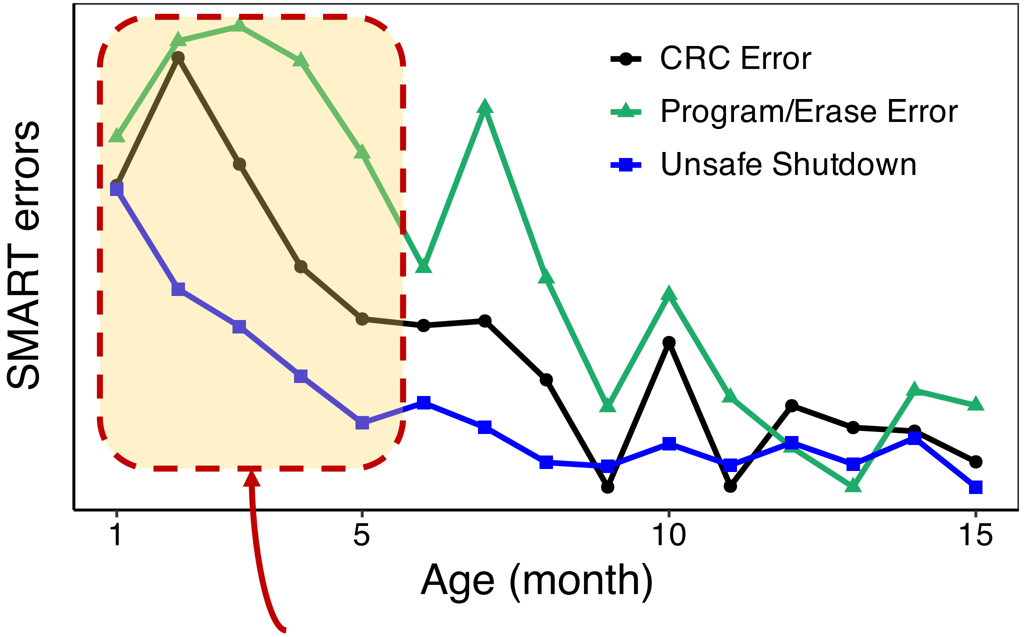
- (SAS/SATA) SSD would experience a drawn-out period of infant mortality.

Infant Mortality

- **What about in NVMe SSD?**



Infant mortality not notable.



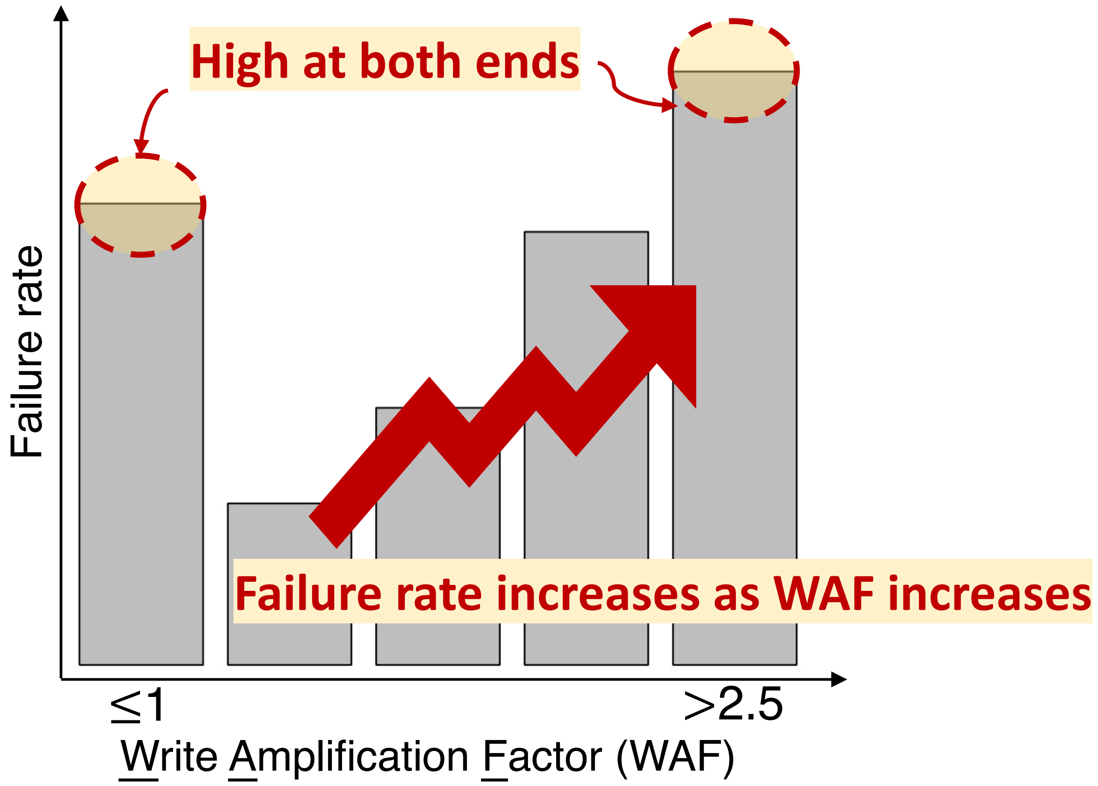
Device errors still prevalent.

- **Possibly due to improvement in FTL error handling.**

- NVMe SSD does not exhibit high failure rates during early deployment.

Write Amplification

- **How does write amplification affect SSD reliability?**
  - Write Amplification Factor (WAF)
    - $WAF = \frac{NAND\ writes}{Logical\ writes}$
    - Usually above 1 (e.g., due to GC)
    - Data compression $\Longrightarrow$ WAF≤ 1
  - Microsoft' 2016[1] on SATA SSD

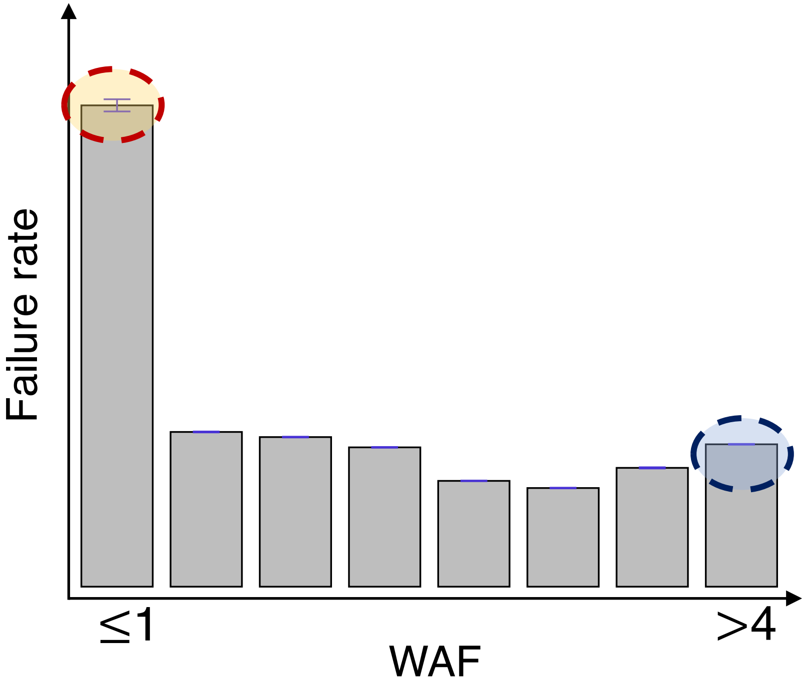


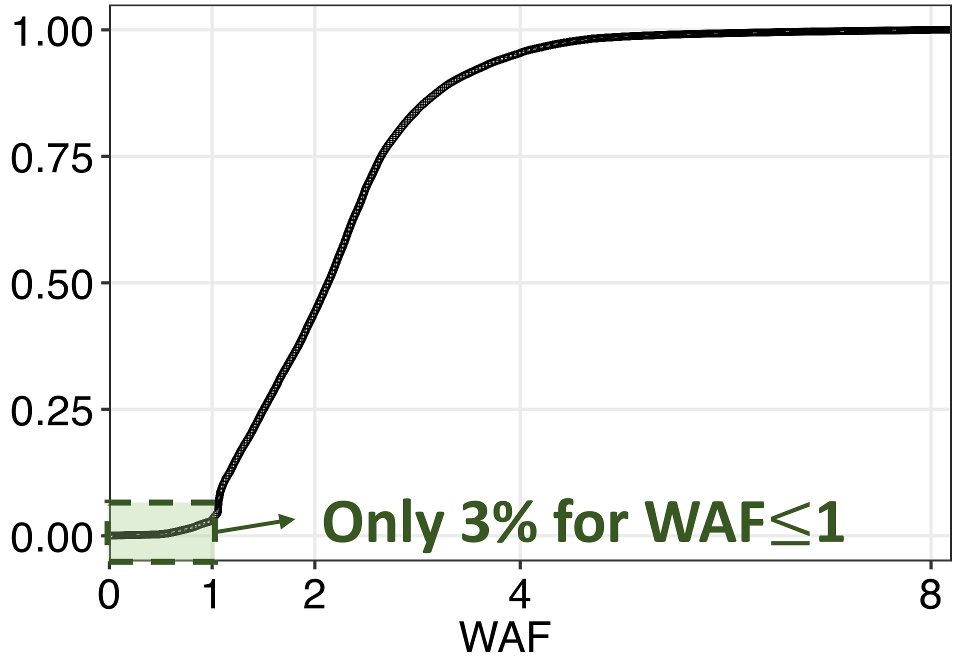
[1] Narayanan et al. *SSD failures in datacenters: What? When? And Why?*

- SATA SSD would experience high failure rates at both low(≤1) and high(>2.5) WAF levels.

Write Amplification

- **What about in NVMe SSD?**





2.19X higher than average

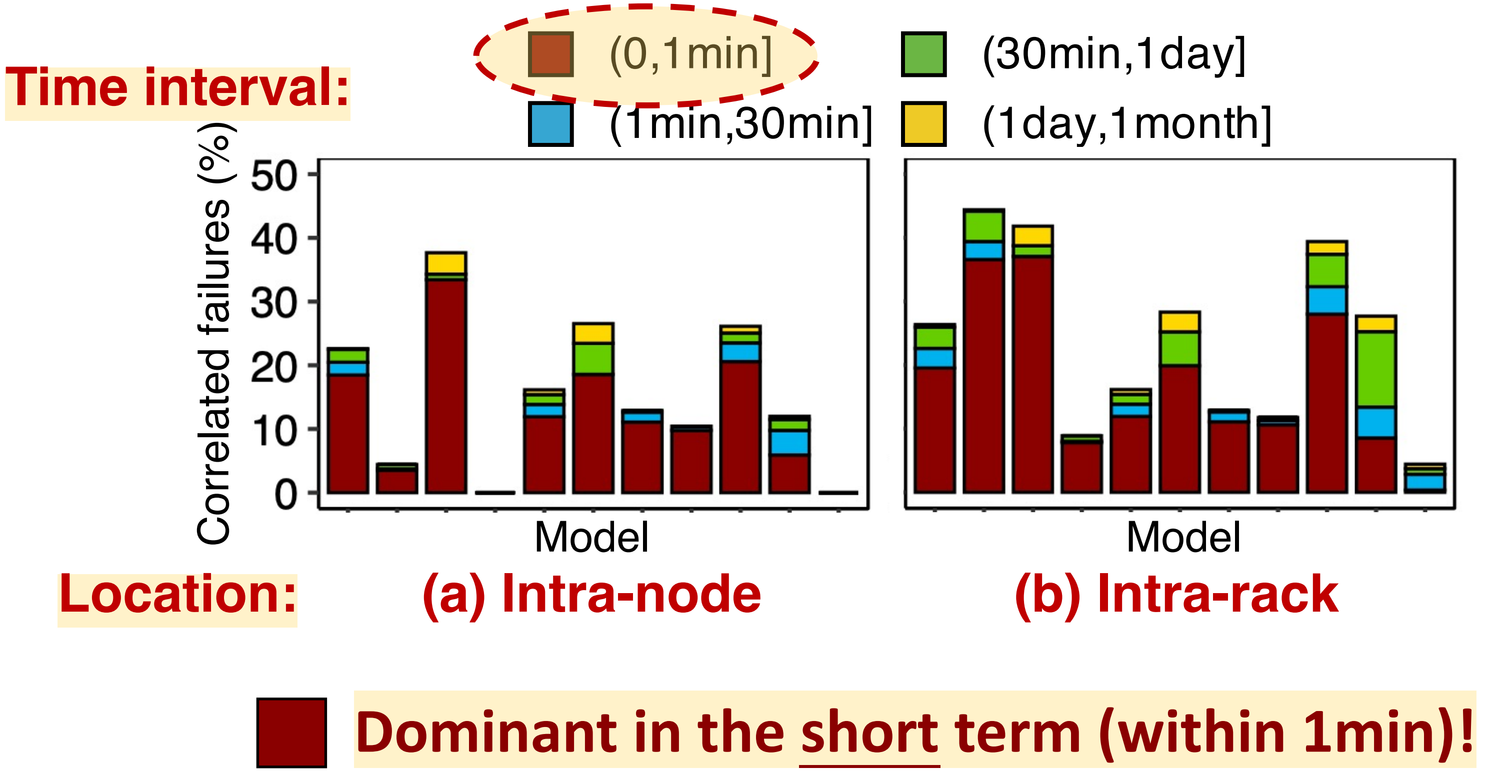
Extremely low WAF is rare

Heavy write amplification ⇏ high failure rate

- NVMe SSD only has notably high failure rates at low WAF levels (i.e., rare but deadly).

Correlated Failures

- **What is the distribution of SSD correlated failures?**
  - Spatially correlated
    - From the same node/rack
  - Temporally correlated
    - Similar time to failure
  - Alibaba' 21[1] on SATA SSD



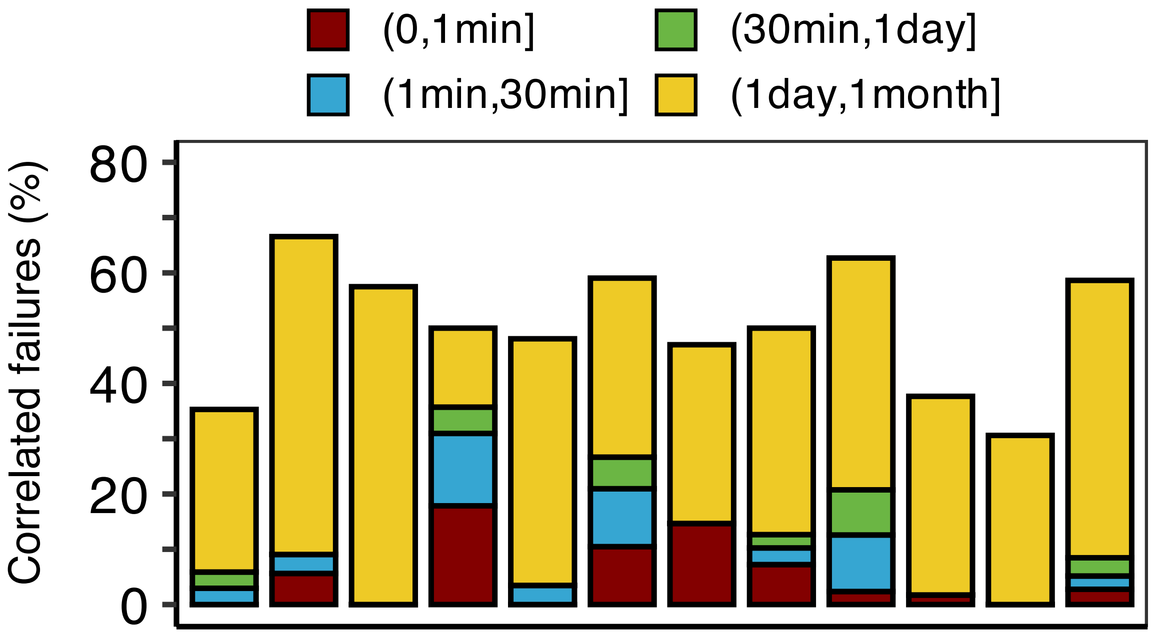
[1] Han et al. *An in-depth study of correlated failures in production SSD-based data centers.*

- For SATA SSD, spatially correlated failures are temporally correlated in the short-term span.

USENIX ATC '22

Correlated Failures

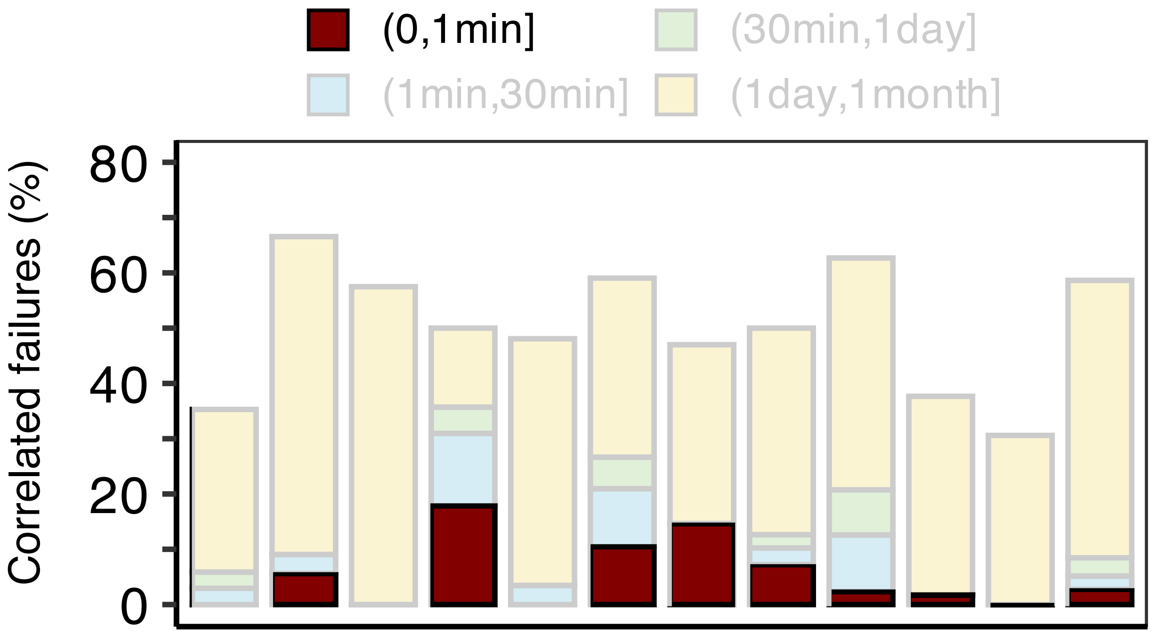
- **What about in NVMe SSD?**



(a) Intra-node

Correlated Failures

- **What about in NVMe SSD?**



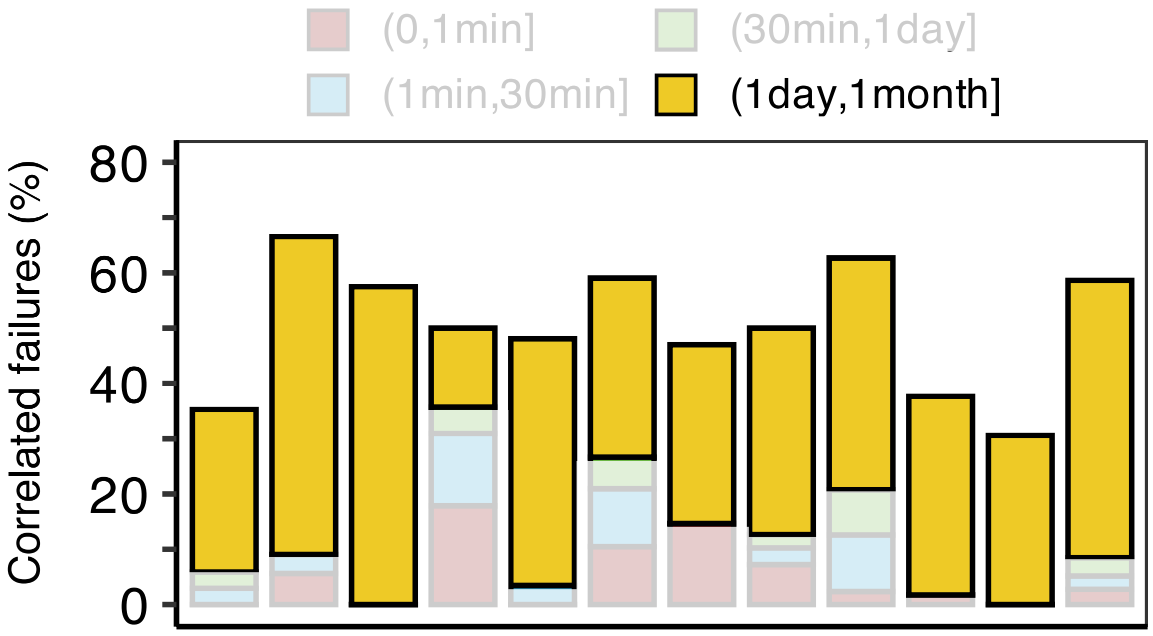
(a) Intra-node

No longer prevalent within 1 minute

USENIX ATC '22

Correlated Failures

- What about in NVMe SSD?



(a) Intra-node

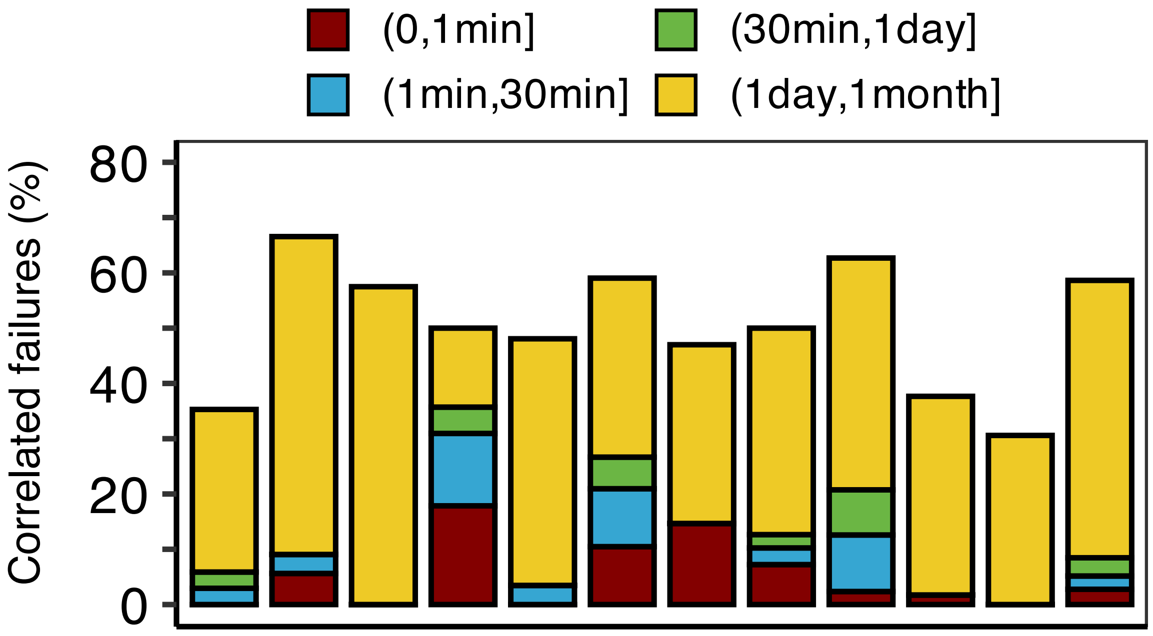
No longer prevalent within 1 minute

Dominant in the long term (> 1 day)

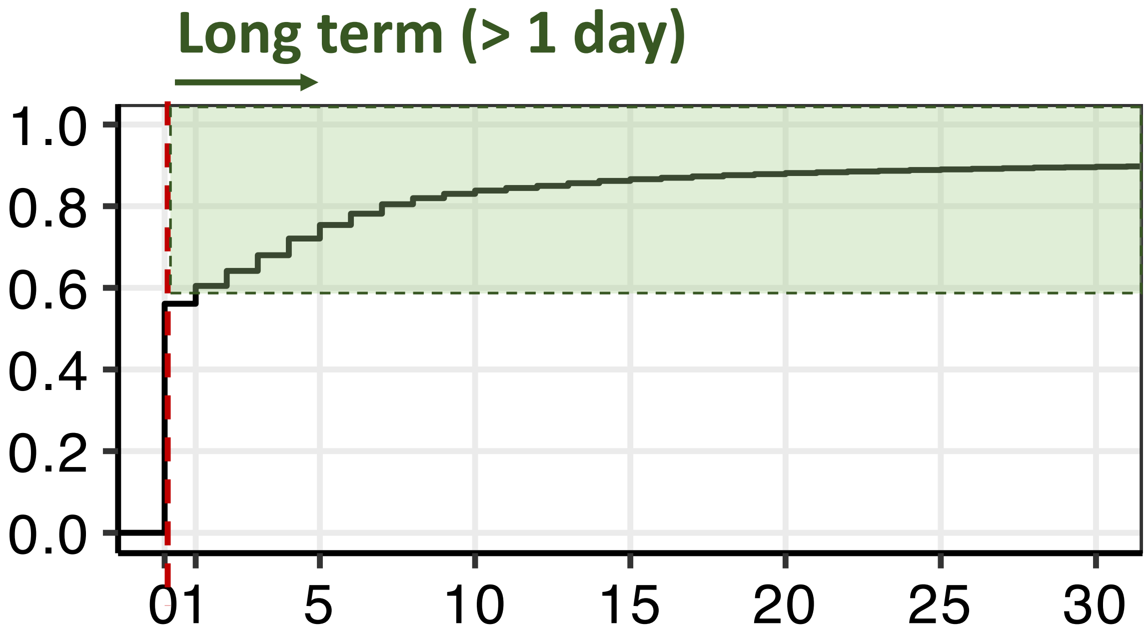
Correlated Failures

USENIX ATC '22

- **What about in NVMe SSD?**



(a) Intra-node



(b) Repair time (days)

No longer prevalent within 1 minute

Dominant in the long term (> 1 day)

~43.90% repaired after 1 day

- Spatially correlated failures are temporally correlated only in the long-term span.

Outline

~~INTRODUCTION~~

~~DATASET~~

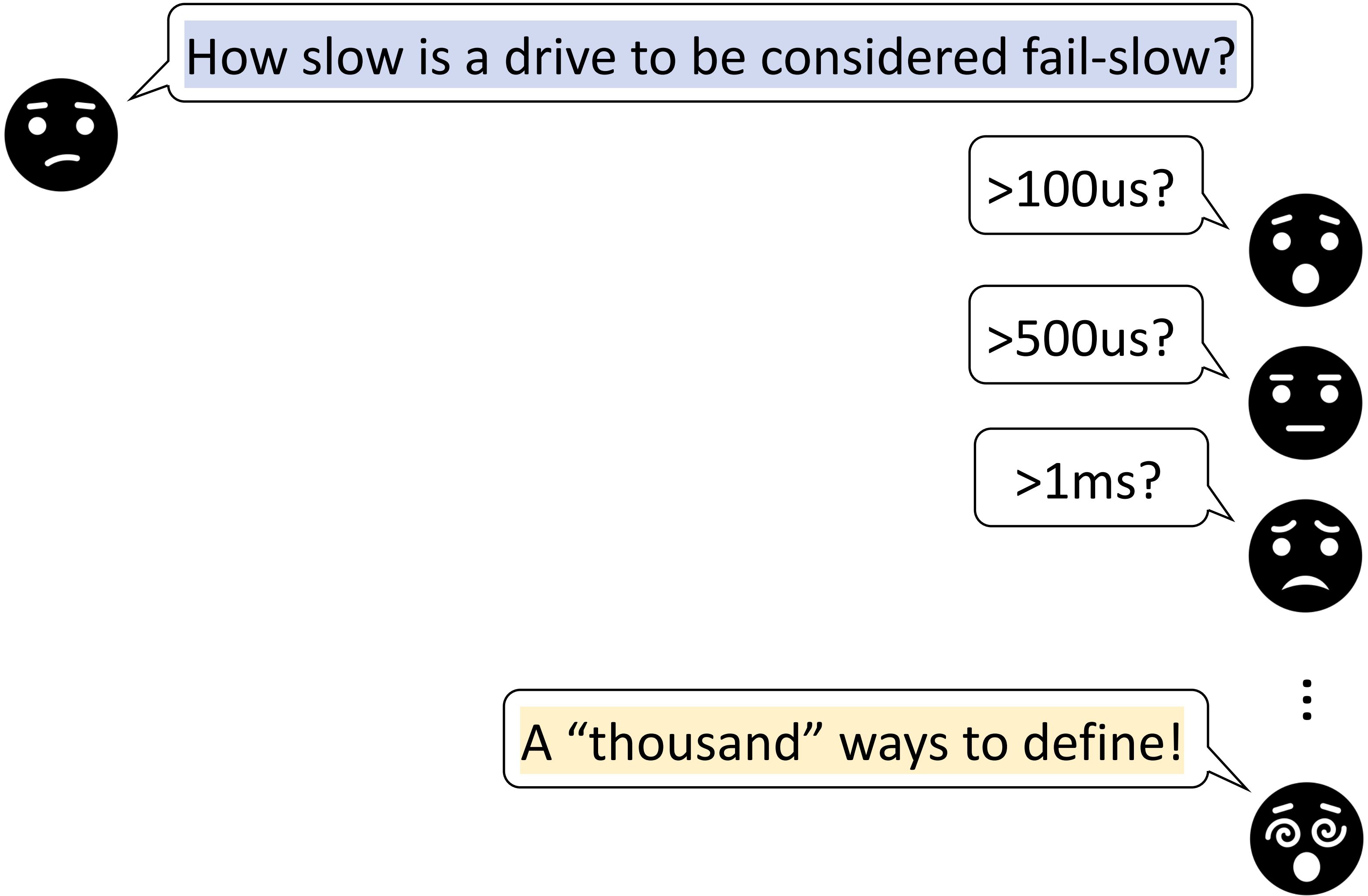
~~FAIL-STOP~~

FAIL-SLOW

SUMMARY & TAKE-AWAY POINTS

How to identify fail-slow drives?

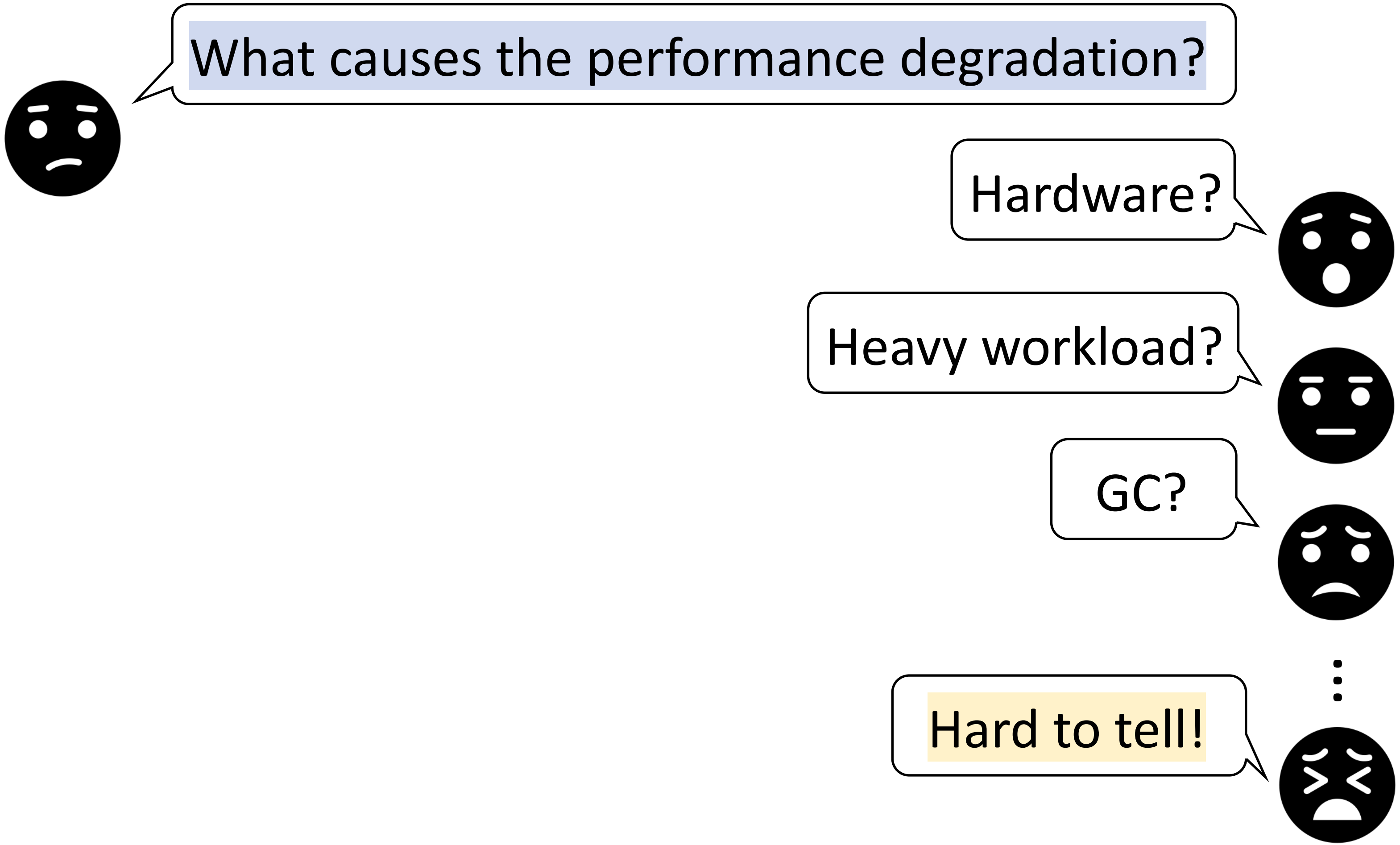
- **No ground truth in <u>identifying fail-slow</u>**



USENIX ATC '22

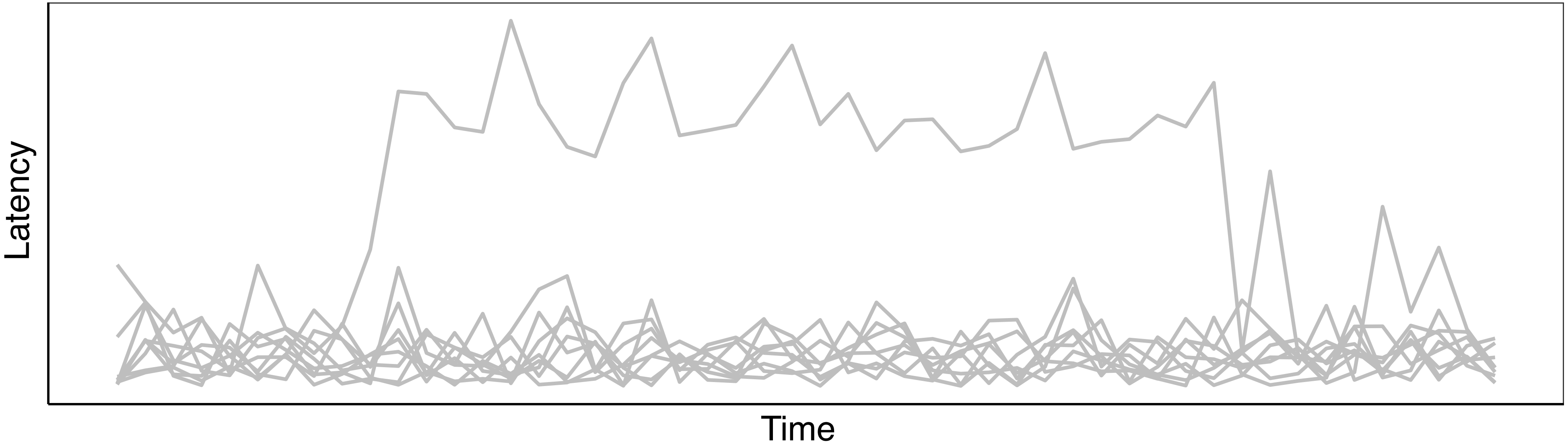
How to identify fail-slow drives?

- No ground truth in identifying fail-slow
- No ground truth in root causes



How to identify fail-slow drives?

- **(I) Identify suspicious drives**

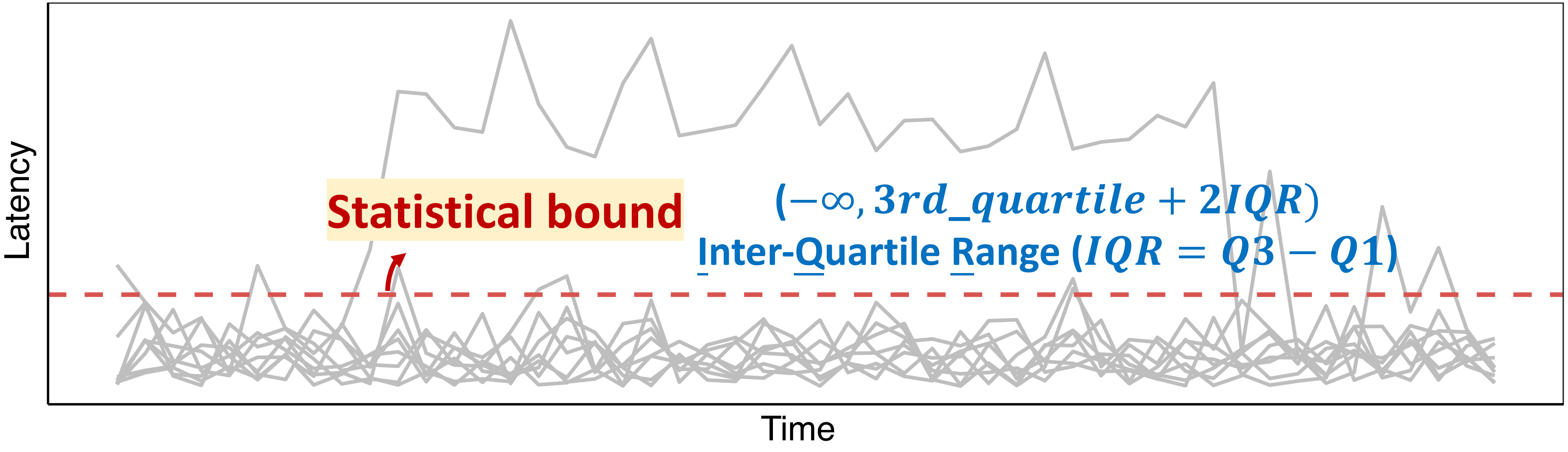


- Our solution: Peer-evaluating drives from the same node to identify the fail-slow.

How to identify fail-slow drives?

USENIX ATC '22

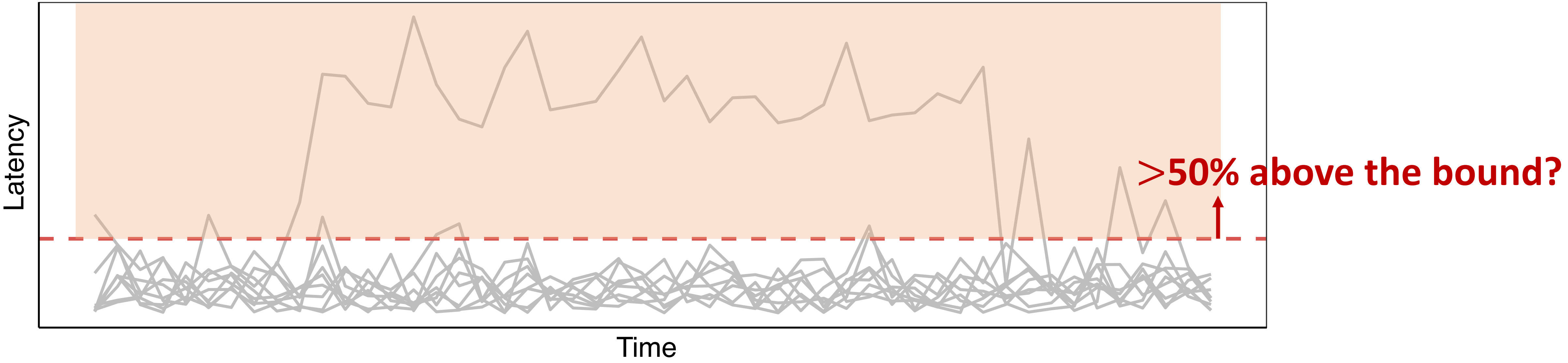
- **(I) Identify suspicious drives**



- Our solution: Peer-evaluating drives from the same node to identify the fail-slow.

How to identify fail-slow drives?

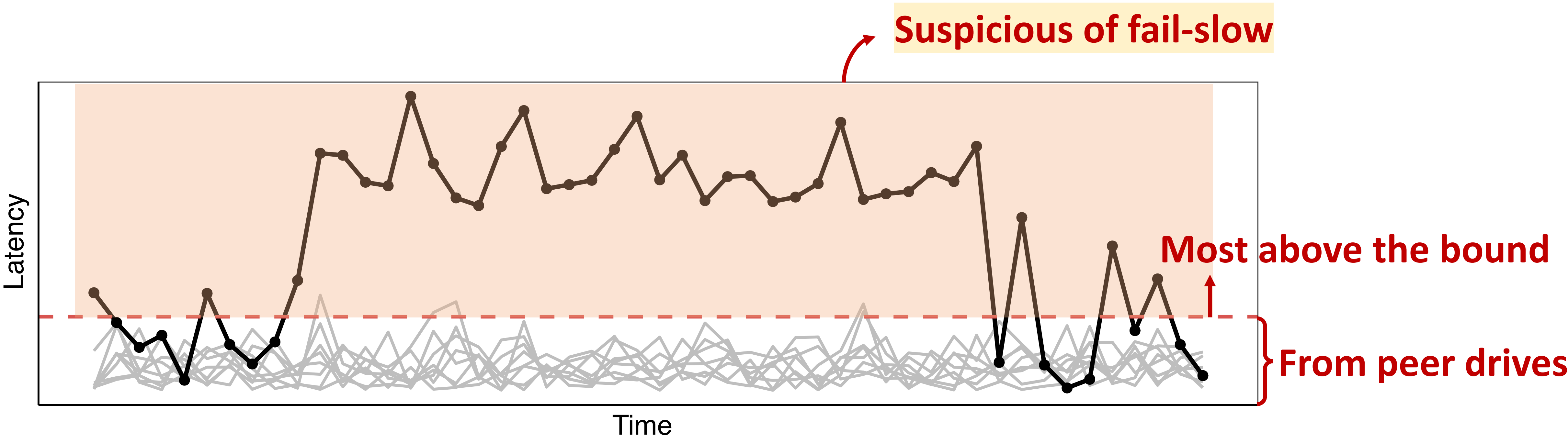
- **(I) Identify suspicious drives**



- Our solution: Peer-evaluating drives from the same node to identify the fail-slow.

How to identify fail-slow drives?

- **(I) Identify suspicious drives**

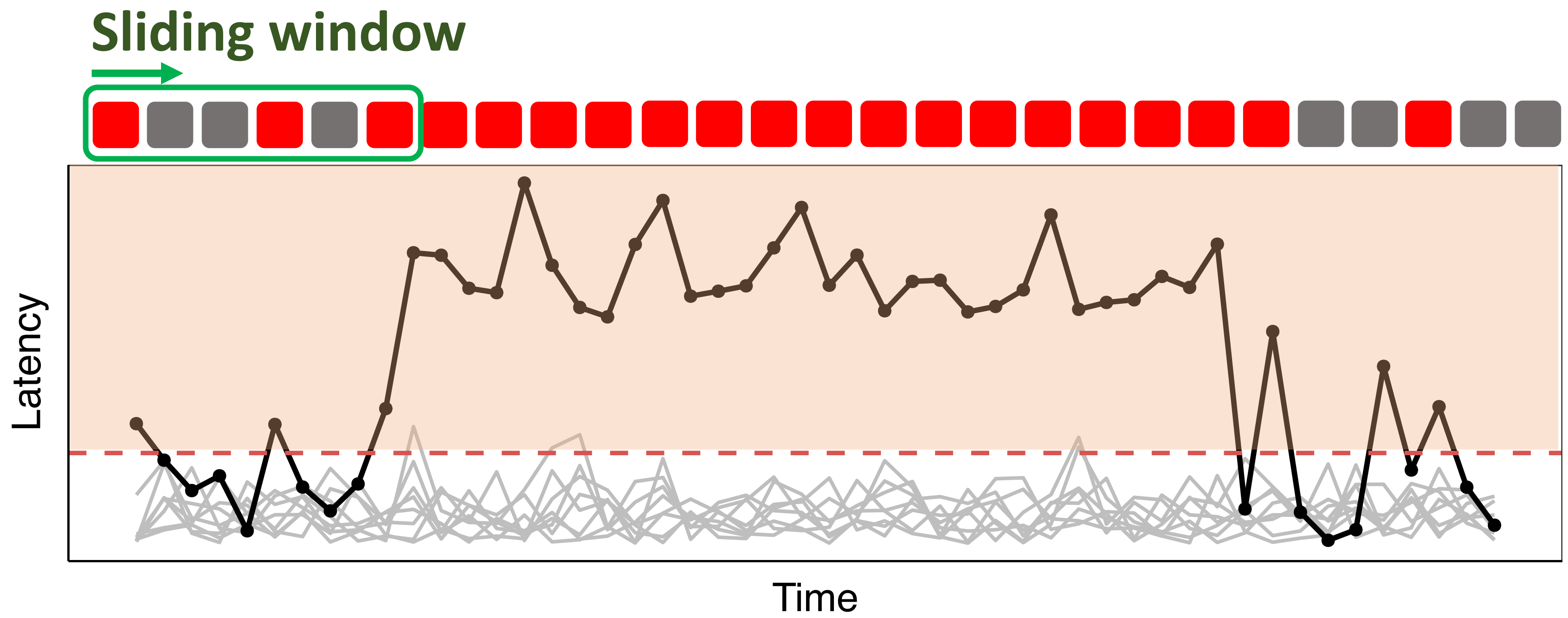


- Our solution: Peer-evaluating drives from the same node to identify the fail-slow.

USENIX ATC '22

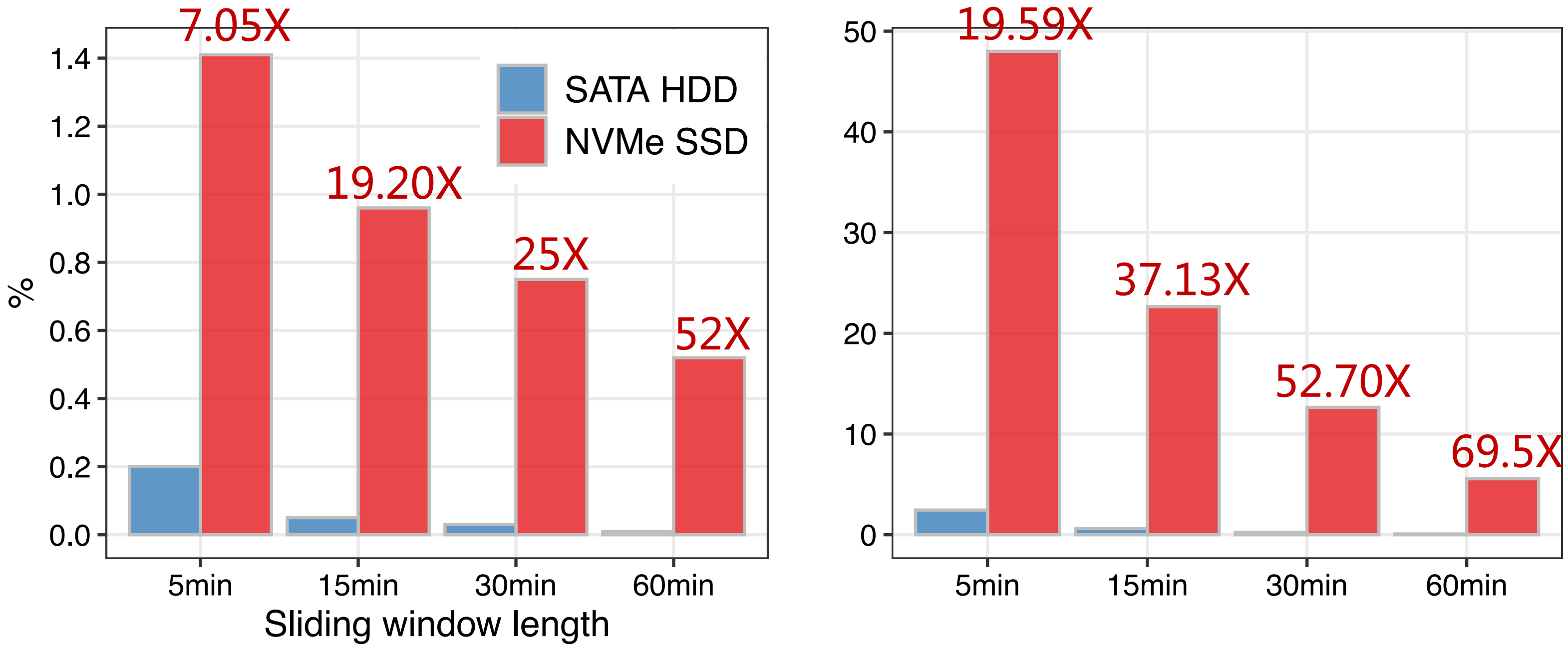
How to identify fail-slow drives?

- (I) Identify suspicious drives
- **(II) Identify slowdown events**



- Our solution: Peer-evaluating drives from the same node to identify the fail-slow.

A widespread concern



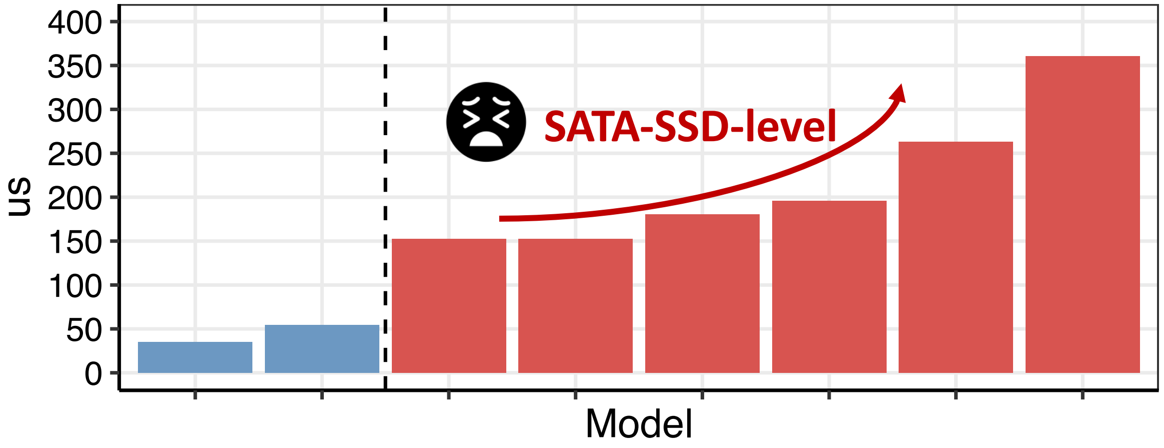
(a) Fail-slow drive (%)

(b) Event frequency

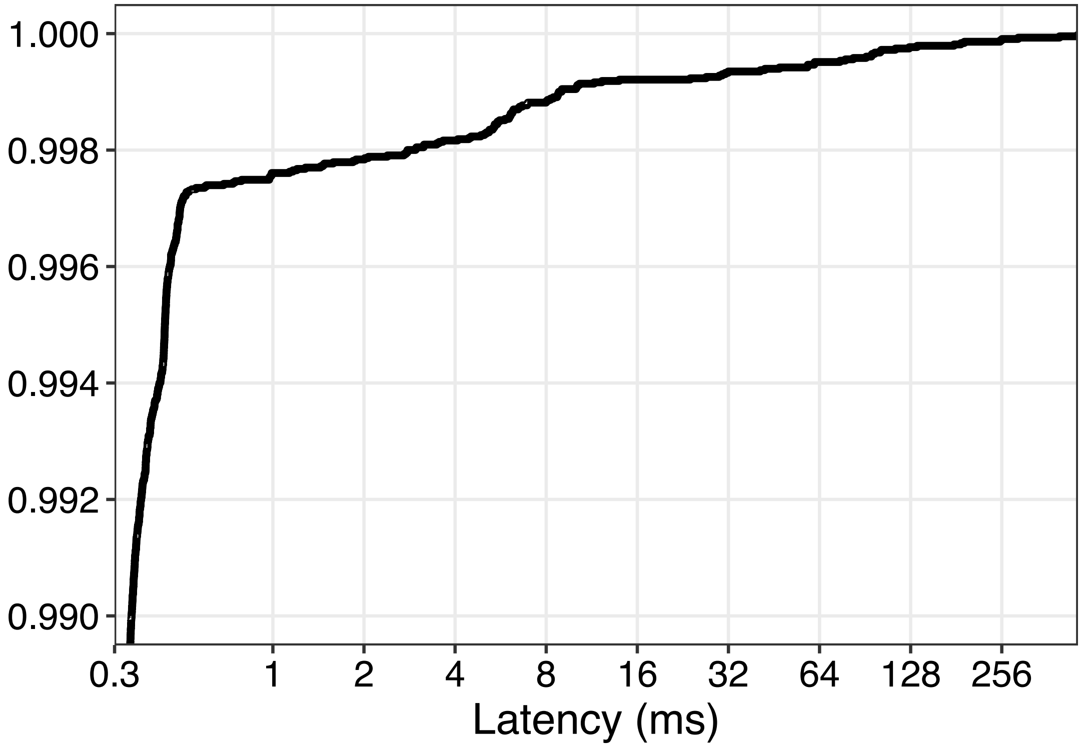
- Compared to HDD, fail-slow failure in NVMe SSD is much more widespread and frequent.

USENIX ATC '22

A severe problem



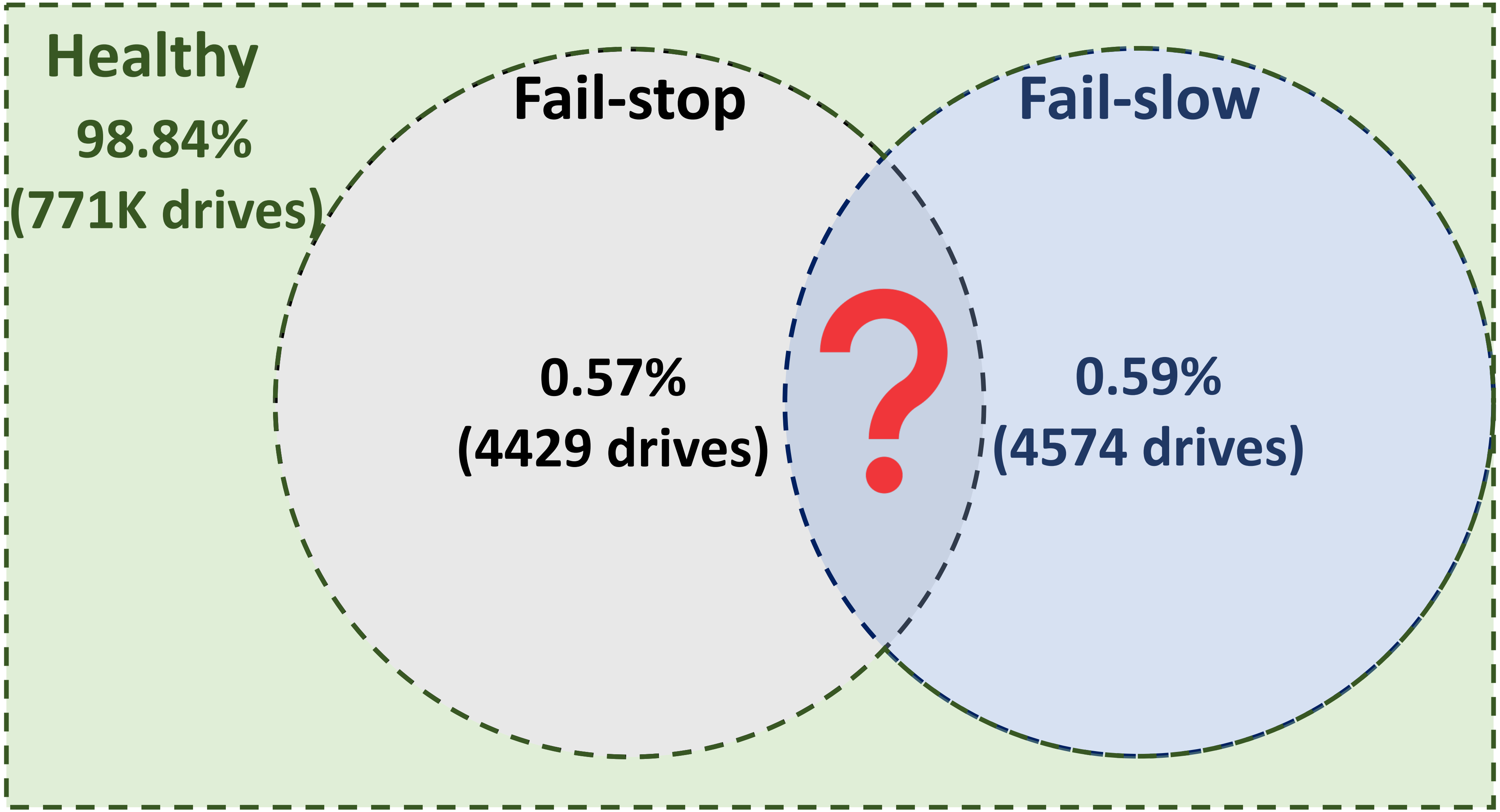
(a) Average event latency



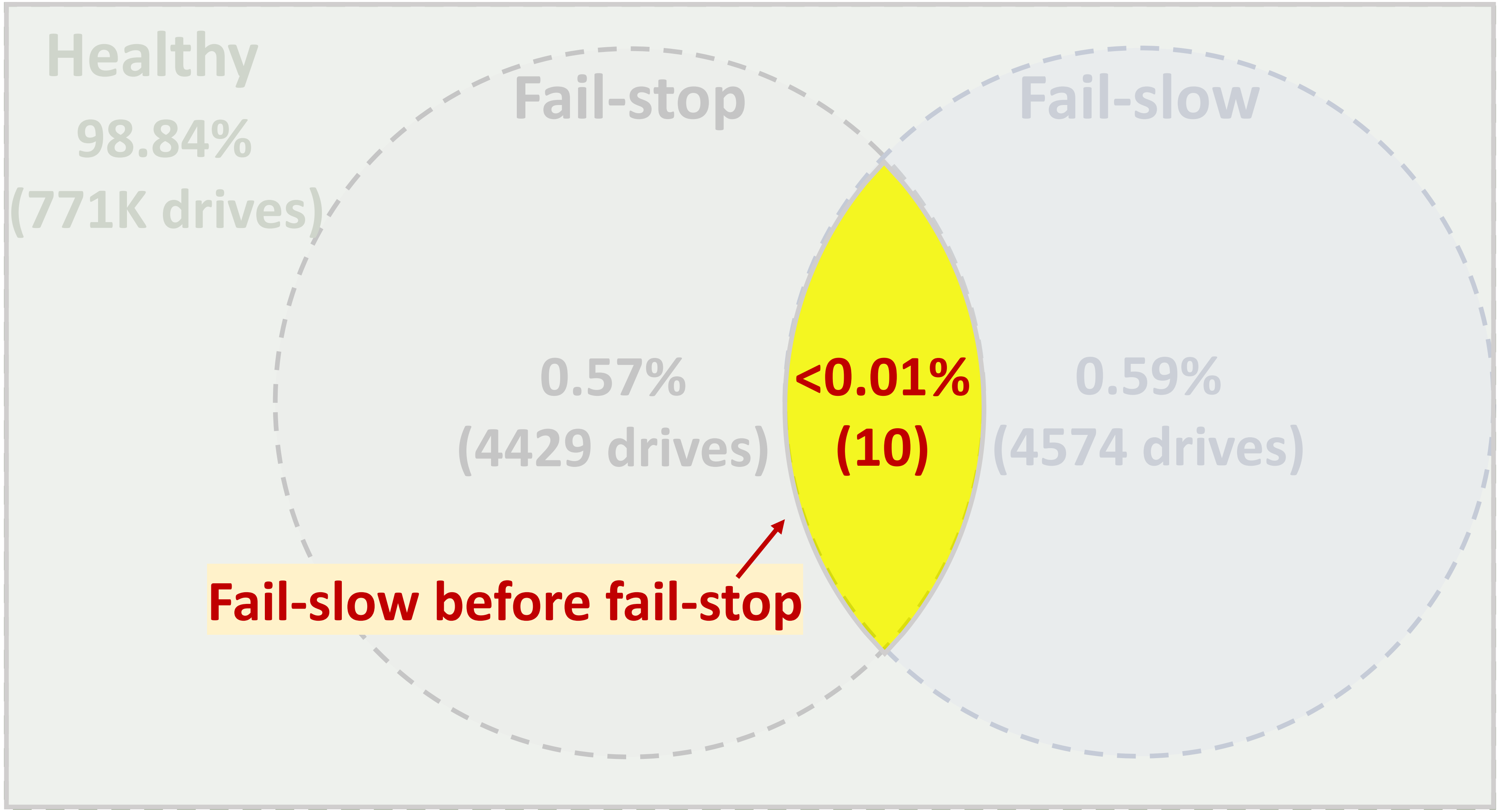
(b) Top 1% slowest event

- Fail-slow NVMe SSD could degrade to SATA SSD or even HDD performance.

Transition to (fail-stop) failures



Transition to (fail-stop) failures



- The transition from fail-slow to fail-stop is rarely observed (i.e., at least not within 5 months).

Other Findings

- **Reoccurrences of slowdown events (§5.2.3)**
- **Impact factors**
  - Manufacturer (§5.2.3)
  - Drive age (§5.3.1)
  - Workload (§5.3.2)
- **SMART attributes are not good indicators of fail-slow (§5.3.3)**

More details in the paper!

Outline

~~INTRODUCTION~~

~~DATASET~~

~~FAIL-STOP~~

~~FAIL-SLOW~~

SUMMARY & TAKE-AWAY POINTS

USENIX ATC '22

Fail-stop failures

Major reliability changes in NVMe SSD (compared to SAS/SATA SSD):

- Infant mortality is not notable in NVMe SSD
- Write amplification
  - NVMe SSD becomes more robust to high write amplification (WAF>2)
  - Low write amplification (WAF≤1) is still rare-but-deadly (i.e., high failure rates)
- Spatially correlated failures (intra-node/rack)
  - Are temporally correlated in the long-term span (i.e., 1 day to 1 month)
  - Are no longer prevalent in the short-term span (i.e., 0 to 1 minute)

USENIX ATC '22

Fail-slow failures

The first large-scale study on fail-slow failures in storage devices.

- Fail-slow failure is widespread and severe in NVMe SSD
  - (Widespread) 1.41% infected within 4-month monitoring (up to 51X higher than HDD)
  - (Severe) Could degrade to SATA SSD or even HDD performance
- Impact factors
  - Manufacturer
  - Drive age
  - Workload
- SMART attributes exhibit negligible correlation with fail-slow metrics
- Fail-slow failures rarely transit to fail-stop failures (at least not within 5 months)

Thank you!

NVMe SSD Failures in the Field:

the Fail-Stop and the Fail-Slow

Ruiming Lu, Erci Xu, Yiming Zhang, Zhaosheng Zhu, Mengtian Wang, Zongpeng Zhu, Guangtao Xue, Minglu Li, Jiesheng Wu

Contact email: lrm318@sjtu.edu.cn