USENIX
ATC '19

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

Erci Xu, Mai Zheng, Feng Qin, Yikang Xu, Jiesheng Wu

淘宝网
Taobao.com
天猫 TMALL.COM

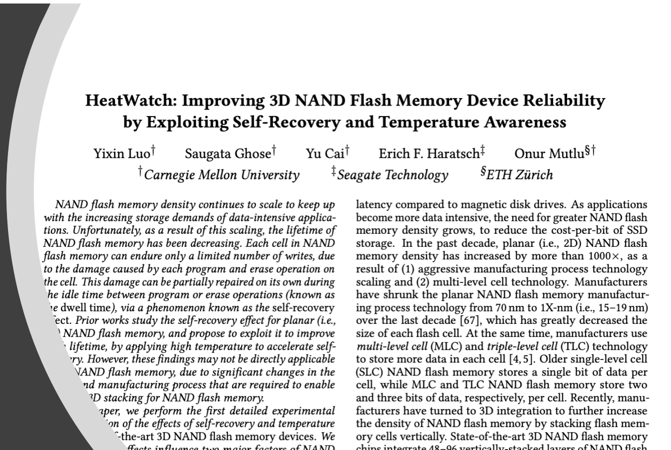
twitter

APACHE Spark

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



HeatWatch: Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness

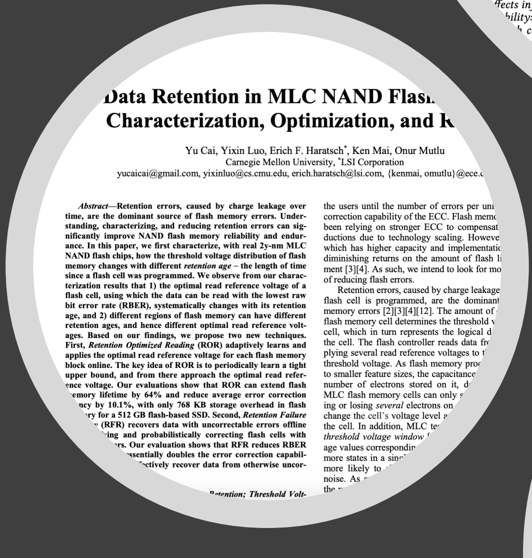
Yixin Luo[†] Saugata Ghose[†] Yu Cai[†] Erich F. Haratsch[‡] Onur Mutlu[§†]
[†]*Carnegie Mellon University* [‡]*Seagate Technology* [§]*ETH Zürich*

NAND flash memory density continues to scale to keep up with the increasing storage demands of data-intensive applications. Unfortunately, as a result of this scaling, the lifetime of NAND flash memory has been decreasing. Each cell in NAND flash memory can endure only a limited number of writes, due to the damage caused by each program and erase operation on the cell. This damage can be partially repaired on its own during *he idle time between program or erase operations (known as* *e* dwell time*), via a phenomenon known as the* self-recovery *fect. Prior works study the self-recovery effect for planar (i.e.,* *NAND flash memory, and propose to exploit it to improve* *lifetime, by applying high temperature to accelerate self-* *ry. However, these findings may not be directly applicable* *NAND flash memory, due to significant changes in the* *nd manufacturing process that are required to enable* *D stacking for NAND flash memory.*

aper, we perform the first detailed experimental *ion of the effects of self-recovery and temperature* *-the-art 3D NAND flash memory devices. We* *ffects influence two major factors of NAND* *bility: (1)* retention loss speed *(i.e., the* *cell leaks charge), and (2)* program *e in programming speed across flash* *very and temperature affect 3D* *ently than they affect planar* *r models of self-recovery* *flash memory. Using* *model for 3D* *etention,*

latency compared to magnetic disk drives. As applications become more data intensive, the need for greater NAND flash memory density grows, to reduce the cost-per-bit of SSD storage. In the past decade, planar (i.e., 2D) NAND flash memory density has increased by more than 1000×, as a result of (1) aggressive manufacturing process technology scaling and (2) multi-level cell technology. Manufacturers have shrunk the planar NAND flash memory manufacturing process technology from 70 nm to 1X-nm (i.e., 15–19 nm) over the last decade [67], which has greatly decreased the size of each flash cell. At the same time, manufacturers use *multi-level cell* (MLC) and *triple-level cell* (TLC) technology to store more data in each cell [4, 5]. Older single-level cell (SLC) NAND flash memory stores a single bit of data per cell, while MLC and TLC NAND flash memory store two and three bits of data, respectively, per cell. Recently, manufacturers have turned to 3D integration to further increase the density of NAND flash memory by stacking flash memory cells vertically. State-of-the-art 3D NAND flash memory chips integrate 48–96 vertically-stacked layers of NAND flash memory [23, 34, 36, 54, 61, 66].

This rapid increase in NAND flash memory density has come at the cost of reduced reliability [4, 5, 11, 44, 50, 58]. NAND flash memory has a limited *lifetime*, which is defined as the number of program and erase operations (known as *P/E cycles*) that can be reliably performed on each flash cell whil avoiding data loss for a minimum data retention time as anteed by vendors [4, 5, 11]. As the manufacturin technology scales, the lifetime has reduced cycles for 70 nm planar NAND fl



Data Retention in MLC NAND Flas Characterization, Optimization, and R

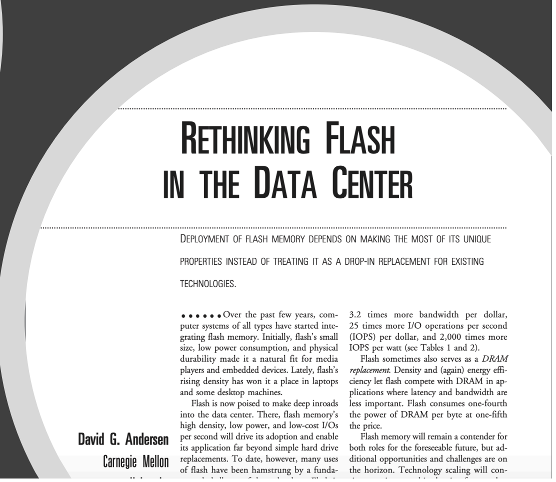
Yu Cai, Yixin Luo, Erich F. Haratsch[*], Ken Mai, Onur Mutlu
Carnegie Mellon University, [*]LSI Corporation
yucaicai@gmail.com, yixinluo@cs.cmu.edu, erich.haratsch@lsi.com, {kenmai, omutlu}@ece.c

***Abstract*—Retention errors, caused by charge leakage over time, are the dominant source of flash memory errors. Understanding, characterizing, and reducing retention errors can significantly improve NAND flash memory reliability and endurance. In this paper, we first characterize, with real 2y-nm MLC NAND flash chips, how the threshold voltage distribution of flash memory changes with different *retention age* – the length of time since a flash cell was programmed. We observe from our characterization results that 1) the optimal read reference voltage of a flash cell, using which the data can be read with the lowest raw bit error rate (RBER), systematically changes with its retention age, and 2) different regions of flash memory can have different retention ages, and hence different optimal read reference voltages. Based on our findings, we propose two new techniques. First, *Retention Optimized Reading* (ROR) adaptively learns and applies the optimal read reference voltage for each flash memory block online. The key idea of ROR is to periodically learn a tight upper bound, and from there approach the optimal read reference voltage. Our evaluations show that ROR can extend flash emory lifetime by 64% and reduce average error correction ncy by 10.1%, with only 768 KB storage overhead in flash ry for a 512 GB flash-based SSD. Second, *Retention Failure* (RFR) recovers data with uncorrectable errors offline ing and probabilistically correcting flash cells with rs. Our evaluation shows that RFR reduces RBER ssentially doubles the error correction capabil- ctively recover data from otherwise uncor-**

etention; Threshold Volt-

the users until the number of errors per un correction capability of the ECC. Flash memo been relying on stronger ECC to compensat ductions due to technology scaling. Howeve which has higher capacity and implementati diminishing returns on the amount of flash l ment [3][4]. As such, we intend to look for mo of reducing flash errors.

Retention errors, caused by charge leakage flash cell is programmed, are the dominant memory errors [2][3][4][12]. The amount of flash memory cell determines the threshold v cell, which in turn represents the logical d the cell. The flash controller reads data fr plying several read reference voltages to t threshold voltage. As flash memory pro to smaller feature sizes, the capacitance number of electrons stored on it, d MLC flash memory cells can only s ing or losing *several* electrons on change the cell's voltage level a the cell. In addition, MLC te *threshold voltage window* age values correspondin more states in a sin more likely to noise. As the



RETHINKING FLASH IN THE DATA CENTER

DEPLOYMENT OF FLASH MEMORY DEPENDS ON MAKING THE MOST OF ITS UNIQUE PROPERTIES INSTEAD OF TREATING IT AS A DROP-IN REPLACEMENT FOR EXISTING TECHNOLOGIES.

David G. Andersen
Carnegie Mellon

Over the past few years, computer systems of all types have started integrating flash memory. Initially, flash's small size, low power consumption, and physical durability made it a natural fit for media players and embedded devices. Lately, flash's rising density has won it a place in laptops and some desktop machines.

Flash is now poised to make deep inroads into the data center. There, flash memory's high density, low power, and low-cost I/Os per second will drive its adoption and enable its application far beyond simple hard drive replacements. To date, however, many uses of flash have been hamstrung by a funda-

3.2 times more bandwidth per dollar, 25 times more I/O operations per second (IOPS) per dollar, and 2,000 times more IOPS per watt (see Tables 1 and 2).

Flash sometimes also serves as a *DRAM replacement*. Density and (again) energy efficiency let flash compete with DRAM in applications where latency and bandwidth are less important. Flash consumes one-fourth the power of DRAM per byte at one-fifth the price.

Flash memory will remain a contender for both roles for the foreseeable future, but additional opportunities and challenges are on the horizon. Technology scaling will con-

Previous Large Scale SSD Studies

- E.g.:
  - Failure rate curve
    - not bathtub
  - FTL impact
    - Thermal Throttling
  - Uncorrectable errors

Flash Reliability in Production: The Expected and the Unexpected

Bianca Schroeder
University of Toronto
Toronto, Canada

Raghav Lagisetty
Google Inc.
Mountain View, CA

Arif Merchant
Google Inc.
Mountain View, CA

Abstract

As solid state drives based on flash technology are becoming a staple for persistent data storage in data centers, it is important to understand their reliability characteristics. While there is a large body of work based on experiments with individual flash chips in a controlled lab environment under synthetic workloads, there is a dearth of information on their behavior in the field. This paper provides a large-scale field study covering many millions of drive days, ten different drive models, different flash technologies (MLC, eMLC, 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 tes (RBER) grow at a much slower rate with wear-out the exponential rate commonly assumed and, more tly, they are not predictive of uncorrectable er- error modes. The widely used metric UBER error rate) is not a meaningful metric, tion between the number of reads ble errors. We see no evi- re reliable than

bility in controlled lab experiments (such as accelerated life tests), using a small population of raw flash chips under synthetic workloads. There is a dearth of studies that report on the reliability of flash drives and their failure characteristics in large-scale production use in the field.

This paper provides a detailed field study of flash reliability based on data collected over 6 years of production use in Google's data centers. The data spans many millions of drive days [1], ten different drive models, different flash technologies (MLC, eMLC and SLC) and feature sizes (ranging from 24nm to 50nm). We use this data to provide a better understanding of flash reliability in production. In particular, our contributions include a d tailed analysis of the following aspects of flash reliab in the field:

1. The different types of errors experienc drives and their frequency in the field
2. Raw bit error rates (RBER), ho by factors such as wear-out their relationship wit tion 4).
3. U

ge-Scale Study of Flash Memory Failures in

Justin Meza
negie Mellon University
meza@cmu.edu

Qiang Wu
Facebook, Inc.
qwu@fb.com

Sanjeev Kumar
Facebook, Inc.
skumar@fb.com

Onur Mu
Carnegie Mellon U
onur@cmu.e

BSTRACT

ervers use flash memory based solid state drives (SSDs) as a high-performance alternative to hard disk drives to store persistent data. Unfortunately, recent increases in flash density have also brought about decreases in chip-level reliability. In a data center environment, flash-based SSD failures can lead to downtime and, in the worst case, data loss. As a result, it is important to understand flash memory reliability characteristics over flash lifetime in a realistic production data center environment running modern applications and system software.

This paper presents the first large-scale study of flash-based SSD reliability in the field. We analyze data collected across majority of flash-based solid state drives at Facebook data enters over nearly four years and many millions of operational urs in order to understand failure properties and trends of h-based SSDs. Our study considers a variety of SSD char- ristics, including: the amount of data written to and read lash chips; how data is mapped within the SSD address he amount of data copied, erased, and discarded by the troller; and flash board temperature and bus power. our field analysis of how flash memory errors man- nning modern workloads on modern SSDs, this rst to make several major observations: (1) do *not* increase monotonically with *flash* hey go through several distinct periods ailures emerge and are subsequently f read disturbance errors are *not* logical data layout across an n-contiguous data), as

Categories and Subject Descriptors

B.3.4. [**Hardware**]: Memory Structures—*Reliability, ing, and Fault-Tolerance*

Keywords

flash memory; reliability; warehouse-scale data centers

1. INTRODUCTION

Servers use flash memory for persistent storage due to low access latency of flash chips compared to hard disk dr Historically, flash capacity has lagged behind hard disk capacity, limiting the use of flash memory. In the past de however, advances in NAND flash memory technology increased flash capacity by more than 1000×. This ra crease in flash capacity has brought both an increase memory use and a decrease in flash memory reliab example, the number of times that a cell can be re grammed and erased before wearing out and faili from 10,000 times for 50 nm cells to only 2,000 ti cells [28]. This trend is expected to continue erations of flash memory. Therefore, if we the operational lifetime and reliability of f devices, we must first fully understand t istics.

In the past, a large body of pric characteristics of flash cells i small numbers e.g., tens) 27, 22, 25, 16, 33, 1 10, 7, 13, 9,

SSD Failures in Datacenters: What? When? and

Iyswarya Narayanan[*], Di Wang[†], Myeongjae Jeon[†], Bikash Sharma[†], Laura Caulfie
Anand Sivasubramaniam[*], Ben Cutler[†], Jie Liu[†], Badriddine Khessib[†], Kushagra Vai
[*]The Pennsylvania State University, [†]Microsoft Corporation
[*]{iun106,anand}@cse.psu.edu,
[†]{wangdi,myeoje,bsharma,laura.caulfield,bcutler,jie.liu,bkhessib,kushagra.vaid}@microsoft.com

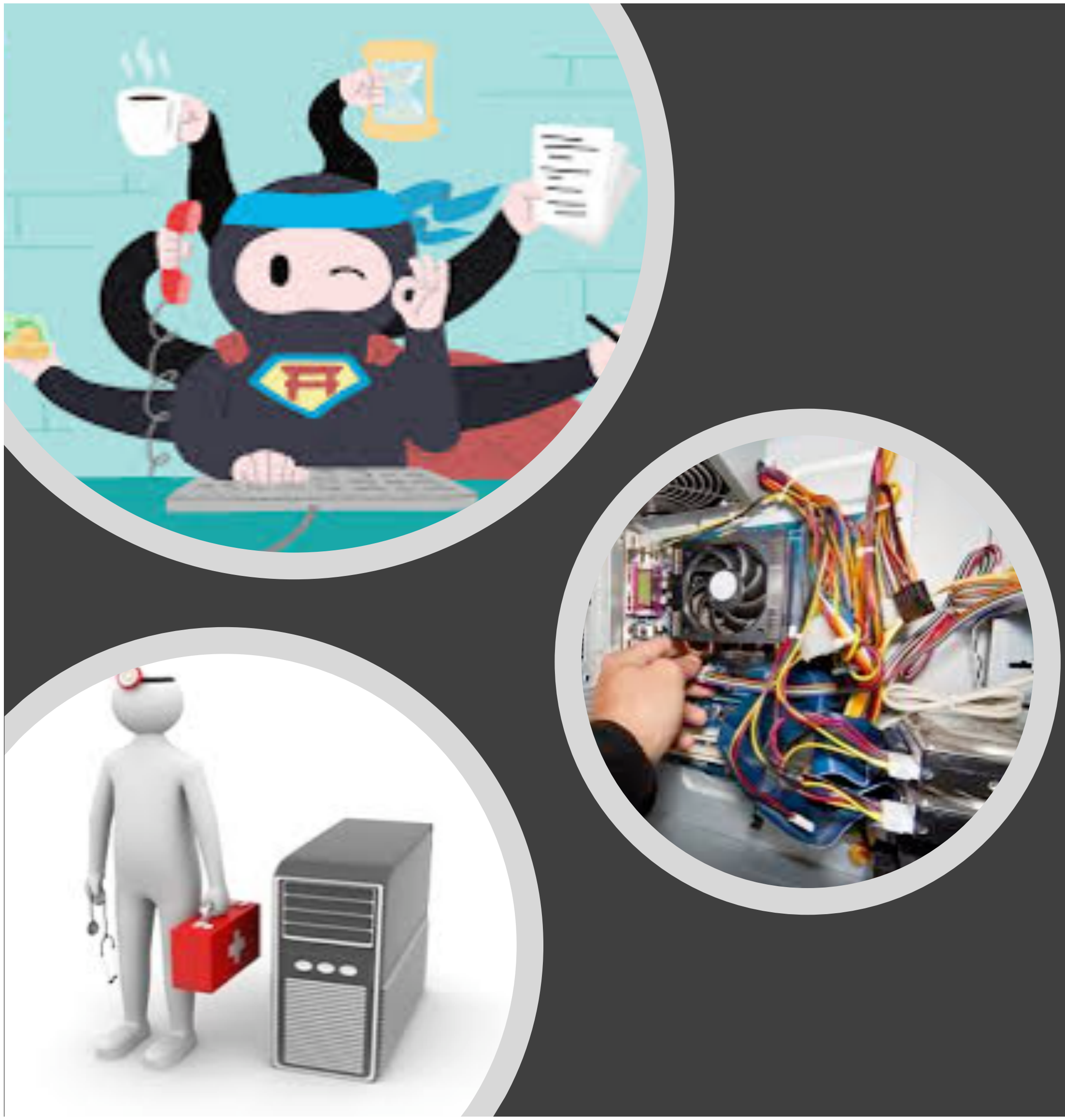
Abstract

Despite the growing popularity of Solid State Disks (SSDs) in the datacenter, little is known about their reliability characteristics in the field. The little knowledge is mainly vendor supplied, and such information cannot really help understand how SSD failures can manifest and impact the operation of production systems, in order to take appropriate remedial measures. Besides actual failure data and the symptoms exhibited by SSDs before failing, a detailed characterization effort requires wide set of data about factors influencing SSD failures, right from provisioning factors to the operational ones. This paper presents an extensive SSD failure characterization by analyzing a wide spectrum of data from over half a million SSDs that span multiple generations spread across several datacenters which host a wide spectrum of workloads over nearly 3 years. By studying the diverse set of design, provisioning and operational factors on failures, and their symptoms, our work provides the first comprehensive analysis of the what, when and why characteristics of SSD failures in production datacenters.

the associated downtime to fix the problem and/or replac the device. It can even take several days to repair/replace storage component after its failure, with associated serve being unusable during this period. To account for this downtime, datacenters resort to over-provisioning (which can add significant cost) in order to meet the desired application availability Service Level Agreements (SLAs).

In the storage stack, SSDs are obviously at an advantage compared to HDDs in terms of failure rates. However, (i) SSDs are between 4X-40X costlier per GB than HDDs, depending on their grade (neutralizing, and in fac out-weighing the lower failure rate advantage); and (ii) a SSD-related failure ticket in our dataset results in a replac ment 79% of the time compared to 11% for HDD-rela tickets (i.e. SSD related failure tickets are more critica the datacenter). These factors, together with rapid adoption[3, 13], motivate us to understand SSD reliab

The current knowledge on SSD failure rate is ily vendor supplied, based on accelerated lab testi controlled conditions. In addition to the paramete tested for, numerous other factors in a product



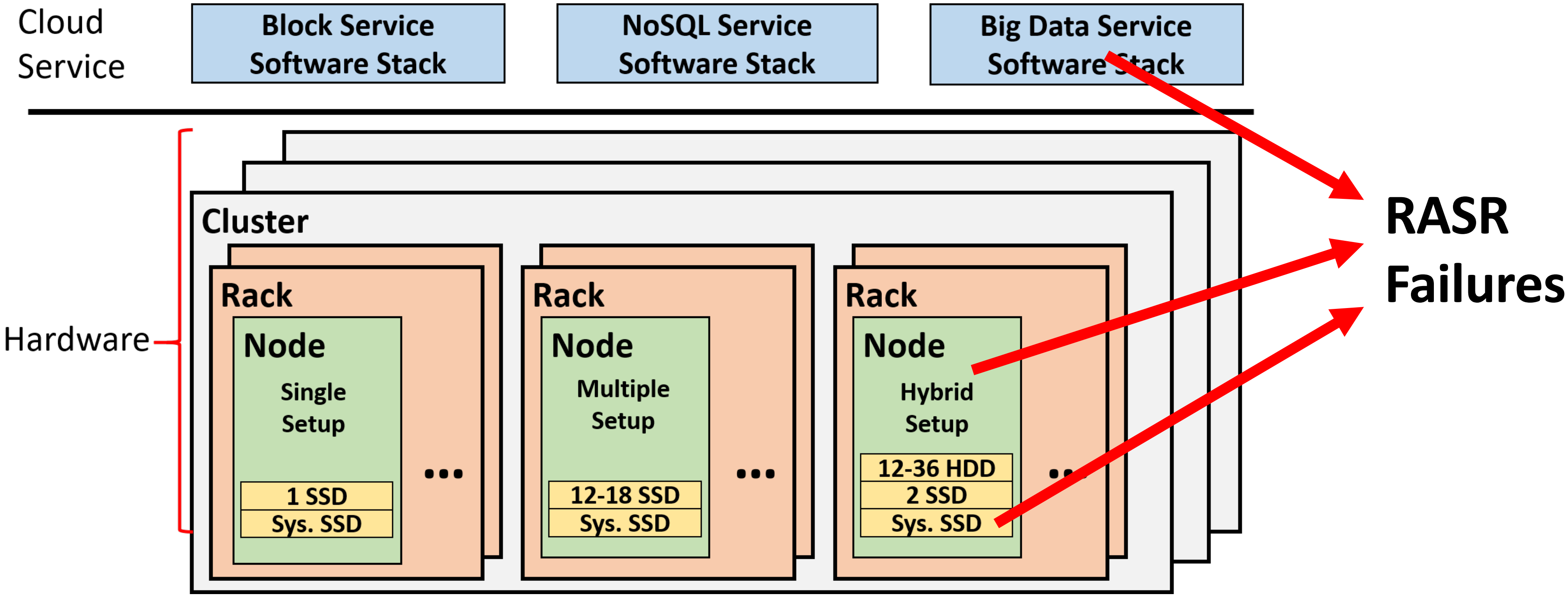
Our Study

- 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

Hardware

Cluster

Rack

Node

Single Setup

1 SSD

Sys. SSD

…

Rack

Node

Multiple Setup

12-18 SSD

Sys. SSD

…

Rack

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

Model	Capacity	Lithography	Age
M1	480GB	20nm	2-3 yrs
M2	800GB	20nm	2-3 yrs
M3	480GB	16nm	1-2 yrs
M4	480GB	20nm	2-3 yrs
M5	480GB	20nm	1-2 yrs

different SSD models

Service	Function
Block Service	Journaling
	Persistence
NoSQL	Journaling
	Persistence
Big Data	Temporary

different SSD usages

Outline

~~INTRODUCTION~~

~~SYSTEM ARCHITECTURE & DATASET~~

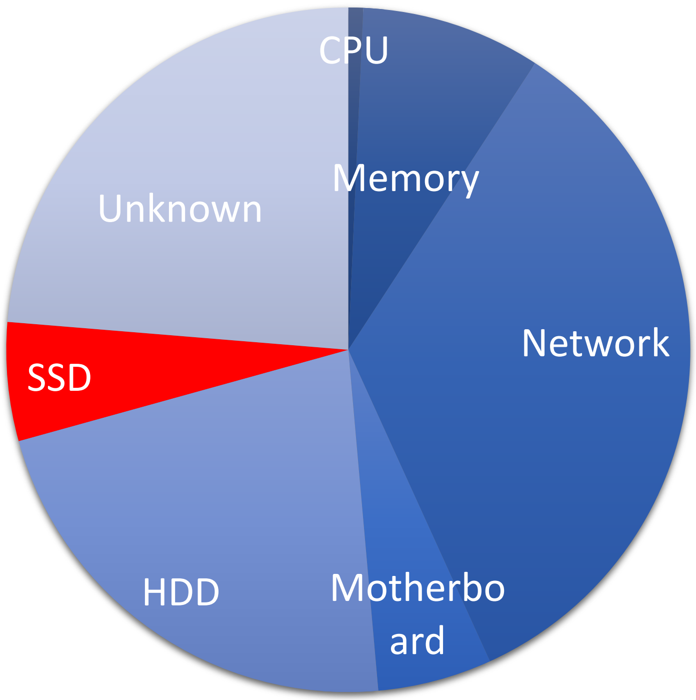
RASR FAILURES OVERVIEW

LESSONS AND ACTIONS

CONCLUSIONS & FUTURE WORK

RASR Failures Overview

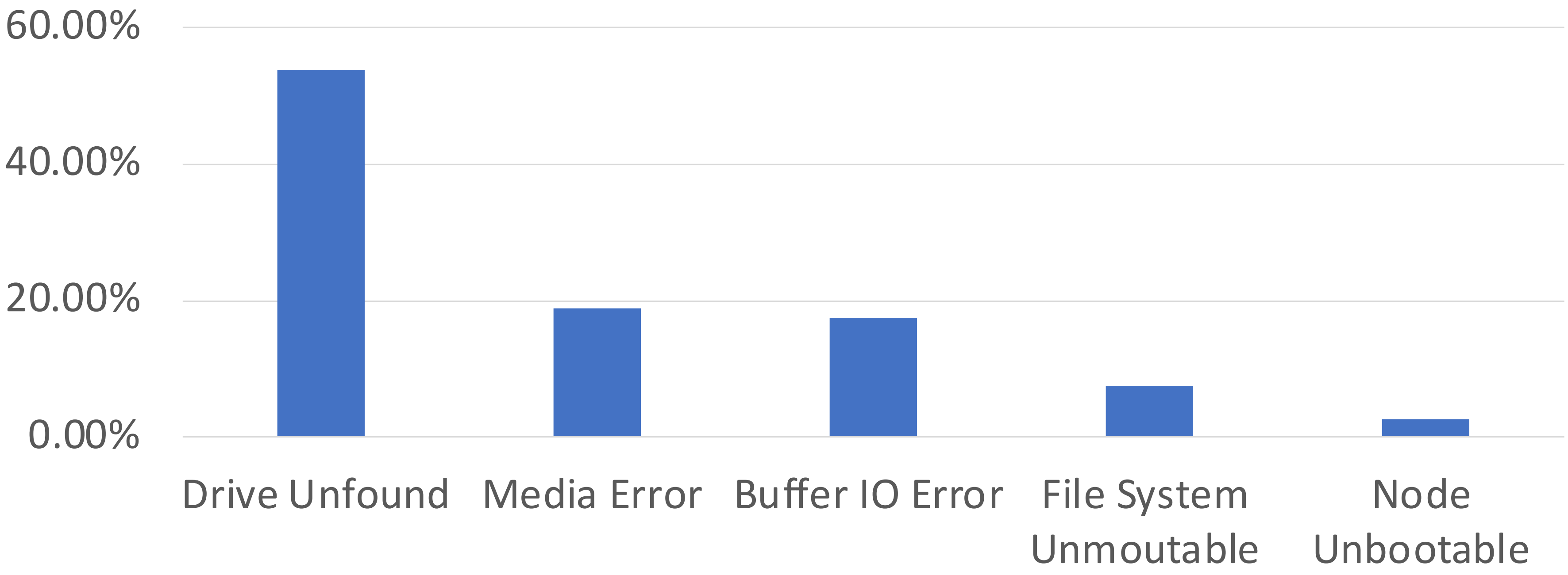
- We have collected around 130K failure tickets over 3 years
- Around 6% of them is RASR Failures. Around 10K events.



RASR Failures

RASR Failures Overview (cont.)

- 5 Symptoms of RASR Failures



Outline

~~INTRODUCTION~~

~~SYSTEM ARCHITECTURE & DATASET~~

~~RASR FAILURES OVERVIEW~~

LESSONS AND ACTIONS

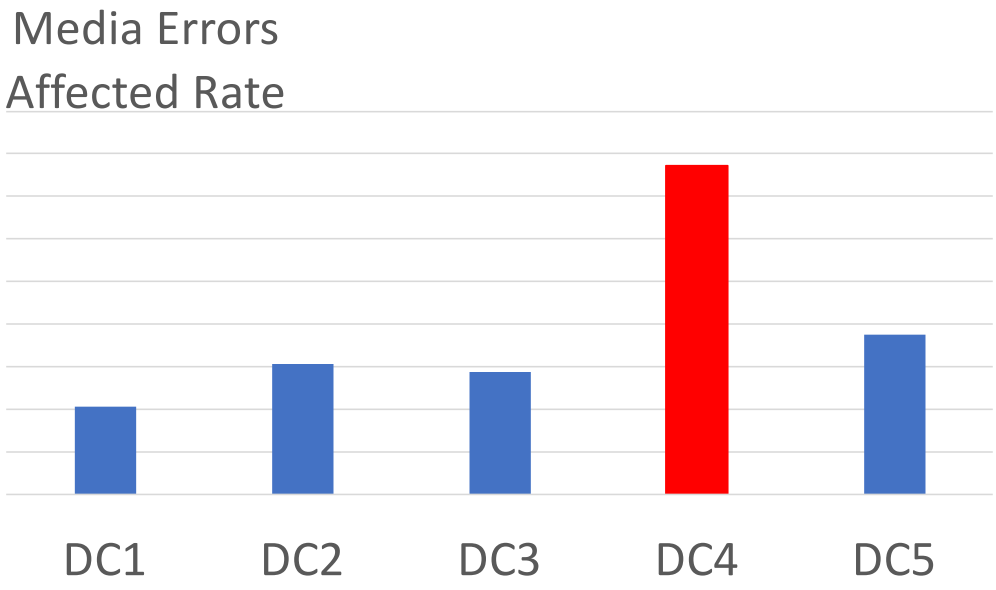
CONCLUSIONS & FUTURE WORK

L&A for Hardware Architects

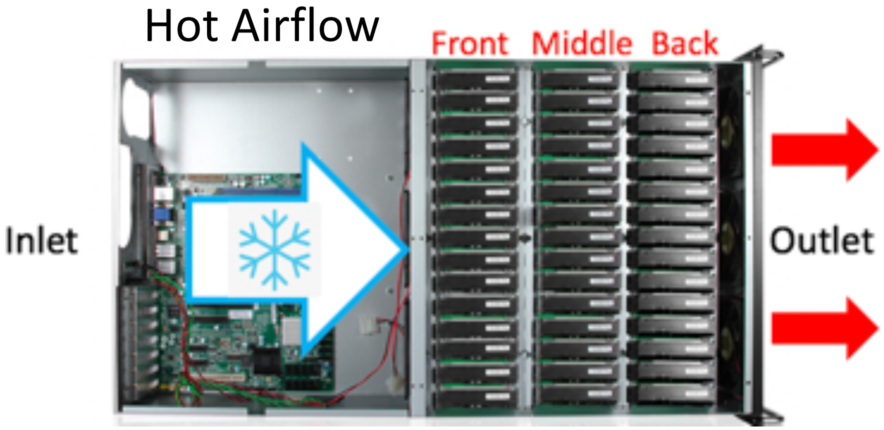
- One DC in our deployment has higher-than-usual Media Error affected rate.

  Under same drive model

  Under same cloud service

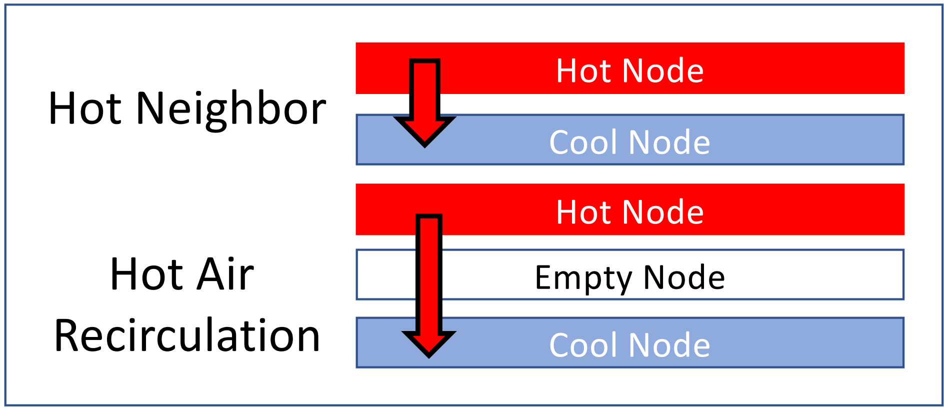


Passive Heating in Hardware Architecture



Intra-node Stacking

SSD Storage Rack


Hot Neighbor

Hot Node

Cool Node

Hot Air Recirculation

Hot Node

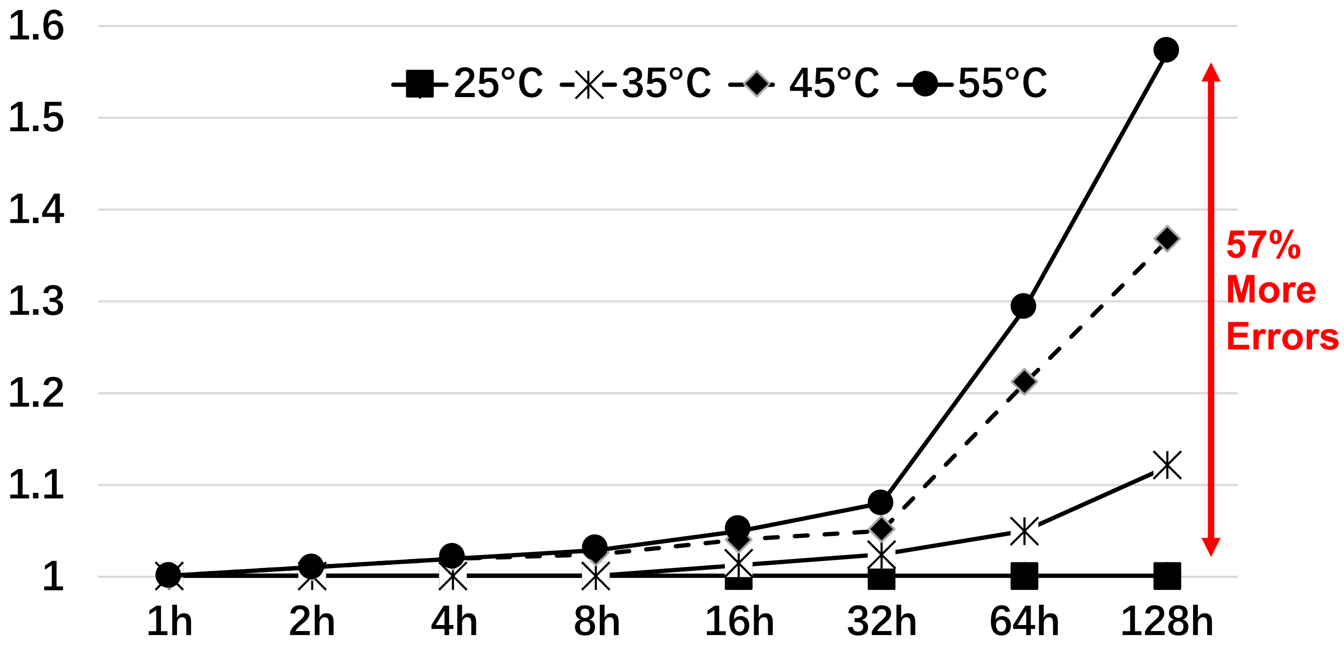
Empty Node

Cool Node

Inter-node Stacking

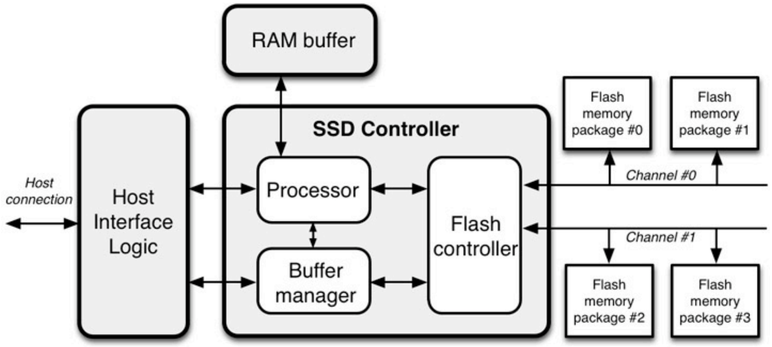
Passive Heating: Heating on *idle* SSDs by neighboring active SSDs

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

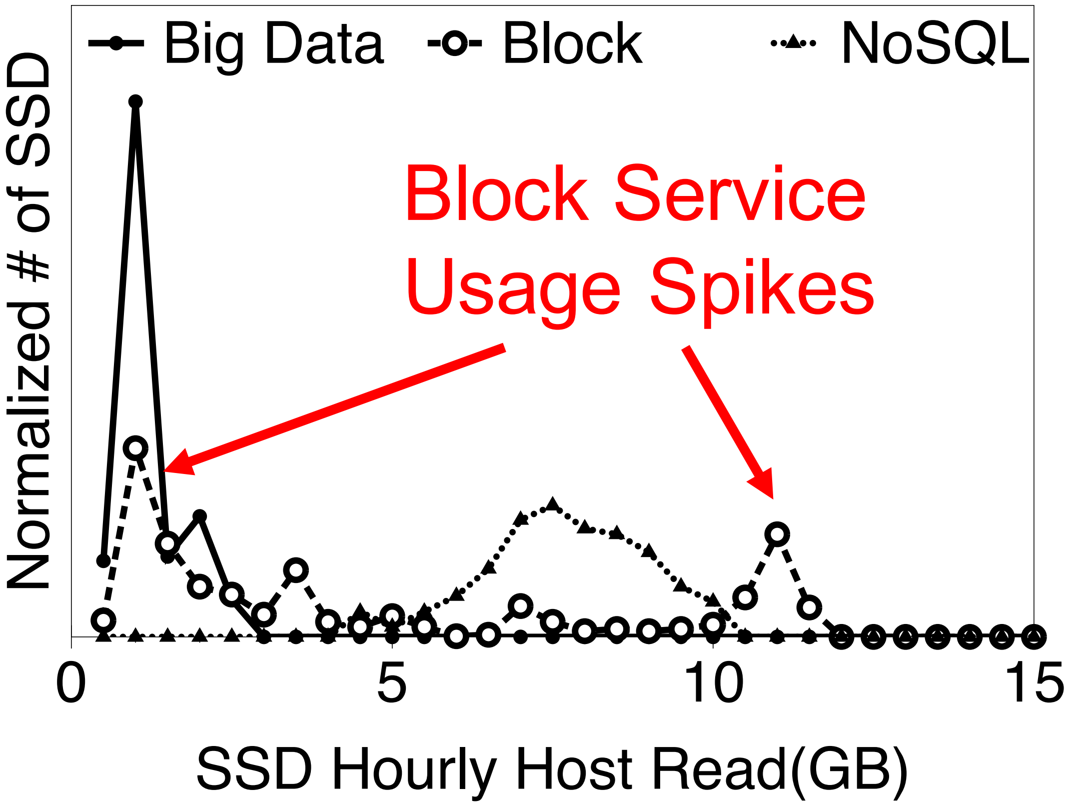
- Certain cloud services may cause unbalanced usage of SSDs

	service	Host Read	Host Write
Average Value Per Hour	Block	7.69GB	6.56GB
	Big Data	1.57GB	1.22GB
	NoSQL	6.10GB	5.28GB
Coefficient of Variance	Block	35.5%	24.9%
	Big Data	1.8%	3.7%
	NoSQL	3.2%	6.2%

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

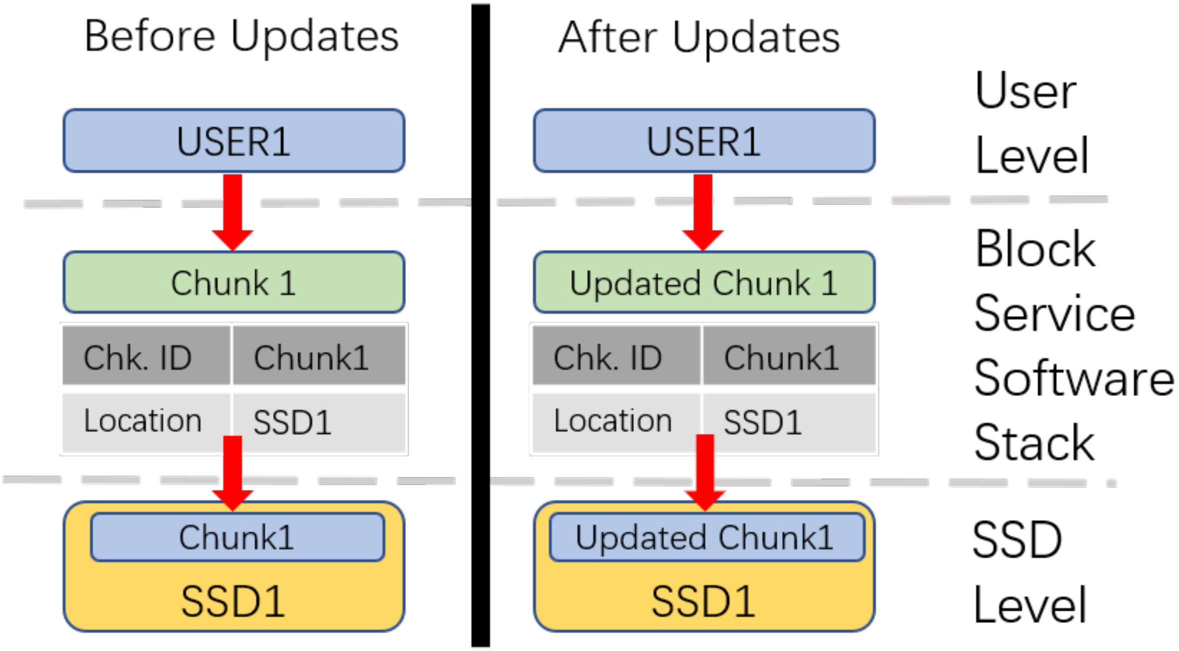
Service Imbalance

- 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.



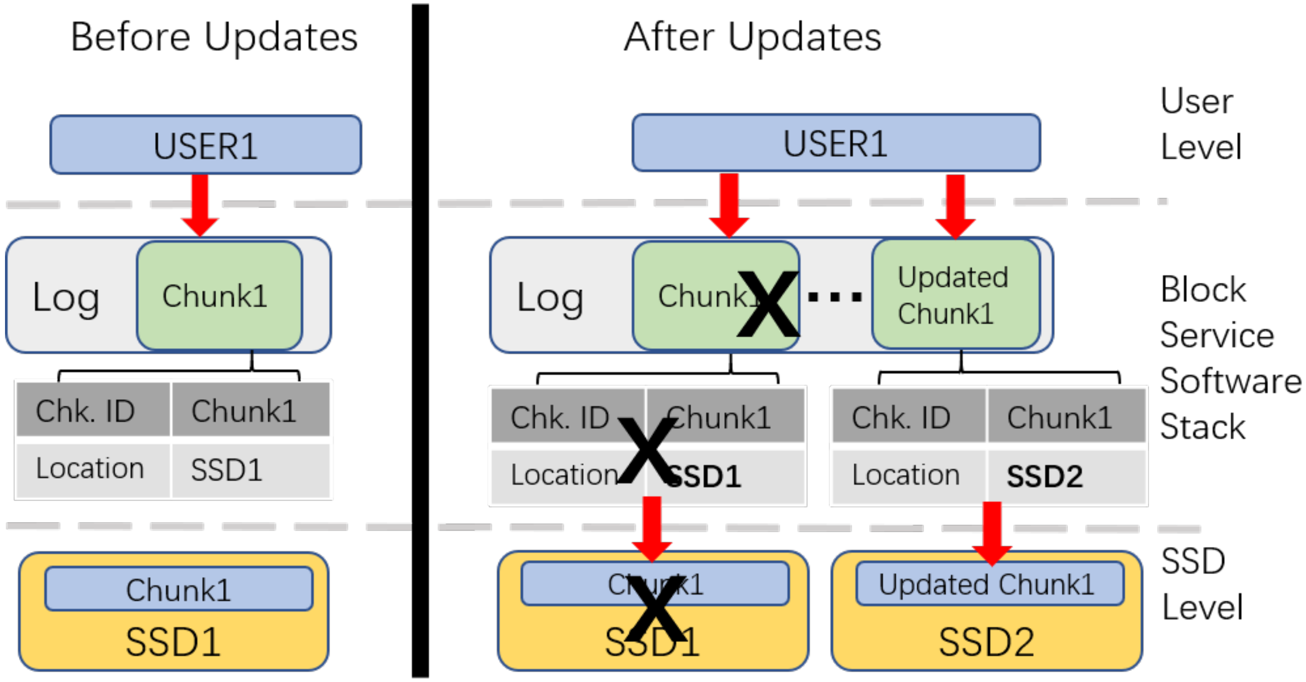
Block Service Usage Spikes

Root Causes: In-place Update Scheme



The updated chunk always write back to the same SSD.

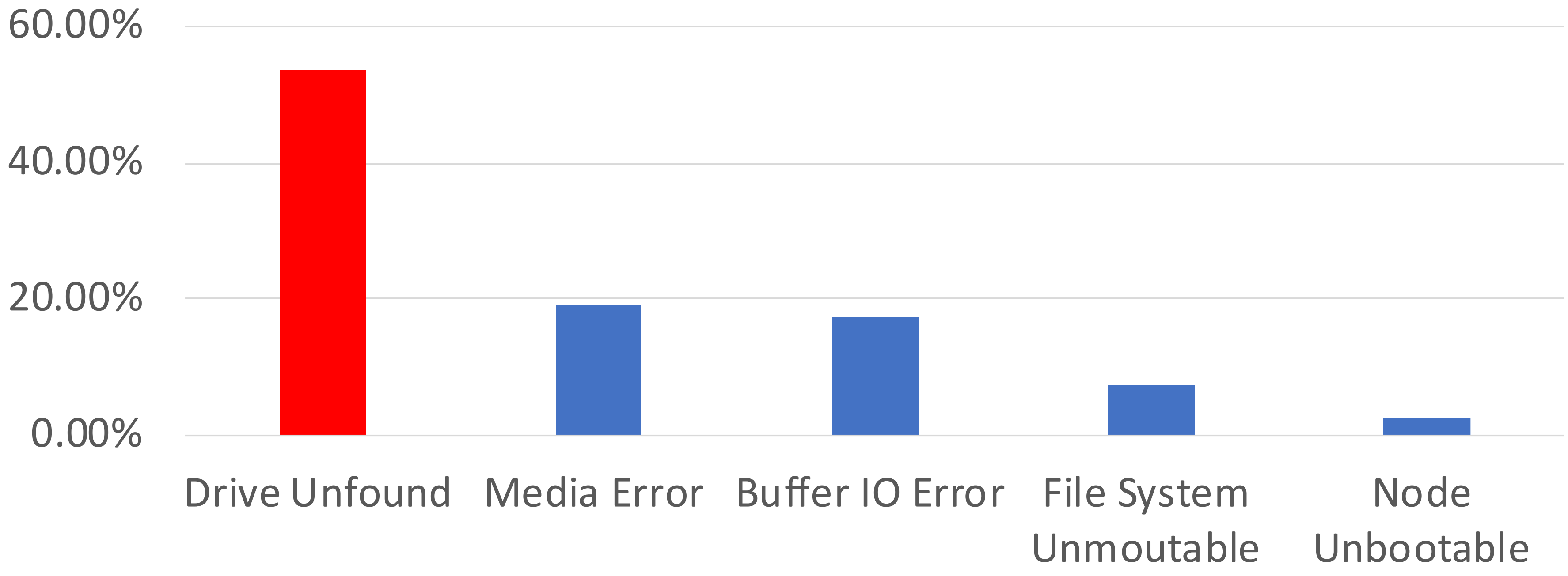
Solutions: Share-log Design



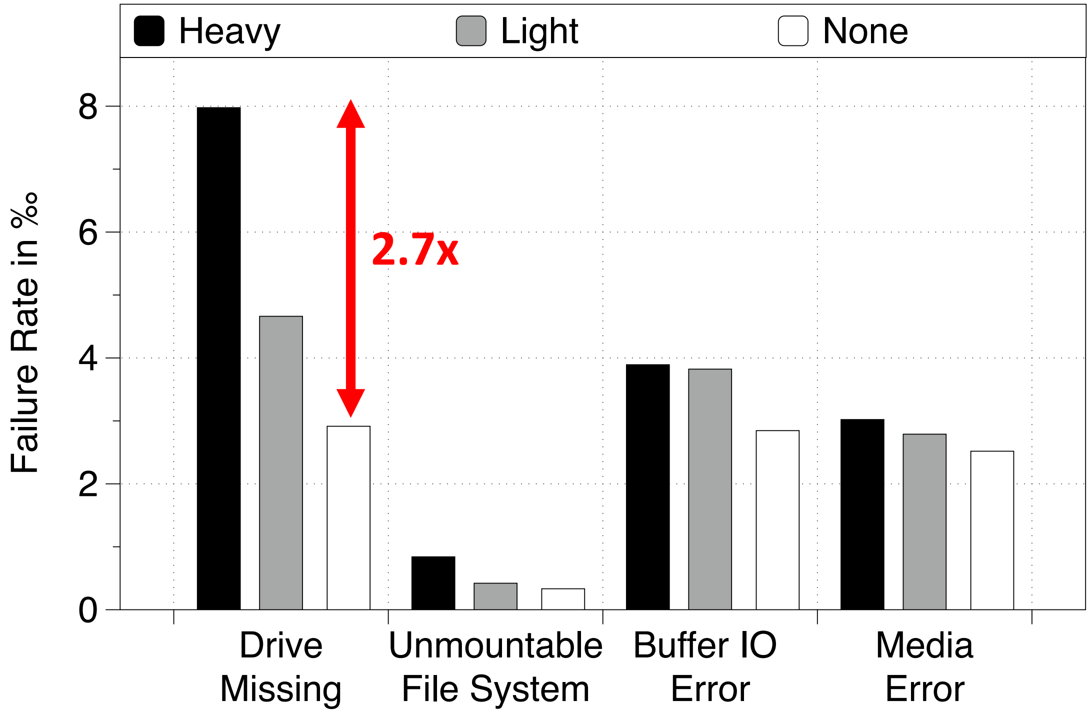
The updated chunk is re-allocated to a *new SSD*.

L&A for System Admins: Part I

- 5 Symptoms of RASR Failures



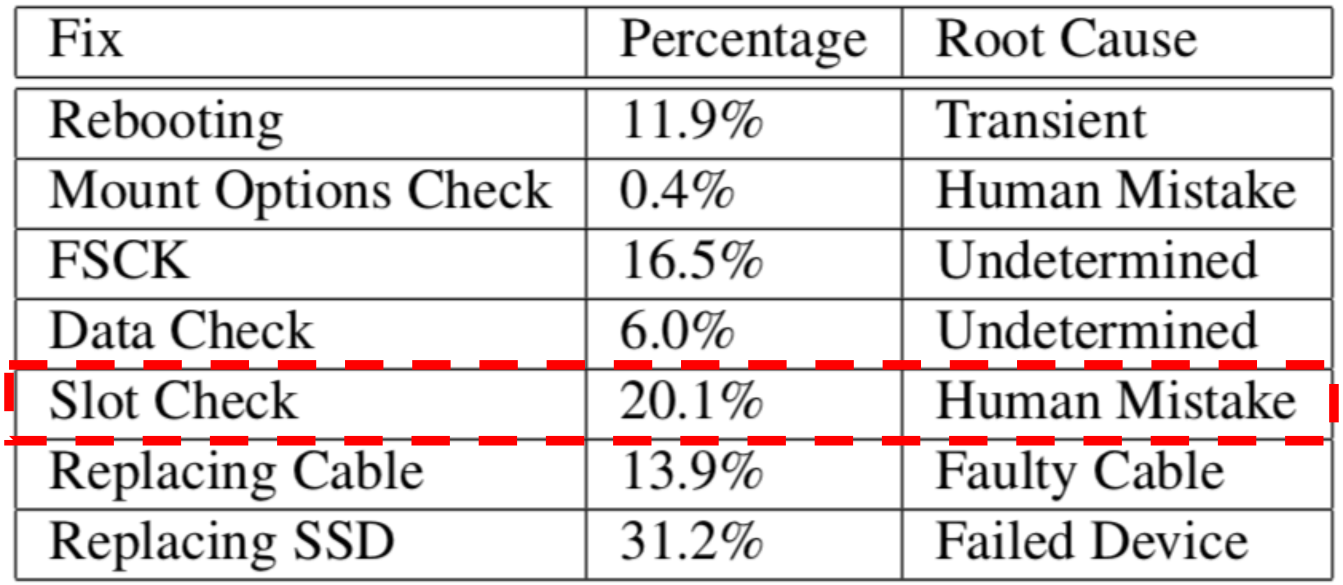
UCRC errors indicate bad cables



SSDs with heavy UCRC errors are 2.7X more likely to lead to "Drive Unfound" failures

L&A for System Admins: Part II

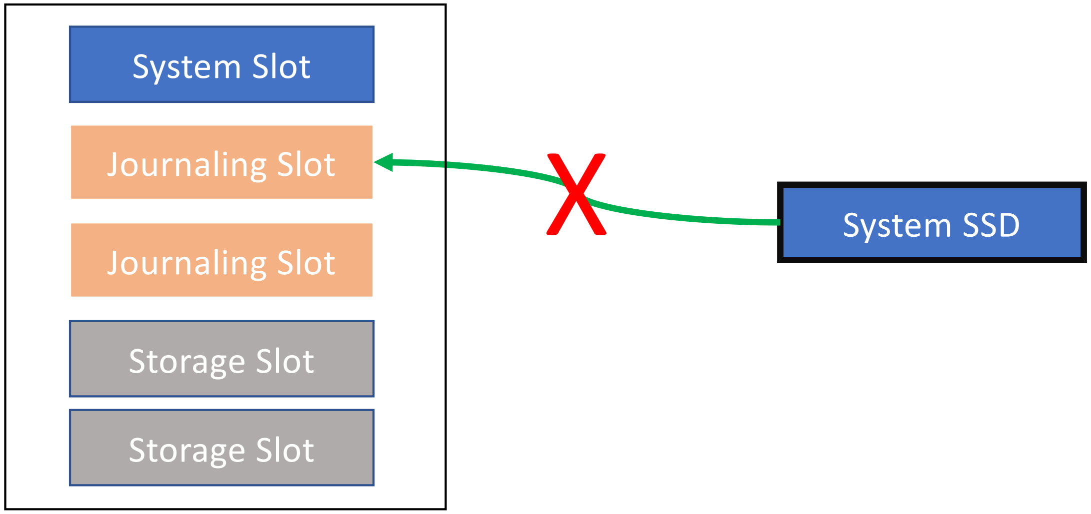
- How to quickly identify root cause of failures?



Fix	Percentage	Root Cause
Rebooting	11.9%	Transient
Mount Options Check	0.4%	Human Mistake
FSCK	16.5%	Undetermined
Data Check	6.0%	Undetermined
Slot Check	20.1%	Human Mistake
Replacing Cable	13.9%	Faulty Cable
Replacing SSD	31.2%	Failed Device

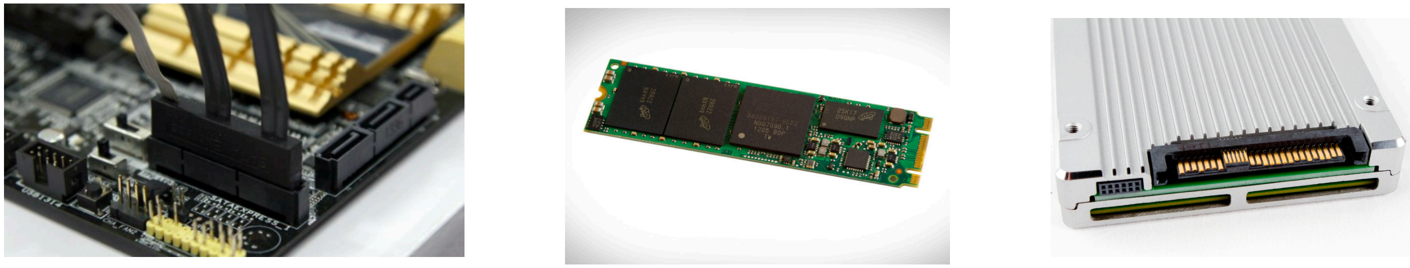
L&A for System Admins: Part II

- 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/

Outline

~~INTRODUCTION~~

~~SYSTEM ARCHITECTURE & DATASET~~

~~RASR FAILURES OVERVIEW~~

~~LESSONS AND ACTIONS~~

CONCLUSIONS & FUTURE WORK

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