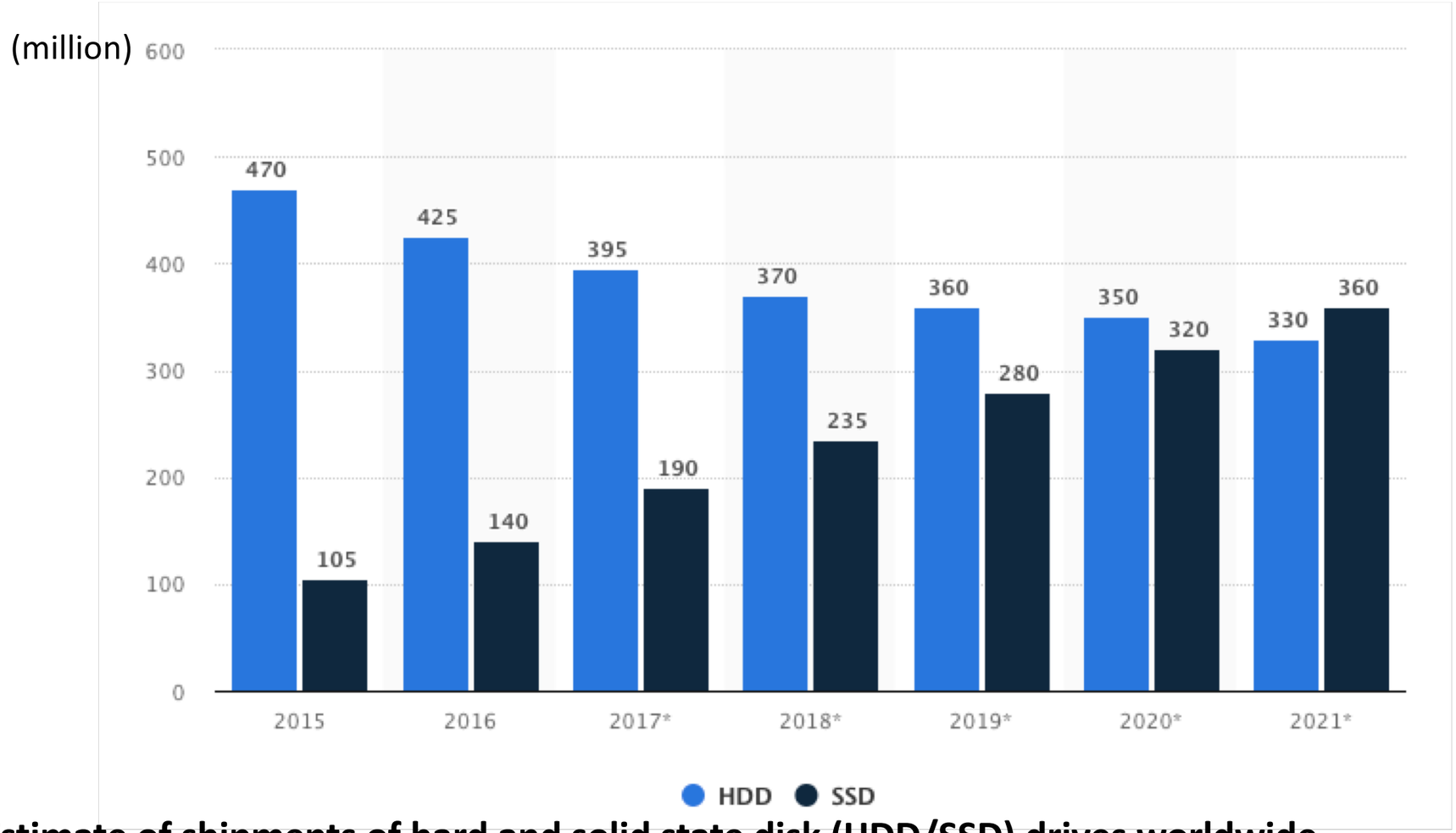
pdsw-discs

Understanding SSD Reliability in Large-Scale Cloud Systems

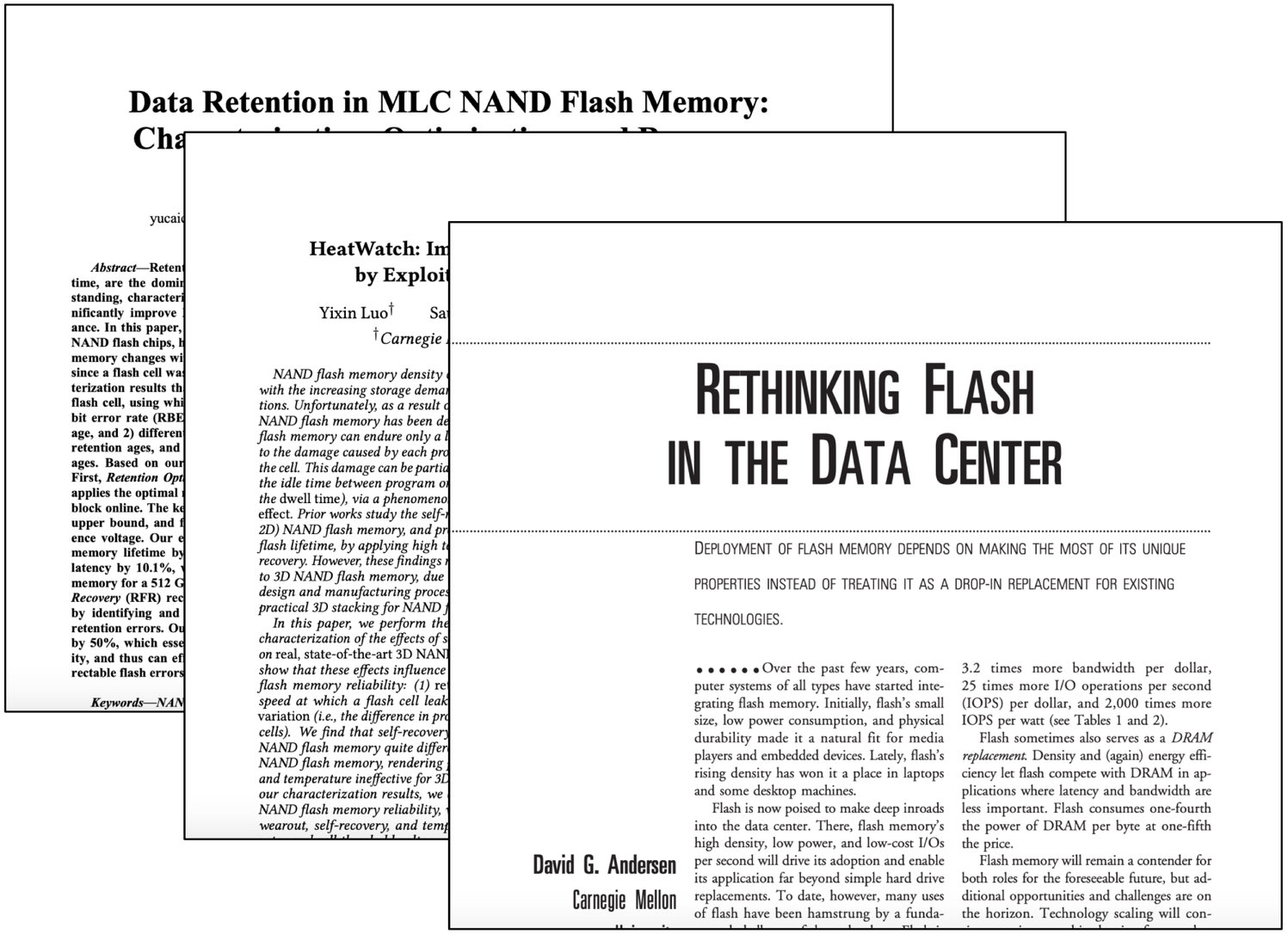
Erci Xu	**Mai Zheng**	Feng Qin	Yikang Xu	Jiesheng Wu
Ohio State University	**Iowa State University**	Ohio State University	Aliyun Alibaba	Aliyun Alibaba

Flash-Based Solid-Stata Drives (SSDs) are more and more popular



Estimate of shipments of hard and solid state disk (HDD/SSD) drives worldwide
https://www.statista.com/statistics/285474/hdds-and-ssds-in-pcs-global-shipments-2012-2017/

Concerns of SSD Reliability



Data Retention in MLC NAND Flash Memory:

yucai

Abstract—Reten
time, are the domi
standing, characteri
nificantly improve
ance. In this paper,
NAND flash chips, b
memory changes wi
since a flash cell wa
terization results th
flash cell, using whi
bit error rate (RBE
age, and 2) differen
retention ages, and
ages. Based on our
First, *Retention Opt*
applies the optimal
block online. The ke
upper bound, and f
ence voltage. Our e
memory lifetime by
latency by 10.1%,
memory for a 512 G
Recovery (RFR) rec
by identifying and
retention errors. Ou
by 50%, which esse
ity, and thus can ef
rectable flash errors

Keywords—NAN

HeatWatch: Im
by Exploit

Yixin Luo† Sa
†*Carnegie*

NAND flash memory density
with the increasing storage dema
tions. Unfortunately, as a result
NAND flash memory has been de
flash memory can endure only a
to the damage caused by each pro
the cell. This damage can be partia
the idle time between program o
the dwell time*), via a phenomeno*
effect. *Prior works study the self-*
2D) NAND flash memory, and pr
flash lifetime, by applying high t
recovery. However, these findings
to 3D NAND flash memory, due
design and manufacturing proces
practical 3D stacking for NAND
In this paper, we perform the
characterization of the effects of s
on real, state-of-the-art 3D NAN
show that these effects influence
flash memory reliability: (1) re
speed at which a flash cell leak
variation *(i.e., the difference in pr*
cells). We find that self-recovery
NAND flash memory quite differ
NAND flash memory, rendering
and temperature ineffective for 3D
our characterization results, we
NAND flash memory reliability,
wearout, self-recovery, and temp

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-

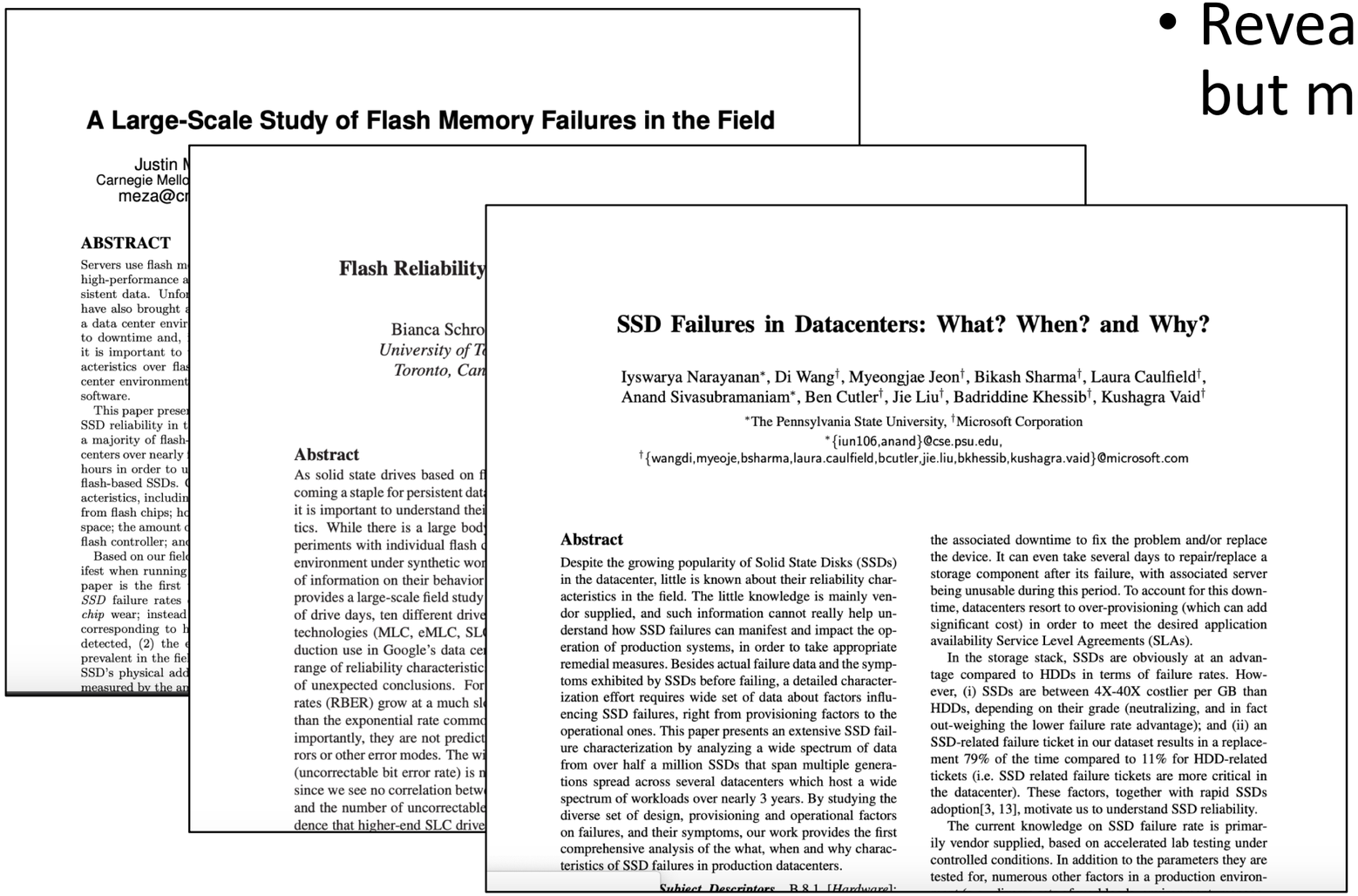
3.2 times more bandwidth per dollar, 25 times more I/O operations per second (IOPS) per dollar, and 2,000 times more IOPS per watt (see Tables 1 and 2).

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-

- Wear out
  - Limited Program/Erase Cycles
- New failure modes
  - Program/Erase Error
  - Metadata corruption
- Sensitive to environment
  - NAND in heated air

Previous Large Scale SSD Studies



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

Justin M
Carnegie Mello
meza@cm

ABSTRACT

Servers use flash m
high-performance a
sistent data. Unfor
have also brought a
a data center envir
to downtime and,
it is important to
acteristics over flas
center environment
software.

This paper preser
SSD reliability in t
a majority of flash-
centers over nearly
hours in order to u
flash-based SSDs.
acteristics, includin
from flash chips; ho
space; the amount
flash controller; and

Based on our field
ifest when running
paper is the first
SSD failure rates
chip wear; instead
corresponding to h
detected, (2) the
prevalent in the fie
SSD's physical add
measured by the an

Flash Reliability

Bianca Schro
University of T
Toronto, Can

Abstract

As solid state drives based on fl
coming a staple for persistent dat
it is important to understand thei
tics. While there is a large body
periments with individual flash d
environment under synthetic wor
of information on their behavior
provides a large-scale field study
of drive days, ten different drive
technologies (MLC, eMLC, SL
duction use in Google's data ce
range of reliability characteristic
of unexpected conclusions. For
rates (RBER) grow at a much sl
than the exponential rate commo
importantly, they are not predict
rors or other error modes. The wi
(uncorrectable bit error rate) is n
since we see no correlation betw
and the number of uncorrectable
dence that higher-end SLC drive

SSD Failures in Datacenters: What? When? and Why?

Iyswarya Narayanan*, Di Wang†, Myeongjae Jeon†, Bikash Sharma†, Laura Caulfield†,
Anand Sivasubramaniam*, Ben Cutler†, Jie Liu†, Badriddine Khessib†, Kushagra Vaid†

*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.

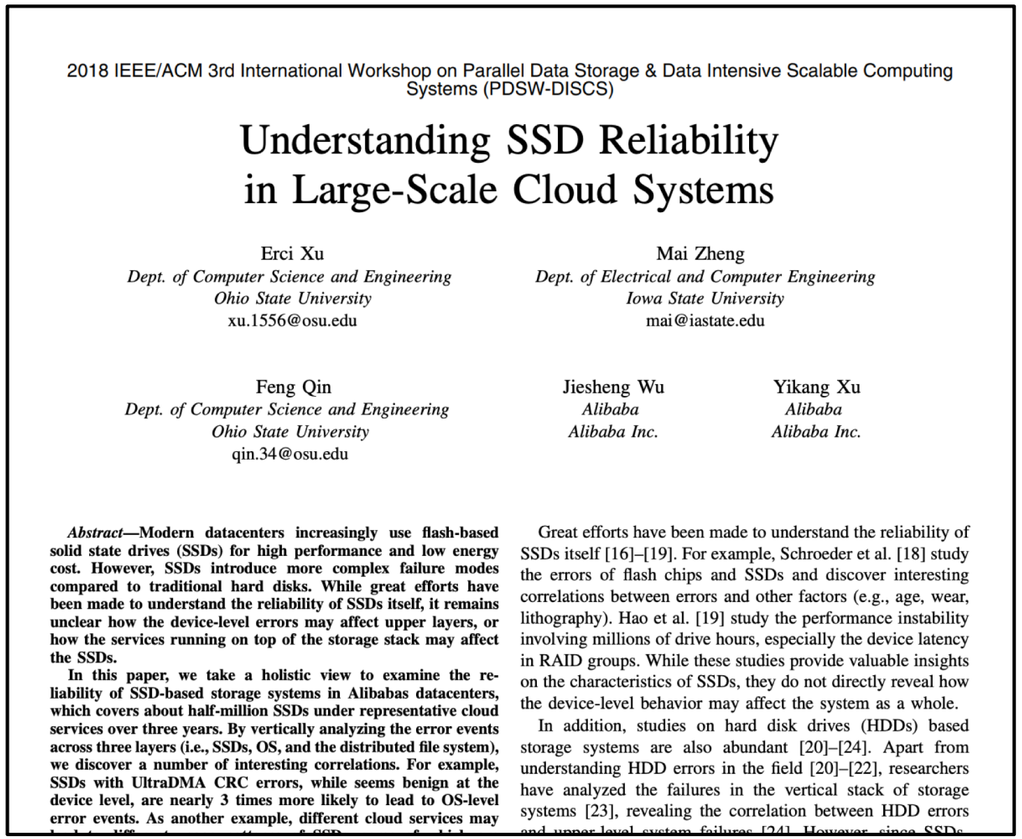
the associated downtime to fix the problem and/or replace the device. It can even take several days to repair/replace a storage component after its failure, with associated server being unusable during this period. To account for this downtime, datacenters resort to over-provisioning (which can add significant 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 fact out-weighing the lower failure rate advantage); and (ii) an SSD-related failure ticket in our dataset results in a replacement 79% of the time compared to 11% for HDD-related tickets (i.e. SSD related failure tickets are more critical in the datacenter). These factors, together with rapid SSDs adoption[3, 13], motivate us to understand SSD reliability.

The current knowledge on SSD failure rate is primarily vendor supplied, based on accelerated lab testing under controlled conditions. In addition to the parameters they are tested for, numerous other factors in a production environ-

- Reveal important characteristics, but mostly only at device level
  - E.g.:
    - Failure rate curve
      - not bathtub
    - FTL impact
      - Thermal Throttling
    - Uncorrectable errors

Our Study:
A holistic view of SSD-related error events



2018 IEEE/ACM 3rd International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISCS)

Understanding SSD Reliability in Large-Scale Cloud Systems

Erci Xu
Dept. of Computer Science and Engineering
Ohio State University
xu.1556@osu.edu

Mai Zheng
Dept. of Electrical and Computer Engineering
Iowa State University
mai@iastate.edu

Feng Qin
Dept. of Computer Science and Engineering
Ohio State University
qin.34@osu.edu

Jiesheng Wu
Alibaba
Alibaba Inc.

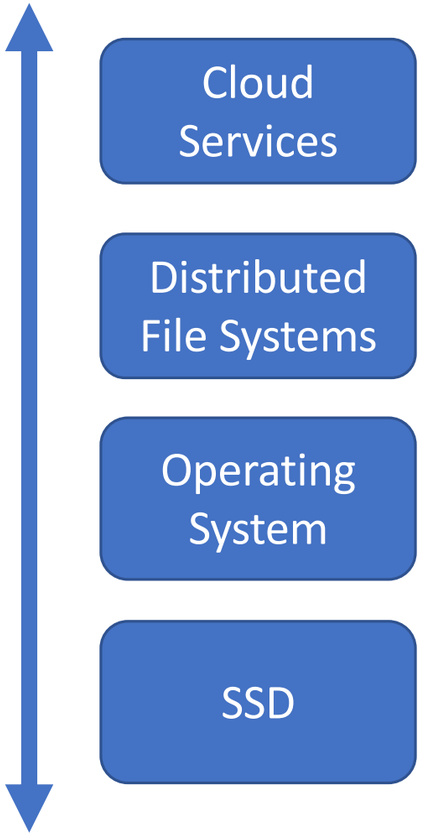
Yikang Xu
Alibaba
Alibaba Inc.

***Abstract*—Modern datacenters increasingly use flash-based solid state drives (SSDs) for high performance and low energy cost. However, SSDs introduce more complex failure modes compared to traditional hard disks. While great efforts have been made to understand the reliability of SSDs itself, it remains unclear how the device-level errors may affect upper layers, or how the services running on top of the storage stack may affect the SSDs.**

In this paper, we take a holistic view to examine the reliability of SSD-based storage systems in Alibabas datacenters, which covers about half-million SSDs under representative cloud services over three years. By vertically analyzing the error events across three layers (i.e., SSDs, OS, and the distributed file system), we discover a number of interesting correlations. For example, SSDs with UltraDMA CRC errors, while seems benign at the device level, are nearly 3 times more likely to lead to OS-level error events. As another example, different cloud services may

Great efforts have been made to understand the reliability of SSDs itself [16]–[19]. For example, Schroeder et al. [18] study the errors of flash chips and SSDs and discover interesting correlations between errors and other factors (e.g., age, wear, lithography). Hao et al. [19] study the performance instability involving millions of drive hours, especially the device latency in RAID groups. While these studies provide valuable insights on the characteristics of SSDs, they do not directly reveal how the device-level behavior may affect the system as a whole.

In addition, studies on hard disk drives (HDDs) based storage systems are also abundant [20]–[24]. Apart from understanding HDD errors in the field [20]–[22], researchers have analyzed the failures in the vertical stack of storage systems [23], revealing the correlation between HDD errors



System admin

Outline

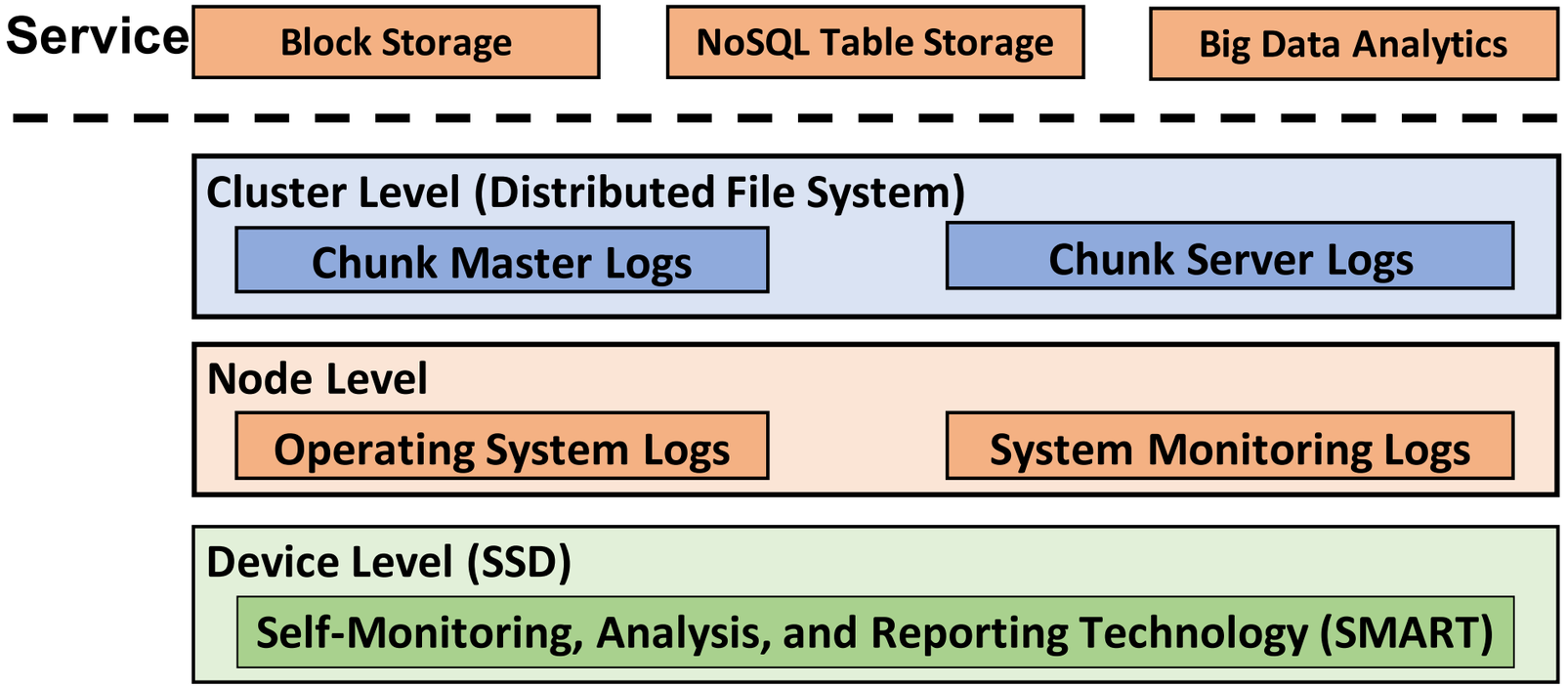
- ~~Introduction~~

- **System Architecture & Dataset**

- Findings
  - Human Mistake
  - Service Unbalance
  - Transmission Error

- Conclusions & Future Work

System Architecture



SSD Fleet in Our Study

- Near half million SSDs from 3 vendors spanning over 3 years deployment

Model	Capacity	Lithography	Age
1-B	480GB	20nm	2-3 yrs
1-C	800GB	20nm	2-3 yrs
1-L	480GB	16nm	1-2 yrs
2-V	480GB	20nm	2-3 yrs
3-V	480GB	20nm	1-2 yrs

different SSD models

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

different SSD usages

Dataset Collected

Level	Event	Definition
DFS	Read Error	DFS cannot read the requested data on time
	Write Error	DFS cannot finish writing with replication on time
Node	Buffer IO Error	A failed read/write from file system to SSD
	Media Error	Software detected actual data corruption
	File System Unmountable	Unable to load the file system on a SSD
	Drive Missing	OS unable to find a plugged SSD
	Wrong Slot	SSD has been plugged to the Wrong SATA slot
Device	Host Read	Total amount of LBA read from the SSD
	Host Write	Total amount of LBA write from the SSD
	Program Error	Total # of errors in NAND write operations
	Raw Bit Error Rate	Total bits corrupted divided by total bits read
	End-to-End Error	Total # of parity check failures between interfaces
	Uncorrectable Error	Total # of data corruption beyond ECC's ability
	UDMA CRC Error	Total # of CRC check failures during Ultra-DMA(UDMA)

Events above SSDs

Outline

- ~~Introduction~~

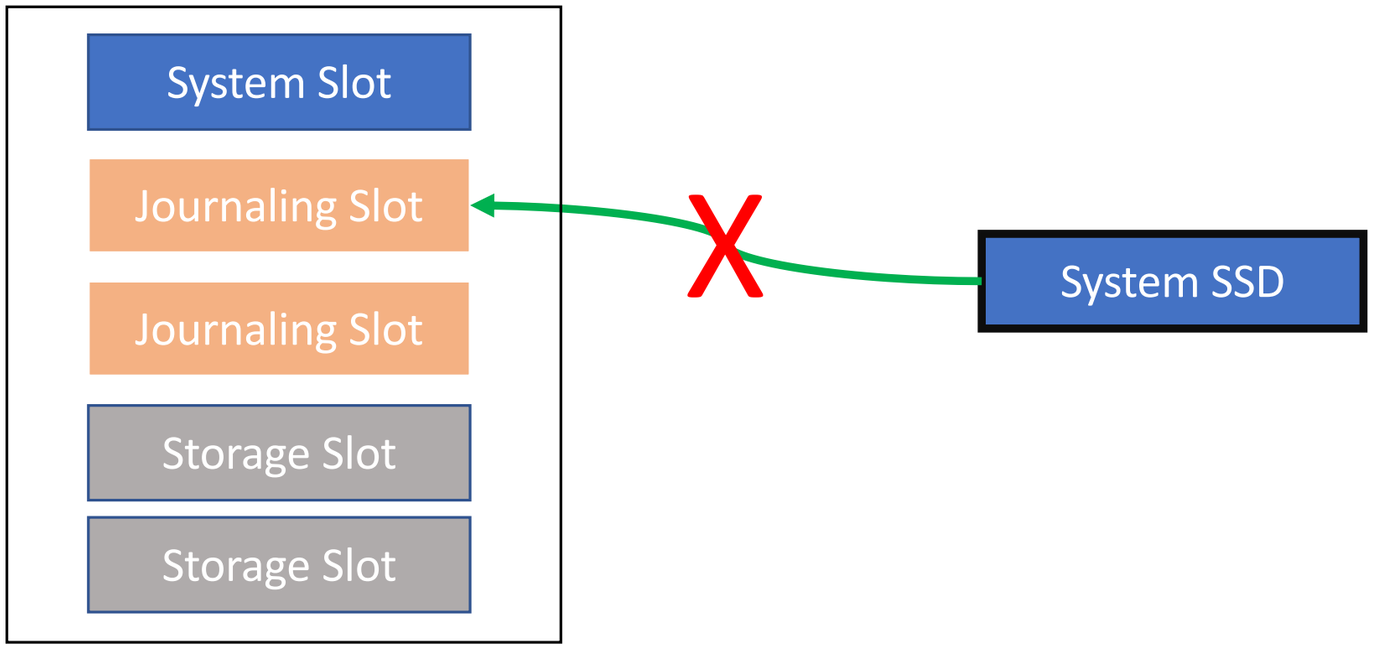
- ~~System Architecture & Dataset~~

- Findings
  - **Human Mistake**
  - Service Unbalance
  - Transmission Error

- Conclusions & Future Work

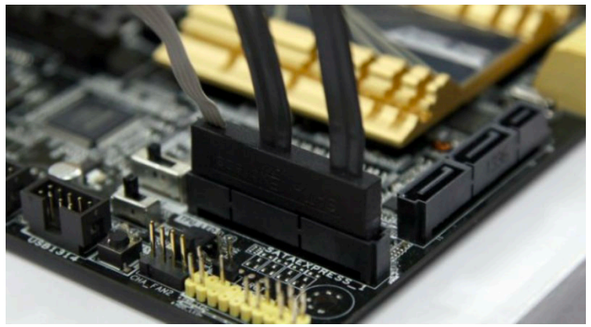
Human Mistakes

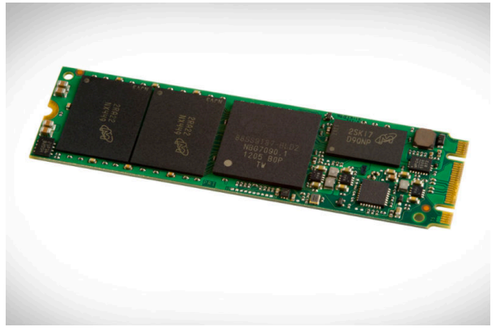
- Over 20% of SSD-related OS-level error events are caused by incorrect manual operations
  - “Wrong Slot” is a dominant case: an SSD is plugged into an incorrect slot.



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~~

- Findings
  - ~~Human Mistake~~
  - **Service Unbalance**
  - Transmission Error

- Conclusions & Future Work

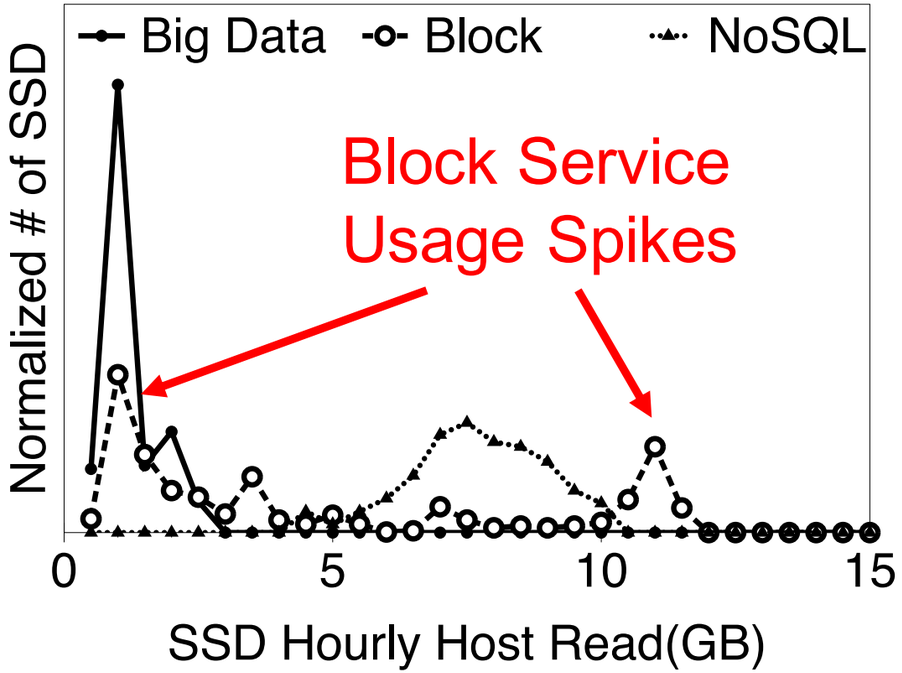
Service Unbalance

- 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 Unbalance



- Each dot in the line equals the cumulative count of SSDs that have hourly host read 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).
- 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.

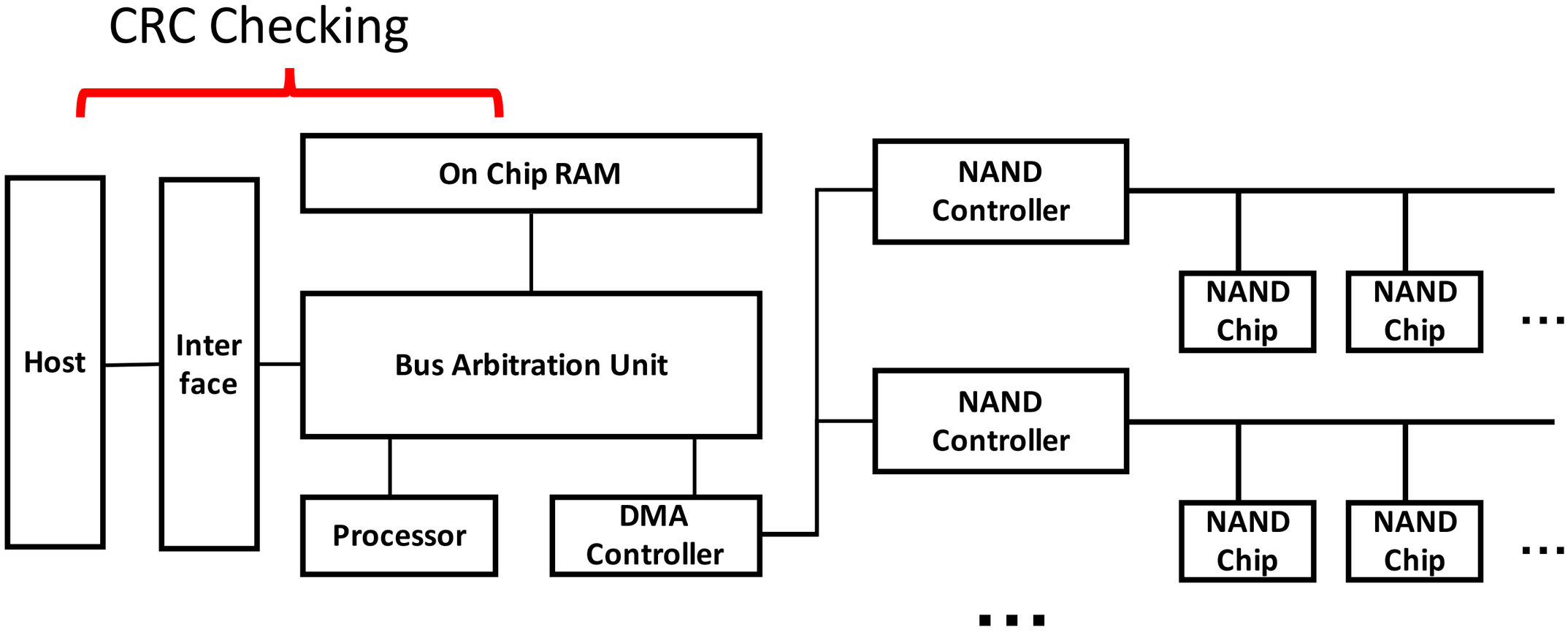
Service Unbalance

- Root cause of the unbalanced usage
  - Block Storage Service tends to map user's logical blocks to SSDs on a limited number of nodes; each node hosts relatively few users' data
  - the I/O patterns of different users vary a lot
- Our solution
  - Shared log structure: users' data are more evenly allocated across SSDs.
  - Usage difference reduced to less than 5% among drives on a test cluster

Outline

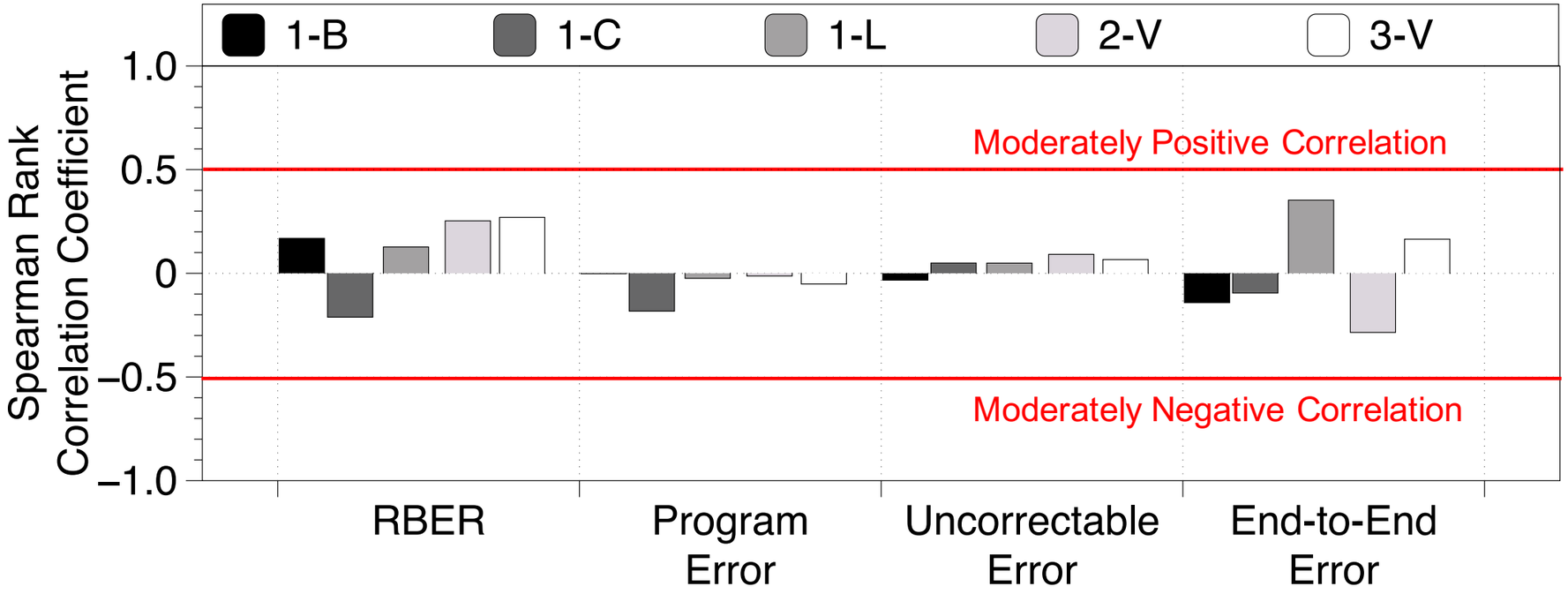
- ~~Introduction~~
- ~~System Architecture & Dataset~~
- Findings
  - ~~Human Mistake~~
  - ~~Service Unbalance~~
  - **Transmission Error**
- Conclusions & Future Work

Transmission Error: UltraDMA CRC (UCRC) error

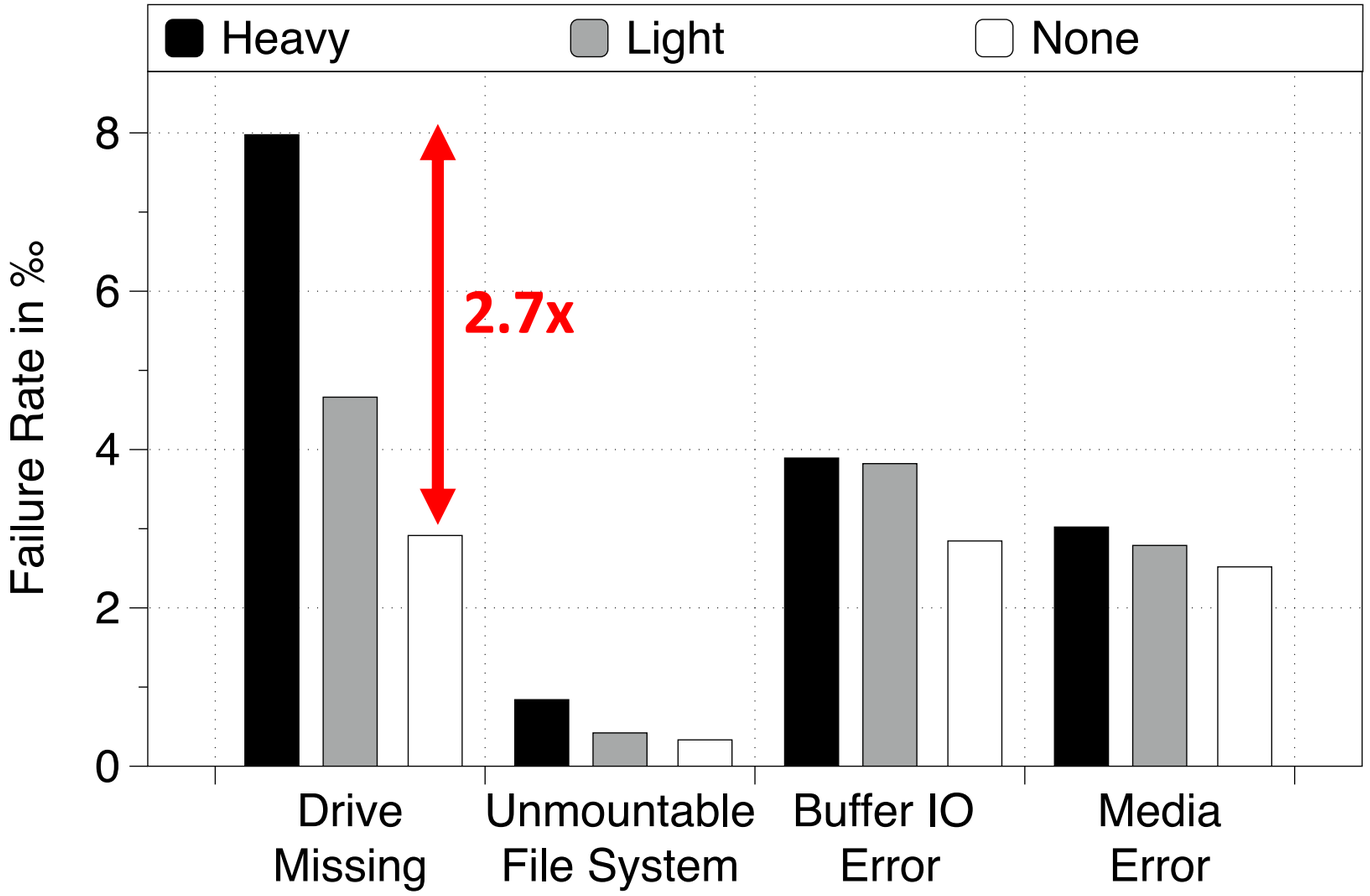


Transmission Error occurs when data fails to pass the CRC checking after SSD-to-Host transmission and would trigger an automatic retry.

UCRC errors are not correlated w/ other device-level errors



UCRC errors are NOT necessarily benign



SSDs with heavy UCRC errors are 2.7X more likely to lead to “Drive Missing” failures

Outline

- ~~Introduction~~
- ~~System Architecture & Dataset~~
- ~~Findings~~
  - ~~Human Mistake~~
  - ~~Service Unbalance~~
  - ~~Transmission Error~~
- **Conclusions & Future Work**

Conclusions & Future Work

- A holistic view of SSD-related error events
  - Human Mistake
    - Plugging an SSD into a wrong slot
    - Mitigated by “One Interface One Purpose”
  - Service Unbalance
    - 15-20% of SSDs are overly used under block storage service
    - Mitigated by shared log structure
  - Transmission Error
    - UCRC error is independent from other device errors
    - UCRC is not necessarily benign

- Next steps
  - more errors, more failure symptoms
  - casual relationship & error propagation paths
  - Predicting device errors or system failures

pdsw-discs

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 Alibaba

Jiesheng Wu
Aliyun Alibaba