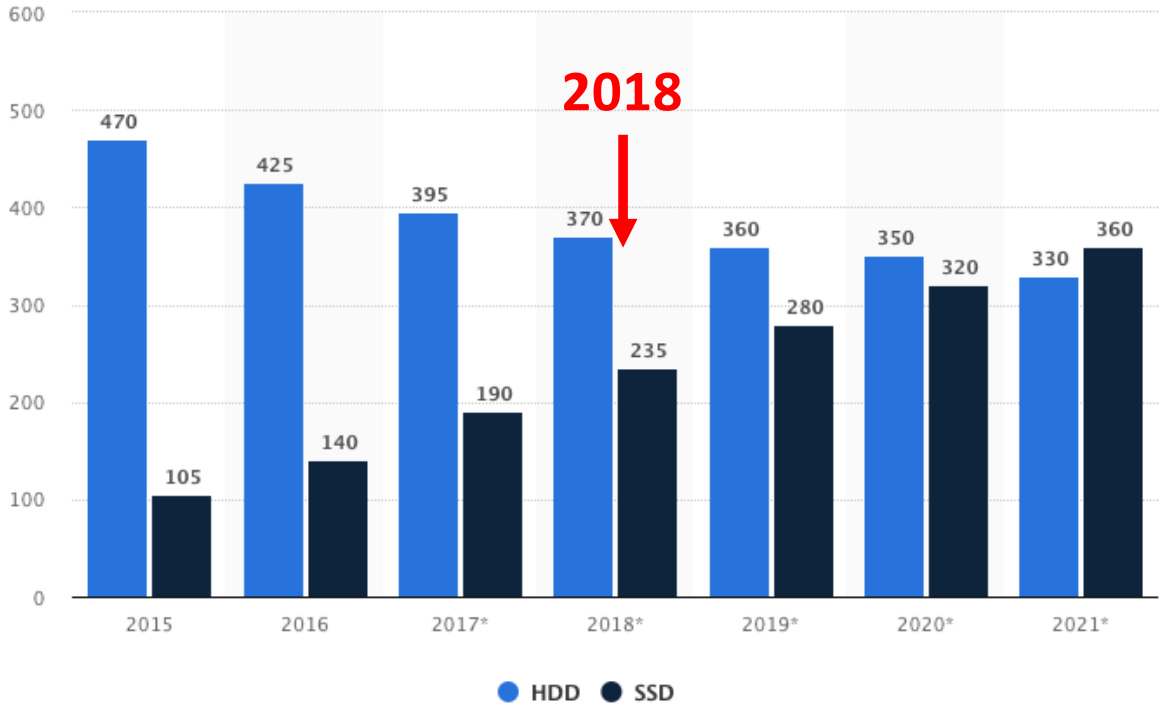
pdsw-discs



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-Memory-Based Solid-Stata Drives is popular in today's DC



- Enterprises and Individuals are increasingly use SSD as storage media
- 2018 Q1: 45.46 million units
- Estimated Enterprise SSDs: 30 million units in 2018, up from 20 million units in 2016

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/

SSD Reliability Concerns

- Wear out issue
  - Limited Life Cycles
- Complex failure modes
  - Program/Erase Error
- Sensitive to environment
  - NAND in heated environment

Previous Large Scale SSD Studies

- Reveal important aspects of flash-based drives in the field

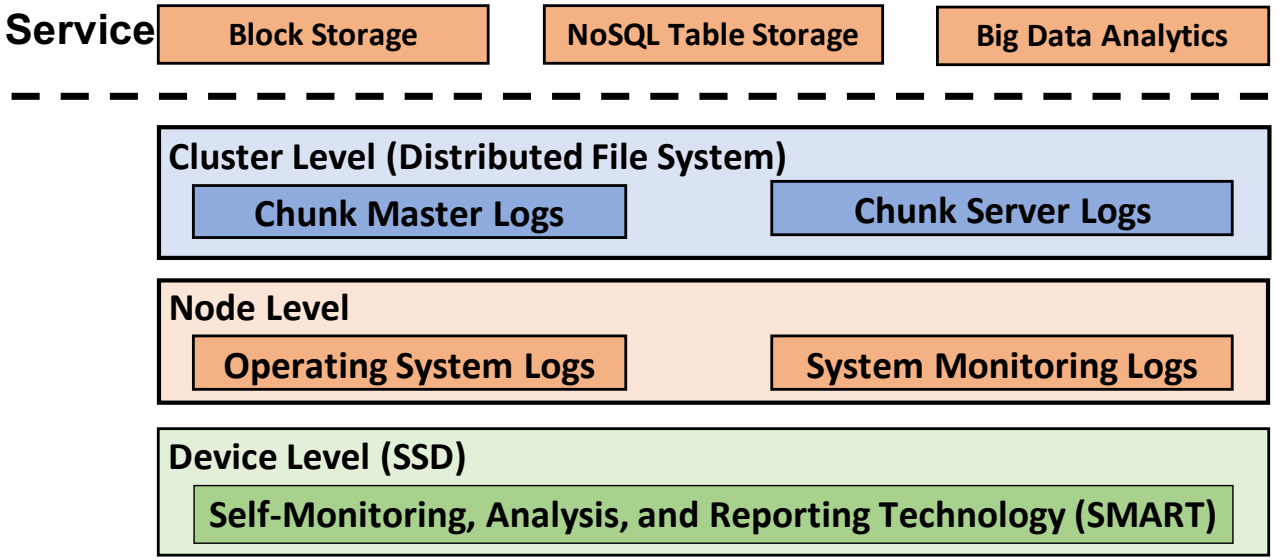
- Life Curve
  - Not Bathtub
- Uncorrectable Errors are popular
  - Program/Erase Error
  - NAND in heated environment

SSD Reliability in the Cloud Systems are more than the device

Outline of Our Talk

- 1. Architecture Overview
- 2. Human Errors
- 3. Service Imbalance
- 4. Transmission Errors
- 5. Conclusions & Future Works

Architecture Overview



SSD Fleet in Our Research

Model	Capacity	Lith.	Age	Rationale
1-**B**	480GB	20nm	2-3yrs	**B**aseline
1-**C**	800GB	20nm	2-3yrs	**C**apacity
1-**L**	480GB	16nm	1-2yrs	**L**ith.
2-**V**	480GB	20nm	2-3yrs	**V**endor
3-**V**	480GB	20nm	1-2yrs	**V**endor

Table 1: **SSD Models**. ***Lith.***: *Lithography; Each model is named as (Vendor id)-Rationale.*

Targeting different cloud products to achieve general findings

Service	Model	Setup	Function
Block	1-B,1-C,1-L 2-V,3-V	Hybrid	Journaling
		Multiple	Persistence
NoSQL	1-B,1-L 2-V,3-V	Hybrid	Journaling
		Multiple	Persistence
Big Data	1-B,1-C,2-V	Single	Temporary

Table 2: Service Workloads. **Hy**: Hybrid Setup, **Mul**: Multiple Setup, **Pers**: Persistent storage, **Jour**: Journaling, **Temp**: Temporarily storing intermediate data.

Targeting representative services with different functionalities and avoiding biased results from different workloads

Datasets in Our Research: Errors and Failures

Level	Event	Definition	Freq.
Node	Buffer IO Error	A failed read/write from file system to SSD	Event
	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	Daily
	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)	

Table 3: **Events Collected in the Target Storage System.** ***Freq.****: Frequency, event logs can be updated daily ("Daily") or upon new events ("Event");*

Including Node level failure to build correlation and understand impact

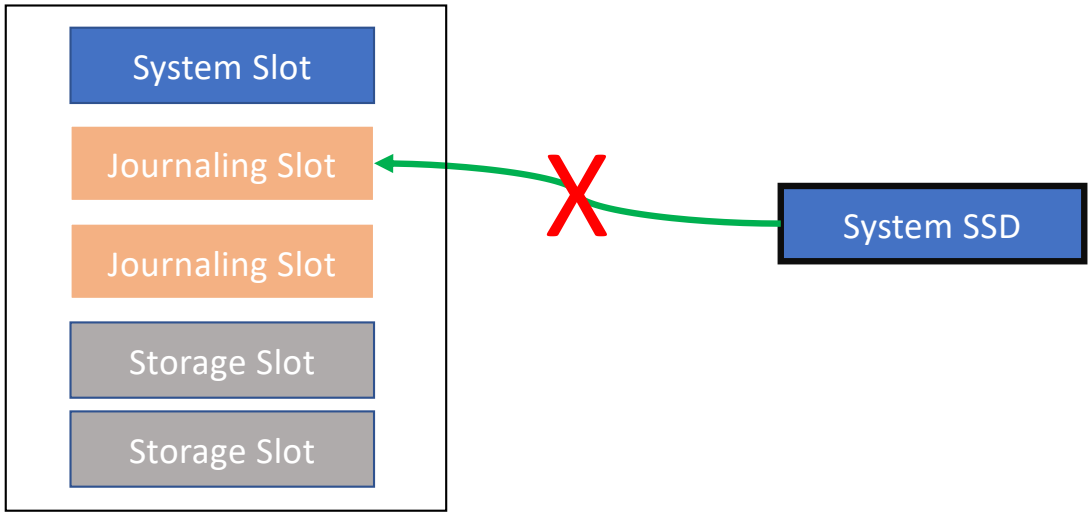
Outline of Our Talk

- 1. Architecture Overview
- 2. Human Errors
- 3. Service Imbalance
- 4. Transmission Errors
- 5. Conclusions & Future Works

Human Errors

- Observation: Over 20% of OS-level error events are caused by incorrect manual operations

- Cause: Wrong Slot is a dominant cause. A SSD is plugged into an incorrect slot.



Wrong Slot

- Root cause: the device is mapped to the SATA slot in 1-to-1 fashion

- Possible solutions: adding an indirect translation layer between file system mounting points and hardware slots

- Drawbacks:
  1. Global scale changes
  2. Kernel Level Modifications

Wrong Slot: Our Solution

- OIOP: One Interface One Purpose
- New interfaces, such as M.2 and U.2
- We switch the mapping from mounting-to-SATA to mounting-to-interface
- Example: In the hybrid setup, the system drive is plugged to the M.2 interface while storage SSDs still use the SATA interface

Outline of Our Talk

- 1. Architecture Overview
- 2. Human Errors
- 3. Service Imbalance
- 4. Transmission Errors
- 5. Conclusions & Future Works

Service Imbalance

- Observation: Certain cloud services may cause unbalanced usage of SSDs

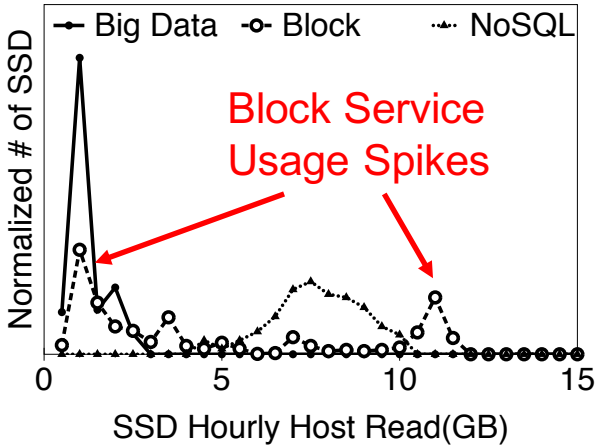
		Host Read	Host Write
Avg. Value /Hour	Block	7.69GB	6.56GB
	BigData	1.57GB	1.22GB
	NoSQL	6.10GB	5.28GB
CV	Block	35.5%	24.9%
	BigData	1.8%	3.7%
	NoSQL	3.2%	6.2%

Table 5: **Host Read and Host Write Comparison between Services**. ***CV****: Coefficient of Variance, the ratio of standard deviation to mean.*

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

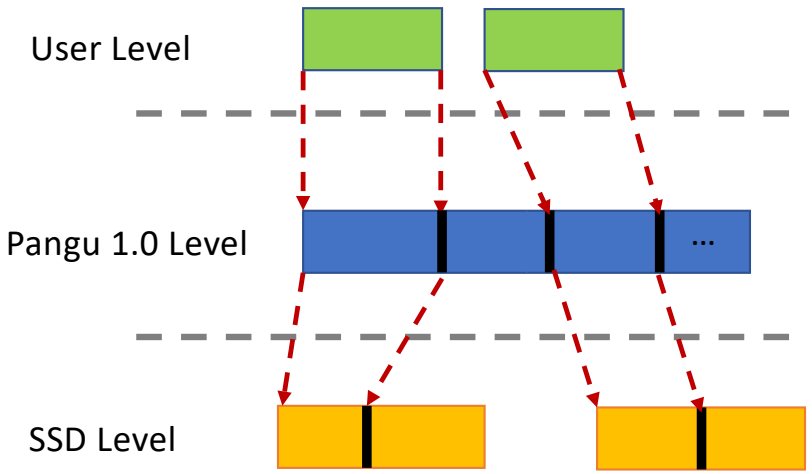
Service Imbalance: Further Breakdown

SSD Hourly Host Read



- Each dot in the line equals the cumulative count of SSDs that have hourly host read (or write) 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).
- On the other hand, 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 Imbalance: Further Breakdown

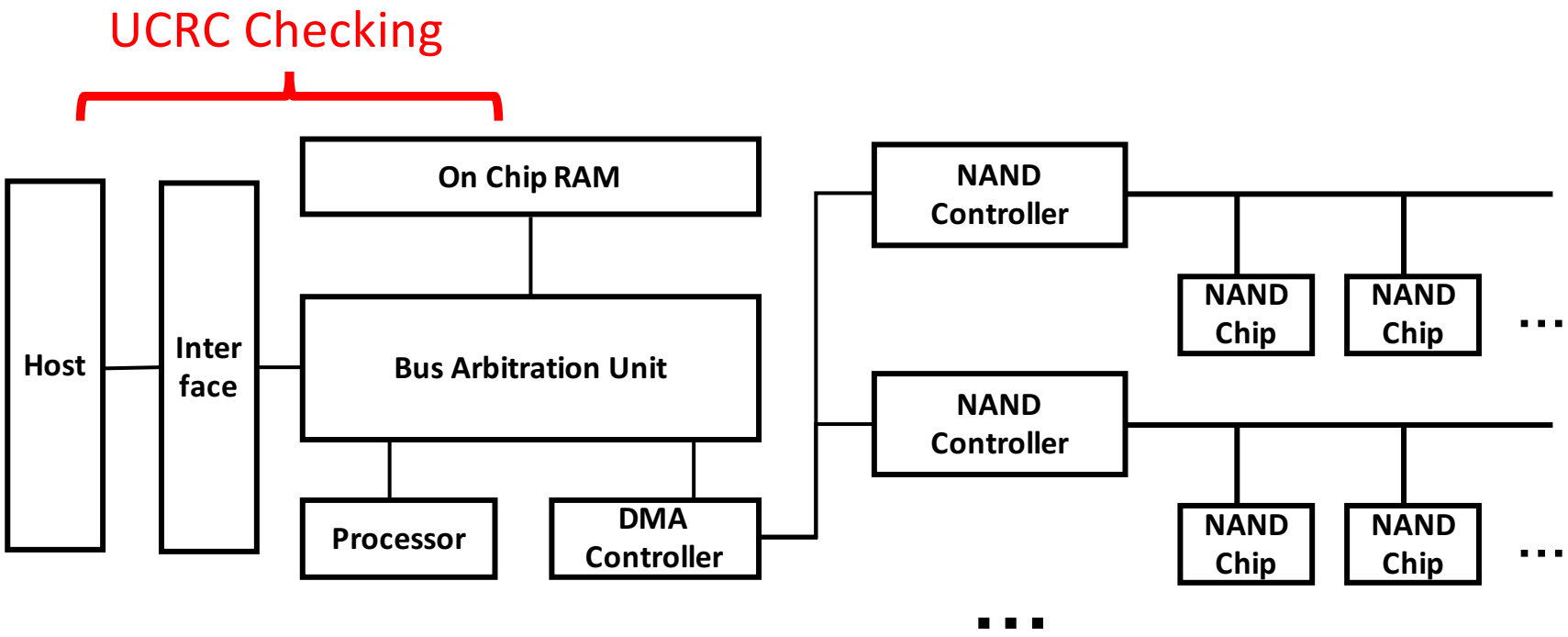


- Pangu 1.0 Data Layout for Block storage
  - In-place updates
- User can generate imbalance of workload to SSDs
- 15%-20% of SSDs have higher workload

Outline of Our Talk

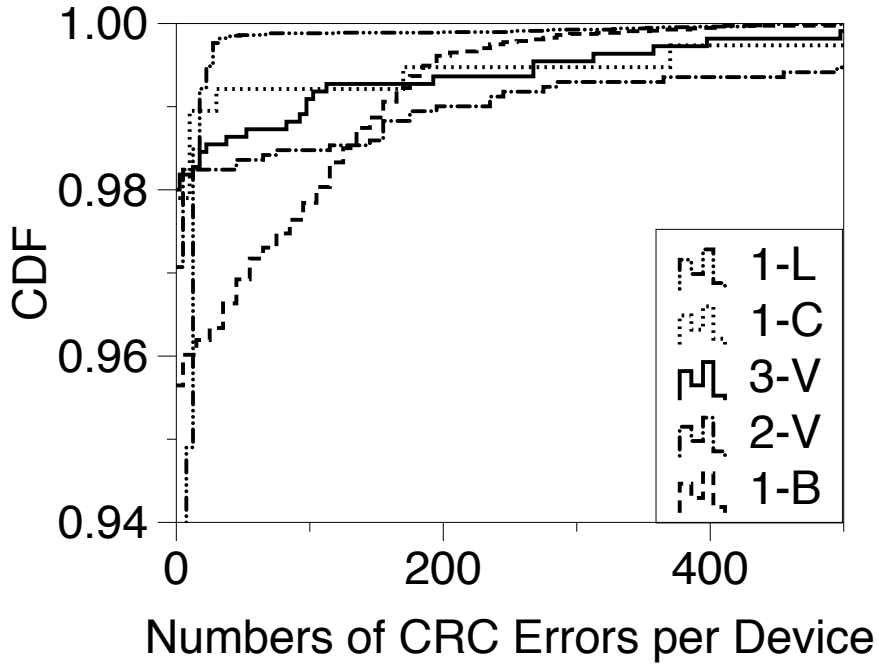
- 1. Architecture Overview
- 2. Human Errors
- 3. Service Imbalance
- 4. Transmission Errors
- 5. Conclusions & Future Works

Transmission Error



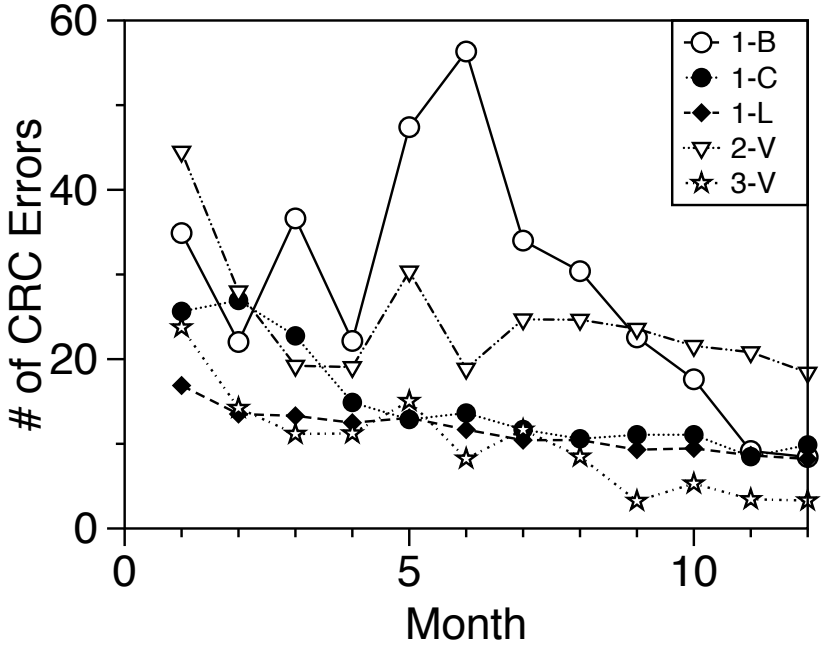
Transmission Error occurs when data fails to pass the CRC checking after SSD-to-Host transmission and would trigger an automatic retry.

Transmission Error: CRC Errors by SSD Model



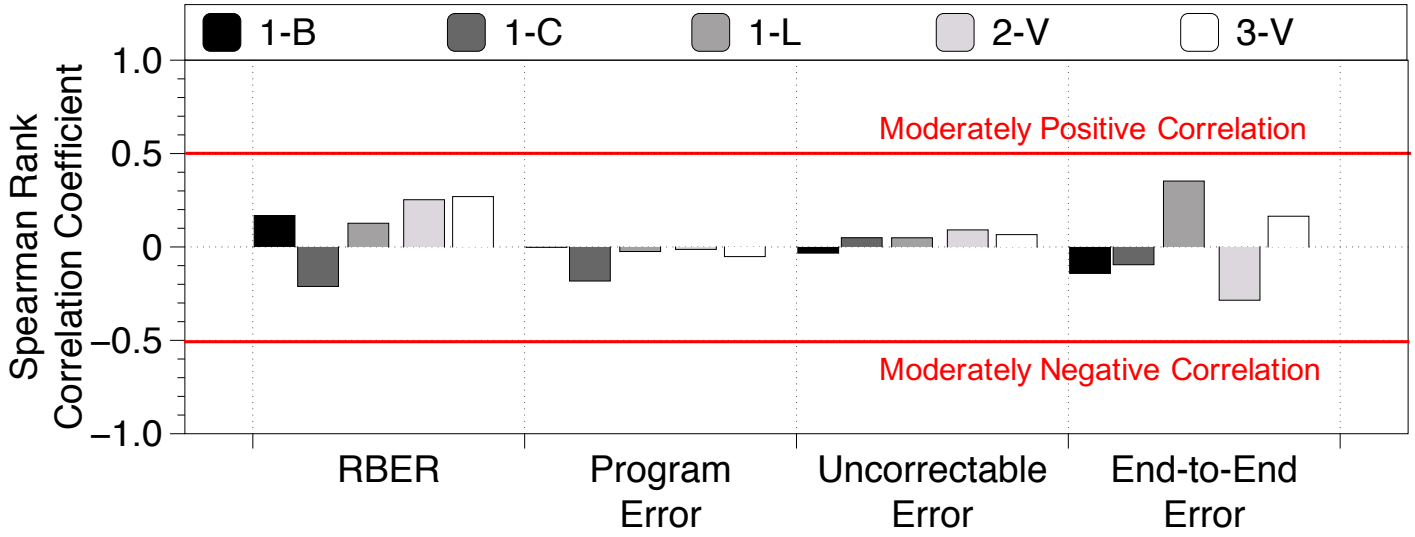
Observation: CRC errors are concentrated around 6% of all drives

Transmission Error: CRC Errors with Age



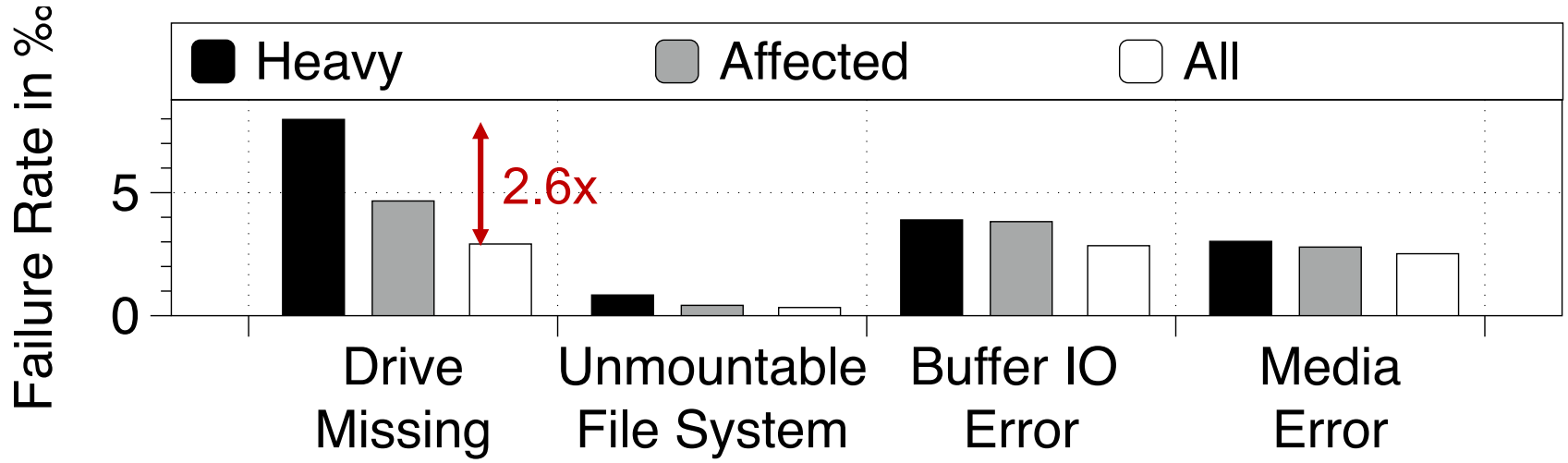
- Observation: errors are stably generated
- The biased distribution and stable generation indicate there can be non-transient factors behind the generation of CRC errors

Transmission Errors: CRC Errors vs. Other Errors



- Results: SPCC results with device level events
- Observation: UCRC errors not correlated with device-level errors. UCRC is an independent type of errors

Transmission Errors: CRC Errors vs. Failures



- Results: Failure rate results with OS-level error events
- Observation: UCRC error can lead to a higher rate of OS-level error events

Outline of Our Talk

- 1. Architecture Overview
- 2. Human Errors
- 3. Service Imbalance
- 4. Transmission Errors
- 5. Conclusions & Future Works

Field Study Summary

- SSD failure is more than just data errors
- Human Error
  - Issue: Wrong Slot Plugging
  - Root Cause: 1-to-1 fashion mapping
- Service Imbalance
  - Issue: 15-20% of SSDs have higher workload
  - Root cause: in-place update
- Impact of Transmission Errors
  - Issue: UCRC error is not necessarily benign in the long run
  - Root cause: UCRC error is a useful indicator for faulty interconnection

pdsw-discs



Thank You!
Q&A

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