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

Motivation

- **Correlated failures**
  - Challenge high storage reliability
  - Complicate design of redundancy protection schemes
- Solid-state drives (SSDs) become the mainstream storage media in modern data centers
- *What are the characteristics of correlated failures of SSDs?*
- *What are implications of correlated failures of SSDs on storage reliability in production environments?*

Our Contribution

- An in-depth data-driven analysis on correlated failures of SSDs from spatial and temporal perspectives
- We provide 15 findings on correlated failures
  - *Intra-node* and *intra-rack failures*
  - Impact of drive characteristics, SMART attributes, and applications
  - Trace-driven simulation on reliability of redundancy schemes under correlated failures
- We release our dataset and source code for public use

Dataset

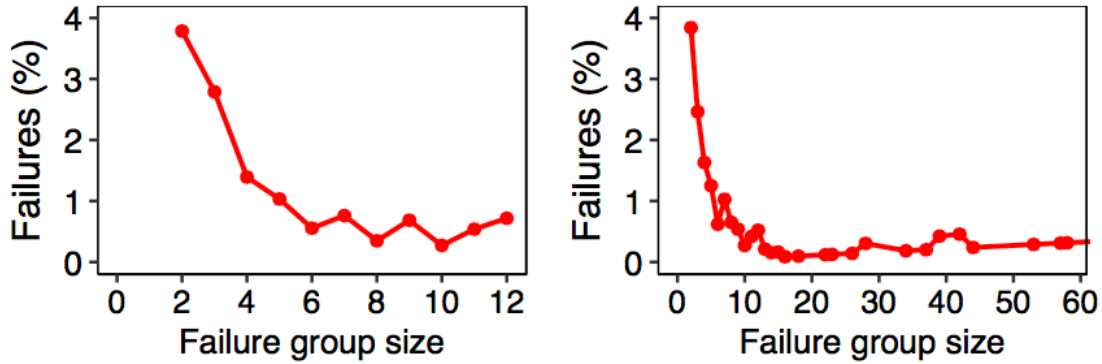
- Collect nearly 1 million SSDs of 11 drive models from 3 vendors over two-year span (Jan. 2018 – Dec. 2019) at Alibaba
- Data types: SMART logs, trouble tickets, locations (e.g., nodes and racks), and applications
- Two main types of SSD failures in trouble tickets:
  - *Whole drive failures:* an SSD either cannot be accessed or loses all data that is unrecoverable
  - *Partial drive failure:* part of the data in an SSD either cannot be accessed and is unrecoverable

Analysis Methodology

- Spatial and temporal properties
  - *Intra-node (intra-rack) failures*: failures co-occurring within a node (rack)
  - *Intra-node (intra-rack) failure time interval*: 30 minutes by default
  - *Intra-node (intra-rack) failure group*: a sequence of intra-node (intra-rack) failures
- Correlation properties
  - Spearman's Rank Correlation Coefficient (SRCC)
  - Measure correlations between correlated failures and SMART attributes

Correlations Among Failures

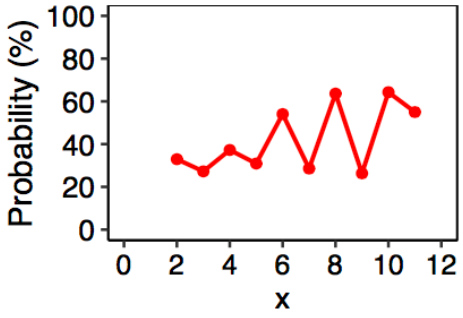
- **Finding 1**: A non-negligible fraction of SSD failures belong to intra-node and intra-rack failures
  - 12.9% (18.3%) of failures are intra-node (intra-rack) failures
  - Intra-node (intra-rack) failure group size can exceed the tolerable limit of some redundancy protection schemes



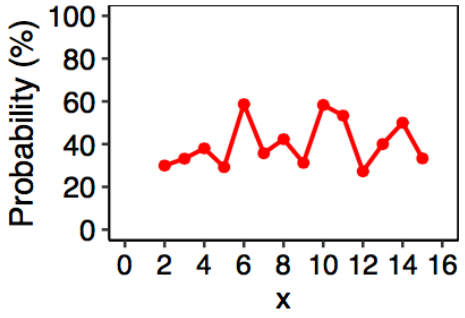
(a) Intra-node failures (b) Intra-rack failures

Correlations Among Failures

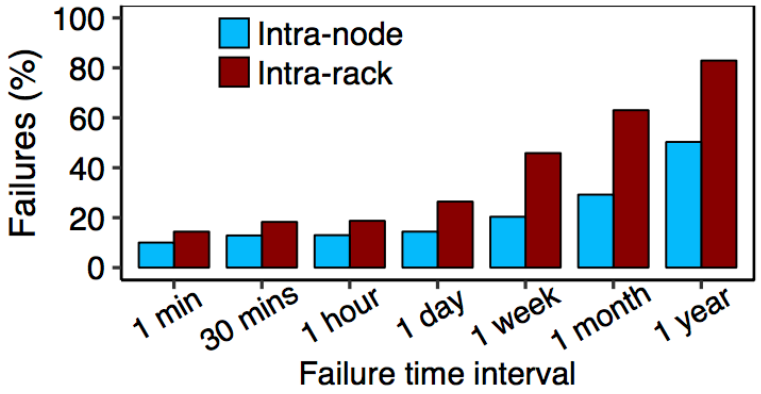
- **Finding 2**: Likelihood of having an additional intra-node (intra-rack) failure depends on existing intra-node (intra-rack) failures
- **Finding 3**: A non-negligible fraction of intra-node and intra-rack failures occur within a short period of time, even within one minute



(a) Intra-node failures

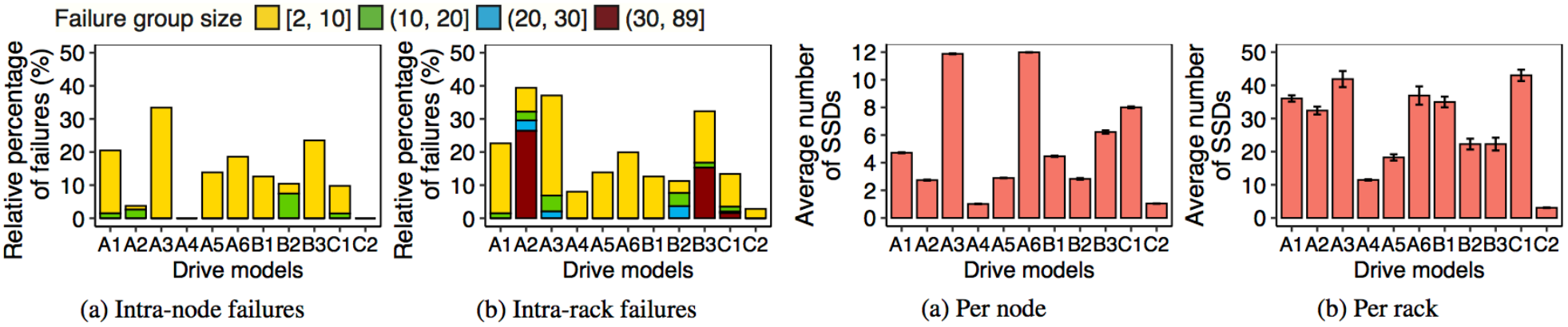


(b) Intra-rack failures



Impact of Drive Models

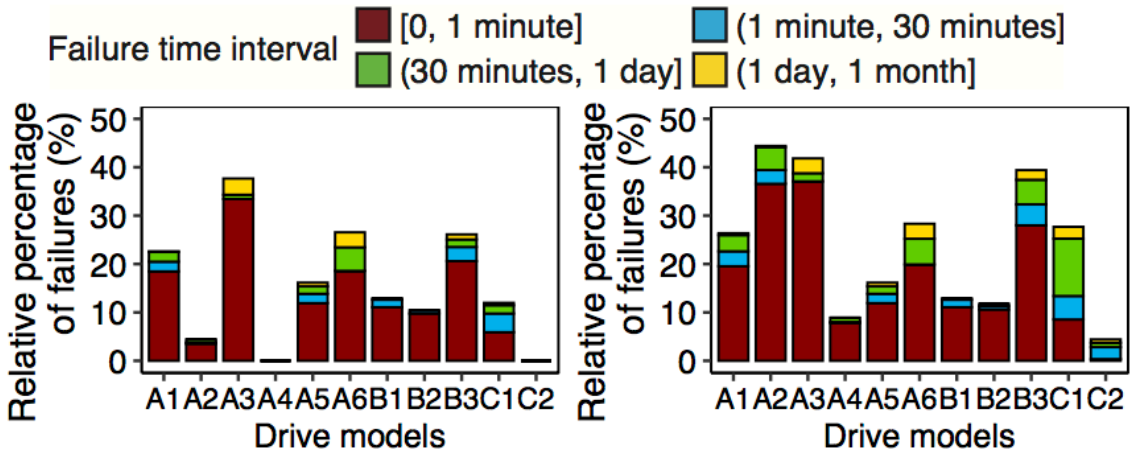
- **Finding 4**: Relative percentages of intra-node (intra-rack) failures vary across drive models
  - Putting too many SSDs from the same drive model in the same nodes (racks) leads to a high percentage of intra-node (intra-rack) failures
  - AFR and environmental factors (e.g., temperature)



(a) Intra-node failures (b) Intra-rack failures (a) Per node (b) Per rack

Impact of Drive Models

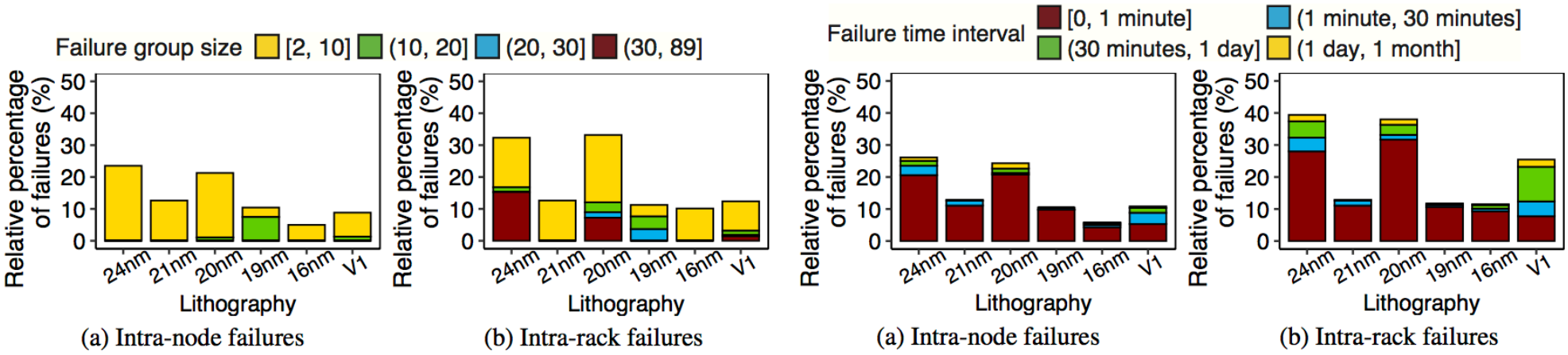
- **Finding 5**: Non-negligible fractions of intra-node and intra-rack failures with a short failure time interval (e.g., one minute) for most drive models
  - 3.5-33.4% (7.8-37.1%) of intra-node (intra-rack) failures except A4 and C2 (C2)



(a) Intra-node failures (b) Intra-rack failures

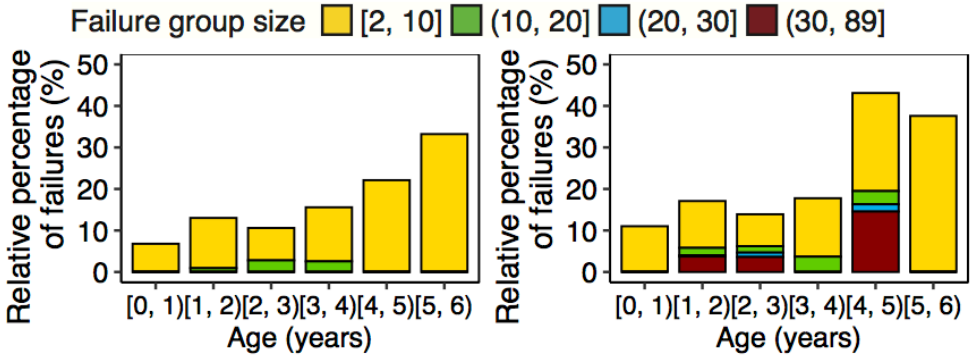
Impact of Lithography

- **Finding 6**: MLC SSDs with higher densities generally have lower relative percentages of intra-node and intra-rack failures
  - For MLC SSDs, a smaller lithography implies a higher density
  - 3D-TLC SSDs have higher densities than those of MLC SSDs

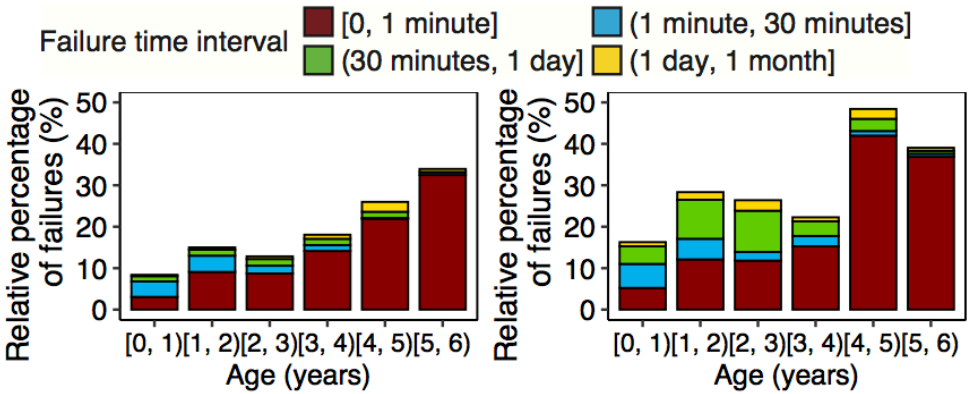


Impact of Age

- **Finding 7**: Relative percentages of intra-node and intra-rack failures increase with age
  - Intra-node and intra-rack failures at an older age are more likely to occur within a short time due to the increasing rated life used



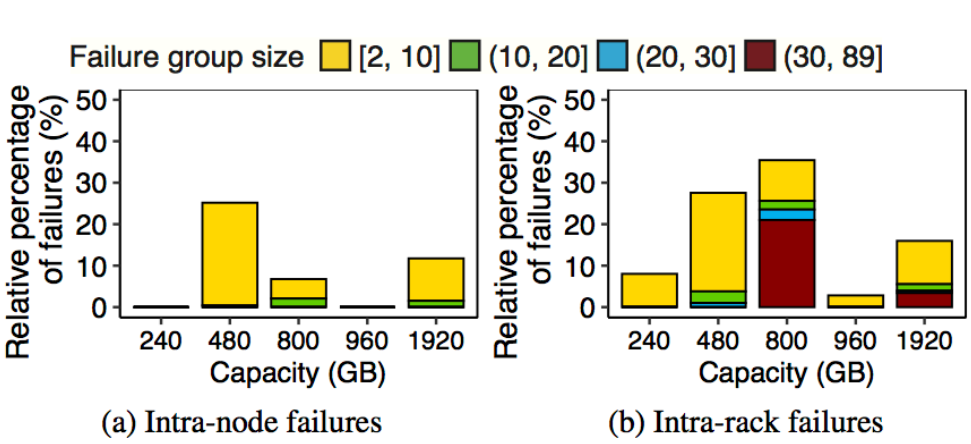
(a) Intra-node failures (b) Intra-rack failures



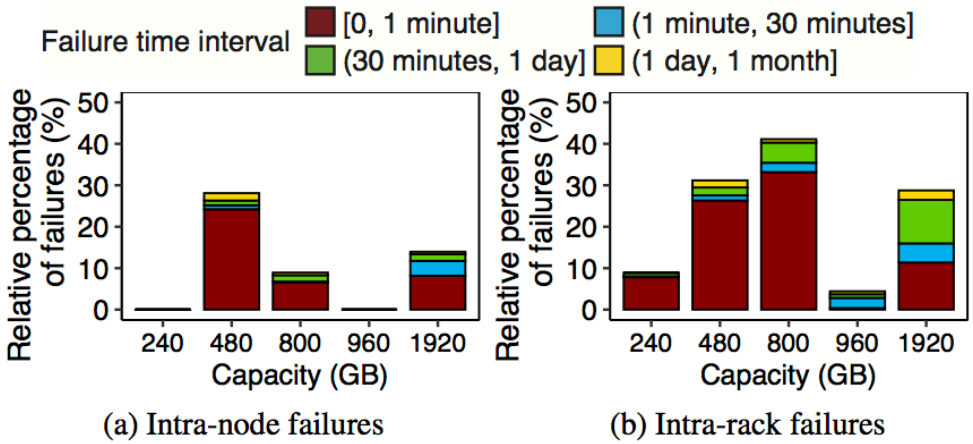
(a) Intra-node failures (b) Intra-rack failures

Impact of Capacity

- **Finding 8**: Relative percentages of intra-node and intra-rack failures vary significantly across the capacity
  - No clear trend between relative percentages of intra-node (or intra-rack) failures for different thresholds of failure time intervals and capacity



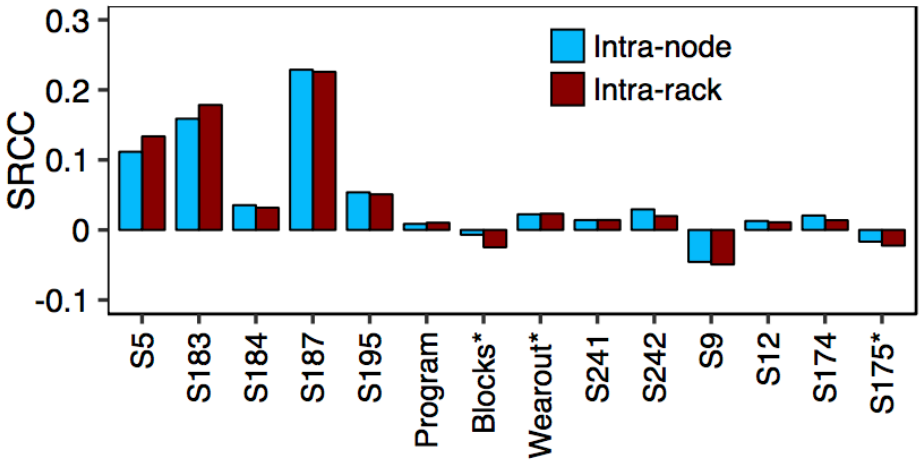
(a) Intra-node failures (b) Intra-rack failures



(a) Intra-node failures (b) Intra-rack failures

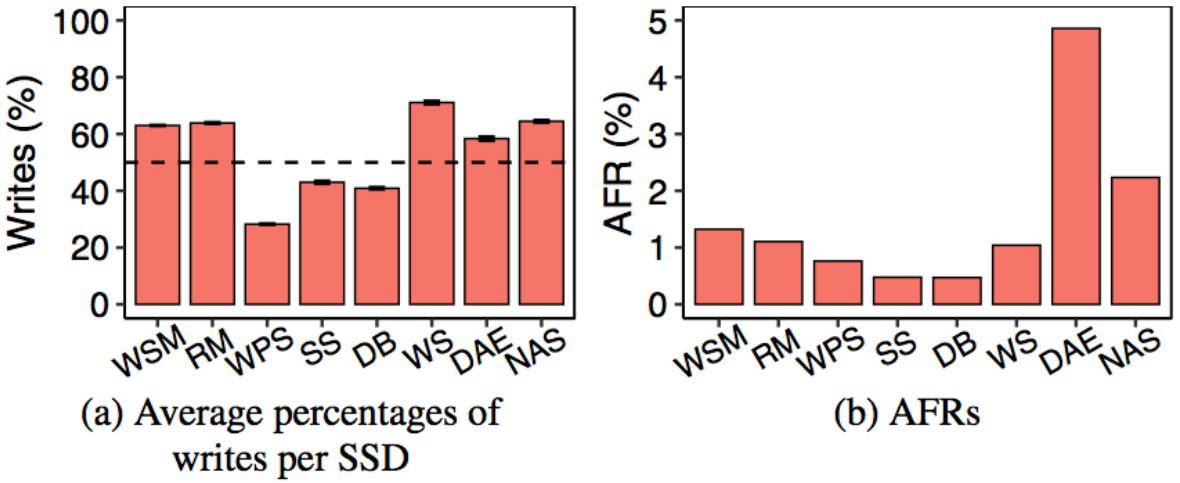
Impact of SMART Attributes

- **Finding 9**: SMART attributes have limited correlations with intra-node and intra-rack failures
  - The highest SRCC values (from S187) are only 0.23 for both intra-node and intra-rack failures
  - SMART attributes are not good indicators for detecting the existence of intra-node and intra-rack failures
  - Intra-node and intra-rack failures have no significant difference of absolute values of SRCC for each SMART attribute



Impact of Applications

- **Finding 10**: Write-dominant workloads lead to more SSD failures overall, but are not the only factors on AFRs
  - Other factors, such as drive models, can affect AFRs

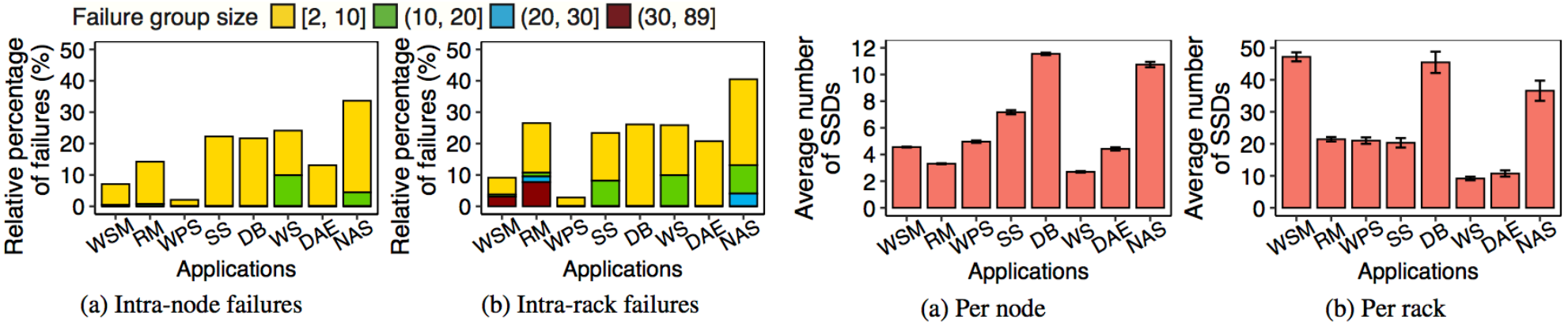


(a) Average percentages of writes per SSD

(b) AFRs

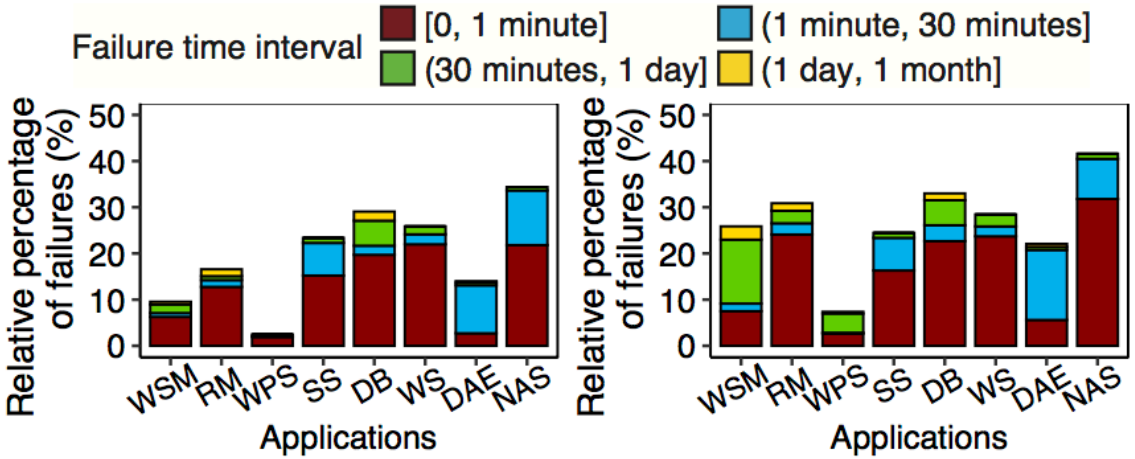
Impact of Applications

- **Finding 11**: Applications with more SSDs per node (rack) and write-dominant workloads tend to have a high percentage of intra-node (intra-rack) failures



Impact of Applications

- **Finding 12**: For the applications, intra-node and intra-rack failures at an older age and with more write-dominant workloads tend to occur in a short time



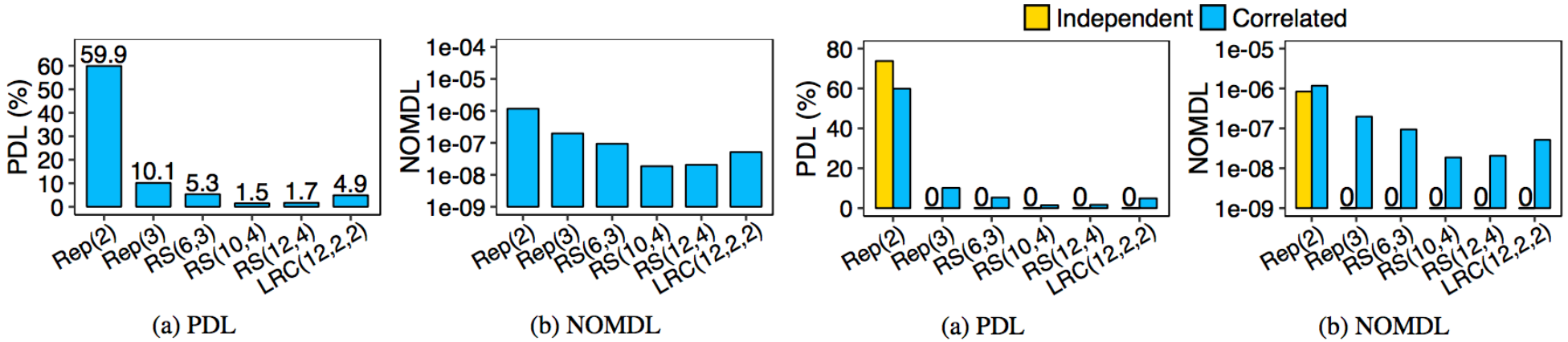
(a) Intra-node failures (b) Intra-rack failures

Case Study: Redundancy Protection

- Redundancy schemes
  - *r-way replication* (Rep(r)): Rep(2) and Rep(3)
  - *Reed-Solomon coding* (RS(*k, m*)): RS(6,3), RS(10,4), and RS(12,4)
  - *Local Reconstruction Coding* (LRC(*k*, *l*, *g*)): LRC(12,2,2)
  - Eager recovery vs. lazy recovery
- Simulator
  - Extend C++ discrete-event simulator SimEDC[TPDS'19 Zhang]
- Metrics
  - Probability of data loss (PDL)
  - Normalized magnitude of data loss (NOMDL)

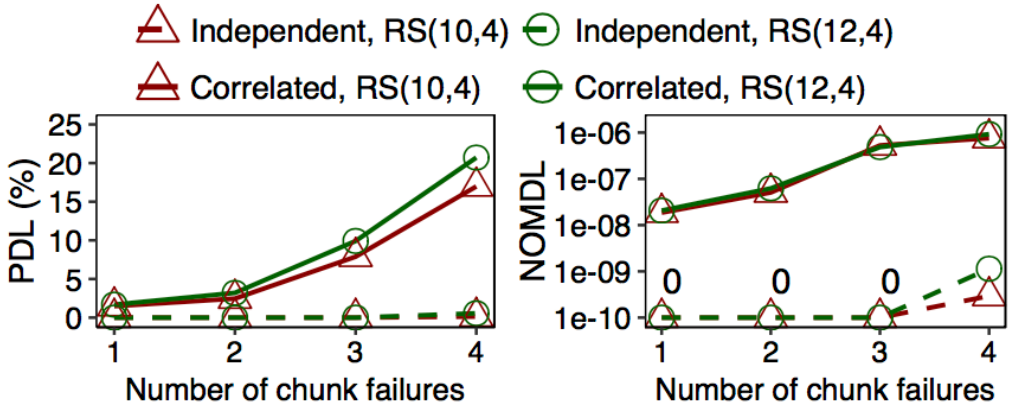
Simulation Results

- **Finding 13**: Erasure coding shows higher reliability than replication based on failure patterns in our dataset
- **Finding 14**: Redundancy schemes that are sufficient for tolerating independent failures may be insufficient for correlated failures



Simulation Results

- **Finding 15**: Lazy recovery is less suitable than eager recovery for tolerating correlated failures in our dataset
  - Eager recovery: a threshold of one
  - High reliability under only independent failures
  - Degrading reliability under correlated failures as the threshold increases



Conclusion

- We report 15 findings on correlated failures of SSDs based on large-scale dataset at Alibaba
  - Spatial and temporal correlations of SSD failures
  - Impact of different factors on correlated failures
  - Trace-driven simulation on reliability of various redundancy schemes under correlated failures
- Dataset and source code:
  - **https://github.com/alibaba-edu/dcbrain/tree/master/ssd_open_data**
  - **http://adslab.cse.cuhk.edu.hk/software/ssdanalysis**

Thank You!
Q & A