SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2015

RAIDShield: Characterizing, Monitoring, and Proactively Protecting Against Disk Failures

Ao Ma

Advanced Development Group

CTO office of EMC Core Technology Division

Pervasive RAID Protection

Disk failures are commonplace

- Whole-disk failure
- Partial failure

RAID is widely deployed

- Protect data against failures with redundancy

RAID Overview

Storage system is evolving

- Escalated use of less reliable drives causes more whole-disk failures
- Increasing disk capacity results in more sector errors

Existing solution

- Add extra redundancy (RAID5, RAID6, …)
  - Ensure data reliability at the cost of storage efficiency

Is adding extra redundancy an efficient solution?

We know tires can wear out or explode

What we did

Analyzed 1 million SATA disks and revealed

- Failure modes degrading RAID reliability
- Reallocated sectors reflect disk reliability deterioration
- Disk failure is predictable

Built RAIDShield, an active defense mechanism

- Reconstruct failing disk before it's too late!
- PLATE: single-disk proactive protection
  - Deployment eliminates 70% of RAID failures
- ARMOR: disk group proactive protection
  - Recognize vulnerable RAID groups

Outline

Background

Disk failure analysis

RAIDShield

- Identify failure indicator
- Reallocated Sector (RS) characterization
- Single disk proactive protection
- Disk group proactive protection

Whole-disk Failure Definition

Disk failure does not follow a fail-stop model

The production systems studied define failure as

- Connection is lost
- An operation exceeds the timeout threshold
- Write fails

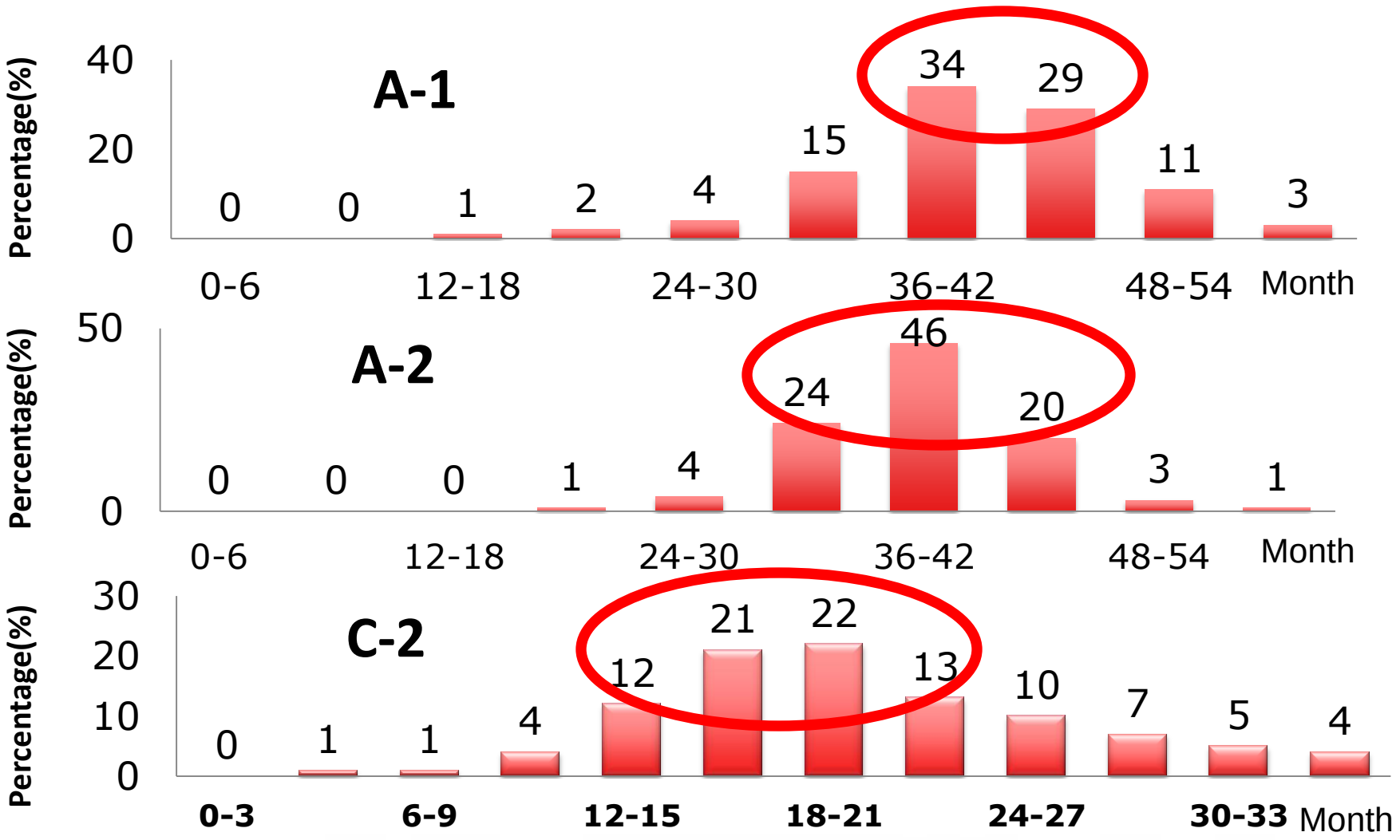
Disk Data Collection

Disk Model	Population (Thousands)	First Deployment	Log Length (Months)
A-1	34	06/2008	60
A-2	165	11/2008	60
B-1	100	06/2008	48
C-1	93	10/2010	36
C-2	253	12/2010	36
D-1	384	09/2011	21

- Each disk drive model is denoted as <family-capacity>
- Relative sizes within a family are ordered by the capacity number
  - E.g. A-2 is larger than A-1

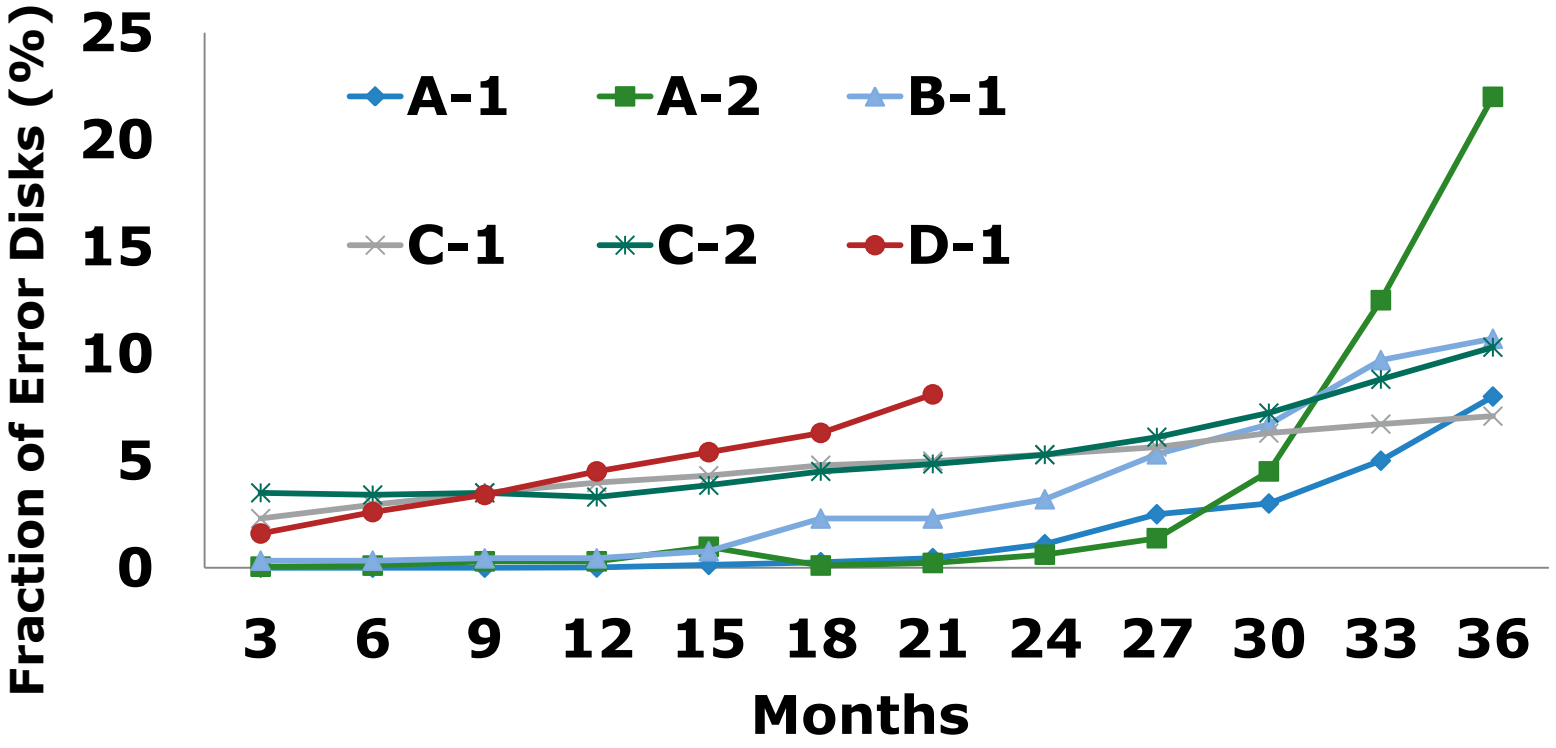
What Do Real Disk Failures Look Like?

Distribution of Lifetime of Failed Drives



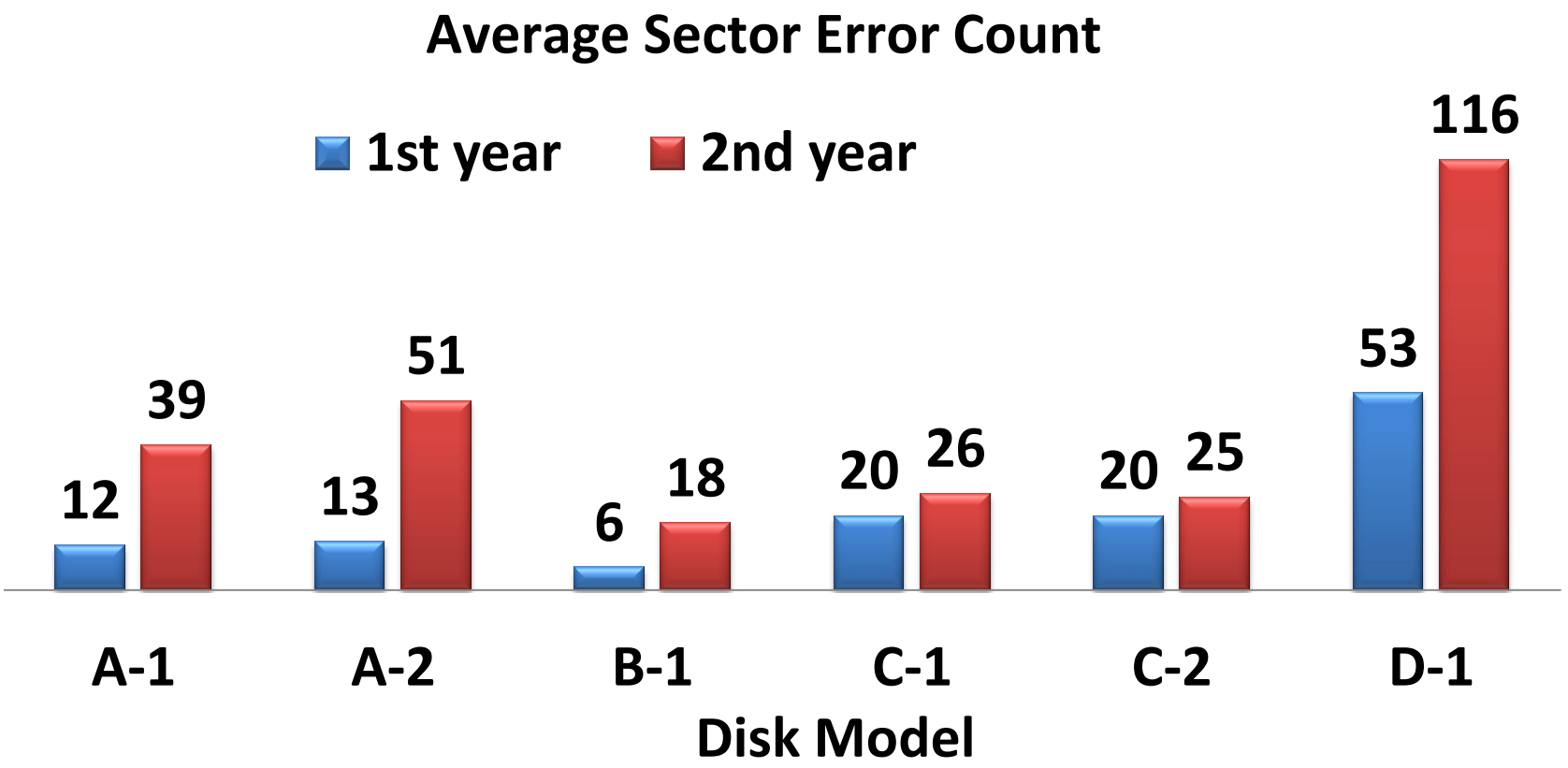
A large fraction of failed drives are found at a similar age

Increasing Frequency of Sector Errors



The number of affected disks keep growing

Increasing Frequency of Sector Errors



Sector error numbers increases continuously

Passive Redundancy is Inefficient

Drive failing at a similar age

- Failure rate is not constant
- A high risk of multiple simultaneous failures

Increasing frequency of sector errors

- Exacerbate risk of reconstruction failures

Ensuring reliability in the worst case requires adding considerable extra redundancy, making it unattractive from a cost perspective

RAIDShield, The Proactive Protection

Intuition

- Disk failing at similar time indicates a hidden pattern

Motivation

- Exploring hidden pattern helps proactively recognize impending failures and migrate vulnerable data in advance
- Ensure data safety with minimal redundancy

Methodology

- Identify indicator of impending failure
- Indicator characterization
- Proactive protection

Identify Failure Indicator

Potential indicators

- Various disk errors

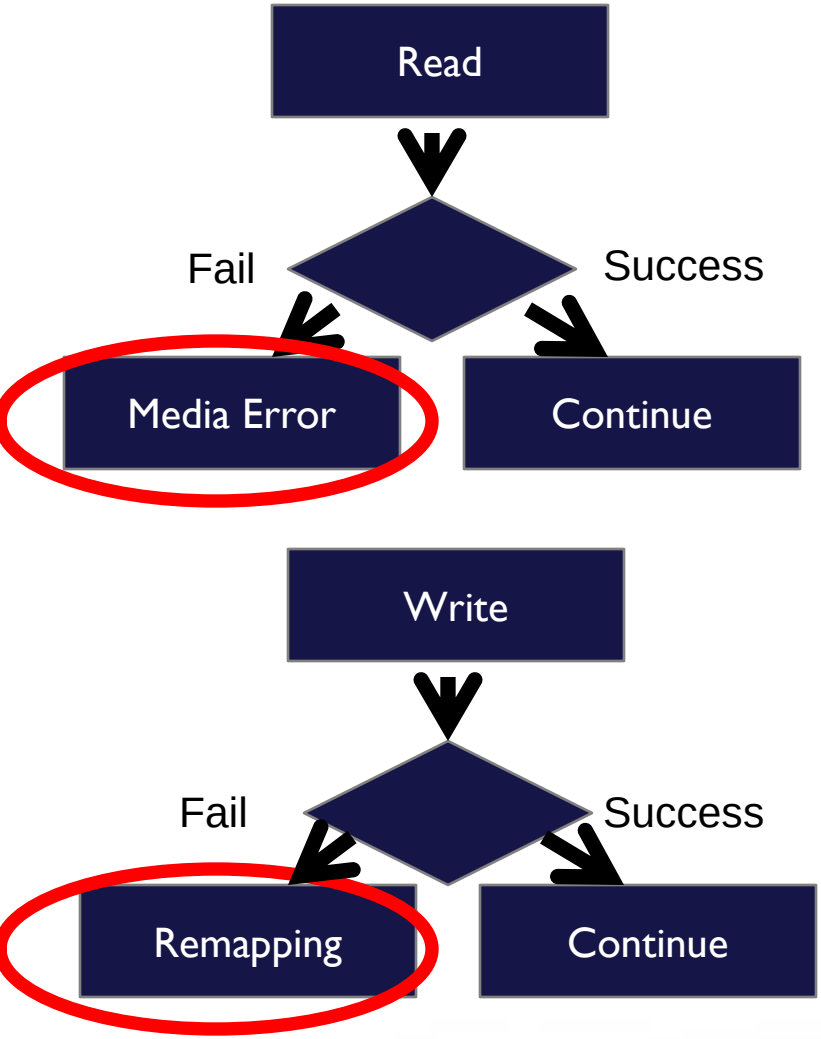
Criteria of a good indicator

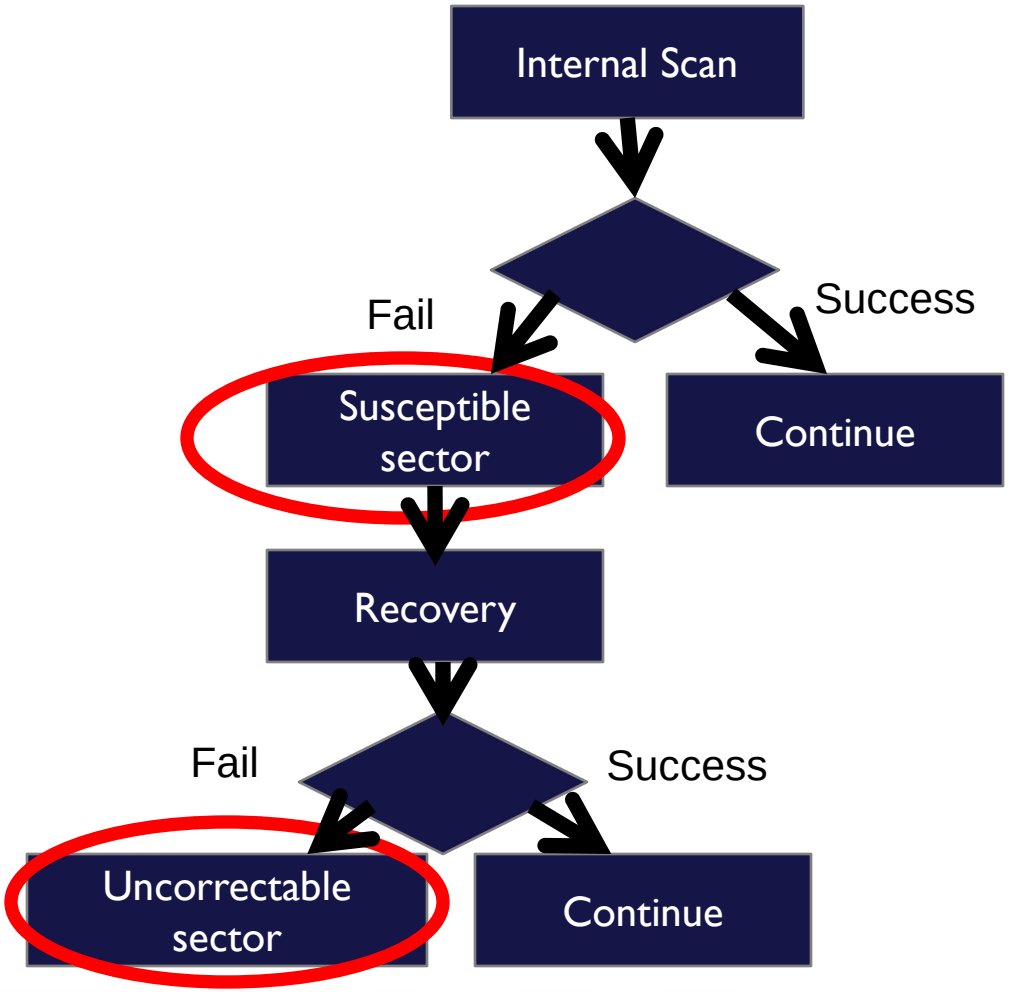
- It happens much more frequently on failed disks rather than working disks

Approach

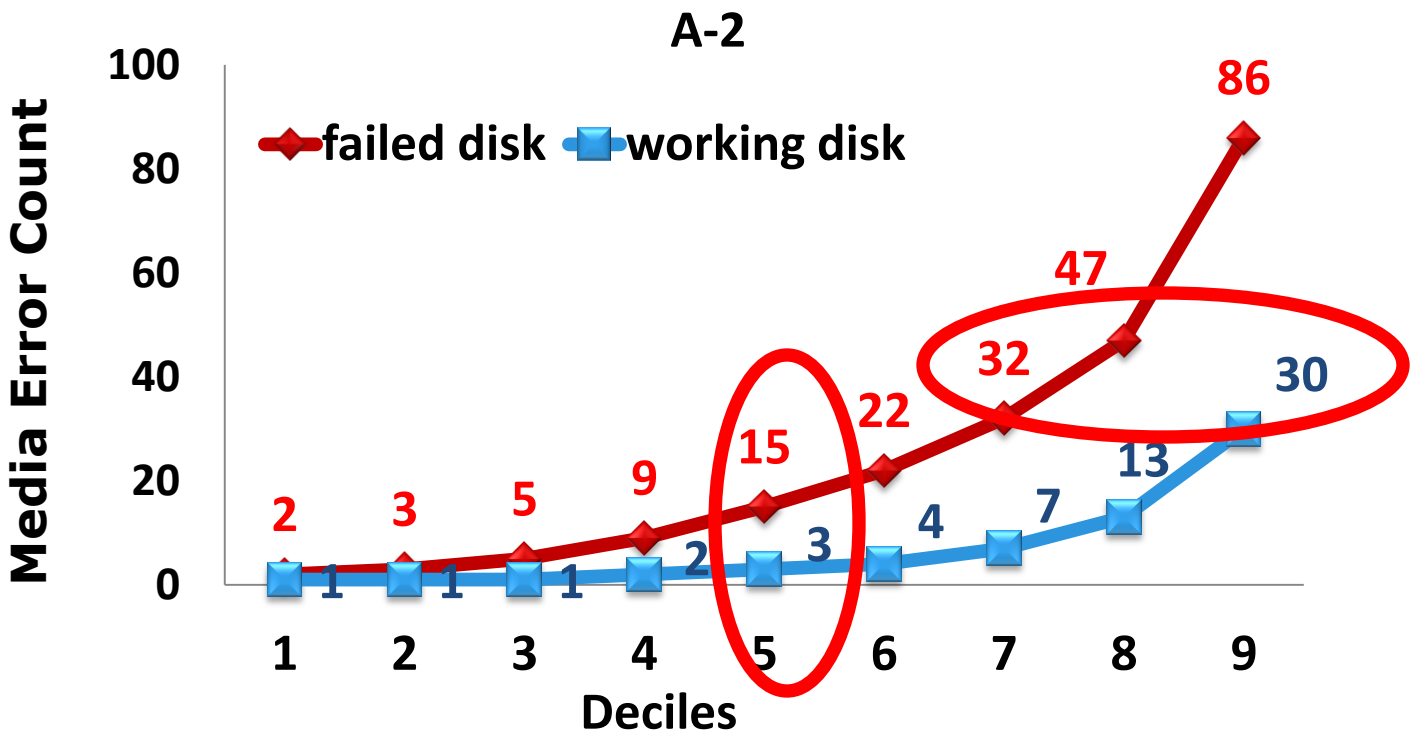
- Quantify the discrimination between error value on failed disks and working ones
  - Deciles comparison is used

Typical Disk Faults and Potential Indicators





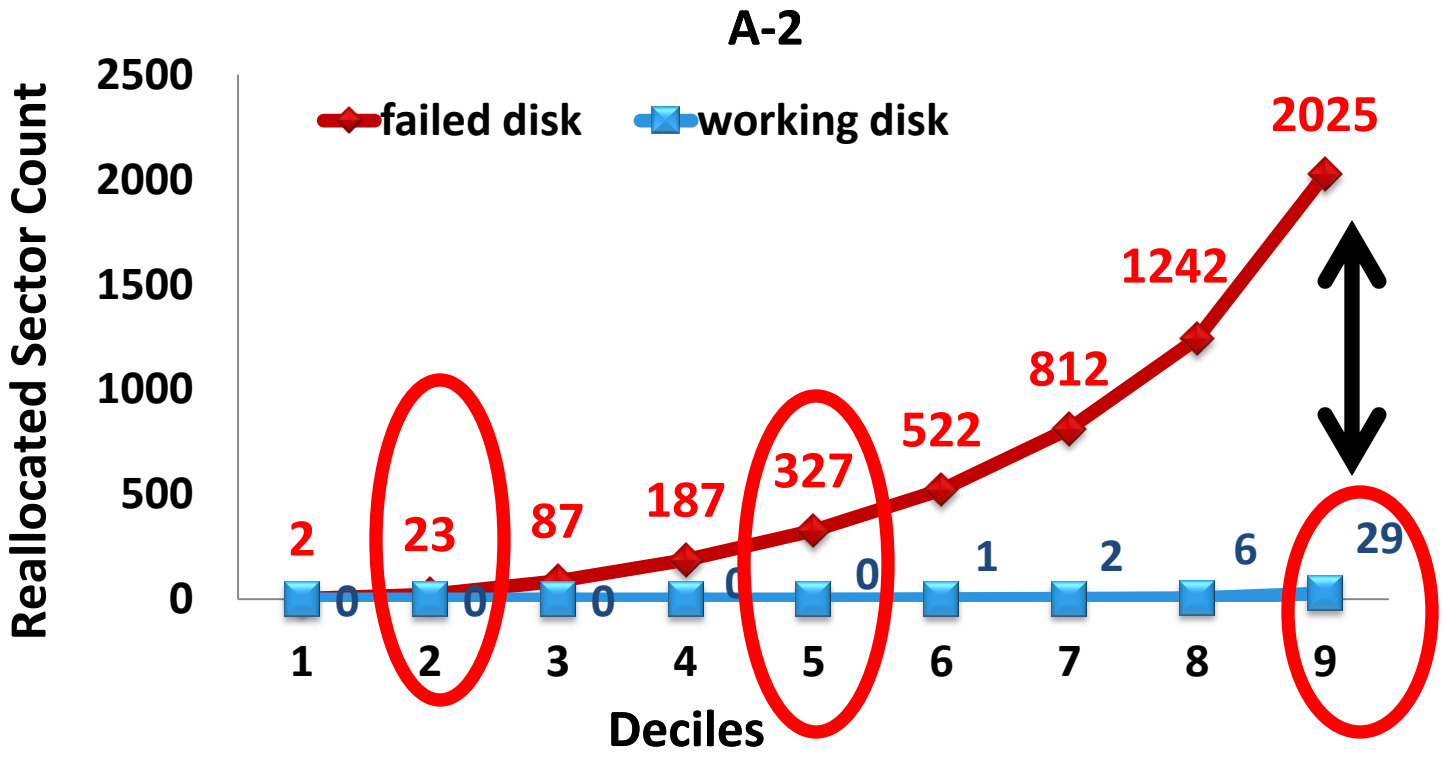
Media Error Comparison



Failed disks have more media errors than working ones

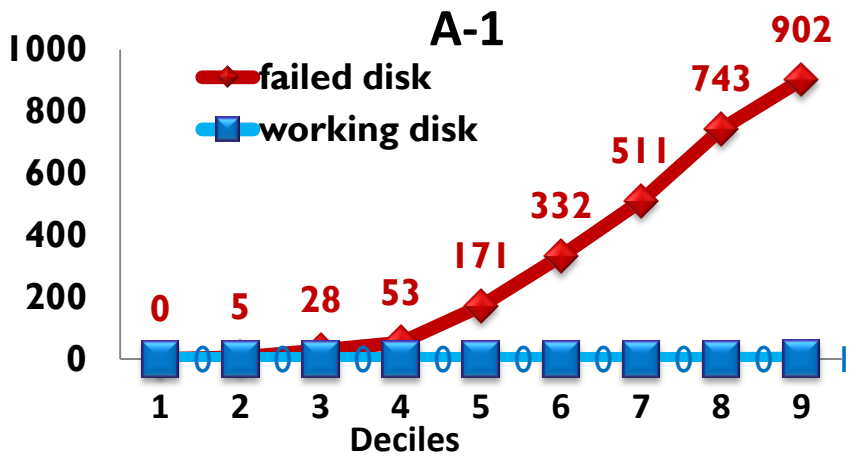
The discrimination is not significant enough

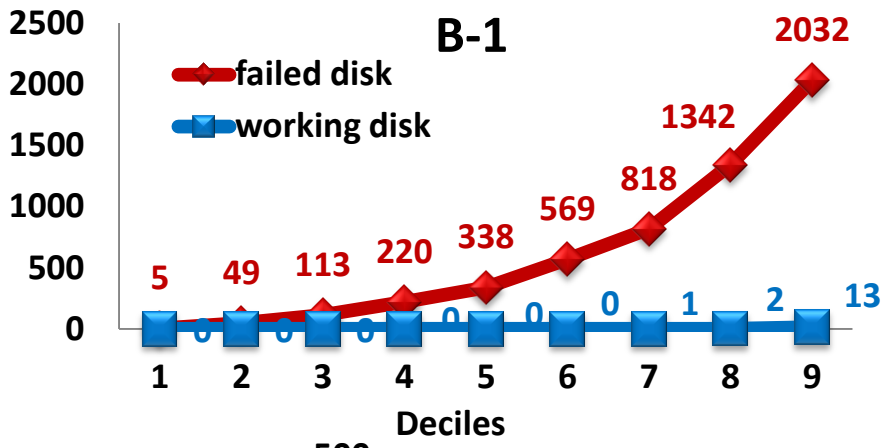
Reallocated Sector (RS) Comparison

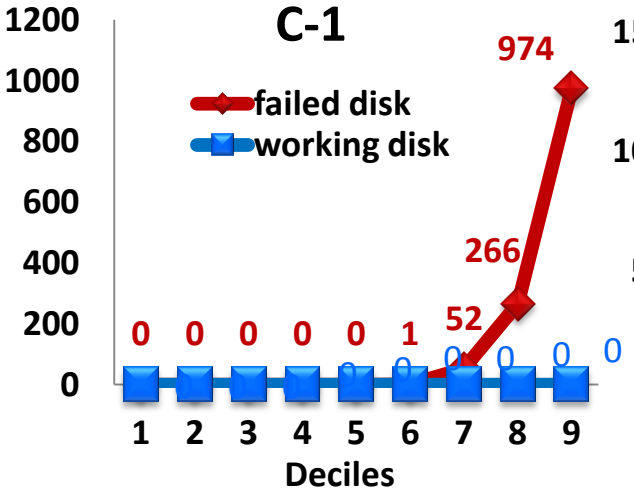


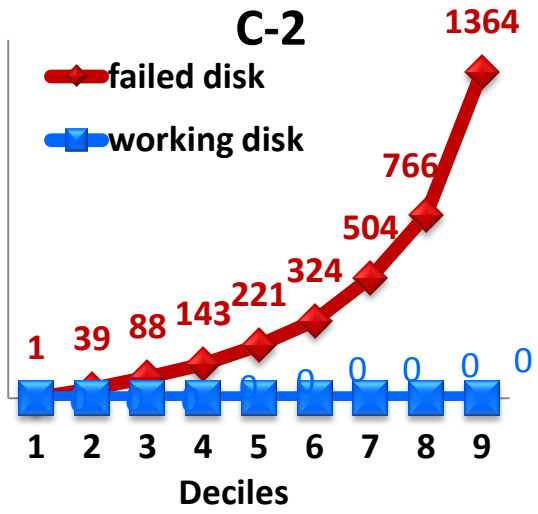
RS is strongly correlated with disk failures

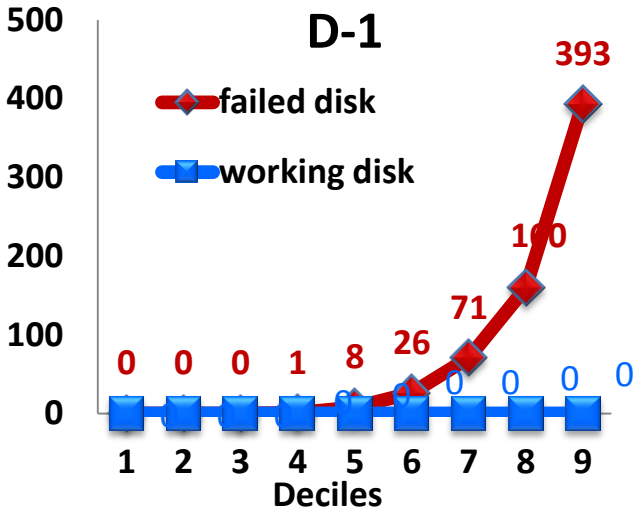
Reallocated Sector (RS) Comparison











All drives studied demonstrate similar correlation

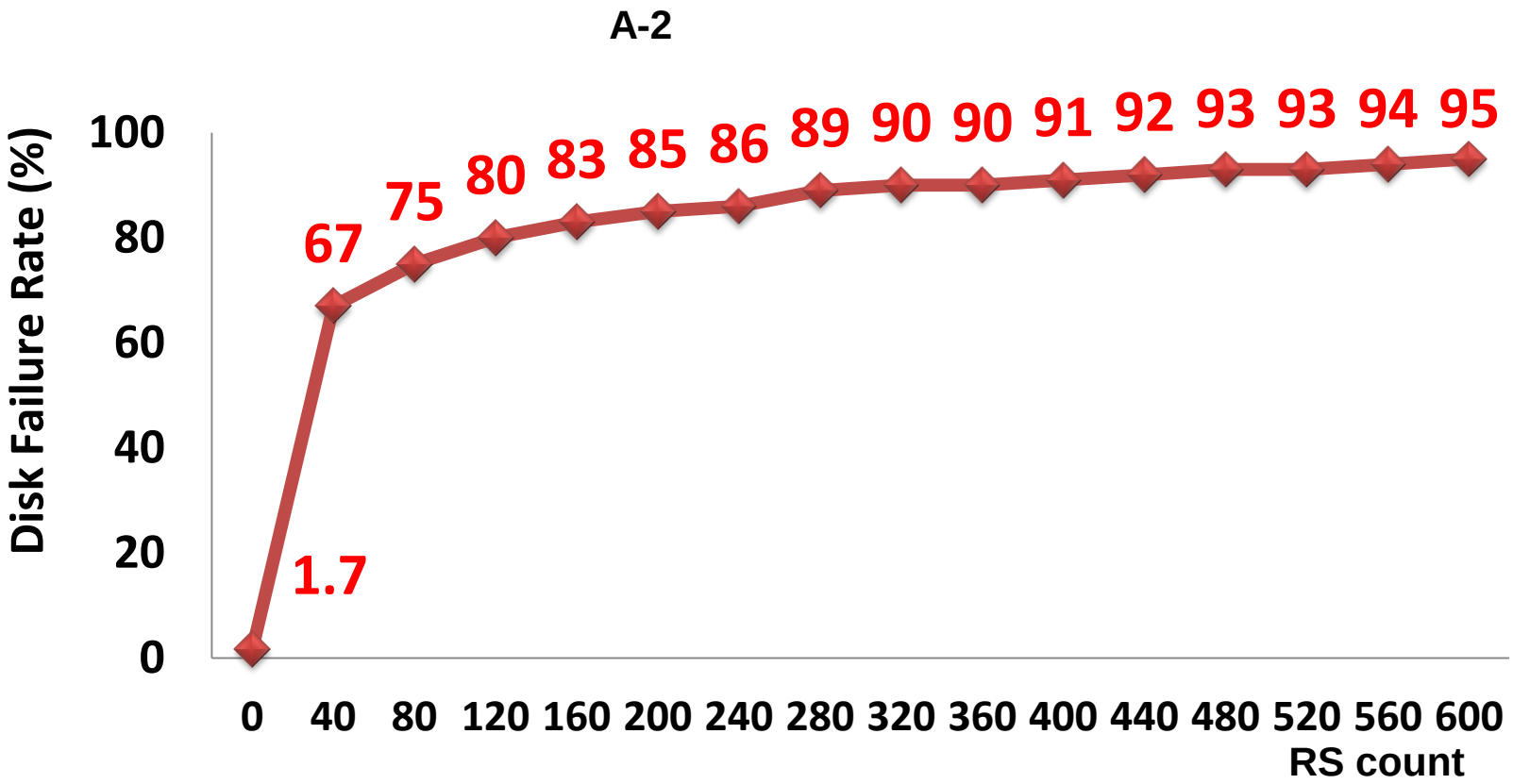
Correlation Between Sector Errors And Whole-disk Failure

Most failed drives tend to have a larger number of RS than working ones

RS is strongly correlated with whole-disk failures, followed by media errors, pending sector errors and uncorrectable sector errors

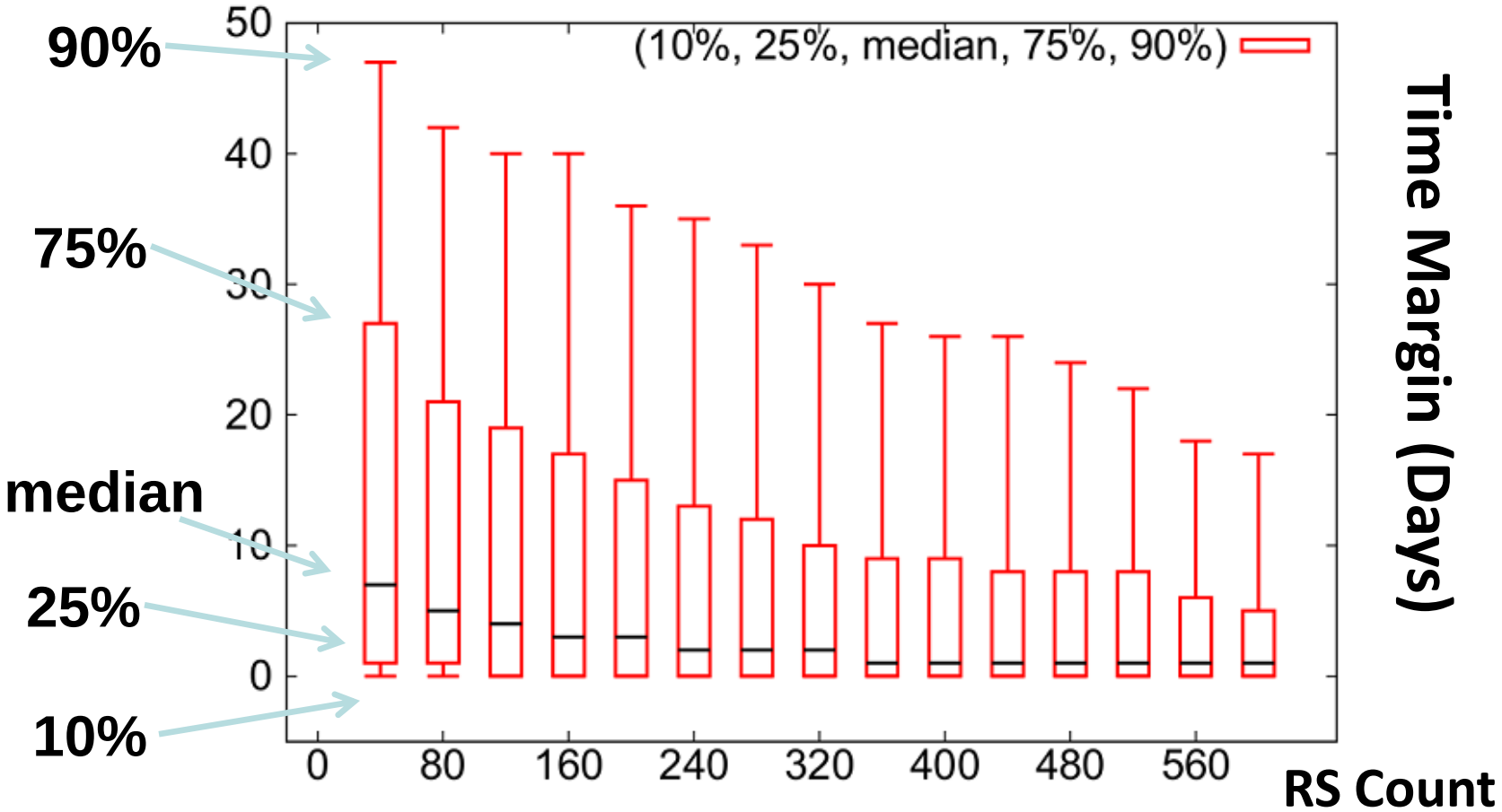
RS is a strong indicator of impending disk failure

RS Characterization (1)
Disk Failure Rate Given Different RS Count



Larger RS count implies higher failure rate in two-month window

RS Characterization (2)
Disk Failure Time Given Different RS Count



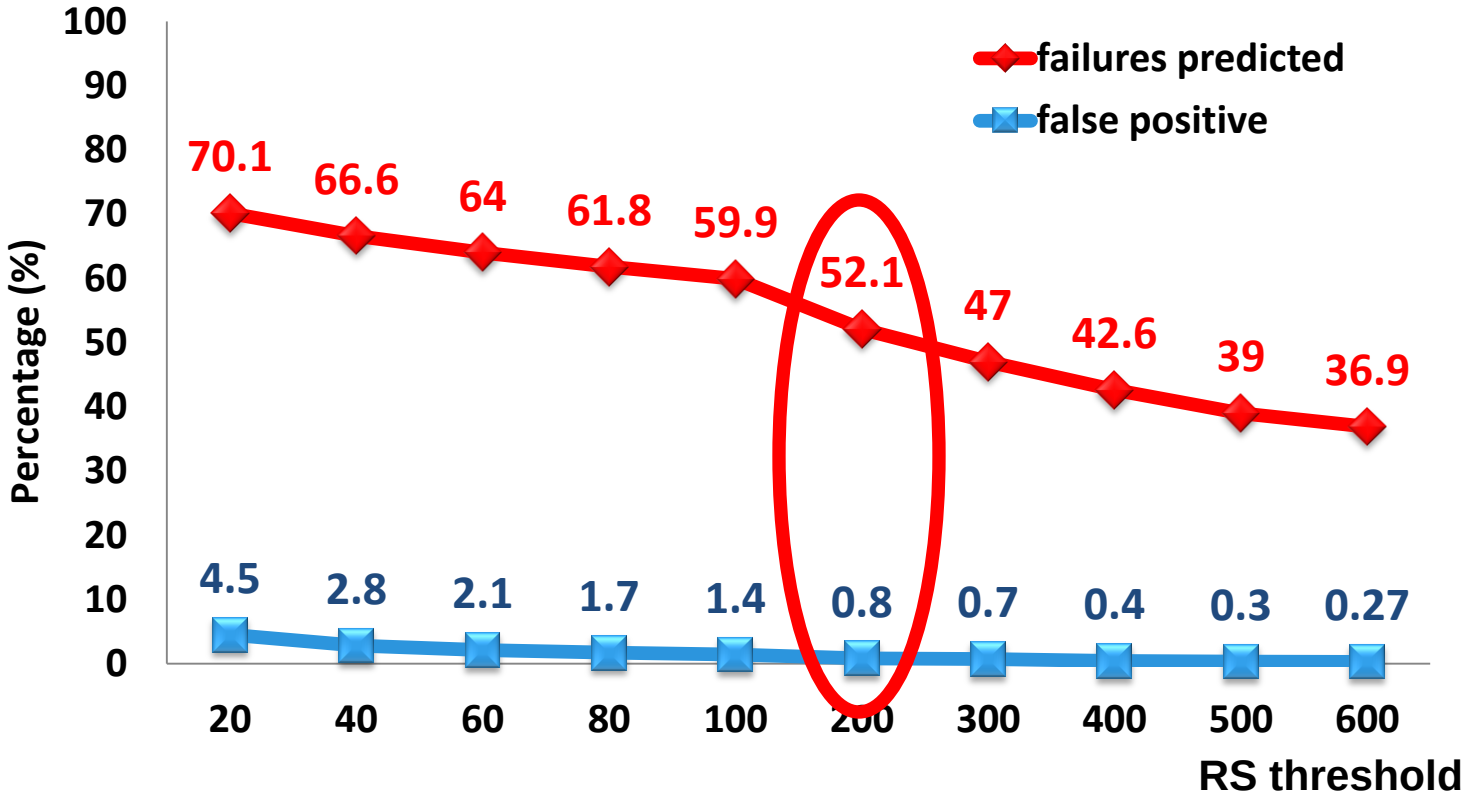
Larger RS count, faster to fail

PLATE: Single Disk Proactive Protection

RS count indicates the degree of disk reliability deterioration

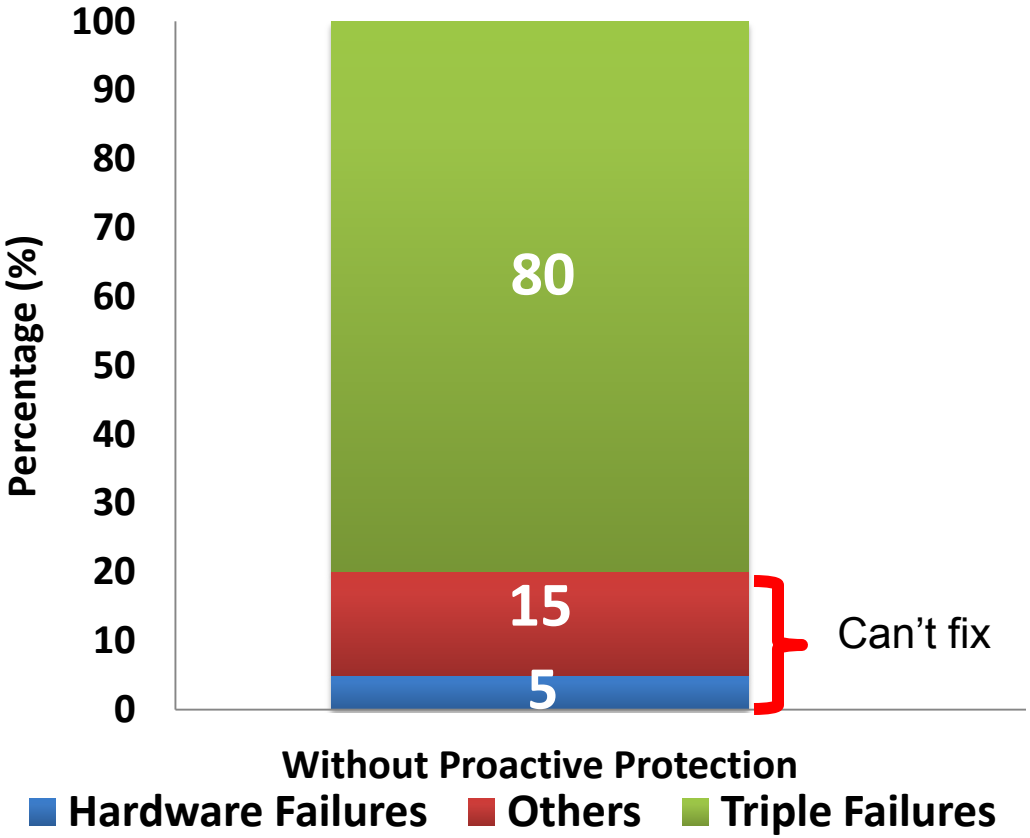
Use the RS count to predict impending disk failure in advance

Simulation Result: Failures Captured Rate Given Different RS Threshold



Both the predicted failure and false positive rates decrease as the threshold increases

PLATE Deployment Result: Causes of Recovery Incidents



Single proactive protection reduces about 70% of RAID failures, equivalent to 88% of the triple-disk failures

Motivation of ARMOR: The RAID Group Proactive Protection

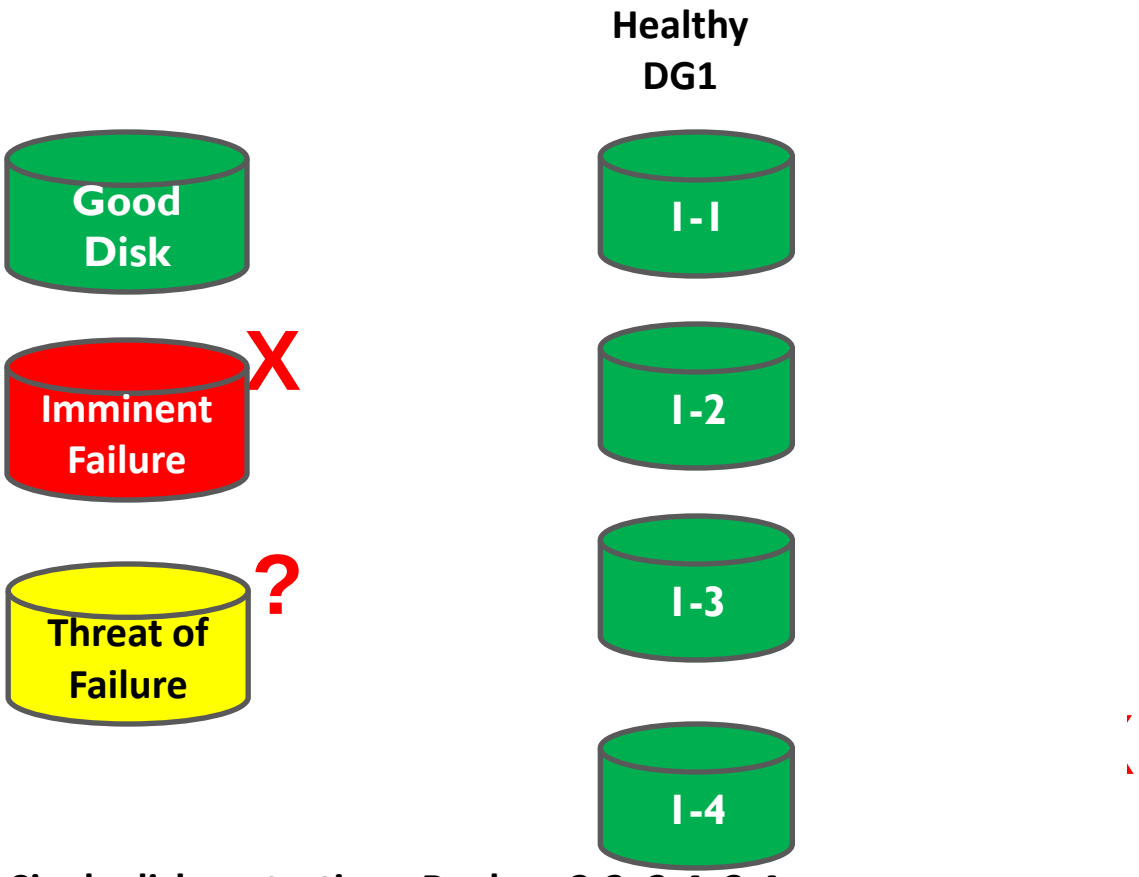
10% remaining triple failures

- PLATE misses RAID failures caused by multiple less reliable drives, whose RS counts haven't exceed the threshold

Triage

- Prioritize disk groups with highest risk

Disk Group Protection Example



Single disk protection: Replace 2-3, 2-4, 3-4
(PLATE) Can't identify DG4 nor the difference between DG2 and DG3

Group protection: Replace DG4 or increase redundancy
(ARMOR) Protect DG4 and recognize the difference between DG2 and DG3

ARMOR Methodology

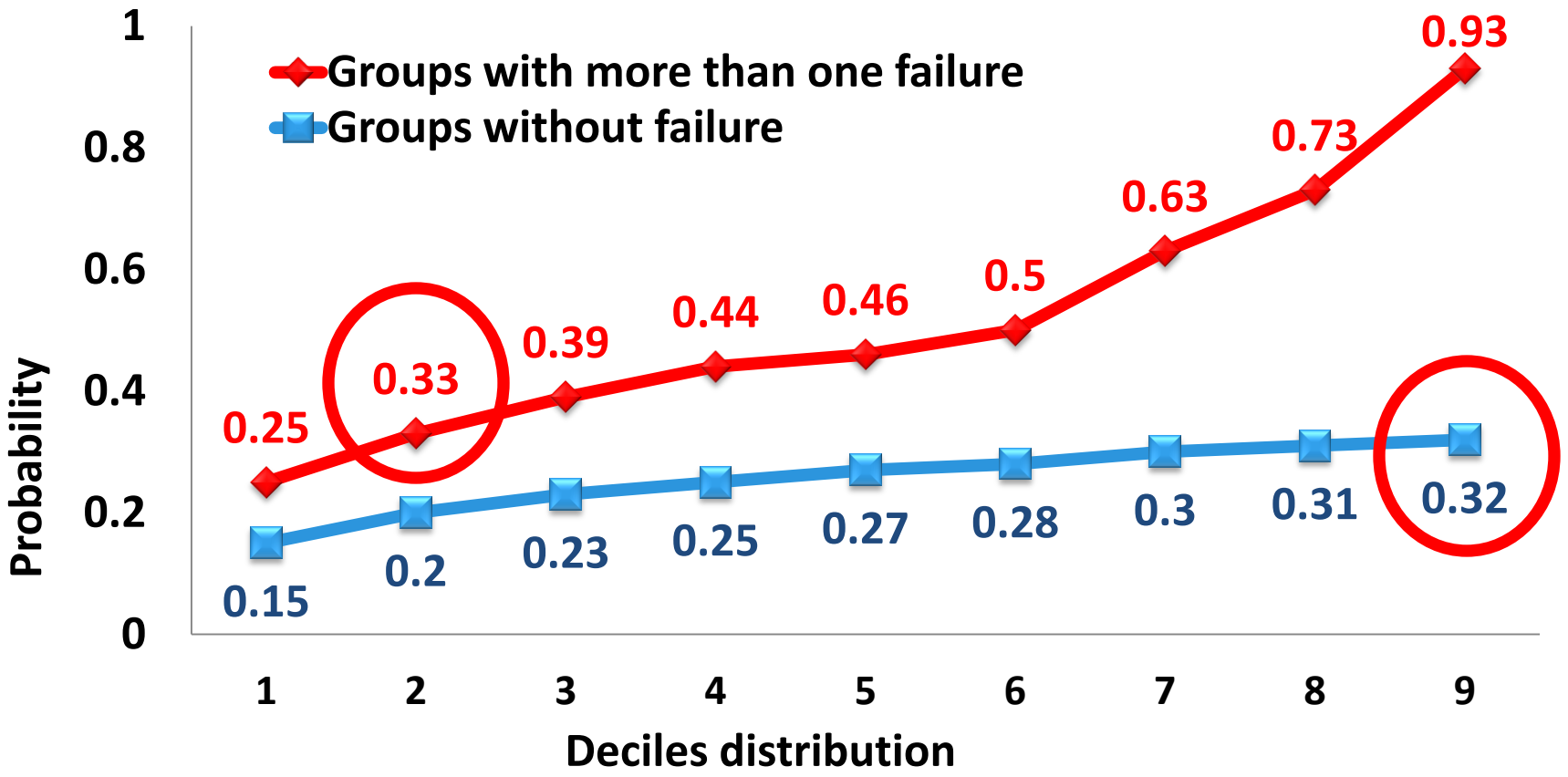
Calculate the single disk failure probability

- Conditional probability through Bayes Theorem

Calculate the probability of a vulnerable RAID

- Combination of those single disk probabilities through joint probability

Evaluation



The discrimination shows ARMOR is effective to recognize endangered DGs

In practice, it identifies most DG failures that are not predicted by PLATE

Related Work

Google reports SMART metrics such as reallocated sector strongly suggest an impending failure, but they also determine that half of the failed disks show no such errors [Pinheiro'07]

- Different workload and RAID rewrite

Disk failure prediction

- Backblaze reports similar indicators found in SMART
- Average maximum latency [Goldszmidt'12]
- SMART failure prediction [Murray'05, Hughes'02]

Summary

We analyzed 1 million SATA drives

- Observe failure modes degrading RAID reliability
- Reveal RS count reflects the disk reliability deterioration
- Disk failure is predictable

We built RAIDSHIELD, an active defense mechanism

- PLATE: single disk proactive protection
  - Deployment eliminates 70% of RAID failures
- ARMOR: disk group proactive protection
  - Recognize vulnerable RAID groups
  - Hope to deploy in future

Is adding extra redundancy an efficient solution?

- Use as much redundancy as needed to ensure availability
- Proactive replacement should decrease the level needed

RAIDShield

Questions?

Contact: ao.ma@emc.com

Joint work with:
Fred Douglis, Guanlin Lu, Darren Sawyer, Surendar Chandra, Windsor Hsu

Acknowledgements
- Andrea Arpaci-Dusseau and Remzi Arpaci-Dusseau
- Data Domain engineering team
- Members of AD and CTO office, Stephen Manley