LAKSHMI BAIRAVASUNDARAM, GARTH GOODSON, BIANCA SCHROEDER, ANDREA ARPACI-DUSSEAU, AND REMZI ARPACI-DUSSEAU

data corruption in the storage stack: a closer look

Lakshmi N. Bairavasundaram is a PhD student in Computer Sciences at the University of Wisconsin, Madison, working with advisors Prof. Andrea Arpaci-Dusseau and Prof. Remzi Arpaci-Dusseau. He received his BE from Anna University, India, and his MS from the University of Wisconsin, Madison.

laksh@cs.wisc.edu

Garth Goodson is a researcher at NetApp, Inc., with interests in virtualization, distributed systems, and new memory technologies. He received his PhD in 2004 from Carnegie Mellon University under the supervision of Greg Ganger.

Garth.Goodson@netapp.com

Bianca Schroeder is an Assistant Professor in the Department of Computer Science at the University of Toronto. Before coming to Toronto, Bianca completed her PhD and a two-year postdoc at Carnegie Mellon University. Her research focuses on computer systems and has earned her multiple best paper awards.

bianca@cs.toronto.edu

Andrea Arpaci-Dusseau is an associate professor of Computer Sciences at the University of Wisconsin, Madison. She received her BS from Carnegie Mellon University and her MS and PhD from the University of California, Berkeley, under advisor David Culler.

dusseau@cs.wisc.edu

Remzi Arpaci-Dusseau is an associate professor of Computer Sciences at the University of Wisconsin, Madison. He received his BS from the University of Michigan and his Master's and PhD from the University of California, Berkeley, under advisor David Patterson.

remzi@cs.wisc.edu

ONE OF THE BIGGEST CHALLENGES IN designing storage systems is providing the reliability and availability that users expect. A serious threat to reliability is silent data corruption (i.e., corruption not detected by the disk drive). In order to develop suitable protection mechanisms against corruption, it is essential to understand its characteristics. In this article, we present the results from the first large-scale field study of data corruption. We analyze over 400,000 corruption instances recorded in production storage systems containing a total of 1.53 million disk drives, over a period of 41 months.

One primary cause of data loss is disk drive unreliability. It is well known that hard drives are mechanical, moving devices that can suffer from mechanical problems, leading to drive failure and *latent sector errors* (detected by the disk's ECC). Less well known, however, is that current hard drives and controllers consist of hundreds of thousands of lines of low-level firmware code. Bugs in this firmware code can cause a more insidious type of disk error: silent data corruption, where the data is silently corrupted with no indication from the drive that an error has occurred.

Silent data corruptions could lead to data loss more often than latent sector errors, since, unlike latent sector errors, they cannot be detected or repaired by the disk drive itself. Worse, basic protection schemes such as RAID may also be unable to detect these problems, thereby returning corrupt data.

The most common technique used in storage systems to detect data corruption is the addition of a higher-level checksum for each disk block, which is validated on each disk block read. However, checksums do not protect against all forms of corruption. Therefore, in addition to checksums, NetApp storage systems also use filesystem-level disk block identity information to detect previously undetectable corruptions.

In order to improve the handling of corruption errors, we need to develop a thorough understanding of data corruption characteristics. Although recent studies provide information on whole disk failures [4, 5, 7] and latent sector errors [1], very little is known about data corruption, its prevalence, and its characteristics. This article summarizes the re-

sults of our study of data corruption first published in the 2008 USENIX FAST conference [2].

Detecting Data Corruption

The data we analyze is from tens of thousands of production and development NetApp storage systems from hundreds of customer sites. These storage systems are designed to detect and handle a wide range of disk-related errors, including silent data corruption. Data corruption may be caused by both hardware and software errors. Hardware bugs include bugs in the disk drive or the disk shelf firmware, bad memory, and adapter failures. Typically, it is not possible to identify the root cause of a corruption error. However, our storage system has several mechanisms in place to detect when data corruption occurs, to prevent propagation of corrupt data. We briefly describe two of those mechanisms.

DATA INTEGRITY SEGMENT

In order to detect corruptions, the system stores extra information along with each disk block. For every 4KB file system block written, the storage controller writes a 64-byte data integrity segment along with the disk block.

One component of the data integrity segment is a checksum of the entire 4KB filesystem block. The checksum is validated by the RAID layer whenever the data is read. Once a corruption has been detected, the original block can usually be restored through RAID reconstruction. We refer to corruptions detected by RAID-level checksum validation as *checksum mismatches*.

A second component of the data integrity segment is the block identity information. The identity information refers to where the block resides within the file system (e.g., this block belongs to inode 5 at offset 100). This identity is cross-checked at file read time to ensure that the block being read belongs to the file being accessed. If, on file read, the identity does not match, the data is reconstructed from parity. We refer to corruptions that are not detected by checksums, but detected through filesystem identity validation, as *identity discrepancies*.

DATA SCRUBBING

In order to proactively detect errors, the RAID layer periodically *scrubs* all disks. A data scrub issues read operations for each physical disk block, computes a checksum over its data, and compares the computed checksum to the checksum located in its data integrity segment. If the checksum comparison fails (i.e., a checksum mismatch), the data is reconstructed from other disks in the RAID group, after those checksums are also verified.

We refer to these cases of mismatch between data and parity as *parity inconsistencies*. Note that data scrubs are unable to validate the extra filesystem identity information stored in the data integrity segment, since this information only has meaning to the file system.

CHECKSUM MISMATCHES

As just described, corruption events are classified into three classes: checksum mismatches, identity discrepancies, and parity inconsistencies. In this article we focus on checksum mismatches, since we find that they occur

with the highest frequency. Checksum mismatches can result from (i) data content corrupted by components within the data path, or (ii) a torn write, wherein only a portion of the data block is written successfully, or (iii) a misdirected write, wherein the data is written to either the wrong disk or the wrong location on disk, thus overwriting and corrupting data [3, 6].

Our study focuses on the characteristics of checksum mismatches, such as their frequency, the factors that affect the development of checksum mismatches, and the statistical properties of checksum mismatches. In our analysis we refer to a 4KB file system block with a checksum mismatch as a *checksum mismatch block*. We call a disk drive a *corrupt disk* if it has at least one checksum mismatch block.

DATA COLLECTION

The data we collected covers a period of 41 months starting in January 2004 and includes tens of thousands of NetApp storage systems containing a total of 1.53 million disk drives. The data was collected by a built-in, low-overhead mechanism called AutoSupport. AutoSupport is included in every NetApp storage system and logs system events back to a central repository.

Our disk drive sample is not only large but also diverse. The disks belong to 14 different *disk families*. Each disk family refers to one particular disk drive product. Typically, disks in the same family only differ in the number of platters and/or heads. The drives come from 31 distinct *disk models*. A disk model is the combination of a disk family and a particular disk size. Finally, the drives cover two different *disk classes*: an enterprise class of Fibre Channel disks and a nearline class of SATA disks.

Result Synopsis

During the 41-month period covered by our data we observed a total of about 400,000 checksum mismatches. Of the total sample of 1.53 million disks, 3855 disks developed checksum mismatches: 3088 of the 358,000 SATA disks (0.86%) and 767 of the 1.17 million Fibre Channel disks (0.065%). This indicates that SATA disks may be more susceptible to corruption leading to checksum mismatches than Fibre Channel disks. On average, each disk developed 0.26 checksum mismatches. By considering only corrupt disks, the mean number of mismatches per disk is 104, the median is 3, and the mode (i.e., the most frequently observed value) is 1 mismatch per disk. The maximum number of mismatches observed for any single drive was 33,000.

DISK CLASS, MODEL, AGE, AND SIZE

We start by examining the dependence of checksum mismatches on factors such as disk class, disk model, and disk age. A disk's age is its time in the field since its ship date.

Figures 1 and 2 shows the cumulative distribution function of the time in the field until the first checksum mismatch occurs for SATA and Fibre Channel disks, respectively.

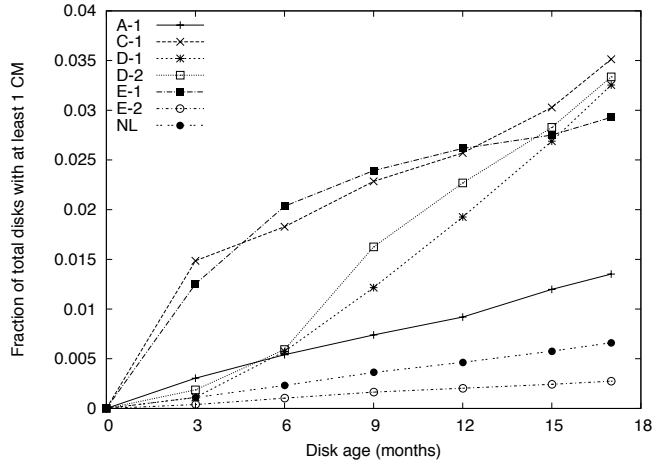
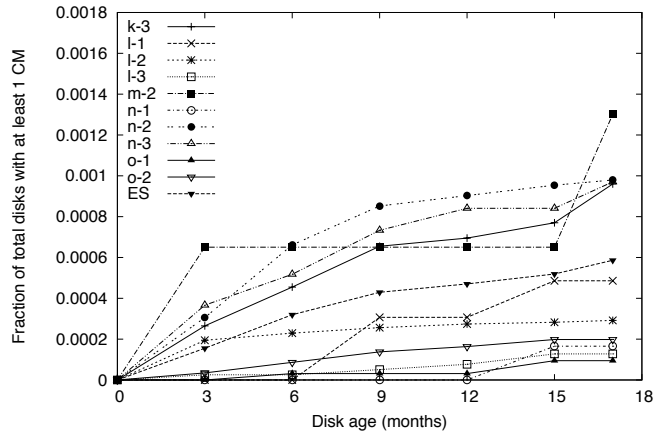


FIGURE 1: CUMULATIVE DISTRIBUTION FUNCTION OF THE TIME IN THE FIELD UNTIL THE FIRST CHECKSUM MISMATCH OCCURS FOR SATA DISKS

FIGURE 2: CUMULATIVE DISTRIBUTION FUNCTION OF THE TIME IN THE FIELD UNTIL THE FIRST CHECKSUM MISMATCH OCCURS FOR FIBRE CHANNEL DISKS



Observation: SATA disks have an order of magnitude higher probability of developing checksum mismatches than Fibre Channel disks.

We find that 0.66% of SATA disks develop at least one mismatch during the first 17 months in the field, whereas only 0.06% of Fibre Channel disks develop a mismatch during that time.

Observation: The probability of developing checksum mismatches varies significantly across different disk models within the same disk class.

We see that there is an order of magnitude difference for developing at least one checksum mismatch after 17 months between the two most extreme SATA disk models: 3.5% for one model vs. 0.27% for the other.

Observation: Age affects different disk models differently with respect to the probability of developing checksum mismatches.

On average, as SATA disks age, the probability of developing a checksum mismatch is fairly constant, with some variation across the models. As Fibre Channel disks age, the probability of developing the first checksum mismatch decreases after about 6–9 months and then stabilizes.

Observation: There is no clear indication that disk size affects the probability of developing checksum mismatches.

Since the impact of disk size on the fraction of disks that develop checksum mismatches is seen in only 7 out of 10 families, we conclude that disk size does not necessarily impact the probability of developing checksum mismatches.

CHECKSUM MISMATCHES PER CORRUPT DISK

Observation: The number of checksum mismatches per corrupt disk varies greatly across disks. Most corrupt disks develop only a few mismatches each. However, a few disks develop a large number of mismatches.

A significant fraction of corrupt disks develop only one checksum mismatch. However, a small fraction of disks develop several thousand checksum mismatches (i.e., 1% of the corrupt disks produce more than half of all mismatches recorded in the data).

Observation: On average, corrupt Fibre Channel disks develop many more checksum mismatches than corrupt SATA disks.

Within 17 months, 50% of corrupt disks develop about 2 checksum mismatches for SATA disks but almost 10 for Fibre Channel disks. Given that very few Fibre Channel disks develop checksum mismatches in the first place, it might make sense to replace the Fibre Channel disk when the first mismatch is detected.

Observation: Checksum mismatches within the same disk are not independent.

We found that the conditional probability of developing further checksum mismatches, given that a disk has at least one mismatch, is higher than the probability of developing the first mismatch. We also found that one particular SATA disk model is particularly aberrant: Around 30% of its corrupt disks develop more than 1000 checksum mismatches.

DEPENDENCE BETWEEN DISKS IN THE SAME SYSTEM

Observation: The probability of a disk developing a checksum mismatch is not independent of that of other disks in the same storage system.

Although most systems with checksum mismatches have only one corrupt disk, we do find a considerable number of instances where multiple disks develop checksum mismatches within the same storage system. In fact, one of the systems in the study that used SATA disks had 92 disks develop checksum mismatches. The probability of 92 disks developing errors independently is less than 10^{-12}, much less than 10^{-5}, the approximate fraction of systems represented by one system.

SPATIAL LOCALITY

We measure spatial locality by examining whether each checksum mismatch block has another checksum mismatch block (a *neighbor*) within progressively larger regions (*locality radius*) around it on the same disk. For example, if in a disk, blocks numbered 100, 200, and 500 have checksum mismatches, then blocks 100 and 200 have one neighbor at a locality radius of 100, and all blocks (100, 200, and 500) have at least one neighbor at a locality radius of 300.

Observation: Checksum mismatches have very high spatial locality. Much of the observed locality is due to consecutive disk blocks developing corruption. Beyond consecutive blocks, the mismatches show very little spatial locality.

For more than 50% of the checksum mismatch blocks in SATA disks and more than 40% of the checksum mismatch blocks in Fibre Channel disks, the immediate neighboring block also has a checksum mismatch (on disks with between 2 and 10 mismatches). These percentages indicate very high spatial locality.

It is interesting to examine how many consecutive blocks have mismatches. We find that, among drives with at least 2 checksum mismatches, on average 3.4 consecutive blocks are affected. In some cases, the length of consecutive runs can be much higher. About 3% of drives with at least 2 mismatches see one or more runs of 100 consecutive blocks with mismatches, and 0.7% of drives with at least 2 mismatches see one or more runs of 1000 consecutive mismatches.

TEMPORAL LOCALITY

Observation: Most checksum mismatches are detected within one minute of a previous detection of a mismatch.

Observation: Checksum mismatches also exhibit temporal locality over larger time windows and beyond the effect of detection time.

The first observation might not be surprising, since it could just be an artifact of the manner in which the detection takes place (by scrubbing). In order to remove the impact of detection time, we examined temporal locality over larger time windows. For each drive, we first determined the number of checksum mismatches experienced in each two-week time window that the drive was in the field and then computed the autocorrelation function (ACF) on the resulting time series. The ACF can be used to determine whether the number of mismatches in one two-week period of our time series is correlated with the number of mismatches observed in two-week periods later.

If checksum mismatches in different two-week periods were independent (no temporal locality on bi-weekly and larger time scales), the autocorrelation would be close to zero at all time lags. Instead, we observe strong autocorrelation even for large lags in the range of up to 10 months.

DISCOVERY

The severity of a data corruption event depends on when it is discovered. If a checksum mismatch is encountered during RAID reconstruction, data loss can result if the system is not configured to handle simultaneous disk failures.

Figure 3 (on p. 12) shows the distribution of requests that detect checksum mismatches. There are five types of requests that discover checksum mismatches: (i) file system reads (*FS Read*); (ii) writes by the RAID layer (*Write*); (iii) reads for disk copy operations (*Non-FS Read*); (iv) reads for scrubbing (*Scrub*); and (v) reads for RAID reconstruction (*Reconstruction*).

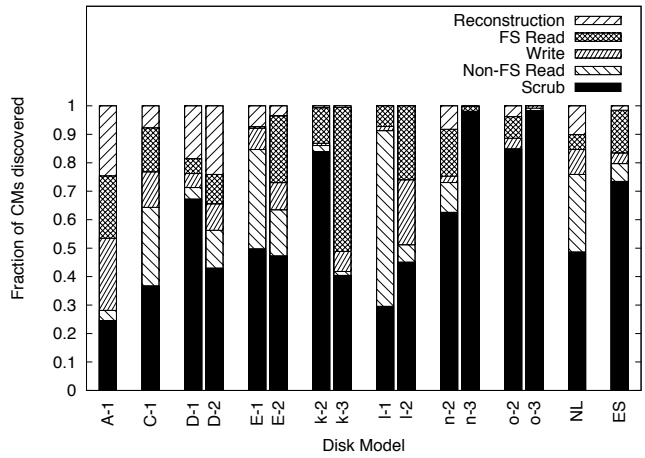


FIGURE 3: DISTRIBUTION OF REQUESTS THAT DETECT CHECKSUM MISMATCHES

Observation: RAID reconstruction encounters a non-negligible number of checksum mismatches.

We see that, on average, data scrubbing discovers about 49% of the checksum mismatches in SATA disks and 73% of the checksum mismatches in Fibre Channel disks. Despite the use of data scrubbing, we find that RAID reconstruction discovers about 8% of the checksum mismatches in SATA disks. For some models more than 20% of checksum mismatches were detected during RAID reconstruction. This observation implies that (i) data scrubbing should be performed more aggressively and (ii) systems should consider protection against double disk failures.

COMPARISON TO LATENT SECTOR ERRORS

When comparing checksum mismatches to latent sector errors we find some interesting similarities and differences:

- Frequency: The probability of developing checksum mismatches is about an order of magnitude smaller than that for latent sector errors.
- Disk model: For both error types, the development of errors depends on the disk model. Interestingly, the SATA disk model with the highest percentage of disks developing latent sector errors also had the lowest percentage of disks developing checksum mismatches.
- Disk class: For both error types, Fibre Channel disks are less likely to develop an error than SATA disks. Surprisingly, however, in both cases, once an error has developed, Fibre Channel disks develop a higher number of errors than SATA disks.
- Spatial locality: Both latent sector errors and checksum mismatches show high spatial locality. However, the locality radius is significantly larger for latent sector errors.

We also found a weak positive correlation between checksum mismatches and latent sector errors. The conditional probability of a latent sector error, given that a disk has checksum mismatch, is about 1.4 times higher than the unconditional probability of a latent sector error for SATA disks and about 2.2 times higher for Fibre Channel disks. We also verified the existence of a correlation between the two error types by performing a chi-square test for independence.

Lessons Learned

We present some of the lessons learned from the analysis for corruption-proof storage system design.

- Albeit not as common as latent sector errors, data corruption does happen. For some drive models as many as 4% of drives develop checksum mismatches during the time examined. Even though rare, identity discrepancies and parity inconsistencies do occur. Therefore, the protection offered by checksums and block identity information is critical to protect against data corruption.
- A significant number (8% on average) of corruptions are detected during RAID reconstruction, creating the possibility of data loss. In this case, protection against double disk failures is necessary to prevent data loss.
- Although the probability of developing a corruption is lower for enterprise-class drives, once they develop a corruption, many more are likely to follow. Therefore, replacing an enterprise-class drive on the first detection of a corruption might make sense.
- Strong spatial locality suggests that redundant data structures should be stored at a distance from each other.
- The high degree of spatial and temporal locality may suggest that corruptions occur at the exact same time, perhaps as part of the same disk request. Thus, important or redundant data structures should be written as part of different write requests spaced over time.
- Strong spatial and temporal locality (over long time periods) suggests that it is worth investigating how the locality can be leveraged for smarter, targeted scrubbing (e.g., trigger a scrub before its next scheduled time) or selective scrubbing of an area of the drive that's likely to be affected.
- Failure prediction algorithms in systems should take into account the correlation of corruption with other errors such as latent sector errors.

Conclusion

We have analyzed data corruption detected in 1.53 million disks used in production storage systems. During a 41-month period we observed more than 400,000 instances of checksum mismatches, 8% of which were discovered during RAID reconstruction, creating the possibility of real data loss.

We identified various characteristics of checksum mismatches, including: (i) the probability of developing the first checksum mismatch is almost an order of magnitude higher for SATA disks than for Fibre Channel disks; (ii) checksum mismatches are not independent occurrences—both within a disk and within different disks in the same storage system—and the number of mismatches per disk follows a heavy-tailed distribution; and (iii) checksum mismatches also show high spatial and temporal locality, encouraging system designers to develop schemes that spread redundant data with respect to both the on-disk location and the time written.

REFERENCES

[1] L.N. Bairavasundaram, G.R. Goodson, S.Pasupathy, and J. Schindler, "An Analysis of Latent Sector Errors in Disk Drives," in *Proceedings of the International Conference on Measurements and Modeling of Computer Systems (SIGMETRICS '07)*, San Diego, California, June 2007.

[2] L.N. Bairavasundaram, G.R. Goodson, B. Schroeder, A.C. Arpaci-Dusseau, and R. Arpaci-Dusseau, "An Analysis of Data Corruption in the Storage Stack," in *Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST '08)*, San Jose, California, Feb. 2008.

[3] W. Bartlett and L. Spainhower, "Commercial Fault Tolerance: A Tale of Two Systems," *IEEE Transactions on Dependable and Secure Computing*, 1(1):87–96 (Jan. 2004).

[4] W. Jiang, C. Hu, Y. Zhou, and A. Kanevsky, "Are Disks the Dominant Contributor for Storage Subsystem Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics," in *Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST '08)*, San Jose, California, Feb. 2008.

[5] E. Pinheiro, W.-D. Weber, and L.A. Barroso, "Failure Trends in a Large Disk Drive Population," in *Proceedings of the 5th USENIX Symposium on File and Storage Technologies (FAST '07)*, San Jose, California, Feb. 2007.

[6] V. Prabhakaran, L.N. Bairavasundaram, N. Agrawal, H.S. Gunawi, A.C. Arpaci-Dusseau, and R.H. Arpaci-Dusseau, "IRON File Systems," in *Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP '05)*, pp. 206–220, Brighton, United Kingdom, Oct. 2005.

[7] B. Schroeder and G.A. Gibson, "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" in *Proceedings of the 5th USENIX Symposium on File and Storage Technologies (FAST '07)*, San Jose, California, Feb. 2007.