WEIHANG JIANG, CHONGFENG HU,
YUANYUAN ZHOU, AND ARKADY KANEVSKY

don't blame disks for every storage subsystem failure

Weihang Jiang is a PhD candidate in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He is particularly interested in system mining that applies data mining techniques to improve system performance, dependability, and manageability.

wjiang3@uiuc.edu

Chongfeng Hu spent his college life at Peking University, China, and is currently a graduate student in the Department of Computer Science at the University of Illinois at Urbana-Champaign. His interests include operating systems, software reliability, storage systems reliability, and data mining.

chu7@cs.uiuc.edu

Yuanyuan Zhou is an associate professor at the University of Illinois, Urbana-Champaign. Her research spans the areas of operating system, storage systems, software reliability, and powesr management.

yyzhou@uiuc.edu

Arkady Kanevsky is a senior research engineer at NetApp Advanced Technology Group. Arkady has done extensive research on RDMA technology, storage resiliency, scalable storage systems, and parallel and distributed computing. He received a PhD in Computer Science from the University of Illinois in 1987. He was a faculty member at Dartmouth College and Texas A&M University prior to joining the industry world. Arkady has written over 60 publications and is a chair of DAT Collaborative and MPI-RT standards.

arkady@netapp.com

DISKS ARE THE KEY COMPONENTS OF storage systems. Researchers at CMU and NetApp had demonstrated a trend of increasing gap between the size of individual disks and disk access time [11], and hence the probability that a secondary failure happens during RAID reconstruction becomes too high for comfort. This led to RAID-6 [4] and RAID-DP [5]. Interestingly, even though disk reliability is critical to storage systems reliability, we found that disks themselves were not the component most likely to fail in storage systems.

In this work, we looked at other components in storage systems beyond disks to answer the following questions: Are disk failures the main source of storage system failures? Can enterprise disks help to build a more reliable storage system than SATA disks? Are disk failures independent? What about other storage component failures? Are techniques that went into RAID design, such as redundant interconnect between disks and storage controllers, really helpful in increasing the reliability of storage systems?

Reliability is a critically important issue for storage systems because storage failures not only can cause service downtime but can also lead to data loss. Building reliable storage systems becomes increasingly challenging as the complexity of modern storage systems grows to an unprecedented level. For example, the EMC Symmetrix DMX-4 can be configured with up to 2400 disks [6], the Google File System cluster is composed of 1000 storage nodes [7], and the NetApp FAS6000 series can support more than 1000 disks per node, with up to 24 nodes in a system [9].

To make things even worse, disks are not the only component in storage systems. To connect and access disks, modern storage systems also contain many other components, including shelf enclosures, cables and host adapters, and complex software protocol stacks. Failures in these components can lead to downtime and/or data loss of the storage system. Hence, in complex storage systems, component failures are very common and critical to storage system reliability.

Although we are interested in failures of a whole storage system, this study concentrates on the core part of it—*the storage subsystem*, which contains

disks and all components providing connectivity and usage of disks to the entire storage system.

We analyzed the NetApp AutoSupport logs collected from about 39,000 storage systems commercially deployed at various customer sites. The data set covers a period of 44 months and includes about 1,800,000 disks hosted in about 155,000 storage shelf enclosures. Our study reveals many interesting findings, providing useful guidelines for designing reliable storage systems. Some of our major findings include:

- Physical interconnect failures make up the largest part (27%–68%) of storage subsystem failures, and disk failures make up the second largest part (19%–56%). Choices of disk types, shelf enclosure models, and other components of storage subsystems contribute to the variability.
- Each individual storage subsystem failure type and storage subsystem failure as a whole exhibit strong self-correlations.
- Storage subsystems configured with redundant interconnects experience 30%–40% lower failure rates than those with a single interconnect.

Data on latent sector errors from the same AutoSupport Database was first analyzed by Bairavasundaram et al. [2], and data on data corruptions was further analyzed by Bairavasundaram et al. [3].

Background

In this section, we detail the typical architecture of storage systems we study in NetApp, the definitions and terminology used in this article, and the source of the data studied in this work.

STORAGE SYSTEM ARCHITECTURE

Figure 1 shows the typical architecture of a NetApp storage system node. A NetApp storage system can be composed of several storage system nodes.

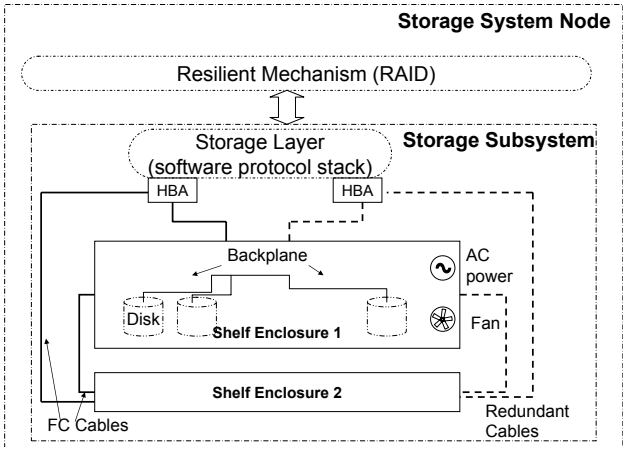


FIGURE 1: STORAGE SYSTEM NODE ARCHITECTURE

From the customer's perspective, a storage system is a virtual device that is attached to customers' systems and provides customers with the desired storage capacity with high reliability, good performance, and flexible management.

Looking from inside, we see that a storage system node is composed of storage subsystems, resiliency mechanisms, a storage head/controller, and other

higher-level system layers. The storage subsystem is the core part of a storage system node and provides connectivity and usage of disks to the entire storage system node. It contains various components, including disks, shelf enclosures, cables and host adapters, and complex software protocol stacks. Shelf enclosures provide a power supply, a cooling service, and a prewired backplane for the disks mounted in them. Cables initiated from host adapters connect one or multiple shelf enclosures to the network. Each shelf enclosure can be optionally connected to a secondary network for redundancy. In the Results section we will show the impact of this redundancy mechanism on failures of the storage subsystem.

Usually, on top of the storage subsystem, resiliency mechanisms, such as RAID, are used to tolerate failures in storage subsystems.

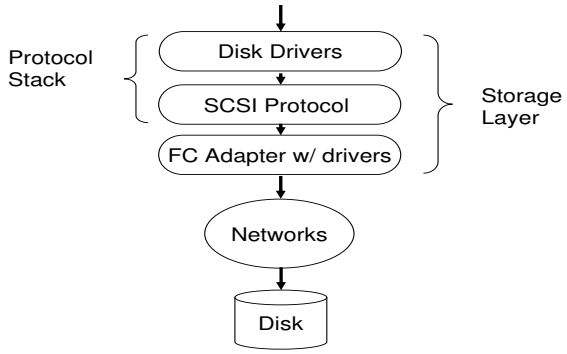


FIGURE 2: I/O REQUEST PATH IN STORAGE SUBSYSTEM

TERMINOLOGY

We use the followings terms in this article:

- *Disk family:* A particular disk product. The same product may be offered in different capacities. For example, "Seagate Cheetah 10k.7" is a disk family.
- *Disk model:* The combination of a disk family and a particular disk capacity. For example, "Seagate Cheetah 10k.7 300 GB" is a disk model. For disk family and disk model, we use the same naming convention as in Bairavasundaram et al. [2, 3]. (See also "Data Corruption in the Storage Stack," p. 6 in this issue.)
- *Failure types:* Refers to the four types of storage subsystem failures—disk failure, physical interconnect failure, protocol failure, and performance failure.
- *Shelf enclosure model:* A particular shelf enclosure product. All shelf enclosure models studied in this work can host at most 14 disks.
- *Storage subsystem failure:* Refers to failures that prevent the storage subsystem from providing storage service to the whole storage system node. However, not all storage subsystem failures are experienced by customers, since some of the failures can be handled by resiliency mechanisms on top of storage subsystems (e.g., RAID) and other mechanisms at higher layers.
- *Storage system class:* Refers to the capability and usage of storage systems. There are four storage system classes studied in this work: nearline systems (mainly used as secondary storage), low-end, mid-range, and high-end (mainly used as primary storage).
- Other terms in the article are used as defined by SNIA [12].

DEFINITION AND CLASSIFICATION OF STORAGE SUBSYSTEM FAILURES

Figure 2 shows the steps and components that are involved in fulfilling an I/O request in a storage subsystem. As shown in Figure 2, for the storage layer to fulfill an I/O request, the I/O request will first be processed and transformed by protocols and then delivered to disks through networks initiated by host adapters. *Storage subsystem failures* are the failures that break the I/O request path; they can be caused by hardware failures, software bugs, and protocol incompatibilities along the path.

To better understand storage subsystem failures, we categorize them into four types along the I/O request path:

- *Disk failure:* This type of failure is triggered by failure mechanisms of disks. Imperfect media, media scratches caused by loose particles, rotational vibration, and many other factors internal to a disk can lead to this type of failure. Sometimes the storage layer proactively fails disks based on statistics collected by on-disk health monitoring mechanisms (e.g., a disk has experienced too many sector errors [1]). These incidences are also counted as disk failures.
- *Physical interconnect failure:* This type of failure is triggered by errors in the networks connecting disks and storage heads. It can be caused by host adapter failures, broken cables, shelf enclosure power outages, shelf backplanes errors, and/or errors in shelf FC drivers. When physical interconnect failures happen, the affected disks appear to be missing from the system.
- *Protocol failure:* This type of failure is caused by incompatibility between protocols in disk drivers or shelf enclosures and storage heads and software bugs in the disk drivers. When this type of failure happens, disks are visible to the storage layer but I/O requests are not correctly responded to by disks.
- *Performance failure:* This type of failure happens when the storage layer detects that a disk cannot serve I/O requests in a timely manner while none of the previous three types of failures are detected. It is mainly caused by partial failures, such as unstable connectivity or when disks are heavily loaded with disk-level recovery (e.g., broken sector remapping).

The occurrences of these four types of failures are recorded in AutoSupport logs collected by NetApp.

Results

FREQUENCY OF STORAGE SUBSYSTEM FAILURES

As we categorize storage subsystem failures into four failure types based on their root causes, a natural question is therefore, "What is the relative frequency of each failure type?" To answer this question, we study the NetApp AutoSupport logs collected from 39,000 storage systems.

Figure 3 presents the breakdown of the average failure rate (AFR) for storage subsystems based on failure types, for all four system classes studied in this work.

Finding (1): Physical interconnect failures make up the largest part (27%–68%) of storage subsystem failures; disk failures make up the second largest part (20%–55%). Protocol failures and performance failures both make up noticeable fractions.

Implications: Disk failures are not always a dominant factor in storage subsystem failures, and a reliability study for storage subsystems cannot focus only on disk failures. Resilient mechanisms should target all failure types.

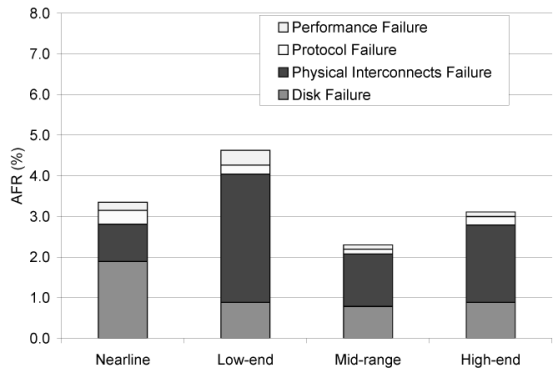


FIGURE 3: AFR FOR STORAGE SUBSYSTEMS IN FOUR SYSTEM CLASSES AND THE BREAKDOWN BASED ON FAILURE TYPES

As Figure 3 shows, across all system classes, disk failures do not always dominate storage subsystem failures. For example, in low-end storage systems, the AFR for storage subsystems is about 4.6%, whereas the AFR for disks is only 0.9%, about 20% of the overall AFR. However, physical interconnect failures account for a significant fraction of storage subsystem failures, ranging from 27% to 68%. The other two failure types, protocol failures and performance failures, contribute 5%–10% and 4%–8% of storage subsystem failures, respectively.

Finding (2): For disks, nearline storage systems show higher (1.9%) AFR than low-end storage systems (0.9%). But for the whole storage subsystem, nearline storage systems show lower (3.4%) AFR than low-end primary storage systems (4.6%).

Implications: Disk failure rate is not indicative of the storage subsystem failure rate.

Figure 3 also shows that nearline systems, which mostly use SATA disks, experience about 1.9% AFR for disks, whereas for low-end, mid-range, and high-end systems, which mostly use FC disks, the AFR for disks is under 0.9%. This observation is consistent with the common belief that enterprise disks (FC) are more reliable than nearline disks (SATA).

However, the AFR for storage subsystems does not follow the same trend. Storage subsystem AFR of nearline systems is about 3.4%, lower than that of low-end systems (4.6%). This indicates that other factors, such as shelf enclosure model and network configurations, strongly affect storage subsystem reliability. The impacts of these factors are examined next.

IMPACT OF SYSTEM PARAMETERS ON STORAGE SUBSYSTEM FAILURES

As we have seen, storage subsystems of different system classes show different AFRs. Although these storage subsystems are architecturally similar, the characteristics of their components, such as shelves, and their redundancy mechanisms, such as multipathing, differ. We now explore the impact of these factors on storage subsystem failures.

SHELF ENCLOSURE MODEL

Shelf enclosures contain power supplies, cooling devices, and prewired backplanes that carry power and I/O bus signals to the disks mounted in

them. Different shelf enclosure models are different in design and have different mechanisms for providing these services; therefore, it is interesting to see how shelf enclosure model affects storage subsystem failures.

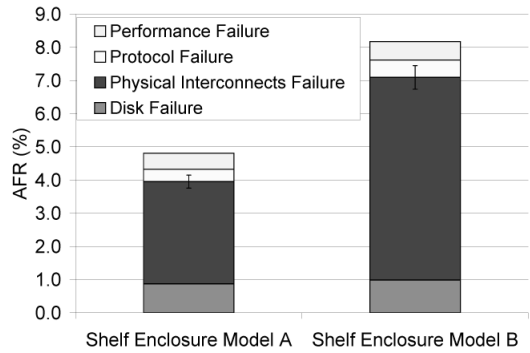


FIGURE 4: AFR FOR STORAGE SUBSYSTEMS BY SHELF ENCLOSURE MODELS USING THE SAME DISK MODEL. THE ERROR BARS SHOW 99.9% CONFIDENCE INTERVALS FOR PHYSICAL INTERCONNECT FAILURES.

Finding (3): The shelf enclosure model has a strong impact on storage subsystem failures.

Implications: To build a reliable storage subsystem, hardware components other than disks (e.g., shelf enclosure) should also be selected carefully.

Figure 4 shows AFR for storage subsystems when configured with different shelf enclosure models but the same disk models. As expected, shelf enclosure model primarily impacts physical interconnect failures, with little impact on other failure types.

To confirm this observation, we tested the statistical significance using a T-test [10]. As Figure 4 shows, the physical interconnect failures with different shelf enclosure models are quite different (3.08±0.20% versus 6.11 ±0.35%). A T-test shows that this is significant at the 99.9% confidence interval, indicating that the hypothesis that physical interconnect failures are impacted by shelf enclosure models is very strongly supported by the data.

NETWORK REDUNDANCY MECHANISM

As we have seen, physical interconnect failures contribute to a significant fraction (27%–68%) of storage subsystem failures. Since physical interconnect failures are mainly caused by network connectivity issues in storage subsystems, it is important to understand the impact of network redundancy mechanisms on storage subsystem failures.

For the mid-range and high-end systems studied in this work, FC drivers support a network redundancy mechanism, commonly called *active/passive multipathing*. This network redundancy mechanism connects shelves to two independent FC networks, redirecting I/O requests through the redundant FC network when one FC network experiences network component failures (e.g., broken cables).

To study the effect of this network redundancy mechanism, we look at the data collected from mid-range and high-end storage systems, and we group them based on whether the network redundancy mechanism is turned on. Owing to the space limitation, we show results only for the mid-range storage systems here. As we observed from our data set, about 1/3 of storage subsystems are utilizing the network redundancy mechanism, whereas the

other 2/3 are not. We call these two groups of storage subsystems *dual path* systems and *single path* systems, respectively.

Finding (4): Storage subsystems configured with network redundancy mechanisms experience much lower (30%–40% lower) AFR than other systems. AFR for physical interconnects is reduced by 50%–60%.

Implications: Network redundancy mechanisms such as multipathing can greatly improve the reliability of storage subsystems.

Figure 5 shows the AFR for storage subsystems in mid-range systems. As expected, secondary path reduces physical interconnect failures by 50%–60% (1.82±0.04 % versus 0.91±0.09 %), with little impact on other failure types. Since physical interconnect failure is just a subset of all storage subsystem failures, AFR for storage subsystems is reduced by 30%–40%. This indicates that multipathing is an exceptionally good redundancy mechanism that delivers reduction of failure rates as promised. As we applied a T-test on these results, we found that the observation is significant at the 99.9% confidence interval, indicating that the data strongly supports the hypothesis that physical interconnect failures are reduced by multipathing configuration.

However, the observation also tells us that there is still further potential in network redundancy mechanism designs. For example, given that the probability for one network to fail is about 2%, the idealized probability for two networks to both fail should be a few orders of magnitude lower (about 0.04%). But the AFR we observe is far from the ideal number.

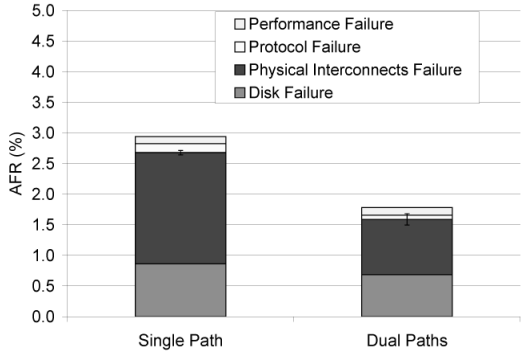


FIGURE 5: AFR FOR STORAGE SUBSYSTEMS BROKEN DOWN BY THE NUMBER OF PATHS. THE ERROR BARS SHOW 99.9% CONFIDENCE INTERVALS FOR PHYSICAL INTERCONNECT FAILURES.

CORRELATIONS BETWEEN FAILURES

In this subsection, we will study the statistical property of storage subsystem failures both from a shelf perspective and from a RAID group perspective.

Our analysis of the correlation between failures is composed of two steps:

1. Derive the theoretical failure probability model based on the assumption that failures are independent.
2. Evaluate the assumption by comparing the theoretical probability against empirical results.

Next, we describe the statistical method we use for deriving the theoretical failure probability model.

STATISTICAL METHOD

We denote the probability for a shelf enclosure (including all mounted disks) to experience one failure during time T as $P(1)$ and denote the probability for it to experience two failures during T as $P(2)$. The relationship between $P(1)$ and $P(2)$ is as follows:

For a complete proof, refer to our conference paper [8].

$$P(2) = \frac{1}{2} P(1)^2$$

Next, we will compare this theoretically derived model against the empirical results collected from NetApp AutoSupport logs.

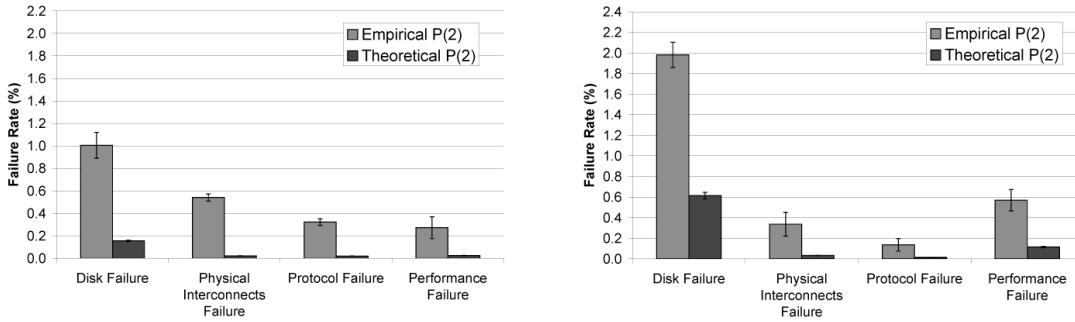
CORRELATION RESULTS

To evaluate the theoretical relation between $P(1)$ and $P(2)$ shown in equation 1, we first calculate *empirical* $P(1)$ and *empirical* $P(2)$ from NetApp AutoSupport logs. Empirical $P(1)$ is the percentage of shelves (RAID groups) that have experienced exactly one failure during time T (where T is set to one year), and empirical $P(2)$ is the percentage that have experienced exactly two failures during time T. Only storage systems that have been in the field for one year or more are considered.

Finding (5): For each failure type, storage subsystem failures are not independent. After one failure, the probability of additional failures (of the same type) is higher.

Implications: The probability of storage subsystem failures depends on factors shared by all disks in the same shelf enclosures (or RAID groups).

Figure 6a shows the comparison between *empirical* $P(2)$ and *theoretical* $P(2)$, which is calculated based on empirical $P(1)$. As we can see in the figure, empirical $P(2)$ is higher than theoretical $P(2)$. More specifically, for disk failure, the observed empirical $P(2)$ is higher than theoretical $P(2)$ by a factor of 6. For other types of storage subsystem failures, the empirical probability is higher than the theoretical correspondences by a factor of 10–25. Furthermore, T-tests confirm that the theoretical $P(2)$ and the empirical $P(2)$ are statistically different with 99.5% confidence intervals.



(A) SHELF ENCLOSURE FAILURES (B) RAID GROUP FAILURES

FIGURE 6: COMPARISON BETWEEN THEORETICAL MODEL [WITH P(2) CALCULATED FROM EQUATION 1] AND EMPIRICAL RESULTS; THE ERROR BARS SHOW 99.5%+ CONFIDENCE INTERVALS.

These statistics provide a strong indication that when a shelf experiences a storage subsystem failure, the probability for it to have another storage subsystem failure increases. In other words, storage subsystem failures from the same shelves are not independent.

Figure 6b shows an even stronger trend for failures from the same RAID groups. Therefore, the same conclusion can be made for storage subsystem failures from the same RAID groups.

Conclusion

In this article we have presented a study of NetApp storage subsystem failures, examining the contribution of different failure types, the effect of some factors on failures, and the statistical properties of failures. Our study is based on AutoSupport logs collected from 39,000 NetApp storage systems, which contain about 1,800,000 disks mounted in about 155,000 shelf enclosures. The studied data covers a period of 44 months. The findings of our study provide guidelines for designing more reliable storage systems and developing better resiliency mechanisms.

Although disks are the primary components of storage subsystems, and disk failures contribute 19%–56% of storage subsystem failures, other components such as physical interconnect and protocol stacks also account for significant percentages (27%–68% and 5%–10%, respectively) of storage subsystem failures. The results clearly show that the other storage subsystem components cannot be ignored when designing a reliable storage system.

One way to improve storage system reliability is to select more reliable components. As the data suggests, storage system reliability is highly dependent on shelf enclosure model. Another way to improve reliability is to employ redundancy mechanisms to tolerate component failures. One such mechanism studied in this work is multipathing, which can reduce AFR for storage systems by 30%–40% when the number of paths is increased from one to two.

We also found that the storage subsystem failure and individual storage subsystem failure types exhibit strong self-correlations. This finding motivates revisiting resiliency mechanisms, such as RAID, that assume independent failures.

A preliminary version of this article was published at FAST '08 [8]. Because of limited space, neither this article nor our FAST paper includes all our results. Readers who are interested in a complete set of results should refer to our NetApp technical report [13].

ACKNOWLEDGMENTS

We wish to thank Rajesh Sundaram and Sandeep Shah for providing us with insights on storage failures. We are grateful to Frederick Ng, George Kong, Larry Lancaster, and Aziz Htite for offering help on understanding and gathering NetApp AutoSupport logs. We appreciate useful comments from members of the Advanced Development Group, including David Ford, Jiri Schindler, Dan Ellard, Keith Smith, James Lentini, Steve Byan, Sai Susarla, and Shankar Pasupathy. Also, we would like to thank Lakshmi Bairavasundaram for his useful comments. Finally, we appreciate our shepherd, Andrea Arpaci-Dusseau, for her invaluable feedback and precious time, and the anonymous reviewers for their insightful comments for our conference paper [8]. This research has been funded by NetApp under the "Intelligent Log Mining" project at CS UIUC. Work of the first two authors was conducted in part as summer interns at NetApp.

REFERENCES

[1] Bruce Allen, "Monitoring Hard Disks with SMART," *Linux Journal*, 2004(117):9 (2004).

[2] Lakshmi N. Bairavasundaram, Garth R. Goodson, Shankar Pasupathy, and Jiri Schindler, "An Analysis of Latent Sector Errors in Disk Drives," *SIGMETRICS Perform. Eval. Rev.* 35(1):289–300 (2007).

[3] Lakshmi N. Bairavasundaram, Garth R. Goodson, Bianca Schroeder, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau, "An Analysis of Data Corruption in the Storage Stack," in *FAST '08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*, San Jose, CA, February 2008.

[4] Mario Blaum, Jim Brady, Jehoshua Bruck, and Jai Menon, "Evenodd: An Efficient Scheme for Tolerating Double Disk Failures in RAID Architectures," *IEEE Transactions on Computing*, 44:192–202 (1995).

[5] Peter Corbett, Bob English, Atul Goel, Tomislav Grcanac, Steven Kleiman, James Leong, and Sunitha Sankar, "Row-diagonal Parity for Double Disk Failure Correction," in *FAST '04: Proceedings of the 3rd USENIX Conference on File and Storage Technologies*, pp. 1–14 (2004).

[6] EMC Symmetrix DMX-4 Specification Sheet (July 2007): http://www.emc.com/collateral/hardware/specification-sheet/c1166-dmx4-ss.pdf.

[7] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System," in *SOSP '03: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, New York, pp. 29–43 (2003).

[8] Weihang Jiang, Chongfeng Hu, Yuanyuan Zhou, and Arkady Kanevsky, "Are Disks the Dominant Contributor for Storage Failures?—A Comprehensive Study of Storage Subsystem Failure Characteristics," in *FAST '08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*, San Jose, CA, February 2008.

[9] FAS6000 Series Technical Specifications: http://www.netapp.com/products/filer/fas6000_tech_specs.html.

[10] A.C. Rosander, *Elementary Principles of Statistics* (Princeton, NJ: Van Nostrand, 1951).

[11] Bianca Schroeder and Garth A. Gibson, "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" in *FAST '07: Proceedings of the 5th USENIX Conference on File and Storage Technologies*, Berkeley, CA, USA, 2007.

[12] Storage Networking Industry Association Dictionary: http://www.snia.org/education/dictionary/.

[13] Weihang Jiang, Chongfeng Hu, Yuanyuan Zhou, and Arkady Kanevsky, "Don't Blame Disks for Every Storage Subsystem Failure—A Comprehensive Study of Storage Subsystem Failure Characteristics," NetApp Research Paper, April 2008: http://media.netapp.com/documents/dont-blame-disks-for-every-storage-subsystem-failure.pdf.