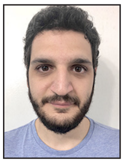
A Study of SSD Reliability in Large Scale Enterprise Storage Deployments

STATHIS MANEAS, KAVEH MAHDAVIANI, TIM EMAMI, AND BIANCA SCHROEDER



Stathis Maneas is a PhD candidate in the Department of Computer Science at the University of Toronto. Prior to that, he obtained his MSc and BSc degrees at the National and Kapodistrian University of Athens. His main research interests include the design and implementation of computer systems, especially storage and file systems, and distributed systems. His current research focuses on the reliability aspect of systems. smaneas@cs.toronto.edu

Kaveh Mahdaviani is a Postdoctoral Fellow in the Department of Computer Science at the University of Toronto. Before that, he completed his PhD, MSc, and BSc programs at the University of Toronto, University of Alberta, and Isfahan University of Technology, respectively. His research focuses on distributed systems, information theory, and coding theory. mahdaviani@cs.toronto.edu

Tim Emami is a Senior Technical Director and the Storage Media subject matter expert at NetApp. During his over 13-year tenure, he helped pioneer ONTAP's transition from disk to flash. Prior to NetApp, he worked in R&D roles at Maxtor, Quantum, WDC, and StorCard. He studied control systems at California Polytechnic State Institute, San Luis Obispo.
Tim.Emami@netapp.com

We present the first large-scale field study of NAND-based SSDs deployed in enterprise storage systems. Our study is based on field data, collected over 2.5 years, for a sample of almost 1.4 million drives from the total SSD population of a major enterprise storage vendor (NetApp). The data allows us to study a large number of factors that were not included in prior work, such as the effect of firmware versions, the reliability of TLC NAND, and correlations between drives within a RAID group. Our analysis provides insight into flash reliability, along with a number of practical implications.

System reliability is arguably one of the most important aspects of a storage system, and, as such, a large body of work exists on the topic of storage device reliability. Much of the older work is focused on hard disk drives (HDDs) [1, 5–7], but as more data is being stored on solid state drives (SSDs), the focus has recently shifted to the reliability of SSDs. In addition to a large amount of work on SSDs in lab conditions under controlled experiments, the first field studies reporting on SSD reliability in deployed production systems have recently appeared [3, 4, 8, 10]. These studies are based on data collected at datacenters at Facebook, Microsoft, Google, and Alibaba, where drives are deployed as part of large distributed storage systems. However, we observe that there still are a number of critical gaps in the existing literature that this work is striving to bridge:

- There were no studies that focus on *enterprise storage systems*. The drives, workloads, and reliability mechanisms in these systems can differ significantly from those in cloud datacenters. For example, the drives used in enterprise storage systems include high-end drives, and reliability is ensured through (single, double, or triple parity) RAID, instead of replication or distributed storage codes.
- We also observe that existing studies do not cover some of the most important characteristics of failures that are required for building realistic failure models, in order to compute metrics such as the mean time to data loss. This includes, for example, a breakdown of the reasons for drive replacements, including the scope of the underlying problem and the corresponding repair action (RAID reconstruction versus draining the drive), and most importantly, an understanding of the correlations between drive replacements in the same RAID group.

In this article, we present some selected findings of our work. For detailed results, please see our USENIX FAST '20 paper [2].

Reasons for Replacements

SSD replacement can be triggered for various reasons, and different subsystems in the storage hierarchy can detect issues that trigger the replacement of drives. For example, issues might be reported by the drive itself, the storage layer, or the file system. Table 1 describes the different *reason types* that can trigger a drive replacement, along with their frequency, the recovery action taken by the system, and the scope of the problem. We group the different reason types behind SSD replacements into four *categories*, labeled A to D, based on their severity.

Bianca Schroeder is an Associate Professor and Canada Research Chair in the Computer Science Department at the University of Toronto. She completed her PhD and a post-doc at Carnegie Mellon University under the guidance of Mor Harchol-Balter and Garth Gibson, respectively. She is an Alfred P. Sloan Research Fellow and the recipient of the Outstanding Young Canadian Computer Science Prize of the Canadian Association for Computer Science, an Ontario Early Researcher Award, five best paper awards and a Test of Time award. bianca@cs.toronto.edu

Category	Type	Pct.	Annual Repl. Rate (%)	Recovery Action	Scope
A	SCSI Error	32.78	0.055	RAID Reconstruction	Full
	Unresponsive Drive	0.60	0.001		
B	Lost Writes	13.54	0.023	RAID Reconstruction	Partial
C	Aborted Commands	13.56	0.023	RAID Reconstruction	Partial
	Disk Ownership I/O Errors	3.27	0.005		
	Command Timeouts	1.81	0.003		
D	Predictive Failures	12.78	0.021	Disk Copy	Zero
	Threshold Exceeded	12.73	0.020		
	Recommended Failures	8.93	0.015		

Table 1: Description of reason types that can trigger a drive replacement. Disk copy operations are performed only where possible (e.g., a spare disk must be available).

The most benign category is category D, which relates to replacements that were triggered by logic either inside the drive or at higher levels in the system, which predicts future drive failure, based on, for example, previous errors, timeouts, and a drive's SMART statistics [9]. The most severe category is category A, which comprises those situations where drives become completely unresponsive, or where the SCSI layer detects a hardware error (reported by the drive) severe enough to trigger immediate replacement and RAID reconstruction of its data.

Category B refers to drive replacements that are taking place when the system suspects the drive to have *lost a write*, e.g., because it did not perform the write at all, wrote it to a wrong location, or otherwise corrupted the write. The root cause could be a firmware bug in the drive, although other layers in the storage stack could be responsible as well. As there are many potential causes, a heuristic is used to decide whether to trigger a replacement or not.

Finally, in category C most of the reasons for replacements are related to commands that were aborted or timed out. For instance, a command can be aborted when the host has sent some write commands to the device, but the actual data never reached the device due to connection issues. Ownership errors are related to the subsystem that keeps track of which node owns a drive; if an error occurs during the communication with this subsystem, the drive is marked as *failed*.

When examining the frequency of each individual type, we observe that SCSI errors are the most common type, responsible for ~33% of all replacements and, unfortunately, also one of the most severe reason types. On the other hand, drives rarely become completely unresponsive (0.60% of all replacements). Fortunately, one-third of all drive replacements are merely preventative (category D), using predictions of future drive failures, and are hence unlikely to have severe impact on system reliability. Finally, the two remaining categories (B and C) are roughly equally common, and both have the potential of partial data loss if RAID reconstruction of the affected data should turn out unsuccessful.

Finding 1: *One-third of replacements are associated with one of the most severe reason types (i.e., SCSI errors); on the other hand, one-third of drive replacements are merely preventative, based on predictions.*

Factors Impacting Replacement Rates

We evaluate how different factors impact the replacement rates of the SSDs in our data set. We make use of the annual replacement rate (ARR) metric, which is commonly used to report failure frequency [4, 5, 7] and is defined as follows:

$$ARR = \frac{\text{Total failed devices}}{\text{Total device years}} \; in\,\%$$

Usage and Age

It is well known that usage, and the wear-out of flash cells that comes with it, affects the reliability of flash-based SSDs; drives are guaranteed to remain functional for only a certain number of program/erase (PE) cycles. In our data set, SLC drives have a PE cycles limit of 100K, whereas the limit of most cMLC, eMLC, and 3D-TLC drives is equal to 10K cycles, with the exception of a few eMLC drive families with a 30K PE cycles limit.

Each drive reports the number of PE cycles it has experienced as a percentage of its PE cycle limit (denoted as *rated life used*), allowing us to study how usage affects replacement rates. Unfortunately, the rated life used is only reported as a truncated integer, and a significant fraction of drives report a zero for this metric, indicating less than 1% of their rated life has been used. Therefore, our first step is a comparison of the ARR of drives that report less than 1% versus more than 1% of their rated life used. The results for eMLC and 3D-TLC drives are shown in Figure 1, which includes both overall replacement rates ("All") and rates broken down by their replacement category (A to D). Throughout the article, error bars refer to 95th percentile confidence intervals; we also exclude two outlier models, i.e., I-C and II-C, with unusually high replacement rates to not obscure trends except for graphs involving individual drive families.

Figure 1 provides evidence for effects of infant mortality. For example, eMLC drives, the drives with less than 1% rated life used, are more likely (1.25x) to be replaced than those with more than 1% of rated life used. When further breaking results down by reason category, we find that drives with less usage consistently experience higher replacement rates for all categories.

Making conclusive claims for the 3D-TLC drives is harder due to limited data on drives above 1% of rated life used, resulting in wide confidence intervals. However, where we have enough data, observations are similar to those for eMLC drives, e.g., we see a significant drop in lost writes for drives above 1% of rated life used.

We also look at replacement rates as a function of a drive's age measured by its total months in the field. Figure 2 shows the conditional probability of a drive being replaced in a given month of its life, i.e., the probability that the drive will fail in month x given that it has survived up to the end of month x-1.

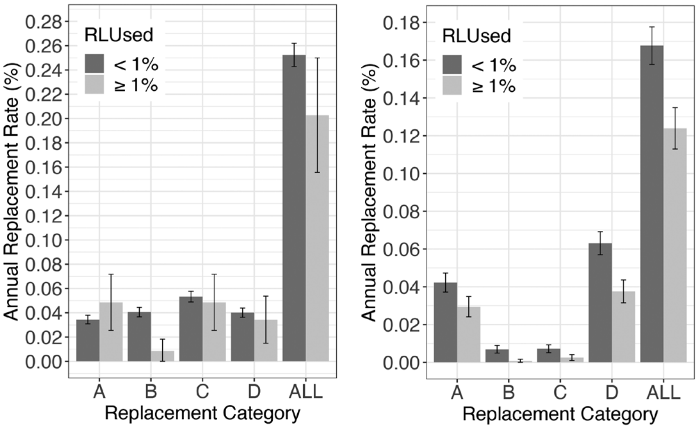


Figure 1: Annual replacement rate per flash type based on the drives' "rated-life-used" percentage

We observe an unexpectedly long period of infant mortality with a shape that differs from the common "bathtub" model, often used in reliability theory. The bathtub model assumes a short initial period of high failure rates, which then quickly drops. Instead, we observe for both 3D-TLC and eMLC drives, a long period (12–15 months) of increasing failure rates, followed by a lengthy period (another 6–12 months) of slowly decreasing failure rates, before rates finally stabilize. This brings up the question of what could be done to reduce these effects. One might consider, for example, an extended, more intense burn-in period before deployment, where drives are subjected to longer periods of high read and write loads. Given the low consumption of PE cycles that drives see in the field (99% of drives do not even use up 1% of their PE cycle limit), there seems to be room to sacrifice some PE cycles in the burn-in process.

Finally, it might be surprising that we do not observe an increase in ARR for drives towards the end of their life, but the majority of drives, even those deployed for several years, do not experience a large number of PE cycles.

Finding 2: *We observe a very drawn-out period of infant mortality, which can last more than a year, and see failure rates 2–3x larger than later in life.*

Flash and Drive Type

The drive models in our study differ in the type of flash they are based on, i.e., in how many bits are encoded in a single flash cell. For instance, Single-Level Cell (SLC) drives encode only one bit per cell, while Multi-Level Cell (MLC) drives encode two bits in one cell for higher data density and thus a lower total cost, but potentially higher propensity to errors. The most recent generation of flash in our data set is based on Triple-Level Cell (3D-TLC) flash with three bits per cell.

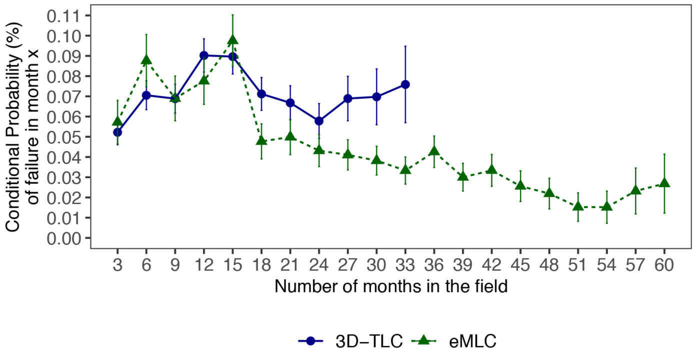


Figure 2: Conditional probability of failure based on a drive's age (number of months in the field) for all drive families

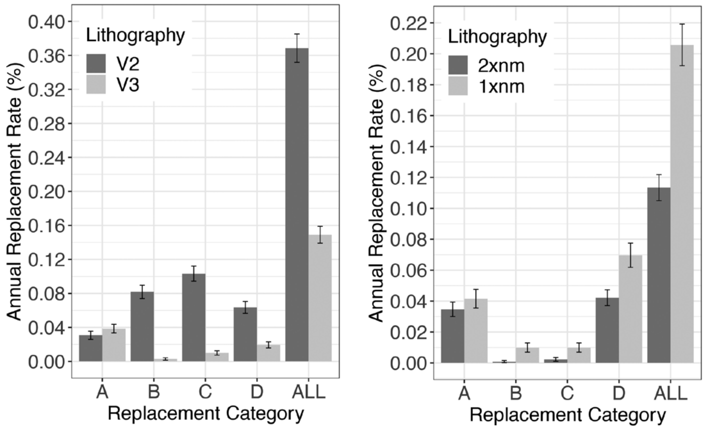


Figure 3: Annual replacement rate per flash type and lithography broken down by replacement category. The 1x nm and 2x nm notations denote any lithography in the range of 10–19 nm and 20–29 nm, respectively.

We turn to Figures 1 and 3 to compare 3D-TLC and eMLC drives, which take usage and lithography into account. Figure 1 indicates that ARRs for 3D-TLC drives are around 1.5x higher than for eMLC drives, when comparing similar levels of usage. Figure 3 paints a more complex picture. While V2 3D-TLC drives have a significantly higher replacement rate than any of the other groups, the V3 3D-TLC drives are actually comparable to 2x nm eMLC drives, and in fact have lower ARR than the 1x nm eMLC drives. So lithography might play a larger role than flash type alone; we take a closer look at lithography below.

When we compare the results for the MLC drives in our data set against previous work, we observe that Narayanan et al. [4] report replacement rates between 0.5–1% for their *consumer* class MLC drives, with the exception of a single enterprise class model, whose replacement rate is equal to 0.1%; however, the authors in [4] consider only fail-stop failures. In our study, we consider different types of failures, and, thus, the reported replacement rates would have been even smaller had we considered only fail-stop failures.

Finding 3: *Overall, the highest replacement rates in our study are associated with 3D-TLC drives. However, no single flash type has noticeably higher replacement rates than the other flash types in this work, indicating that other factors (e.g., lithography) can have a bigger impact on reliability.*

Lithography

Lithography has been shown to be highly correlated with a drive's raw bit error rate (RBER); models with smaller lithography report higher RBERs according to a study based on datacenter drives [8], but not necessarily higher replacement rates. We explore what these trends look like for the drives in enterprise storage systems. To separate the effect of lithography from flash type, we perform the analysis separately for each flash type.

The bar graph in Figure 3 (right) shows the ARR for eMLC drives separated into 2x nm and 1x nm lithographies broken down by failure category, also including one bar for replacements of all categories. The 1x nm and 2x nm notations denote any lithography in the range of 10–19 nm and 20–29 nm, respectively. We observe that the higher density 1x nm drives experience almost twice the replacement rate of 2x nm drives. Also, replacement rates for each of the individual reason types are higher for 1x nm drives than for 2x nm, with the only exception of reason category A, which corresponds to unresponsive drives.

In contrast to eMLC drives, the 3D-TLC drives see higher replacement rates for the lower density V2 drives, which internally have fewer layers than V3. When breaking replacement rates down by failure reason, we observe that consistently with the results for TLC drives, the only reason code that is not affected by lithography is category A, which corresponds to unresponsive drives.

Finding 4: *In contrast to previous work, higher density drives do not always see higher replacement rates. In fact, we observe that, although higher density eMLC drives have higher replacement rates, this trend is reversed for 3D-TLC.*

Firmware Version

Given that bugs in a drive's firmware can lead to drive errors or, in the worst case, to an unresponsive drive, we are interested to see whether different firmware versions are associated with a different ARR. Each drive model in our study experiences different firmware versions over time. We name the first firmware version of a model FV1, the next one FV2, and so on. An individual drive's firmware might be updated to a new version, but we observe that the majority of drives (70%) appear under the same firmware version in all data snapshots.

Figure 4 shows the ARR associated with different firmware versions for each drive family. Considering that firmware varies across drive families and manufacturers, it only makes sense to compare the ARR of different firmware versions within the same drive family. To avoid other confounding factors, in particular

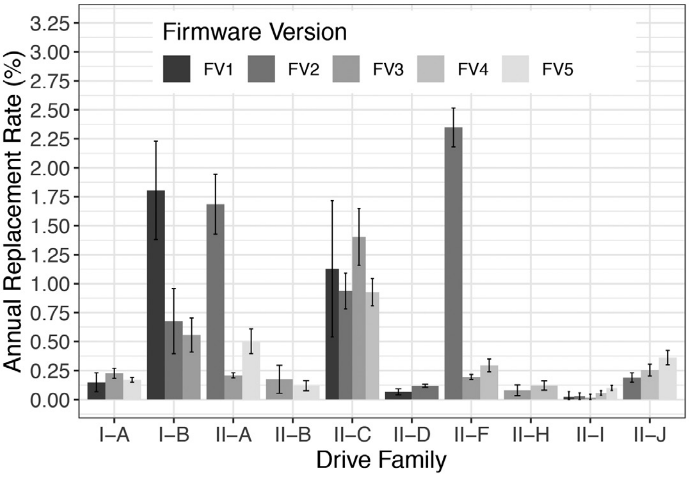


Figure 4: Effect of the firmware version on replacement rates broken down by drive family

age and usage, the graph in Figure 4 only includes drives with rated life used of less than 1% (the majority of drives).

We find that drives' firmware version can have a tremendous impact on reliability. In particular, the earliest versions can have an order of magnitude higher ARR than later versions. This effect is most notable for families I-B (more than 2x decrease in ARR from FV1 to FV2), II-A (8x decrease from FV2 to FV3), and II-F (more than 10x decrease from FV2 to FV3). Finally, we note that this effect persists even if we only include drives whose firmware has never changed in our data snapshots.

Finding 5: *Earlier firmware versions can be correlated with significantly higher replacement rates, emphasizing the importance of firmware updates.*

Correlations between Drive Failures

A key question when deriving reliability estimates—e.g., for different RAID configurations—is how failures of drives within the same RAID group are correlated.

For a detailed understanding of correlations, we consider all RAID groups that have experienced more than one drive replacement over the course of our observation period, and plot in Figure 5 the time between consecutive drive replacements within the same RAID group. We observe that very commonly, the second drive replacement follows the preceding one within a short time interval. For example, 46% of consecutive replacements take place at most one day after the previous replacement, while 52% of all consecutive replacements take place within a week of the previous replacement.

Another important question in RAID reliability modeling is how the chance of multiple failures grows as the number of drives in the RAID group increases. Figure 6 (left) presents, for the most common RAID group sizes, the percentage of RAID groups of

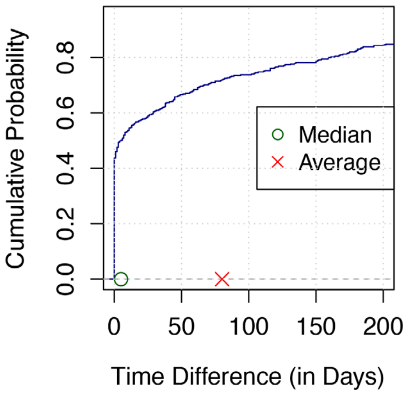


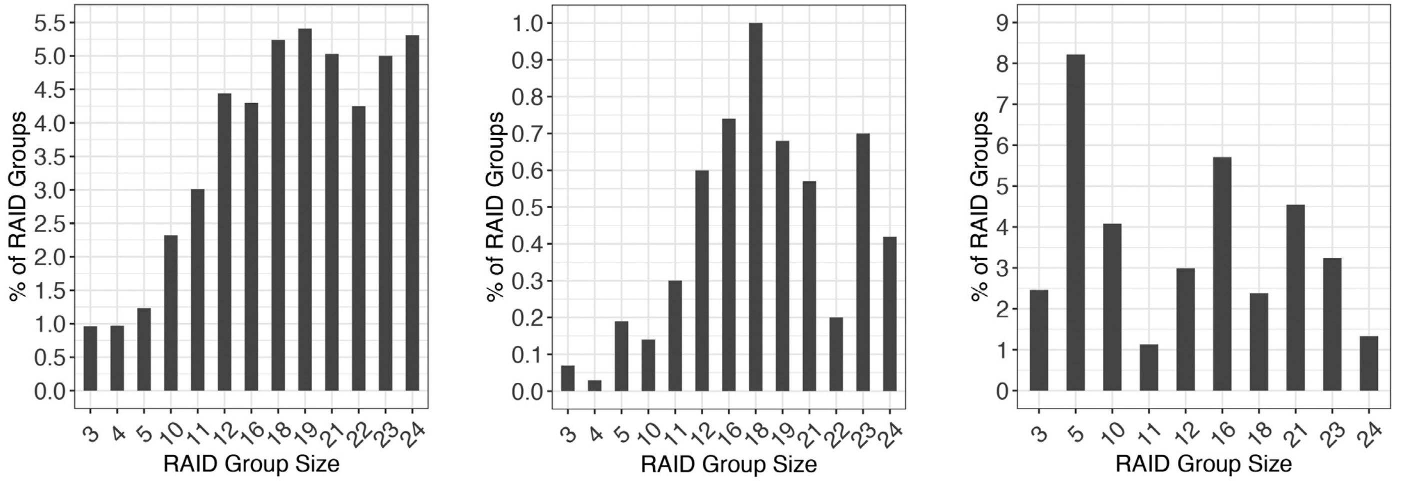
Figure 5: Time difference between successive replacements within RAID groups

that size that experienced at least one drive replacement. As one would expect, larger RAID groups have a higher chance of experiencing a drive replacement; yet, the effect of a RAID group's size on the replacement rates saturates for RAID groups comprising more than 18 drives.

Concerning multiple failures within the same RAID group, we make an interesting observation in Figure 6 (middle). When we look at the percentage of RAID groups that have experienced at least two drive replacements (potential double failure), this does not seem to be clearly correlated with RAID group size. In other words, the largest RAID group sizes do not necessarily seem to have a higher rate of double (or multiple) failures compared to smaller RAID groups.

This observation is confirmed when we look at the conditional probability that a RAID group will experience more replacements, *given that it has already experienced another replacement*, in Figure 6 (right). More precisely, for each RAID group size, we consider those RAID groups that had at least one drive replacement and compute what percentage of them had at least one more replacement within a week. Interestingly, we observe there is no clear trend that larger RAID group sizes have a larger chance of one drive replacement being followed by more replacements. Note that, as already mentioned, the chance of experiencing a drive failure grows with the size of the RAID group (Figure 6 left); however, the chance of *correlated failures* does not show a direct relationship with the group's size.

Finding 6: *While large RAID groups have a larger number of drive replacements, we find no evidence that the rate of multiple failures per group (which is what can create potential for data loss) is correlated with RAID group size. The reason seems to be that the likelihood of a follow-up failure after a first failure is not correlated with RAID group size.*



(a) RAID groups that experience at least one replacement

(b) RAID groups that experience multiple replacements

(c) RAID groups with replacement that experience at least one follow-up replacement within the next week

Figure 6: Statistics on replacements within RAID groups

Conclusion

Previous work has focused on the reliability characteristics of SSDs deployed in distributed datacenter storage systems. Our work presents the first large-scale field study of NAND-based SSDs in *enterprise storage systems* [2]. Below, we summarize some of the most important findings and implications of our work:

- Our observations emphasize the importance of firmware updates, as earlier firmware versions can be correlated with significantly higher failure rates. Yet 70% of drives remain at the same firmware version throughout the length of our study. Consequently, we encourage enterprise storage vendors to make firmware upgrades as easy and painless as possible so that customers apply the upgrades without worries about stability issues.
- We observe significant correlations between failures within RAID groups. This emphasizes the importance of incorporating correlated failures into any analytical models in order to arrive at realistic estimates of the probability of data loss. It also makes a case for more than just single-parity RAID.
- The failure rates in our study do not resemble the "bathtub" shape assumed by classical reliability models. Instead, we observe no signs of increased failure rates at end of life and also a very drawn-out period of infant mortality, which can last for more than a year and see failure rates 2–3x larger than later in life.
- There has been a fear that the limited PE cycles of NAND SSDs can create a threat to data reliability in the later part of a RAID system's life due to correlated wear-out failures, as the drives in a RAID group age at the same rate. Instead, we observe that correlated failures due to infant mortality are likely to be a bigger threat.

Acknowledgments

We would like to acknowledge several people at NetApp for their contributions to this work: Rodney Dekoning, Saumyabrata Bandyopadhyay, and Anita Jindal for their early support and encouragement, and Aziz Htite, who helped cross-validate our data and assumptions along the way. We would also like to thank the internal reviewers within the ATG, ONTAP WAFL, and RAID groups for their careful and detailed feedback. A very special thank you to Biren Fondekar's Active IQ team in Bangalore: Asha Gangolli, Kavitha Degavinti, and finally Vinay N., who spent countless late nights on the phone with us as we cleaned and curated this work's foundational data sets. We also thank our USENIX FAST reviewers and our shepherd, Devesh Tiwari, for their detailed feedback and valuable suggestions.

References

[1] D. N. Bairavasundaram, G. R. Goodson, S. Pasupathy, and J. Schindler, "An Analysis of Latent Sector Errors in Disk Drives," in *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '07)*, pp. 289–300.

[2] S. Maneas, K. Mahdaviani, T. Emami, and B. Schroeder, "A Study of SSD Reliability in Large Scale Enterprise Storage Deployments," in *Proceedings of the 18th USENIX Conference on File and Storage Technologies (FAST '20)*, pp. 137–149: https://www.usenix.org/system/files/fast20-maneas.pdf.

[3] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "A Large-Scale Study of Flash Memory Failures in the Field," in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '15)*, pp. 177–190.

[4] I. Narayanan, D. Wang, M. Jeon, B. Sharma, L. Caulfield, A. Sivasubramaniam, B. Cutler, J. Liu, B. Khessib, and K. Vaid, "SSD Failures in Datacenters: What? When? And Why?" in *Proceedings of the 9th ACM International on Systems and Storage Conference (SYSTOR '16)*, pp. 7:1–7:11.

[5] E. Pinheiro, W.-D. Weber, and L. André Barroso, "Failure Trends in a Large Disk Drive Population," in *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07)*, pp. 17–23: https://www.usenix.org/legacy/event/fast07/tech/full_papers/pinheiro/pinheiro.pdf.

[6] B. Schroeder, S. Damouras, and P. Gill, "Understanding Latent Sector Errors and How to Protect Against Them," *ACM Transactions on Storage (TOS)*, vol. 6, no. 3 (September 2010), pp. 9:1–9:23.

[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 Conference on File and Storage Technologies (FAST '07)*, pp. 1–16: https://www.usenix.org/legacy/event/fast07/tech/schroeder/schroeder.pdf.

[8] B. Schroeder, R. Lagisetty, and A. Merchant, "Flash Reliability in Production: The Expected and the Unexpected," in *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST '16)*, pp. 67–80: https://www.usenix.org/system/files/conference/fast16/fast16-papers-schroeder.pdf.

[9] Wikipedia, "S.M.A.R.T.": https://en.wikipedia.org/wiki/S.M.A.R.T. Accessed: 3/2/20.

[10] E. Xu, M. Zheng, F. Qin, Y. Xu, and J. Wu, "Lessons and Actions: What We Learned from 10k SSD-Related Storage System Failures," in *Proceedings of the 2019 USENIX Annual Technical Conference (USENIX ATC '19)*, pp. 961–976: https://www.usenix.org/system/files/atc19-xu_0.pdf.