Operational Characteristics of SSDs in Enterprise Storage Systems: A Large-Scale Field Study

Stathis Maneas and Kaveh Mahdaviani, *University of Toronto;*
Tim Emami, *NetApp;* Bianca Schroeder, *University of Toronto*

https://www.usenix.org/conference/fast22/presentation/maneas

This paper is included in the Proceedings of the
20th USENIX Conference on File and Storage Technologies.

February 22–24, 2022 • Santa Clara, CA, USA

978-1-939133-26-7

Open access to the Proceedings
of the 20th USENIX Conference on
File and Storage Technologies
is sponsored by USENIX.

Operational Characteristics of SSDs in Enterprise Storage Systems: A Large-Scale Field Study

Stathis Maneas
University of Toronto

Kaveh Mahdaviani
University of Toronto

Tim Emami
NetApp

Bianca Schroeder
University of Toronto

Abstract

As we increasingly rely on SSDs for our storage needs, it is important to understand their operational characteristics in the field, in particular since they vary from HDDs. This includes operational aspects, such as the level of write amplification experienced by SSDs in production systems and how it is affected by various factors; the effectiveness of wear leveling; or the rate at which drives in the field use up their program-erase (PE) cycle limit and what that means for the transition to future generations of flash with lower endurance. This paper presents the first large-scale field study of key operational characteristics of SSDs in production use based on a large population of enterprise storage systems covering almost 2 million SSDs of a major storage vendor (NetApp).

1 Introduction

Solid state drives (SSDs) have become a popular choice for storage systems over the past decade, increasingly replacing hard disk drives (HDDs). The performance and expected lifespan of an SSD are affected by operational characteristics in ways that are fundamentally different than for HDDs. For example, the lifespan is affected by write rates, as flash wears out, while performance is affected by the workload's read/write ratio due to the big differences between read and write latencies. Moreover, SSDs require background work, such as garbage collection and wear leveling, which generates write amplification and affects a drive's performance and lifespan. Usage characteristics, such as workload intensity (in particular write rates), the read/write ratio, and how full a drive is, affect how effectively a drive can manage these housekeeping tasks. Finally, drive specific details, such as whether a drive supports multi-stream writes or the amount of over-provisioned space, are expected to impact lifetime and performance as well.

As we increasingly rely on SSDs for our storage needs, it is important to understand what these operational characteristics look like for drives in the field and how they impact drives. Unfortunately, there are no large-scale field studies providing a comprehensive view of these characteristics for SSDs in the field. While there are a few recent field studies involving large-scale deployments, these have a different focus studying failure characteristics [30, 33, 36, 41, 46], fail-slow faults [13, 38], and performance instabilities [16] associated with SSDs in production.

In this paper, we present the first large-scale field study of several key operational characteristics of NAND-based SSDs in the field, based on NetApp's enterprise storage systems. Our study is based on telemetry data collected over a period of 4+ years for a sample of NetApp's total SSD population, which covers more than one billion drive days in total. Specifically, our study's SSD population comprises almost 2 million drives, which span 3 manufacturers, 20 different families (product batches, see detailed definition in §2), 2 interfaces (i.e., SAS and NVMe), and 4 major flash technologies, i.e., cMLC (*consumer-class*), eMLC (*enterprise-class*), 3D-TLC, and 3D-eTLC. Our data set is very rich, and includes information on usage, such as host reads and writes, total physical device writes, along with information on each drive's wear leveling and write amplification. Furthermore, our data contains each system's configuration, including all its RAID groups and the role of every drive within a RAID group (i.e., data or parity), among a number of other things.

We use this rich data set to answer questions, such as:

- What are the write rates that drives experience in production systems, and how close do drives get to reaching wear-out? What does this mean for future generations of flash with lower endurance limits?
- What are the write amplification factors that drives experience in production systems? How do those numbers compare to those reported in academic work?
- How effective are SSDs in production environments at wear leveling?
- How is write amplification affected by various factors, including FTL-related factors (e.g., drive model, firmware versions, over-provisioned space, support of multi-stream writes) and workload factors (e.g., write rates and read/write ratios, whether the drive is used as a cache or for persistent storage, whether the drive's role is data, parity or partitioned)?

2 Methodology

2.1 System Description

Our study is based on a rich collection of telemetry data from a large population of enterprise storage systems in production, comprising almost 2 million SSDs. The systems employ the WAFL file system [17] and NetApp's ONTAP operating system [37], while they run on custom Fabric-Attached Storage (FAS) hardware and use drives from multiple manufacturers. The systems are general-purpose, multi-tenant, and multi-protocol (NFS, FCP, iSCSI, NVMe_oF, S3), used by thousands of customers for very different applications (sometimes on the same node. In contrast to cloud data centers, which use mostly block- and object-based protocols, the majority of the systems in our data set use NFS (i.e., a file-based protocol). Applications running on our systems include file services, (enterprise) databases, financial technologies (fintechs), retail, electronic design automation (EDA) workloads, media entertainment, data analytics, artificial intelligence, and machine learning. Note though that storage vendors (e.g., NetApp) have no direct insight into what applications a customer is running on their systems, or who are the individual users of each system. Therefore, it is not trivial to break down our analysis by the type of application a system is running.

The operating system uses software RAID to protect against drive failures. Table 2 shows the breakdown of RAID group sizes in our systems, along with the breakdown of RAID schemes per range of RAID group sizes. As we observe, SSDs widely adopt RAID schemes protecting beyond single-device failures, especially with AFFs and larger arrays.

Our data set comprises systems with a wide range of hardware configurations, concerning CPU, memory, and total SSDs. Each system contains a large dynamic random access memory (DRAM) cache. Incoming write data is first buffered into the system's DRAM and then logged to non-volatile memory (NVRAM). Once the buffered (dirty) data is stored into persistent storage, during a consistency point (CP), it is then cleared from NVRAM and is (safely) retained in DRAM until it is overwritten by new data. We refer the reader to prior work for more information on the design and implementation details of NetApp systems [22, 23, 30, 31].

According to their usage, systems are divided into two different *types*: one that uses SSDs as a write-back cache layer on top of HDDs (referred to as **WBC**), and another consisting of flash-only systems, called **AFF** (All Flash Fabric-Attached-Storage (FAS)). An AFF system uses either SAS or NVMe SSDs, and is an enterprise end-to-end all-flash storage array. In WBC systems, SSDs are used as an additional caching layer that aims to provide low read latency and increased system throughput. Still, not all reads and writes are served from the SSD cache. Depending on the cache replacement policy, reads and writes can bypass the SSD cache and get served directly from the underlying HDD layer. For example, sequential user writes will typically get stored from DRAM

	Drive characteristics					Usage Characts.	
Drive Family	**Cap. (GB)**	**Flash Tech.**	**DWPD**	**PE Cycles Limit**	**OP**	**First Deploy-ment**	**Drive Power Years**
I - A	200	eMLC	10	10K	44%	Apr '14	5.69
	400					Apr '14	5.66
	800					Mar '14	5.01
	1600					Mar '14	5.49
I - B	400	eMLC	10	10K	44%	Dec '15	4.44
	800					Jan '16	4.15
	1600					Jan '16	4.25
I - C	400	eMLC	3	10K	28%	Jan '17	3.37
	800					Jan '17	3.06
	1600					Mar '17	3.32
	3800					Dec '16	2.87
I - D	3800	eMLC	1	10K	7%	Jul '17	2.82
I - E	800	3D-eTLC	1	7K	20%	Dec '18	1.67
	960					Dec '18	1.45
	3800					Dec '18	1.12
	7600					Jan '19	1.32
	15000					Jan '19	1.26
II - A	3840	3D-TLC	1	10K	7%	Jan '16	4.39
II - B	3800	3D-TLC	1	10K	7%	Oct '16	3.58
II - C	8000	3D-TLC	1	10K	7%	Sep '17	2.89
	15300					Sep '16	2.99
II - D	960	3D-TLC	1	10K	7%	Oct '16	3.37
	3800					Oct '16	3.57
II - E	400	3D-TLC	3	10K	28%	Dec '16	3.81
	800				28%	Jan '17	3.45
	3800				7%	Dec '16	3.75
II - F	960	3D-TLC	1	10K	7%	Dec '19	0.40
	3800					Mar '20	0.46
II - G	400	3D-TLC	3	10K	28%	Jan '16	4.17
	800					Feb '16	4.32
	1600					Jan '16	4.58
II - H	800	3D-TLC	3	10K	28%	Apr '18	1.91
	960		1		7%	Jan '18	1.77
	3800		1		7%	Jan '18	1.69
	8000		1		7%	May '18	1.63
	15000		1		7%	May '18	1.44
	30000		1		7%	Jul '18	1.43
II - I	800	cMLC	3	10K	28%	Sep '13	6.48
II - J	200	eMLC	10	30K	28%	Sep '13	6.92
	400					Sep '13	6.47
	800					Oct '13	6.74
II - K	400	eMLC	3	30K	28%	May '15	5.07
	800					Jul '15	4.98
	1600					Jun '15	5.11
III-A	960	3D-eTLC	1	7K	20%	Oct '19	0.69
	3800					Oct '19	0.54
	7600					Oct '19	0.57
II - X	3800	TLC	1	10K	7%	Aug '18	1.83
	7600					Jul '18	2.07
II - Y	3800	TLC	1	10K	7%	Jan '19	0.78
	7600					May '19	1.05
	15000					Dec '18	1.08
II - Z	3800	TLC	1	10K	7%	Jul '20	0.25
	7600					Jun '20	0.40
	15000					Jun '20	0.35

Table 1: *Summary statistics describing the key characteristics of the different drive families in our data set. The last three rows involve SSDs with an NVMe storage interface, whereas all the other rows involve SAS drives. The standard deviation in the drives' power-on years ranges from 0.05 to 0.98 for most drive families.*

Distribution of RAID Group Sizes		
Range	**AFF**	**WBC**
[3, 9]	16.21%	56.49%
[10, 19]	36.21%	28.98%
[20, 29]	47.58%	14.53%

Distribution of RAID schemes for AFF systems			
Scheme \ **Range**	**[3, 9]**	**[10, 19]**	**[20, 29]**
RAID-4 [39]	11.55%	0.99%	0.95%
RAID-DP [8]	88.43%	98.62 %	98.21%
RAID-TEC [11]	0.02%	0.39%	0.84%
Distribution of RAID schemes for WBC systems			
Scheme \ **Range**	**[3, 9]**	**[10, 19]**	**[20, 29]**
RAID-4 [39]	61.65%	21.50%	0.62%
RAID-DP [8]	38.18%	77.45%	97.55%
RAID-TEC [11]	0.17%	1.05%	1.83%

Table 2: *The top table shows the breakdown of RAID group sizes per system type, while the bottom table shows the breakdown of RAID schemes per range of RAID group sizes.*

directly to HDDs (as these can be executed efficiently on the HDDs and are also likely to pollute the SSD cache). Similarly, reads that result in an SSD cache miss are brought into DRAM and will be written to the SSD cache as well only if the cache replacement policy determines that the chance of reuse is high and it is worth evicting another block from the SSD cache.

In the remainder of the paper, we make use of the following terms (adapted from [3]):

• **Drive family**: A particular drive product, which may be shipped in various capacities, from one manufacturer, using a specific generation of SSD controller and NAND. Our data set contains 20 different families (denoted by a capital letter A–Z) from three different manufacturers (denoted as I, II, and III). We prepend the manufacturer's symbol to each drive family in order to explicitly associate each family with its manufacturer (e.g., I-A, II-C).

• **Drive model**: The combination of a drive family and a particular capacity. For instance, the I-B drive family comes in three models whose capacity is equal to 400, 800 or 1600GB.

• **Drive age**: The amount of time a drive has been in production since its ship date, rather than its manufacturing date.

The first six columns in Table 1 describe the key characteristics associated with the different drive families in our data set. Specifically, for each drive family, Table 1 includes all the corresponding drive models (in an anonymized form), along with the capacity, flash technology, endurance (specified in Drive Writes Per Day (DWPD) and the program-erase (PE) cycles limit), and over-provisioning (OP) associated with each model. As shown in Table 1, the SSD population in our study spans a large number of configurations that have been common in production settings over the last several years.

2.2 Data Collection and Description

Most systems in the field send telemetry data in the form of NetApp Active IQ® (previously called AutoSupport) bundles, which track a large set of system and device parameters (without containing copies of the customers' actual data). These bundles are collected and used for detecting potential issues.

Device and System Metrics	Sections
Host Write Rates/Read Rates	§3.1.1, §5
Annualized NAND Usage Rate	§3.1.2
Write Amplification Factor (WAF)	§3.2, §4
Avg/Max Erase Operations	§3.3
System Fullness	§3.4

Table 3: *The list of metrics analyzed in this study.*

Our study is based on mining and analyzing this rich collection of messages. Specifically, our data set is organized into 21 snapshots, each of which is generated after parsing the corresponding support messages collected at the following 21 points in time: Jan/Jun '17, Jan/May/Aug/Dec '18, Feb–Dec '19, Jan/Jul/Nov '20, and Mar '21. Each snapshot contains monitoring data for every system (and its drives). Table 3 shows all the metrics that are analyzed in this study.

3 What does SSD overall usage look like?

The performance and endurance of an SSD depend significantly on a number of operational characteristics. In this section, we use our field data to study four of the most important characteristics (described below). To the best of our knowledge, our work is the first to present details on these characteristics for a large-scale population of flash-based production systems.

• We begin by studying *write rates* (§3.1), as experienced by enterprise drives in the field (including both host and physical writes), as they significantly impact the lifetime of an SSD.

• We then examine the *write amplification factors (WAF)* observed by the SSDs in our study (§3.2), as write amplification is another major factor that can reduce a drive's endurance.

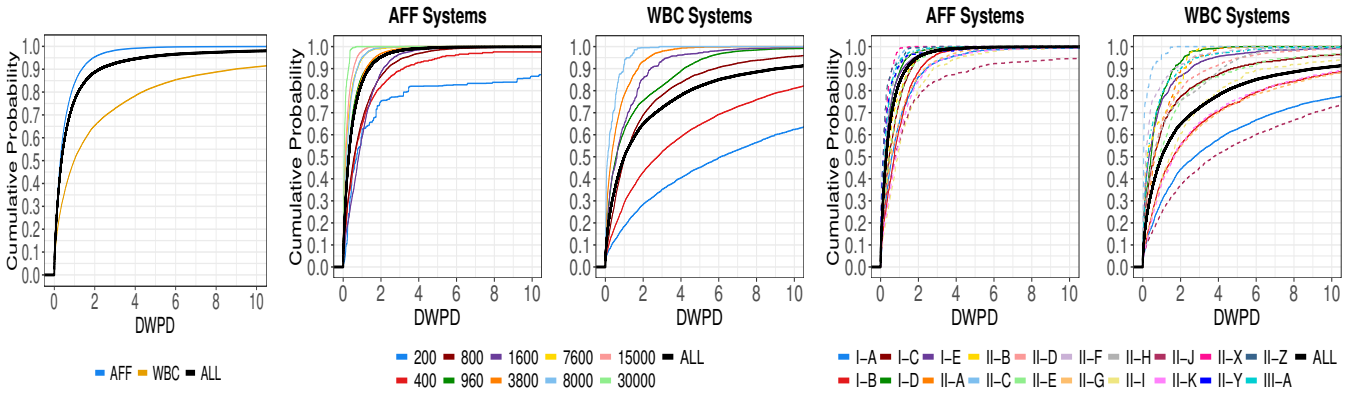
• Next, we look at how efficient drives are at *wear leveling* (§3.3), as it is a key mechanism that prolongs a drive's lifetime by preventing heavily used blocks from premature wear-out.

• Finally, we look at the *fullness* of systems (§3.4), i.e., what fraction of a system's total storage capacity is actually used. Fullness can significantly affect SSD operations, as a full system will trigger garbage collection more frequently and also has less free space to facilitate wear leveling and other housekeeping tasks.

3.1 Host Write Rates and NAND Usage Rates

A major concern when deploying SSDs are the write rates these devices will experience in the field, as erase operations wear out flash cells; therefore, the write intensity of a workload significantly affects the lifetime of flash-based SSDs. This is a particular concern looking forward since future generations of flash are expected to have an order of magnitude lower endurance than today's drives.

The goal of this section is to study a number of important aspects associated with write rates, as experienced by drives in enterprise systems, including how close drives get to their point of wear-out, how write rates vary across systems, differences between host and physical writes seen by a drive,



(a) *Per System Type.* (b) *Per Drive Capacity.* (c) *Per Drive Family.*

Figure 1: *Distribution of the drives' (normalized) host writes broken down by system type (1a), drive capacity (1b), and drive family (1c). In Figure 1c only, each line type corresponds to a different manufacturer.*

and an analysis of how feasible it would be to migrate workloads observed on today's systems to future generations of flash with lower program-erase (PE) cycle limits (i.e., the maximum number of PE cycles each SSD is rated for by its manufacturer).

3.1.1 Host Write Rates

We first look at write rates from the angle of *host writes*, i.e., the writes as they are generated by the applications and the storage stack running on top of the SSDs and measured and collected by the operating system (in contrast to physical NAND writes, which we examine in §3.1.2).

Because of the significance of write rates in the context of SSD endurance, drive manufacturers specify for each model its *Drive Writes Per Day* (DWPD), which is defined as the average number of times a drive's entire capacity can be written per day over its lifetime without wearing out prematurely. Typical DWPD numbers in drive datasheets are 1 and 3, and our population also includes some models with a DWPD of 10 (see Table 1 for all drive models in our study). However, trends associated with recent technologies suggest DWPD will drop below 1 in the future [34].

Understanding host writes is also important in other contexts. For example, when setting up workload generators or benchmarks for experimental system research, it is important to understand what realistic workloads one wants to emulate look like, and write rates are an important aspect of that.

Despite the significance of host writes and the fact that flash-drives have been routinely deployed at large scale for the past decade, we are not aware of any study reporting on host write rates in such systems. The goal of our measurements is to close this gap.

We present our results in Figure 1a. The black solid line in Figure 1a (left) shows the Cumulative Distribution Function (CDF) of the DWPD experienced by the drives across our entire population. In addition, the graph also breaks the results down into AFF (all flash) systems and WBC systems (where the flash is used as a write-back cache).

We make a number of high-order observations:

• The DWPD varies widely across drives: the median DWPD of the population is only 0.36, well below the limit that today's drives can sustain. However, there is a significant fraction of drives that experiences much higher DWPD. More than 7% of drives see DWPD above 3, higher than what many of today's drive models guarantee to support. Finally, 2% of drives see DWPD above 10, pushing the limits even of today's drive models with the highest endurance.

• When separating the data into AFF and WBC systems, we observe (probably not surprisingly) that WBC systems experience significantly higher DWPD. Only 1.8% of AFF drives see DWPD above 3 compared to a quarter of all WBC drives. The median DWPD is 3.4× higher for WBC than AFF, while the the 99th percentile is 10.6× higher.

• We note vast differences in DWPD across the WBC systems, including a long tail in the distribution. While the median is equal to 1, the drives in the 99th and the 99.9th %-ile experience DWPD of 40 and 79, respectively. What that means is that designers and operators of WBC systems need to be prepared for their systems to handle a vast range of DWPD values, including very high ones. It also means that optimally provisioning the drive endurance for a WBC system is much harder due to the huge range of DWPD in such systems.

Next, we perform a more fine-grained analysis of the DWPD experienced by different SSDs, by grouping drives based on their capacity (Figure 1b) and by drive family (Figure 1c). The reasoning is that different customers will purchase drives of different capacities depending on their applications' needs, so drives of different capacities likely see different types of workloads. Similarly, different drive families might be deployed in different types of systems that differ in the workloads that run on them.

• Turning to Figure 1b, we were surprised to see how significantly DWPDs vary depending on *drive capacity*. In particular, there is a very clear trend that smaller capacity drives see larger DWPD. While this trend is consistent for both

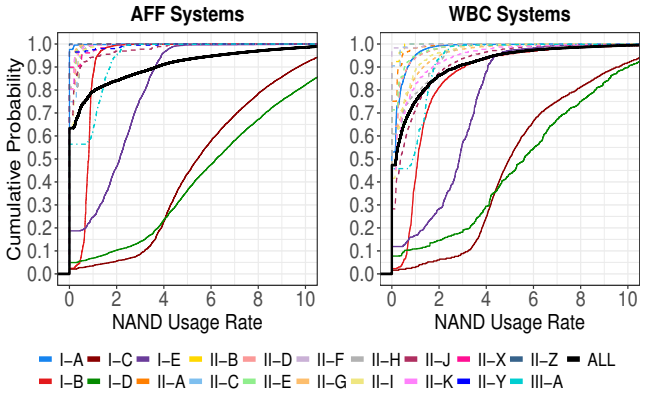


Figure 2: *Distribution of the drives' Annualized NAND Usage Rates broken down by drive family and system type. Each line type corresponds to a different manufacturer.*

AFF and WBC systems, the effect is particularly pronounced for WBC systems: here, the median DWPD for the smallest capacity drives is more than 100× higher than for the largest capacity drives (DWPD of 0.05 compared to 6). We note that this trend also holds when comparing the DWPD of different capacity drives within the same drive family, so the effect can be clearly attributed to drive capacity rather than family.

• Interestingly, we also observe significant differences in DWPD across drive families (Figure 1c). For example, for WBC systems, the median DWPD ranges from 0.04 to 3.75 across drive families. We also observe that for both AFF and WBC systems, it is the same drive families that experience higher DWPD than the average population.

3.1.2 NAND Usage Rates

The second metric associated with write operations focuses on *physical NAND device writes*. Physical device writes are typically higher than the raw host writes due to the device's background operations (e.g., garbage collection, wear leveling). For each drive model, manufacturers specify a limit on the number of physical writes it can tolerate before wearing out, in terms of the program-erase (PE) cycle limit (see Table 1 for the PE cycle limit of the drives in our population).

We are interested in studying the rate at which drives in the field approach their PE cycle limit, a question that is of particular concern as future generations of flash are expected to have significantly lower PE cycle limits [32]. Towards this end, for each drive, we determine the percentage of its PE cycle limit that it uses up per year, on average, a metric that we refer to as *Annualized NAND Usage Rate*:

$$Ann.\ NAND\ Usage\ Rate = \frac{\%\ of\ PE\ Cycle\ Limit\ Used\ So\ Far}{Power\text{-}On\ Years} \quad (1)$$

Figure 2 shows the NAND usage rates for AFF and WBC systems. The black solid line in each graph shows the CDF of the NAND usage rates across all the drives, irrespective of drive family. Since physical writes depend heavily on a drive's FTL (unlike host writes which are mostly driven by the applications), the figure also shows the CDF of NAND usage rates separately for each drive family.

We make the following key observations:

• Annualized NAND Usage Rates are generally low. The majority of drives (60% across the entire population) report a NAND usage rate of zero[1], indicating that they use less than 1% of their PE cycle limit per year. At this rate, these SSDs will last for more than 100 years in production without wearing out.

• There is a huge difference in NAND Usage Rates across drive families. In particular, drive families I-C, I-D, and I-E experience much higher NAND usage rates compared to the remaining population. These drive families do not report higher numbers of host writes (recall Figure 1c), so the difference in NAND usage rates cannot be explained by higher application write rates for those models.

We therefore attribute the extremely high NAND usage rates reported by I-C/I-D drives to other housekeeping operations which take place within the device (e.g., garbage collection, wear leveling, and data rewrite mechanisms to prevent retention errors [6]). We study this aspect in more detail in Section 3.2 and in Section 4, where we consider Write Amplification Factors (WAF).

• There is little difference in NAND usage rates of AFF systems and WBC systems. This is surprising given that we have seen significantly higher host write rates for WBC systems than for AFF systems. At first, we hypothesized that WBC systems more commonly use drives with higher PE cycle limits, so higher DWPD could still correspond to a smaller fraction of the PE cycle limit. However, we observe similar NAND usage rates for WBC systems and AFF systems, even when comparing specific drive families and models with the same PE cycle limit. Interestingly, as we will see in Section 3.2, the reason is that WBC systems experience lower WAF, which compensates for the higher host write rates.

Projections for next generation drives: We can use NAND usage rates to make projections for next generation QLC drives. Considering that endurance is estimated to be reduced for QLC drives [32], we are interested in determining how many SSDs in our data set could be replaced by a QLC SSD without wearing out within a typical drive lifetime of 5 years.

• If we assume that the PE cycle limit of QLC drives drops to 1K, then we find that the vast majority of our population (~95% of drives when excluding the two outlier models I-C and I-D) could have used QLC drives without wearing them out prematurely.

3.2 Write Amplification Factor (WAF)

The write amplification factor (WAF) plays a critical role, as the added writes due to garbage collection, wear leveling, and other SSD-internal housekeeping tasks, can negatively impact

[1] Unfortunately, the % of PE cycle limit used per SSD is reported as a truncated integer.

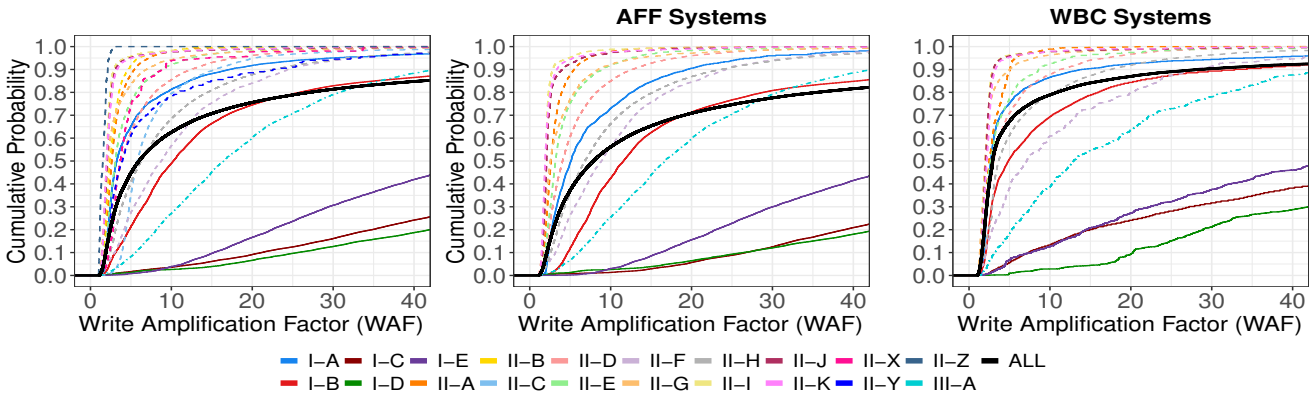


Figure 3: *Distribution of the drives' WAF broken down by drive family (left), along with both drive family and system type (middle and right). Each line type corresponds to a different manufacturer.*

both the endurance and performance of an SSD. It is therefore not surprising that a large body of work has been dedicated to reducing WAF and its impact, for example by optimizing FTLs in different ways [7, 14, 15, 18, 20, 25, 44, 47–49] or by making use of Multi-stream SSDs [4, 21, 40].

Unfortunately, despite the large body of work in industry and academia on WAF, we do not have a good understanding of how effective real drives in production systems are in controlling WAF. To the best of our knowledge, there is no large-scale field study reporting and analyzing WAF in production systems. The existing field studies that mention WAF for production systems are either limited to one particular type of application (financial services) [29] or are based on a small population of one flash technology (3D-TLC) [28]; both studies simply report an average WAF across their systems of 1.3 and 1.5, respectively (without any further analysis).

One goal of this paper is to improve our understanding of WAF in production systems. We begin in this section with some high-level statistics on WAF and then later in Section 4 study in more detail the impact of various factors on WAF.

The black solid line in Figure 3 (left) shows the distribution of WAF across all the drives in our population. In the same graph, we also show WAF broken down by drive family, as a drive's FTL affects WAF.

We make a number of high-level observations:

• For the vast majority of our SSDs, the WAF they experience is higher than the WAF of 1.3 observed in [29] (a field study reporting WAF numbers, but only in the context of financial services applications) and the WAF of 1.5 observed in [28] (based on a sample of 3D-TLC SSDs from Huawei's storage systems). Specifically, 98.8% and 96% of our SSDs observe a WAF larger than 1.3 and 1.5, respectively. This observation underlines the importance of field studies spanning a large range of systems with different applications and devices.

• The drives in our population span a huge range of WAF values. While the 10th percentile is only 2, the 99th percentile is 480. This motivates us to study the effect of several different factors on WAF. We start below with a high-level study of the role of the FTL and workloads, and continue with a more detailed study of factors impacting WAF in Section 4.

WAF and the FTL: As different drive families vary in their firmware, comparing the WAF across drive families provides insights into the relationship between the FTL and WAF.

• Figure 3 (left) shows that some drive families have *drastically* higher WAF than others. In particular, the I-C, I-D, and I-E families experience WAF that is an order of magnitude higher than that for most of the other drive families, with median WAF values of around 100 (!) in the case of I-C and I-D. Note that these drive families do not experience a different host write rate and we have no indication that they are being deployed in systems that tend to run different types of applications, so there is no obvious explanation due to workload characteristics. Also, differences in WAF persist even when we compare with drive families of the same age and capacity.

Upon closer inspection, we found that these particular models perform background work every time the SSD has idle cycles to spare, thereby consuming their PE cycles as a side effect. Interestingly, it seems that this background work is not due to garbage collection or wear leveling (the best studied contributors to WAF), but due to aggressive rewriting of blocks to avoid retention problems, where stored data is (periodically) remapped before the corresponding flash cells accumulate more retention errors than what can be corrected by error correction codes (ECC) [6].

We note that this unexpected effect drives up WAF not only for drives with extremely low utilization, but also for the busiest drives (e.g., top 5%) of the two outlier families.

In summary, the FTL has a huge impact on WAF.

WAF and workload: Our data also provides evidence of the impact of workload characteristics on WAF:

• First, we observe that there is significant variation in WAF even when comparing only drives within the same drive family (rather than across families). The 95th percentile of a drive family's WAF is often 9× larger than the corresponding median. These differences are likely due to different drives within a family being exposed to different workloads.

• Second, when we compare the WAF for AFF and WBC

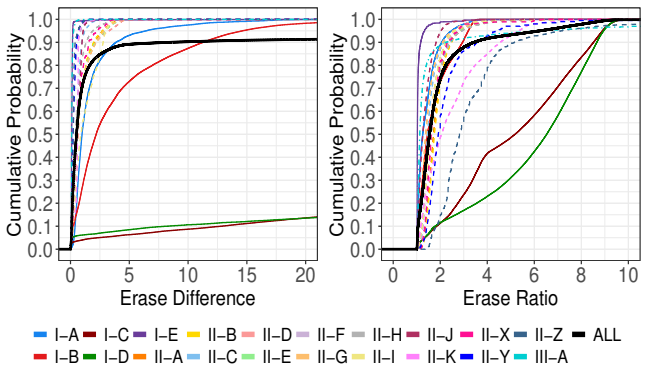


Figure 4: *Distribution of metrics associated with wear leveling, calculated based on the number of erase operations per block. Each line type corresponds to a different manufacturer.*

systems in Figure 3 (two right-most figures), we observe that for the same drive families, WBC systems experience significantly lower WAF than AFF systems, indicating that WBC workloads are more flash friendly. This results in an another interesting observation:

• Thanks to their lower WAF, WBC drives in our systems do not see a higher NAND usage rate than AFF systems, despite their higher DWPD (recall Figure 1c). This observation is significant, because the application of SSDs in caches is considered the most demanding, in terms of endurance requirements, and widely accepted best practices recommend to use only drives with the highest endurance for these applications. Our observations indicate that this might not always be necessary.

Comparison with simulation studies: Due to the dearth of field data on WAF, a number of authors have resorted to trace-driven simulation studies to explore WAF and how it is affected by various factors; therefore, it is interesting to compare the numbers we observe against those studies.

• The values reported for WAF in trace-driven simulation studies [5, 9, 10, 19, 42, 45] are at the low end of the WAF range we observe for AFF production systems, and even the *maximum* values reported in these studies fall only into the mid range (often below median) of the WAF values we observe. For example, the *highest WAF* in [45] is 2, in [42] it is 7, and in [5, 9, 10, 19] it is 12, which correspond to the 9th, 49th, and 62th percentile respectively, of the WAFs in our AFF population.

We draw two possible conclusions from this differences:

• A significant challenge that researchers in our community face is the lack of publicly available I/O traces from SSD-based storage systems. As a result, existing experimental work, including the simulation studies cited above, is based on traces that are i) based on HDD systems and ii) mostly relatively old (more than a decade for some popular traces). These traces might not be representative of today's workloads running on SSD-based systems and also do not cover aspects relevant to SSD-based systems (e.g., the TRIM command). As a community, it is important that we find more relevant traces to use as base to our work.

• The differences in WAF between our production systems and existing simulation studies also indicate that it is very difficult to reproduce all complexities and subtleties of modern FTLs in simulation.

3.3 Wear Leveling

A critical job of an SSD controllers is wear leveling, which aims to spread the erase operations evenly over all blocks on the device. This is important for multiple reasons. First of all, it serves to increase the device's lifetime by preventing frequently used blocks from wearing out prematurely. Second, it can help avoid performance problems, since blocks with higher erasure cycles are associated with higher error rates [12], and retries and other error-correction efforts can significantly add to latency.

Wear leveling is a difficult problem because real-world workloads are rarely uniform and commonly exhibit strong skew in per-block update frequencies. An ideal wear leveling mechanism distributes write operations in such a way so that all blocks within an SSD wear out at the same rate. At the same time, there is a delicate trade-off, as aggressive wear leveling will increase the number of write operations, thereby increasing WAF and overall drive wear-out.

In this section, we explore how effective modern FTLs are at wear leveling. Specifically, our data set contains the average number of times the blocks in a drive have been erased, along with the corresponding maximum value. Based on these values, we calculate two different metrics that characterize how evenly erase operations are distributed across all blocks:

The *Erase Ratio* is the ratio between the maximum and average number of erase operations per SSD:

$$Erase\ Ratio = \frac{Max.\ Erase\ Ops}{Avg.\ Erase\ Ops} \quad (2)$$

The *Erase Difference* is the absolute difference between maximum and the average number of erase operations normalized by the PE cycle limit:

$$Erase\ Difference = \frac{Max.\ Erase\ Ops - Avg.\ Erase\ Ops}{PE\ Cycle\ Limit}(\%) \quad (3)$$

Figure 4 shows the Erase Difference and Erase Ratio across our entire population (black solid line) and broken down by drive family. The ideal value of the Erase Ratio is 1, whereas the ideal value of the Erase Difference is 0. We make the following observations:

• Not surprisingly, wear leveling is not perfect. The median Erase Ratio is 1.55, indicating that the maximum block undergoes 55% more erase operations than the average block. 5% of the drives have an erase ratio larger than 6 meaning their maximum block wears out 6× faster than the average - that means when the maximum block has reached end of life the

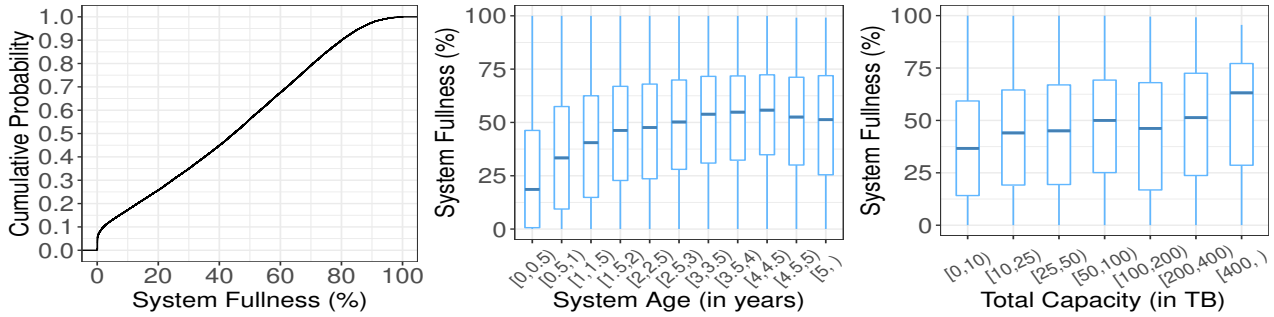


Figure 5: *Distribution of AFF systems' fullness.*

average block has only used 16% of its PE cycle limit.[2].

• There is a striking difference in the wear leveling metrics across drive families. The fact that I-C and I-D drives, for example, report significantly higher wear leveling metrics (despite having similar age, capacity, and DWPD to some other families) indicates that different firmware implementations take vastly different approaches to wear leveling. More generally, it seems that different *manufacturers* follow very different philosophies with respect to wear leveling: when looking at the Erase Difference metric, we see that the four families with the largest Erase Difference all belong to the same manufacturer (i.e., I).

• It is surprising that drive models I-C and I-D do a significantly worse job at wear leveling than other drive models, despite the fact that these two models experience higher WAF (recall §3.2). This means that the additional background work that these drives are performing is not contributing towards better wear leveling. Instead, we believe that the additional background work of those two drive families is because they *rewrite data more aggressively than others in order to avoid retention errors*. This is a very interesting observation, since data rewrite for retention errors has received much less attention than other sources of WAF (e.g., garbage collection, wear leveling). In fact, current SSD simulators and emulators (e.g., FEMU [26]) do not implement data rewrite for retention errors, and therefore do not capture this source of WAF.

3.4 Fullness

Another critical factor in the operation of an SSD-based storage is the system's *fullness*. We define *fullness* as the fraction of the drives' nominal capacity that is filled with valid data, i.e., the fraction of the Logical Block Address (LBA) space currently allocated. Fullness can affect the overall performance of a system, as it can impact the frequency of garbage collections, and also determines how much free room there is for operations like wear leveling. Also, fullness is of practical importance for capacity planning, as systems that run out of available space before the end of their lifetime need to be expanded with additional storage.

On the other hand, from the garbage collection's point of view, fullness denotes what fraction of blocks inside the drive are currently not programmable. This includes blocks containing valid data, but also blocks containing invalidated data which have not been erased yet. In our analysis, we focus only on the (allocated) LBA space.

In this section, we are interested in exploring what fullness looks like for enterprise storage systems, how it changes over a drive's lifetime, and how it varies as a function of factors such as drive capacity. Our study is the first to characterize this important system aspect for flash-based storage systems.

We begin with a high-level view of fullness by considering the CDF of fullness across the entire population of AFF systems, as shown in Figure 5 (left); we consider only AFF systems in our study of fullness, as the concept of fullness does not apply in the same way to WBC systems, which use SSDs only as a cache on top of HDDs.

• We observe that the average system is around 45% full, and the median is also around 45%, i.e., more than half of the storage capacity is free.

• The distribution of fullness across systems is roughly uniform. The CDF flattens only above 80%, i.e., values below 80% are all roughly equally likely, while values above 80% are relatively less common.

Next, in Figure 5 (middle), we look at how fullness changes over a system's lifetime. Understanding this aspect of fullness is relevant, for example, in the context of capacity planning.

• Maybe not surprisingly, system fullness increases with age. (Consider for example the median over time, as indicated by the dark link in the center of each box plot). However, the rate of increase is not uniform: fullness grows relatively fast over the first two years and stabilizes after that.

• Interestingly, despite the fact that generally fullness increases over time, there are some very young systems that are quite full and some old systems that are quite empty: slightly more than 5% of young systems (less than 1 year old) are more than 80% full, whereas 19% of old systems (more than 4 years old) are less than 25% full.

• An interesting observation from a capacity planning point of view is that systems who end up being full at the end of their life are also among the fullest systems early in their life. In other words, if a system has not used a significant amount of its physical space after its first couple of years in

[2]We do not have data on the minimum number of erase operations of a drive's blocks – naturally the difference between the minimum and maximum block would be even more pronounced.

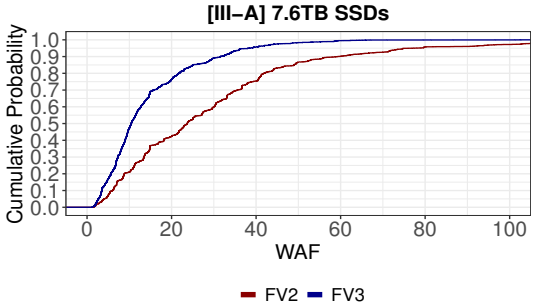


Figure 6: *WAF comparison between two firmware versions within the same drive family.*

production, its fullness will most probably remain (relatively) low in the future.

Given that systems vary hugely in their total capacity, ranging from tens of TBs to a couple of PBs, another interesting question is whether users of larger capacity systems actually make use of the additional capacity of their systems. Towards this end, Figure 5 (right) presents boxplots of system fullness, broken down by system capacity.

• Interestingly, we observe that system with *larger* total capacity tend to be *more* full: the largest systems are 1.7× fuller (in terms of median) than the other systems. This seems to indicate that customers who purchase larger capacity systems do indeed have larger capacity needs and are also better at predicting how much storage capacity they need.

Comparison with fullness reported for other types of systems: A seminal work by Agrawal et al. [1], published more than a decade ago, studied file system characteristics, including fullness, for personal desktop computers at Microsoft. They observed average fullness values ranging from 45–49% and a uniform distribution. This is quite similar to our observations – which is surprising given that their study looks at completely different types of systems (personal desktop computers using HDDs).

The only other work we found that reports on fullness in production systems is by Stokely et al. [43], which studies the usage characteristics of an HDD-based distributed file systems within a private cloud environment. Their results indicate that the fraction of the quota used by an average user of those systems is significantly larger than the levels of fullness we observe: on average users use 55% of their purchased quota; for the largest quota requests this number increases to 69%. The reason might be that it is easier for a user to increase their individual quota in a distributed storage system when running out of space, compared to increasing the physical capacity of an enterprise storage system. Therefore capacity planning for an enterprise storage system has to be more conservative.

4 Which factors impact WAF?

There are a number of factors that are commonly assumed to impact a drive's WAF, including the design of a drive's FTL, usage characteristics, how full the system is, along with the size of the drive's over-provisioned space. In this section, we try to shed more light on the impact of each of these factors on WAF, as experienced in production systems.

4.1 Flash Translation Layer (FTL)

This points to the importance of different design choices made by different FTL implementations. In Section 3.2, we have observed huge differences in WAFs across different drive families, even when controlling for other factors, such as drive capacity and DWPD.

In this section, we attempt a more fine-grained look at the impact of FTLs on WAF. Instead of comparing WAF across different drive families, we now look at different firmware versions within a given drive family.

Firmware version and WAF: We performed this study for several drive families, and discuss the results for drive family III-A, as a representative sample. The most common firmware versions for this drive family are versions FV2 and FV3. We see consistently across all capacities of this drive model that the more recent firmware version FV3 is associated with lower WAF than the earlier FV2 version.

For illustration, we present the CDF of WAFs for firmware versions FV2 and FV3 for the 7.6TB capacity model of drive family III-A in Figure 6. We chose this particular capacity because it offers the cleanest comparison, as the population of FV2 drives and the population of FV3 drives are quite similar with respect to other factors, such as DWPD, deployment time, and system type.

We observe a clear difference between the WAF of drives on firmware version FV2 versus version FV3. For example, both the median and the 90th percentile of WAF are around 2× larger for FV2 than the more recent FV3 version.

4.2 Workload Characteristics

Many aspects of workload characteristics can impact a drive's WAF. Unfortunately, the analysis of many of these aspects (e.g., sequentiality of writes and deletes, skewness of updates across blocks) would require block-level IO-traces, whose collection for production systems at large scale is infeasible.

Instead, we focus on the following five aspects of workload characteristics for which we were able to collect data: the first is write intensity as measured by drive writes per day (DWPD) seen by a drive. The second is the *role* of a drive within a RAID group[3]; we distinguish among *data* and *partitioned* drives, where each partition of the drive is part of a different RAID group and different partitions can play different roles in their RAID groups. We exclude *parity* drives due to insufficient data. The third and fourth factors are the drive capacity and drive interface (SAS vs. NVMe), respectively, as drives of different capacities and different interfaces will be used in different types of systems, which might be used for different types of workloads. The fifth factor is the read/write ratio of

[3] In our data set's RAID systems, parity blocks are not rotated.

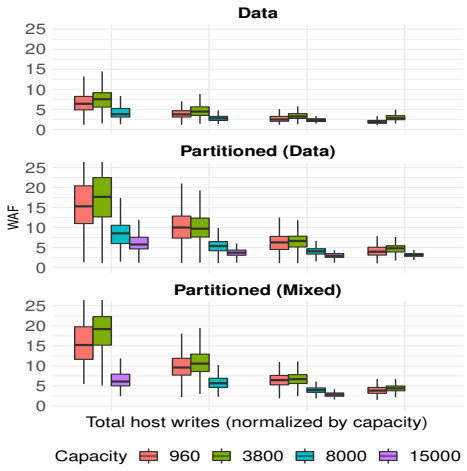


Figure 7: *Impact of different factors on WAF, while focusing on different roles within a single drive family.*

the workload.

Figure 7 shows WAF broken down by the first three aspects described above. We describe our observations below.

Drive Writes Per Day (DWPD) and WAF: We observe that consistently across different capacities and drive roles, *WAF decreases* as the number of *DWPD increases*. Median WAF is up to 4.4× higher for the drive populations with the lowest DWPD (left-most group of bars in Figure 7) than the group with the highest DWPD (right-most group of bars). This could suggest that SSDs operate more efficiently (in terms of background tasks and WAF) under higher write rates. It could also mean that some FTL background work is constant, i.e., not strongly dependent on DWPD; therefore, higher DWPD will reduce the effect of this constant work on the WAF ratio.

Drive role and WAF: We observe a significant difference in WAF depending on the drive *role*. In particular, the median WAF associated with *partitioned* SSDs is *significantly higher* (by up to 3X) than that for *data* SSDs. One possible explanation for the higher WAF of partitioned SSDs might be that they are forced to handle requests coming from different workloads with potentially different characteristics, thus experiencing a mixture of write patterns.

We do note that the difference across roles decreases as the number of (normalized) total host writes increases, suggesting that write rates have a stronger impact on WAF than its role.

Drive capacity and WAF: When we explore the impact of capacity, we observe that *higher-capacity* SSDs (i.e., 8TB and 15TB) experience *lower WAF* compared to the two smaller-capacities, for the same range of total host writes and the same drive role. In particular, their median WAF can be up to 2–3× smaller, with the difference being more pronounced when the amount of total host writes is low. Still, 3.8TB SSDs experience slightly higher WAF compared to 960GB SSDs, suggesting that smaller-capacity SSDs do not necessarily experience higher WAF (i.e., other factors have a stronger impact on WAF).

Drive interface and WAF: The workloads that customers

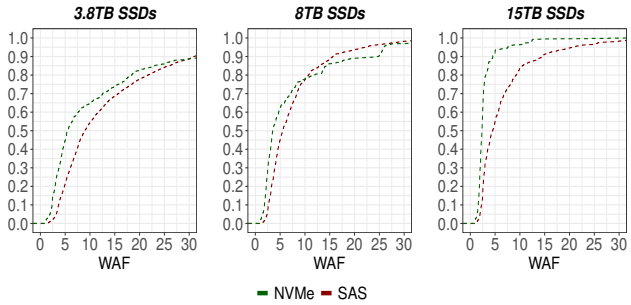


Figure 8: *WAF comparison between SAS and NVMe SSDs for different drive capacities.*

choose to run on NVMe drives tend to experience slightly smaller WAF than those on SAS drives. Results are shown in Figure 8, which compares the distribution of WAF as experienced by SAS and NVMe drives respectively, broken down by three different drive capacities. The populations of SAS and NVMe drives were chosen in such a way as to control for other factors, such as DWPD, total time in production, drive role, and system fullness. We removed the outlier drive families, for which we observed earlier an extremely high WAF, from the SAS population, so they do not bias our results.

Considering that NVMe SSDs make use of a similar FTL compared to SAS drives, we expect differences in WAF to come mostly from them being used differently. For instance, the NVMe technology is still relatively new and as a result, in our data set, the population of NVMe-based systems is smaller than the SAS-based population. The NVMe systems are mostly used by (a small set of) customers who are early adopters of this new technology. These customers, and their workloads, might be different from the average customers across the whole population. Therefore, the workloads experienced by NVMe and SAS drives can be quite different (for now); the difference will likely become less pronounced over time, as more customers move to NVMe-based systems.

Read/write ratios and WAF: We observe a positive correlation between a workload's R/W ratio and WAF. More precisely, we used the buckets of drives we created for Figure 7 (so that we control for capacity, write rates and drive role), and computed for each bucket the Spearman correlation coefficient between the R/W ratio and WAF for the drives in the bucket. The correlation coefficient is between 0.2 and 0.4 for most buckets, indicating a positive correlation.

4.3 Fullness

The fullness of a drive can affect how effectively it can manage its internal housekeeping tasks, such as garbage collection and wear leveling, especially when its total free space becomes (too) low.

We study the effect of fullness on WAF by dividing our population into drives that are more than 80% full and those that are less than 80% full, and comparing their WAF.

Interestingly, we observe *no significant differences* in the WAF experienced by the two sub-populations; in fact, SSDs which are more full experience (slightly) smaller WAF overall,

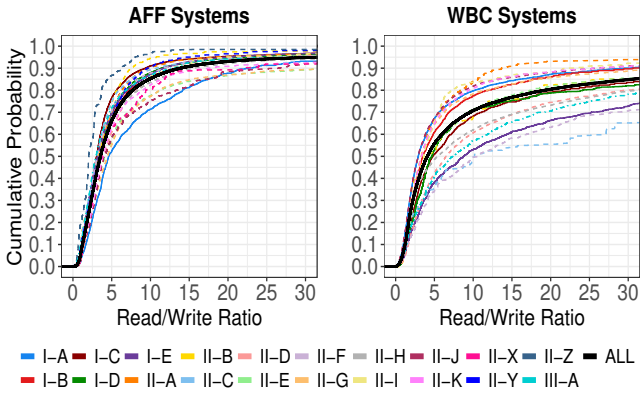


Figure 9: *Distribution of the drives' R/W ratios per system type. Each line type corresponds to a different manufacturer.*

suggesting that the drives' WAF is dominated by other factors than fullness, such as their firmware.

4.4 Over-provisioning (OP)

Another interesting question is whether WAF varies depending on the drives' amount of over-provisioning (OP). OP refers to the fraction of a drive's capacity that is reserved as spare capacity to improve a drive's wear leveling, garbage collection, and random write performance; thus, it is expected to help reduce WAF. In fact, prior work uses simulations to demonstrate the positive effect of higher OP on WAF [10, 19].

Common OP percentages for real drives are 7% and 28% and our SSD population includes drives with both OP percentages, allowing us to compare their effect on WAF.

Surprisingly, we observe that the drives with the *higher OP* (i.e., 28%) actually experience *higher*, rather than lower WAF. One possible explanation could be that many of the drives in our population are not very full (§3.4), and therefore the extra capacity in the OP space does not make much of a difference. We therefore look at the effect of OP for only those drives that are full (more than 80% of capacity in use) and still observe slightly higher WAF for SSDs with higher OP. This suggests that there are likely other factors (e.g., workload characteristics and firmware) that are more dominant than OP. For example, the drives with 7% OP have support for multi-stream writes, which might help lower their WAF. Finally, 7% OP is a younger technology and thus, the corresponding systems can be (potentially) adopted by a different set of customers, whose workload characteristics might be different from the average customers across the 28% OP population.

4.5 Multi-stream Writes

Several drive models in our data set support multi-stream writes (MSW) [21], which can help reduce WAF by allowing the host to dictate the data placement on the SSD's physical blocks. In fact, our analysis of OP showed that drives with 7% OP, all of which have MSW support, report lower WAF. Therefore, we perform a detailed analysis on the impact of MSW, while controlling for other factors (e.g., DWPD, *role*).

We observe relatively clear trends for *data* drives, where populations with MSW have 30-40% lower WAF than comparable populations without MSW.

However, the trend is not clear, and in fact sometimes reversed for *partitioned* drives. It's possible that workload factors (which we previously saw are strong for partitioned drives) dominate those populations. It's also possible that for partitioned drives the streams are mostly used for performance isolation of different partitions, rather than for reducing WAF.

5 Read/Write (R/W) Ratios

In this section, we characterize the read/write (R/W) ratios exhibited by the workloads in our systems. R/W ratios are an interesting aspect of SSD-based systems for multiple reasons:

First, the combination of reads and writes can significantly impact the observed performance of reads in SSDs, as read operations compete internally with (slower) write operations. Second, newer generation of SSDs, such as QLC SSDs, are targeted for read-intensive workloads, as their PE cycle limits are much smaller than previous generations (up to 10× compared to TLC [32]). Therefore, exploring the trends in existing workloads is interesting. Third, providing data on R/W ratios in production systems helps researchers and practitioners to set up more realistic testbeds, as a workload's R/W ratio is a key configuration parameter of existing benchmark tools, such as FIO [2]. The results of our study can be used to parameterize simulation and experimental testbeds with R/W ratios that are representative of production workloads. Finally, in WBC systems, where the SSDs are used as a caching layer on top of HDDs, the read/write ratio can be viewed as a measure of the cache's effectiveness in caching reads.

In this section, we perform the analysis of read/write ratios separately for WBC systems and AFF systems, as read/write ratios have a different interpretation for these two systems. We distinguish between the two system *types*, as customers who buy HDD-based systems tend to use them differently from those who buy SSD-based systems; in our analysis, we characterize the differences.

5.1 R/W ratios and AFF systems

Figure 9 (left) shows the distribution of R/W ratios associated with the SSDs in AFF systems, computed based on host reads and host writes reported by our systems. We begin with a few high-level observations based on this figure:

- We observe that the vast majority of drives, around 94%, experience more reads than writes. The median R/W ratio is 3.6:1 and the 95th percentile is 61:1.
- These R/W ratios are in stark contrast to trace analysis results from HDD-based storage systems, which generally have more writes than reads. For example, the FIU traces [24], the majority of volumes in the Microsoft traces [35], and the recent block I/O traces from Alibaba data centers [27], all experience more writes than reads.
- The significant difference between the R/W rates of SSD-

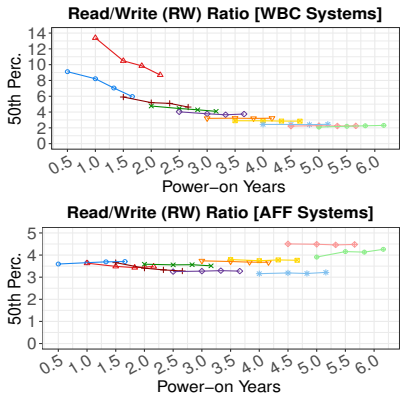


Figure 10: *Evolution of the drives' R/W ratios over time, for WBC (top) and AFF (bottom) systems.*

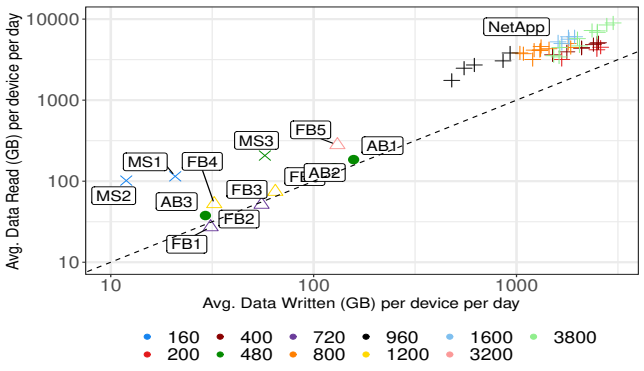


Figure 11: *Comparison of the daily workload experienced by SSDs in data centers versus enterprise storage systems. The dotted line represents equal amount of daily data reads and writes; read-dominant workloads are above the line.*

and HDD-based systems underlines the importance of our earlier observation that replaying traces from HDD systems for experiments or simulations of SSD systems is problematic. Our community needs to find a way to collect and make publicly available block traces from SSD-based systems.

R/W ratios over time: The next question we look at is whether R/W ratios remain stable over time. Towards this end, we group SSDs into cohorts based on their age and monitor each cohort of drives over time. Each cohort spans a 6-month time frame (e.g., months 12–18, representing the first six months of the 2nd year in production). Note that there is no overlap between cohorts; for instance, if an SSD is placed into the cohort corresponding to total deployment time up to 18 months, it is not placed into any other cohort before that, even though at some point in its lifetime it had been deployed for that amount of time. For each cohort, we report its (median) R/W ratio at different points in time.

The results for R/W ratios over time are shown in Figure 10 (bottom). Each line segment in the graph corresponds to one of the cohorts described above. We make two observations:

• R/W ratios in AFF systems remain rather stable over time, suggesting that the characteristics of the corresponding workloads do not drastically change over time.

• The only time in a drive's lifetime when R/W ratios tend to change is towards their end of life. In particular, we see ratios increasing after around 4.5 years in production. This might likely be due to systems being drained before being retired.

R/W ratios and system capacity and fullness: Next, we explore whether R/W ratios look different based on system capacity and system fullness.

We find that systems with smaller capacities are associated with higher R/W ratios; the 50th and 90th percentiles of the R/W rates associated with smaller systems are up to 2× higher than those for larger systems.

When we examine how R/W ratios look like for different levels of fullness, we interestingly observe no significant differences in the R/W ratios among systems which use more than 25% of their total space, suggesting that systems which are more full do not necessarily experience read-dominant workloads only.

Comparison with data center drives: Three recent field studies on data center drives at Facebook, Microsoft, and Alibaba, which mainly focus on failure characteristics, also report some aggregate statistics on the read and write rates associated with these drives [33, 36, 46].

Figure 11 plots the physical NAND read and write rates for those data center drives (except for Alibaba drives which involve host reads/writes), as well as the (host) read and write rates of the SSDs in our enterprise storage systems (which do not involve any requests served from the DRAM cache); we have selected only those SSDs from our data set with a capacity comparable to the data center drives.

We make two observations:

• First, the workloads associated with the SSDs in our data set are significantly more intensive: the corresponding read and write rates are at least one order of magnitude higher than the ones in the other two studies (note the log scale on both axes). Keeping in mind that our rates involve host reads and writes, while those of the two data center studies report physical reads and writes, the actual differences are even larger.

• Second, in contrast to the drives at Facebook [33] and Alibaba [46], which report a comparable number of reads and writes, our systems see a larger number of reads than writes. Still, concerning drives at Facebook, the difference might be due to the fact the we report host writes while that study reports physical writes (which include WAF writes).

The R/W rates of Microsoft drives [36] look comparable to ours in Figure 11, however given that their write rates are physical NAND writes, while our write rates are host writes (not accounting for WAF), the R/W ratios of their applications are likely much higher than the average across our systems.

In summary, the read and write rates and the read/write ratios experienced by SSDs in enterprise storage systems vary significantly from those reported for data center drives, highlighting the differences (in terms of workload characteristics) between enterprise storage systems and data centers.

5.2 R/W ratios and Write-back cache systems

R/W ratios in WBC systems have a different significance than for AFF systems (where they mostly characterize the difference in reads and writes generated by applications running on the systems). SSDs in WBC systems are used as a cache layer that aims to increase performance, while persistent storage is provided by another layer consisting of HDDs. The R/W ratio of accesses to the SSDs can therefore be viewed as one measure of the effectiveness of the cache: the R/W ratio provides some indication of how many cache reads (hits) we get for one write to the cache[4].

As we observe in Figure 9, WBC systems experience higher R/W ratios than AFF systems. Specifically, the median R/W ratio across the entire population is 4.1:1 and the 95th percentile is 150:1 (4.4× higher than for AFF). The high R/W ratios associated with WBC systems suggest that the cache layer is used effectively.

We make some interesting observations regarding R/W ratios of WBC systems over time. Figure 10 (top) again fixes cohorts of drives of similar age and monitors them over time.

• When following an individual cohort of drives over time (i.e., one specific line segment in the graph), we observe a clear drop in R/W ratio over time, particularly in the first half of a drive's life. This indicates the cache is becoming less effective as a read cache over time, likely because the total amount of data stored on the system increases over time and as a result, the (fixed-size) cache can cache increasingly smaller fractions of the total data. In particular, towards the end of a drive's life R/W ratios are quite low, with only two reads for every write.

• We make another interesting observation when comparing the R/W ratios for different line segments against each other, in particular in areas of the x-axis where they overlap. More recently deployed systems tend to have higher R/W ratios, even when comparing them with older systems at the same age. This might either indicate a trend of workloads changing over time towards higher R/W ratios, or customers configuring their storage systems differently (with a larger cache size relative to the amount of data stored for more recent systems).

6 Conclusions

We briefly summarize the key findings of our study in Table 4.

7 Acknowledgements

We would like to acknowledge all the internal reviewers at NetApp whose feedback improved this paper a lot. We owe a debt of gratitude to NetApp's ActiveIQ team; Asha Gangolli, Kavitha Degavinti, and Vinaya Nagaraj, who generated the data sets we used for this paper. We also thank our FAST reviewers and our shepherd, Xiaosong Ma, for their detailed and valuable feedback. This work was supported by an NSERC Discovery Grant and an NSERC Canada Research Chair.

[4]Note that the R/W ratio is not exactly a cache hit rate, as we do not know how many reads bypass the cache and read straight from the HDD layer.

Most Important Findings
§3.1.2: The majority of SSDs in our data set consume PE cycles at a very slow rate. Our projections indicate that the vast majority of the population (~95%) could move toward QLC without wearing out prematurely.
§3.1, 3.2: The host write rates for SSDs used as caches are significantly higher than for SSDs used as persistent storage. Yet, they do not see higher NAND write rates as they also experience lower WAF. It is thus not necessarily required to use higher endurance drives for cache workloads (which is a common practice).
§3.2: WAF varies significantly (orders of magnitude) across drive families and manufacturers. We conclude that the degree to which a drive's firmware affects its WAF can be surprisingly high, compared to other factors also known to affect WAF.
§3.2: We identify as the main contributor to WAF, for those drive families with the highest WAF, the aggressive rewriting of blocks to avoid retention issues. This is surprising, as other maintenance tasks (e.g., garbage collection, wear-leveling) generally receive more attention; common flash simulators and emulators (e.g., FEMU) do not even model rewriting to avoid retention issues.
§3.2: The WAF of our drives is higher than values reported in various academic studies based on trace-driven simulation. This demonstrates that it is challenging to recreate the real-world complexities of SSD internals and workloads in simulation.
§3.3: Wear leveling is not perfect. For instance, 5% of all SSDs report an erase ratio above 6, i.e., there are blocks in the drive which will wear out six times as fast as the average block. This is a concern not only because of early wear-out, but also because those blocks are more likely to experience errors and error correction contributes to tail latencies.
§3.4: AFF systems are on average 43% full. System fullness increases faster during the first couple of years in production, and after that increases only slowly. Systems with the largest capacity are fuller than smaller systems.
§4.3, §4.4: We find that over-provisioning and fullness have little impact on WAF in practice, unlike commonly assumed.
§5: The vast majority of workloads (94%) associated with SSDs in our systems are read-dominant, with a median R/W ratio of 3.62:1, highlighting the differences in usage between SSD-based and HDD-based systems. Many widely-used traces from HDD-based systems see more writes than reads, raising concerns when using these traces for SSD research, as is common in practice.
§5: The read and write rates for the drives in our enterprise storage systems are an order of magnitude higher than those reported for data center drives (comparing same-capacity drives).
§5: The read/write ratio reported by SSDs that act as caches decreases significantly over their lifetime. This might indicate a decreasing effectiveness of the SSD cache over time.
§3.2, §5: The differences between some of our results and those reported based on the analysis of widely used HDD-based storage traces emphasize the importance for us as a community to bring some representative SSD-based traces into the public domain.

Table 4: *The most important findings per section.*

References

[1] Nitin Agrawal, William J Bolosky, John R Douceur, and Jacob R Lorch. A five-year study of file-system metadata. *ACM Transactions on Storage (TOS)*, 3(3):9–40, 2007.

[2] Jens Axboe. Flexible I/O tester. https://github.com/axboe/fio. Accessed: 2020-01-05.

[3] Lakshmi N. Bairavasundaram, Garth R. Goodson, Shankar Pasupathy, and Jiri 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)*, pages 289–300, 2007.

[4] Janki Bhimani, Jingpei Yang, Zhengyu Yang, Ningfang Mi, NHV Krishna Giri, Rajinikanth Pandurangan, Changho Choi, and Vijay Balakrishnan. Enhancing ssds with multi-stream: What? Why? How? In *2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC)*, pages 1–2. IEEE, 2017.

[5] Yu Cai, Saugata Ghose, Erich F Haratsch, Yixin Luo, and Onur Mutlu. Error characterization, mitigation, and recovery in flash-memory-based solid-state drives. *Proceedings of the IEEE*, 105(9):1666–1704, 2017.

[6] Yu Cai, Yixin Luo, Erich F Haratsch, Ken Mai, and Onur Mutlu. Data retention in MLC NAND flash memory: Characterization, optimization, and recovery. In *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, pages 551–563. IEEE, 2015.

[7] Chandranil Chakraborttii and Heiner Litz. Reducing write amplification in flash by death-time prediction of logical block addresses. In *Proceedings of the 14th ACM International Conference on Systems and Storage*, pages 1–12, 2021.

[8] Peter Corbett, Bob English, Atul Goel, Tomislav Grcanac, Steven Kleiman, James Leong, and Sunitha Sankar. Row-diagonal parity for double disk failure correction. In *Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST '05)*, pages 1–14. USENIX Association, 2004.

[9] Peter Desnoyers. Analytic modeling of SSD write performance. In *Proceedings of the 5th Annual International Systems and Storage Conference*, pages 1–10, 2012.

[10] Peter Desnoyers. Analytic models of SSD write performance. *ACM Transactions on Storage (TOS)*, 10(2):1–25, 2014.

[11] Atul Goel and Peter Corbett. RAID triple parity. *ACM SIGOPS Operating Systems Review*, 46(3):41–49, 2012.

[12] Laura M Grupp, Adrian M Caulfield, Joel Coburn, Steven Swanson, Eitan Yaakobi, Paul H Siegel, and Jack K Wolf. Characterizing flash memory: Anomalies, observations, and applications. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 24–33, 2009.

[13] Haryadi S Gunawi, Riza O Suminto, Russell Sears, Casey Golliher, Swaminathan Sundararaman, Xing Lin, Tim Emami, Weiguang Sheng, Nematollah Bidokhti, Caitie McCaffrey, Deepthi Srinivasan, Biswaranjan Panda, Andrew Baptist, Gary Grider, Parks M Fields, Kevin Harms, Robert B Ross, Andree Jacobson, Robert Ricci, Kirk Webb, Peter Alvaro, Birali H Runesha, Mingzhe Hao, and Huaicheng Li. Fail-slow at scale: Evidence of hardware performance faults in large production systems. *ACM Transactions on Storage (TOS)*, 14(3):1–26, 2018.

[14] Sangwook Shane Hahn, Sungjin Lee, and Jihong Kim. To collect or not to collect: Just-in-time garbage collection for high-performance SSDs with long lifetimes. In *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2015.

[15] Longzhe Han, Yeonseung Ryu, and Keunsoo Yim. CATA: A garbage collection scheme for flash memory file systems. In *International Conference on Ubiquitous Intelligence and Computing*, pages 103–112. Springer, 2006.

[16] Mingzhe Hao, Gokul Soundararajan, Deepak Kenchammana-Hosekote, Andrew A Chien, and Haryadi S Gunawi. The Tail at Store: A Revelation from Millions of Hours of Disk and SSD Deployments. In *Proceedings of the 14th USENIX conference on File and Storage Technologies (FAST '16)*, pages 263–276, 2016.

[17] Dave Hitz, James Lau, and Michael A Malcolm. File System Design for an NFS File Server Appliance. In *USENIX Winter*, volume 94, 1994.

[18] Jen-Wei Hsieh, Tei-Wei Kuo, and Li-Pin Chang. Efficient identification of hot data for flash memory storage systems. *ACM Transactions on Storage (TOS)*, 2(1):22–40, 2006.

[19] Xiao-Yu Hu, Evangelos Eleftheriou, Robert Haas, Ilias Iliadis, and Roman Pletka. Write amplification analysis in flash-based solid state drives. In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference*, pages 1–9, 2009.

[20] Jaeyong Jeong, Sangwook Shane Hahn, Sungjin Lee, and Jihong Kim. Lifetime improvement of NAND flash-based storage systems using dynamic program and erase scaling. In *12th USENIX Conference on File and Storage Technologies (FAST '14)*, pages 61–74, 2014.

[21] Jeong-Uk Kang, Jeeseok Hyun, Hyunjoo Maeng, and Sangyeun Cho. The multi-streamed solid-state drive. In *6th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage '14)*, 2014.

[22] Ram Kesavan, Matthew Curtis-Maury, Vinay Devadas, and Kesari Mishra. Countering Fragmentation in an Enterprise Storage System. *ACM Trans. Storage (TOS)*, 15(4), jan 2020.

[23] Ram Kesavan, Jason Hennessey, Richard Jernigan, Peter Macko, Keith A Smith, Daniel Tennant, and VR Bharadwaj. FlexGroup Volumes: A Distributed WAFL File System. In *2019 USENIX Annual Technical Conference (ATC '19)*, pages 135–148. USENIX Association, 2019.

[24] Ricardo Koller and Raju Rangaswami. I/O deduplication: Utilizing content similarity to improve I/O performance. *ACM Transactions on Storage (TOS)*, 6(3):1–26, 2010.

[25] Eunji Lee, Julie Kim, Hyokyung Bahn, Sunjin Lee, and Sam H Noh. Reducing write amplification of flash storage through cooperative data management with NVM. *ACM Transactions on Storage (TOS)*, 13(2):1–13, 2017.

[26] Huaicheng Li, Mingzhe Hao, Michael Hao Tong, Swaminathan Sundararaman, Matias Bjørling, and Haryadi S Gunawi. The CASE of FEMU: Cheap, accurate, scalable and extensible flash emulator. In *16th USENIX Conference on File and Storage Technologies (FAST '18)*, pages 83–90, 2018.

[27] Jinhong Li, Qiuping Wang, Patrick PC Lee, and Chao Shi. An In-Depth Analysis of Cloud Block Storage Workloads in Large-Scale Production. In *2020 IEEE International Symposium on Workload Characterization (IISWC)*, pages 37–47. IEEE, 2020.

[28] Peng Li, Wei Dang, Congmin Lyu, Min Xie, Quanyang Bao, Xiaofeng Ji, and Jianhua Zhou. Reliability Characterization and Failure Prediction of 3D TLC SSDs in Large-scale Storage Systems. *IEEE Transactions on Device and Materials Reliability*, 21(2):224–235, 2021.

[29] Shuwen Liang, Zhi Qiao, Jacob Hochstetler, Song Huang, Song Fu, Weisong Shi, Devesh Tiwari, Hsing-Bung Chen, Bradley Settlemyer, and David Montoya. Reliability characterization of solid state drives in a scalable production datacenter. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3341–3349. IEEE, 2018.

[30] Stathis Maneas, Kaveh Mahdaviani, Tim Emami, and Bianca 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)*, Santa Clara, CA, 2020. USENIX Association.

[31] Stathis Maneas, Kaveh Mahdaviani, Tim Emami, and Bianca Schroeder. Reliability of SSDs in Enterprise Storage Systems: A Large-Scale Field Study. *ACM Trans. Storage (TOS)*, 17(1), jan 2021.

[32] Chris Mellor. WD and Tosh talk up penta-level cell flash. https://blocksandfiles.com/2019/08/07/penta-level-cell-flash/. Accessed: 2021-01-04.

[33] Justin Meza, Qiang Wu, Sanjev Kumar, and Onur 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)*, pages 177–190, 2015.

[34] Micron. Comparing SSD and HDD Endurance in the Age of QLC SSDs. https://www.micron.com/-/media/client/global/documents/products/white-paper/5210_ssd_vs_hdd_endurance_white_paper.pdf. Accessed: 2021-01-03.

[35] Dushyanth Narayanan, Austin Donnelly, and Antony Rowstron. Write off-loading: Practical power management for enterprise storage. *ACM Transactions on Storage (TOS)*, 4(3):1–23, 2008.

[36] Iyswarya Narayanan, Di Wang, Myeongjae Jeon, Bikash Sharma, Laura Caulfield, Anand Sivasubramaniam, Ben Cutler, Jie Liu, Badriddine Khessib, and Kushagra Vaid. SSD Failures in Datacenters: What? When? And Why? In *Proceedings of the 9th ACM International on Systems and Storage Conference (SYSTOR '16)*, pages 7:1–7:11, 2016.

[37] NetApp Inc. Data ONTAP 9. https://docs.netapp.com/ontap-9/index.jsp. Accessed: 2020-10-12.

[38] Biswaranjan Panda, Deepthi Srinivasan, Huan Ke, Karan Gupta, Vinayak Khot, and Haryadi S Gunawi. IASO: A Fail-Slow Detection and Mitigation Framework for Distributed Storage Services. In *2019 USENIX Annual Technical Conference (ATC '19)*, pages 47–62. USENIX Association, 2019.

[39] David A Patterson, Garth Gibson, and Randy H Katz. *A case for redundant arrays of inexpensive disks (RAID)*, volume 17. ACM, 1988.

[40] Eunhee Rho, Kanchan Joshi, Seung-Uk Shin, Nitesh Jagadeesh Shetty, Jooyoung Hwang, Sangyeun Cho, Daniel DG Lee, and Jaeheon Jeong. FStream: Managing flash streams in the file system. In *16th USENIX Conference on File and Storage Technologies (FAST '18)*, pages 257–264, 2018.

[41] Bianca Schroeder, Raghav Lagisetty, and Arif Merchant. Flash Reliability in Production: The Expected and the Unexpected. In *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST '16)*, pages 67–80, Santa Clara, CA, 2016. USENIX Association.

[42] Mansour Shafaei, Peter Desnoyers, and Jim Fitzpatrick. Write amplification reduction in flash-based SSDs through extent-based temperature identification. In *8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage '16)*, 2016.

[43] Murray Stokely, Amaan Mehrabian, Christoph Albrecht, Francois Labelle, and Arif Merchant. Projecting disk usage based on historical trends in a cloud environment. In *Proceedings of the 3rd workshop on Scientific Cloud Computing*, pages 63–70, 2012.

[44] Shunzhuo Wang, You Zhou, Jiaona Zhou, Fei Wu, and Changsheng Xie. An Efficient Data Migration Scheme to Optimize Garbage Collection in SSDs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(3):430–443, 2020.

[45] Guanying Wu and Xubin He. Delta-FTL: improving SSD lifetime via exploiting content locality. In *Proceedings of the 7th ACM European Conference on Computer Systems (EuroSys '12)*, pages 253–266, 2012.

[46] Erci Xu, Mai Zheng, Feng Qin, Yikang Xu, and Jiesheng Wu. Lessons and Actions: What We Learned from 10K SSD-Related Storage System Failures. In *2019 USENIX Annual Technical Conference (ATC '19)*, pages 961–976, Renton, WA, 2019. USENIX Association.

[47] Jing Yang, Shuyi Pei, and Qing Yang. WARCIP: Write amplification reduction by clustering I/O pages. In *Proceedings of the 12th ACM International Conference on Systems and Storage*, pages 155–166, 2019.

[48] Pan Yang, Ni Xue, Yuqi Zhang, Yangxu Zhou, Li Sun, Wenwen Chen, Zhonggang Chen, Wei Xia, Junke Li, and Kihyoun Kwon. Reducing garbage collection overhead in SSD based on workload prediction. In *11th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 19)*, 2019.

[49] Qi Zhang, Xuandong Li, Linzhang Wang, Tian Zhang, Yi Wang, and Zili Shao. Lazy-RTGC: A real-time lazy garbage collection mechanism with jointly optimizing average and worst performance for NAND flash memory storage systems. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 20(3):1–32, 2015.