On the Impact of Garbage Collection on flash-based SSD Endurance

Robin Verschoren and Benny Van Houdt
Dept. Math and Computer Science
University of Antwerp, Belgium

Abstract

Garbage collection has a profound impact on the write amplification in flash-based SSDs, which in turn may significantly reduce its life span. The unequal wear of data blocks further contributes to this reduced life span. In this paper we study two performance measures: the *SSD endurance* which assesses the life span of an SSD and the *PE fairness* which is a measure for the degree of unequal wear.

We demonstrate, using a mean field model and simulation, how these measures are affected by the garbage collection algorithm, spare factor, etc. Numerical results indicate that under uniform random writes there is no need to implement a wear leveling technique. For hot and cold data we see that design choices that lower the PE fairness may still result in a higher SSD endurance, which suggests that one should not emphasize too much on equaling the wear.

1 Introduction

Two of the main difficulties faced by flash-based SSD designers are the inability to perform erase operations on a page level and the limited number of program-erase (PE) cycles that a block can tolerate [5]. While the former is addressed using out-of-place writes, the latter has led to the introduction of various wear leveling techniques to extend the life span of the drive [8]. There has been a lot of work on assessing the write amplification caused by out-of-place writes (e.g., [2, 4, 10, 11]), but far fewer studies exist that focus on the life span of an SSD.

In this paper we introduce two performance measures called the PE fairness and SSD endurance and study how they are affected by various system parameters as well as by the garbage collection (GC) algorithm. The PE fairness indicates to what extent blocks undergo the same number of PE cycles during the life span of the drive. The SSD endurance measures the number of full drive writes that can be performed before any block reaches its maximum number of PE cycles. The SSD endurance is thus a combination of the write amplification and PE fairness.

2 System description

In this paper we focus on an SSD with a page-mapped FTL that contains N physical blocks each holding b pages with a spare factor S_f (that is, with over-provisioning factor $1/(1-S_f)$). We assume the SSD operates using two special blocks, called write frontiers (WFs), to support out-of-place writes: a WF for writes requested by the host, termed the external WF (WFE), and a WF for writes performed by GC, termed the internal WF (WFI). The objective of supporting these two WFs is to achieve a form of data separation without the need to implement a hot/cold data identification technique.

The GC algorithm is invoked whenever the WFE becomes full. Assume that the last $b-j^*$ pages of the WFI are in the erase state when the GC algorithm is invoked, while the first j^* are in the valid/invalid state. Further assume the victim block selected by the GC algorithm contains j valid pages. Consider the following 2 cases:

1. If $j \leq b-j^*$, the j valid pages of the victim block are simply copied to the WFI leaving the last $b-j^*-j$ pages in the erase state. After copying the j valid pages to the WFI, the victim block is erased and becomes the new WFE. Hence, the next b host writes make use of the WFE before the GC algorithm is invoked again.

2. If $j > b-j^*$, $b-j^*$ of the j valid pages are copied to the WFI. The remaining $j-(b-j^*)$ valid pages are copied to RAM and back to the victim block after the victim block has been erased. In this case, the victim block becomes the new WFI and the GC algorithm is immediately invoked again.

We mainly focus on the set of d-choices GC algorithms [10, 11, 12], where $d \geq 1$ is an integer. Under d-choices GC the victim block is selected as follows: d blocks are chosen uniformly at random and the one containing the least number of valid pages among the d chosen blocks becomes the victim block (ties are broken arbitrarily). When $d = 1$ we obtain the Random GC algorithm, while setting d equal to the number of blocks on the SSD results in the Greedy GC algorithm [2, 6, 4]. Under uniform random writes the Greedy GC is known to minimize the write amplification [13], while in case of hot/cold data there exists an optimal finite d when minimizing the write amplification [11].

Note the d-choices GC algorithm does not exploit any information that may be maintained by a potential wear leveling mechanism. Hence, the system under consideration does not rely on any form of wear leveling. One of the questions we do intend to answer is how much room there is left for any wear leveling mechanism to further improve the endurance of the system.

3 Performance measures

In this section we introduce the two main performance measures studied in this paper, for completeness we revisit the well-known write amplification (WA) first. The WA is equal to the ratio between the total number of writes performed on the drive divided by the number of writes requested by the host system. To be mathematically precise, let X_j be the random variable denoting the number of valid pages on the victim block selected during the j-th GC call, then the write amplification *up to the time of the n-th GC call* can be expressed as

$$WA(n) = \frac{bn}{\sum_{j=1}^{n} \sum_{i=0}^{b} (b-i) P[X_j = i]},$$

as selecting a victim block with i valid pages leaves room for $b-i$ writes by the host. Note when talking about the WA one typically refers to $\lim_{n\to\infty} WA(n)$.

To define the first performance measure, called the *PE fairness (PE_f)*, let W_{max} represent the maximum number of PE cycles that a block can tolerate (for simplicity we assume this is a fixed number). The PE fairness is defined as the mean number of PE cycles performed on a block before any block reaches W_{max} PE cycles divided by W_{max}. More formally, let Y_k denote a random variable representing the number of times the GC algorithm is invoked before any block is erased for the k-th time, then the PE fairness is given by

$$PE_f(W_{max}) = \sum_{n\geq 1} P[Y_{W_{max}} = n] \frac{n/N}{W_{max}} = \frac{E[Y_{W_{max}}]}{W_{max} N},$$

as after n GC calls the mean number of PE cycles performed on a block part of a set of N blocks equals n/N. If the PE fairness is close to one, it is clear that there is little to no use in implementing a wear leveling technique. We believe this is an easier to interpret measure for the fairness than Jain's fairness index proposed in [7].

The second measure of interest, termed the *SSD endurance (SSD_e)*, is a measure for the expected total number of host writes performed before any block reaches the predefined maximum number W_{max} of PE cycles. Hence,

$$SSD_e(W_{max}) = \frac{E[\sum_{j=1}^{Y_{W_{max}}} \sum_{i=0}^{b} (b-i) P[X_j = i]]}{bN}.$$

Note the unit used to express the SSD endurance is the total number of Full Drive Writes (FDWs). The SSD endurance is roughly equal to W_{max} times the PE fairness divided by the WA. While it is attractive to have a PE fairness close to one, the main reason for striving for high fairness exists in prolonging the life span of the drive, which is captured by the SSD endurance. Hence, having a PE fairness close to one is nice, but in the end only the SSD endurance truly matters. Finally, we note that this definition of the endurance is the same as the one used in [1] for USB flash drives, which does not take NAND data refresh operations into account that may be needed to guarantee data retention [3].

4 Uniform random writes

In order to study the impact of a number of system parameters on the PE fairness and SSD endurance when subject to uniform random writes, we can extend the mean field model of [10] that analyzed the write amplification only. The generalization exists in setting up drift equations for $m_{i,w}$ which represents the fraction of the total number of blocks holding i valid pages on which exactly w erase operations have been performed.

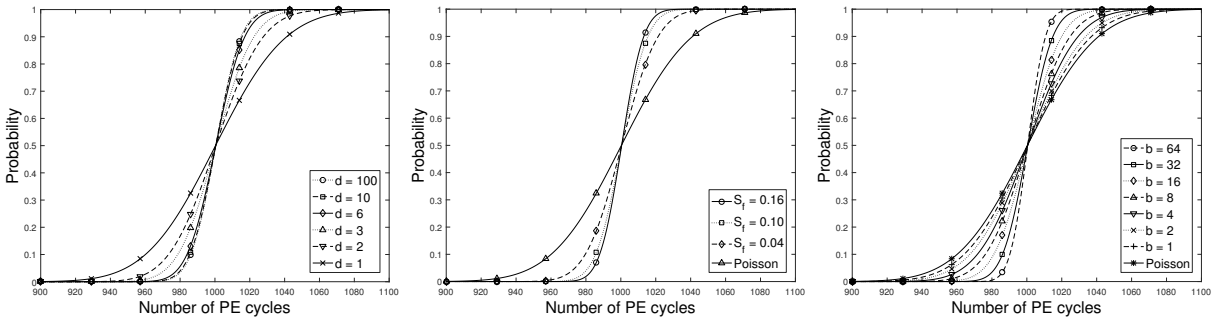
Let $p_{i,w}(\vec{m})$ be the probability that the GC algorithm selects a block with i valid pages that has been erased w times. For instance, if we use the d-choices GC algorithm which does not take the number of PE cycles that occurred into account, we have that $p_{i,w}(\vec{m})$ equals

$$\frac{m_{i,w}}{m_i} \left[\left(\sum_{\ell=i}^{b} \sum_{w\geq 0} m_{\ell,w} \right)^d - \left(\sum_{\ell=i+1}^{b} \sum_{w\geq 0} m_{\ell,w} \right)^d \right]$$

for $m_i = \sum_{w\geq 0} m_{i,w} > 0$ and $p_{i,w}(\vec{m}) = 0$ for $m_i = 0$. Let $p_i(\vec{m}) = \sum_{w\geq 0} p_{i,w}(\vec{m})$.

Without going into detail, the set of ODEs for the generalized mean field model is given by $\frac{d}{dt} m_{i,w}(t) = f_{i,w}(\vec{m}(t))$, where $f_{i,w}(\vec{m})$ equals

$$\frac{(i+1)m_{i+1,w} - i m_{i,w}}{b\rho} \left(\sum_{j=1}^{b} p_{b-j}(\vec{m}) j \right) - p_{i,w}(\vec{m}). \quad (1)$$



(a) $b = 32$, $S_f = 0.1$. (b) $b = 32$, $d = 10$. (c) $d = 100$, $S_f = 0.1$.

Figure 1: The distribution of the number of PE cycles performed on a block under uniform random writes at time $t = 1000$.

for $i < b$ and $f_{b,w}(\vec{m})$ equals

$$\sum_{i=0}^{b} p_{i,w-1}(\vec{m}) - \frac{b m_{b,w}}{b\rho}\left(\sum_{j=1}^{b} p_{b-j}(\vec{m})\, j\right) - p_{b,w}(\vec{m}), \quad (2)$$

where $p_{i,-1}(\vec{m})$ is defined as 0. While the mean field model introduced in [10] considers a system using a single WF, it is not hard to see that the performance of such a system is identical to the 2 WF setting of Section 2 in case of uniform random writes.

Using simulation experiments we found that this model produces highly accurate results for the PE fairness and SSD endurance for systems consisting of several thousand blocks (similar to [10] for the WA), but we do not present these results here due to a lack of space.

Distribution of PE cycles: By relying on (1) and (2) we can determine the distribution of the number of PE cycles after Nt GC calls by numerically solving the ODE up to time t starting with $\sum_{i=0}^{b} m_{i,0}(0) = 1$. The results are depicted in Figure 1 and clearly indicate that the distribution of the number of PE cycles becomes less variable as the number of choices d increases, as the spare factor S_f increases and as the number of pages per block b increases. Thus, the Random GC algorithm (i.e., $d = 1$) performs the worst and the Greedy GC algorithm (i.e., d large) performs best both in terms of the write amplification and PE fairness. Note when we state that the Greedy algorithm has the best PE fairness, we mean within the class of d-choices GC algorithms as FIFO clearly has the best possible overall PE fairness.

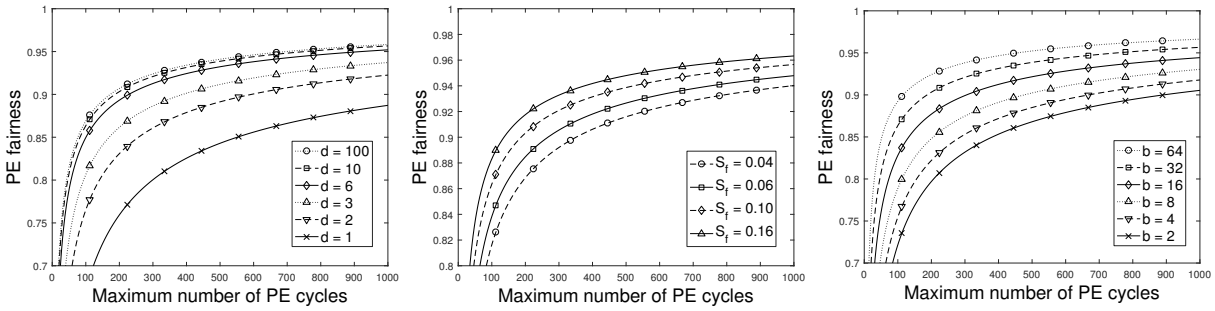
When $d = 1$ it is easy to check that $m_w(t) = \sum_{i=0}^{b} m_{i,w}(t) = t^w e^{-t}/w!$. Hence, the distribution of the number of PE cycles on a block after Nt GC calls converges to a Poisson distribution with parameter t as N tends to infinity (as expected).

PE fairness: To determine the PE fairness via the set of ODEs specified by (1) and (2), we numerically solve the ODE starting with $\sum_{i=0}^{b} m_{i,0}(0) = 1$ up to time t_{max}, where t_{max} is the smallest t such that $\sum_{w \geq W_{max}} \sum_{i=0}^{b} m_{i,w}(t) > 1/N$. The PE fairness is found as t_{max}/W_{max}.

Figure 2 shows the PE fairness as a function of the maximum number of PE cycles W_{max} that a single block can tolerate. It indicates that increasing the number of choices d, number of pages per block b or the spare factor S_f results in an increase in the PE fairness. Also note that under uniform random writes one often observes a PE fairness above 0.95, even when the maximum number of PE cycles is as low as 1000. In other words, by the time that any block reaches 1000 PE cycles the average number of PE cycles that an arbitrary block has endured is above 950. This implies that under uniform random writes there is hardly any room left to improve the PE fairness by implementing some form of wear leveling.

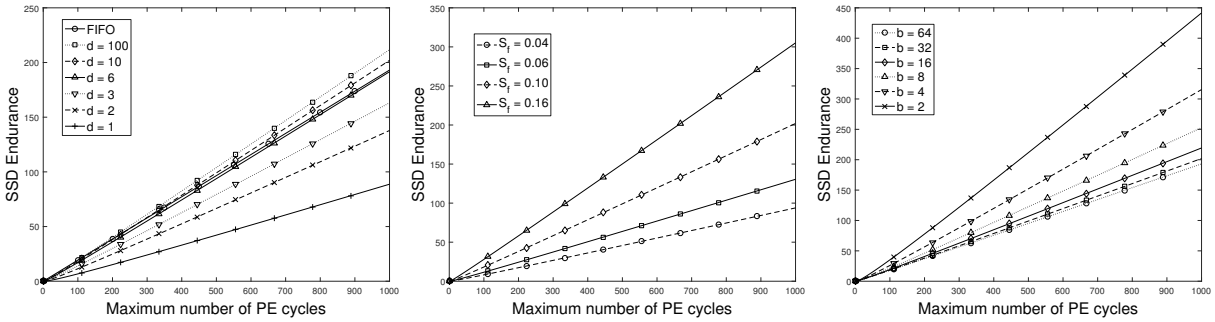
SSD endurance: Figure 3 depicts the SSD endurance in terms of the maximum number of PE cycles W_{max} that a single block can tolerate. The SSD endurance improves as the number of choices d or the spare factor S_f increases, which is in line with the previous results as larger d and S_f values improve the PE fairness and result in a lower write amplification. With respect to the impact of the number of pages b per block, we observe that lower b values result in a higher SSD endurance. The reason is that larger b values cause a higher write amplification, which outweighs the improvement in the PE fairness. Note that while the PE fairness is often above 0.9 the number of FDWs is well below W_{max} due to the (unavoidable) high WA under uniform random writes.

Figure 3 also depicts the SSD endurance of the FIFO GC algorithm. While the FIFO GC has the best PE fairness, its SSD endurance is below that of the Greedy algorithm (i.e., d-choices with d large) due to its somewhat



(a) $b = 32$, $S_f = 0.1$. (b) $b = 32$, $d = 10$. (c) $d = 100$, $S_f = 0.1$.

Figure 2: PE fairness under uniform random writes ($N = 10^4$).



(a) $b = 32$, $S_f = 0.1$. (b) $b = 32$, $d = 10$. (c) $d = 10$, $S_f = 0.1$.

Figure 3: SSD endurance under uniform random writes ($N = 10^4$).

higher WA.

5 Workloads with hot and cold data

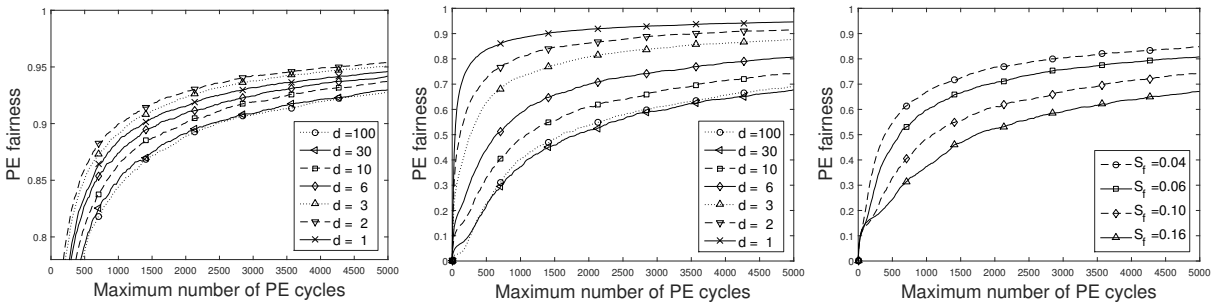
In order to be able to study the impact of data hotness in a structured manner, we focus on synthetic workloads of the Rosenblum type [9]. More specifically, we assume we have two types of logical pages: hot and cold pages. A fraction f of the logical pages is hot and a write request updates a hot (cold) page with probability r $(1-r)$.

Although it is possible to extend the mean field model in [11] in a manner similar to the uniform random writes case, the computation times needed to numerically solve the set of ODEs becomes problematic and we therefore rely on simulations only. All presented simulation results are for a drive consisting of $N = 10,000$ blocks and are averaged over 50 runs.

Before discussing the results, note that (partially) separating hot and cold data has both a positive and negative impact on the SSD endurance. It is well known that data separation results in a lower WA (e.g., [4, 11, 12]), but at the same time the PE fairness may worsen as the blocks holding (mostly) hot data may be subject to more PE cycles than blocks containing (mostly) cold data. Hence the main question is which of these two opposing forces dominates and to what extent does this depend on the GC algorithm.

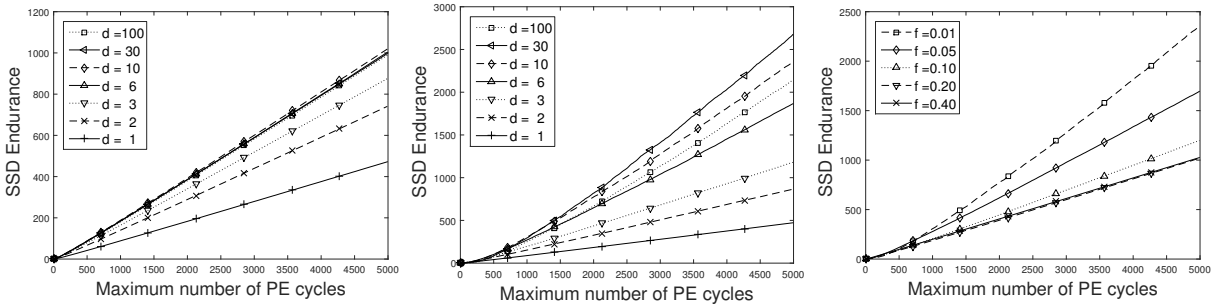
PE fairness: Figure 4 depicts the impact of d and S_f on the PE fairness in the presence of hot and cold data. Figures 2a, 4a and 4b confirm that increasing data hotness leads to a lower PE fairness. The values for the PE fairness also indicate that in case of hot and cold data some form of wear leveling may help to prolong the SSD life span. These figures also show that while large d values gave rise to a better PE fairness under uniform random writes, the reverse happens in case of hot and cold data. This can be understood by noting that when the GC algorithm selects a new WFE, small d values often result in the selection of a block that previously stored (mostly) cold data. The WFE on the other hand will mainly contain hot data when full (on average the WFE contains rb hot and $(1-r)b$ cold pages). Hence, for d small the hot data is less likely to use the same set of blocks for a long period of time leading to a better PE fairness.

Figures 2b and 4c indicate that the impact of the spare factor S_f also changes when we introduce data hotness: with hot and cold data smaller spare factors result in a better PE fairness. This is probably due to the fact that the GC algorithm is less effective in selecting a victim



(a) $S_f = 0.1$, $r = 0.8$, $f = 0.2$. (b) $S_f = 0.1$, $r = 0.99$, $f = 0.01$. (c) $d = 10$, $r = 0.99$, $f = 0.01$.

Figure 4: PE fairness under hot and cold data for $b = 32$ pages per block ($N = 10^4$)



(a) $r = 0.8$, $f = 0.2$. (b) $r = 0.99$, $f = 0.01$. (c) $d = 10, r = 1 - f$.

Figure 5: SSD endurance under hot and cold data for $b = 32$ pages per block and spare factor $S_f = 0.1$ ($N = 10^4$)

block that previously stored mostly hot data when the spare factor is small.

SSD endurance: While it is desirable to have a PE fairness close to one, the main reason for striving for an equal wear lies in improving the SSD endurance. Figure 5 shows the SSD endurance for various d values and hotness values. We first note that while setting d small (e.g., $d \leq 3$) resulted in a higher PE fairness, this is typically not a good choice for the SSD endurance as the WA for very small d is much higher than for large d and this outweighs the better PE fairness. Further, as with the WA (see [11]) there is an optimal finite choice for d for the SSD endurance in case of hot and cold data. For instance, for the case presented in Figure 5a setting $d = 13$ (not shown) minimizes the WA.

More importantly, while Figure 4 showed that the PE fairness reduces significantly when the hot data becomes hotter, Figure 5c clearly indicates that making the hot data hotter is typically beneficial for the SSD endurance. This observation suggests that selecting a GC algorithm that minimizes the WA may lead to a much more profound improvement in the SSD endurance compared to selecting a GC algorithm that puts too much emphasis on the PE fairness, that is, on achieving a more or less equal wear on all blocks.

We do note that as the WA approaches one (it is 1.575 for $f = 0.01$ in Figure 5c) the PE fairness becomes the dominating factor in the SSD endurance. Thus, GC algorithms that do take the wear into account will further increase the SSD endurance as long as the WA is kept equally low.

6 Conclusions and future work

In this paper we introduced the PE fairness and SSD endurance performance measures and studied how these are affected by the GC algorithm, the spare factor, etc. We indicated that under uniform random writes the Greedy GC algorithm has a near optimal SSD endurance as it is known to be optimal with respect to the WA and has a PE fairness close to one. In case of hot and cold data the PE fairness may be well below one, however a lower PE fairness may still result in a higher SSD endurance as the WA tends to have a more profound impact on the SSD endurance.

For future work we intend to look at the impact of data separation techniques on the results presented in this paper. Further, while wear leveling techniques clearly improve the PE fairness, the question remains whether they can significantly improve the SSD endurance as wear leveling typically comes at the cost of an increased WA.

References

[1] BOBOILA, S., AND DESNOYERS, P. Write endurance in flash drives: Measurements and analysis. In *Proceedings of the 8th USENIX Conference on File and Storage Technologies* (Berkeley, CA, USA, 2010), FAST'10, USENIX Association, pp. 9–9.

[2] BUX, W., AND ILIADIS, I. Performance of greedy garbage collection in flash-based solid-state drives. *Perform. Eval. 67*, 11 (Nov. 2010), 1172–1186.

[3] DESNOYERS, P. What systems researchers need to know about nand flash. In *Presented as part of the 5th USENIX Workshop on Hot Topics in Storage and File Systems* (Berkeley, CA, 2013), USENIX.

[4] DESNOYERS, P. Analytic models of SSD write performance. *ACM Trans. Storage 10*, 2 (Mar. 2014), 8:1–8:25.

[5] GRUPP, L. M., DAVIS, J. D., AND SWANSON, S. The bleak future of NAND flash memory. In *Proc. of USENIX Conference on File and Storage Technologies* (2012).

[6] ILIADIS, I. Rectifying pitfalls in the performance evaluation of flash solid-state drives. *Perform. Eval. 79* (2014), 235 – 257. Special Issue: Performance 2014.

[7] LI, Y., LEE, P., AND LUI, J. Stochastic modeling of large-scale solid-state storage systems: Analysis, design tradeoffs and optimization. *ACM SIGMETRICS Perform. Eval. Rev. 41*, 1 (2013), 179–190.

[8] MURUGAN, M., AND DU., D. Rejuvenator: A static wear leveling algorithm for NAND flash memory with minimized overhead. In *Proc. of IEEE MSST* (2011).

[9] ROSENBLUM, M., AND OUSTERHOUT, J. K. The design and implementation of a log-structured file system. *ACM Trans. Comput. Syst. 10*, 1 (Feb. 1992), 26–52.

[10] VAN HOUDT, B. A mean field model for a class of garbage collection algorithms in flash-based solid state drives. *ACM SIGMETRICS Perform. Eval. Rev. 41*, 1 (2013), 191–202.

[11] VAN HOUDT, B. Performance of garbage collection algorithms for flash-based solid state drives with hot/cold data. *Perform. Eval. 70*, 10 (2013), 692–703.

[12] VAN HOUDT, B. On the necessity of hot and cold data identification to reduce the write amplification in flash-based SSDs. *Performance Evaluation 82* (2014), 1 – 14.

[13] YANG, Y., MISRA, V., AND RUBENSTEIN, D. On the optimality of greedy garbage collection for SSDs. *ACM SIGMETRICS Perform. Eval. Rev. 43*, 2 (Sept. 2015), 63–65.