Analysis of write amplification on solid state drives in datacenters

Slađana Đurašević
Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0002-3598-5781

Uroš Pešović
Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0001-8722-6544

Vanja Luković
Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0002-1887-6102

Borislav Đorđević
Institute Mihajlo Pupin
Belgrade, Serbia
ORCID: 0000-0002-6145-4490

***Abstract* — Solid state drives (SSD) becomes the dominant type of secondary memory in personal computer systems, thanks to their superior read and write performance when compared to traditional hard disk drives (HDD). Due to the limitations of NAND flash memory technology SSD has several disadvantages that make them unfavorable for permanent data storage in data centers. One of the critical factors in SSD operation is write amplification which introduces additional system writes that must be performed for each user write, thus increasing cell wear and reducing drive performance. In this paper, we analyzed the write amplification of SSD drives operating in the Backblaze data center.**

Keywords—Solid state drive, NAND FLASH, SMART, Write amplification

I. Introduction

Data storage becomes one of the fastest rising filed in computing due to the exponential increase in the amount of data produced by almost every field of human activity. Starting in 2012, mankind entered so-called the Zettabyte Era, the period when the amount of digital data in the world first exceeded one zettabyte (ZB). According to the Digital Age 2025 study, it is expected that by the end of 2025 global amount of data will exceed 175 ZB. Digital data can be stored on different storage mediums such as hard disk drives, magnetic tape, optical medium, or semiconductor flash medium. Today most dominant mediums for data storage are hard disk drives and flash, which constitute almost 85% of global storage share, 60% share for hard disk drives and 25% share for flash storage. It is expected that by 2025 share of the flash storage will continue to increase to 35% while the share of HDD will decrease to 50% [1].

Hard disk drives are used for computer storage for more than 65 years thanks to the constant evolution of HDD technology in improving the storage density. The basic principle of HDD electromechanical design has not changed thus it represents a major drawback in further increasing the HDD performance in terms of read and write speeds. Flash technology appeared in the late eighties and was primarily used for storing small quantities of data in embedded systems. Thanks to the improvements in miniaturization, flash technology became competitive with hard disk drives in terms of data density and read and write performance. SSDs are capable to achieve much faster read and write speeds when compared to HDDs, due to the lack of moving parts thus SSDs are becoming the dominant type of secondary memory in personal computers. It is to be expected that SSDs are more durable than hard drives, because they do not have mechanical parts, but it is not the case. Although they seem to be a superior solution for the average user compared to hard drives, SSDs have several disadvantages that make them unfavorable for permanent data storage in data centers that rely on HDDs.

The main SSD disadvantage is write mechanism which requires the movement of data to new memory cells which results in a phenomenon known as write amplification. Due to the limited write endurance of the flash medium, the occurrence of write amplification can significantly reduce SSD lifetime. The main design challenge in SSDs is to ensure equal medium wear while minimizing write amplification. In this paper, we will analyze the write amplification of SSDs operating in the Backblaze [2] data center. The primary factor we observe was the average erase count, which expectedly increased with the amount of data written to drive.

II. NAND FLASH TECHNOLOGY

The essence of the data storage in the flash cell is represented by the amount of electric charge, which is trapped between the layers of insulator inside the cell. Based on the number of discrete charge levels used, several types of flash cells can be used as shown in Fig 1. In the case of the SLC (Single Level Cell), a cell single binary bit is represented as the charged state for logical zero or the uncharged state as a logical one [3].

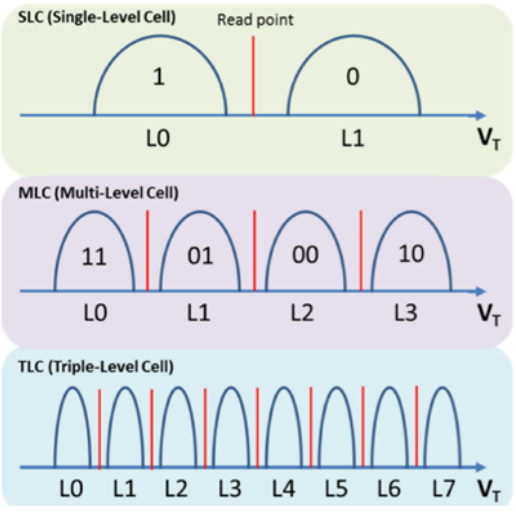

Fig. 1. FLASH cell technologies

This type of cell has low data density per area and could not compete with HDD which provides comparable data densities at fraction of the cost. Therefore, more advanced types of flash cells are in use, such as MLC (Multi-Level Cell) cells that store two bits using four charge levels or TLC (Triple Level Cell) cells that store three bits using eight charge levels. Furthermore, these technologies rely on vertical stacking of flash cells, known as 3D NAND which significantly increases storage density. Currently, QLC (Quad-Level Cell) flash cells, which store 4-bit value per cell, and PLC (Penta-Level Cell) flash cells, which store 5-bit value per cell, are being developed.

Placing more bits per one flash cell increases the capacity of the SSD without increasing the chip size, but also reduces storage reliability due to small differences between charge levels which makes it difficult to write and read bits. Also, a flash memory cell can only be written a certain number of times during its lifetime, with SLC cells allowing 100,000 erase cycles, MLC cells allowing 10,000 erase cycles, and TLC cells about 3,000 erase cycles [4]. Newer QLC and PLC cells allow just form 100 to 1000 erase cycles. Depending on the type of memory cells used in the SSD, the performance and reliability (lifespan) of the SSD are determined. Unlike a hard disk where data can be written to any location at any time and data can be easily overwritten, flash cells in an SSD must first be erased before new content can be written to them. Erase process discharges the flash cells to be able to store new data. Since flash cells have a limited number of erase cycles, it is necessary to ensure that data is cyclically written to all cells to wear them evenly. Flash controller maps logical data addresses to physical addresses on a flash using an FTL (Flash Translation Layer) table, in a process known as wear leveling. When writing data in a logical location, the physical location where the data was located will be erased, and the controller will write the data in a new physical location that has fewer erase cycles than the previous one and will perform mapping of the logical address in the FTL table to a new physical address.

Due to the organization of NAND flash memory, it is not possible to access cells individually to perform read and write operations. A page is the smallest unit of data that can be read or written to memory. The write operation can only be performed on pages that have been previously erased. In case it is necessary to change some data on the page, the content of the page is read to the internal register, the data is updated in register, and the updated version of the page is stored on a free page and this operation is called "read-modify-write". Unlike HDDs, the data is not updated on the spot, because the free page is at a different address from the page that originally contained the data, to ensure even wear. When data is saved on a new page, the original page is marked as obsolete and will remain so until it is deleted. Since an individual page cannot be deleted, the entire block to which the page belongs must be deleted, and the other pages with the correct data are moved to a new free block. By copying data from one block to another block, one lifecycle of that page or block is used [5]. This operation, called the garbage collection, is performed in the background by the SSD controller and is shown in Fig 2. Garbage collection is of great importance for SSDs, as it allows the drive to mitigate the impact of the cycle of erasing and writing data to flash memory. In that case, the SSD controller will have to perform additional data migration when the user wants to write new data, which results in additional data writing to flash memory, which leads to greater memory wear and shortens the life of the SSD. A measure of this phenomenon is defined by WA (Write Amplification) parameter which represents the ratio between an actual amount of data written to the flash memory and amount of data to be written issued by host [6].

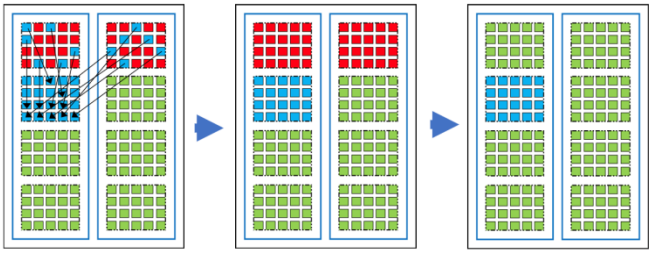

Fig. 2. NAND erase process: left - copy/move valid data (blue pages), middle - invalid blocks (red blocks), right erased block (green blocks).

With a blank SSD, all pages in blocks are free and data can be written to them immediately, so in this case the WA parameter is equal to one, which is the ideal value. However, as the SSD fills, there are fewer and fewer blank pages, and the controller must free up space in blocks, which affects the appearance of increased writing. The higher the WA parameter, the more additional data will be written to flash memory leading to increased flash memory wear, which will reduce SSD life. Also, movement of additional data will take up the bandwidth of flash memory, which affects the performance of the SSD.

The main design challenge in SSDs is focused on efficient flash controllers, which can improve SSD performance and extend the service life. There are many ways how flash controller can reduce WA, such as data compression, over-provisioning and use of sequential writing. Data compression requires high performance hardware, while over-provisioning limits the available free space.

III. Results

The SSD disks analyzed in this paper were used as system disks on the Backblaze data storage servers to store the server operating system. Storage servers contain up to 60 HHDs that permanently store user data, while the one SSD. is used to boot the storage server operating system, and to store log files and temporary files produced by the storage server. Each day, the SSD will read, write and delete files depending on the activity of the storage server itself. SSDs began to be used in the Backblaze data center from 2018, and by the end of 2021, the number of SSDs reached 2,200 on data warehouse servers. In this paper, we analyze two models of SSDs from Seagate, which are the most common in the data center ZA2000CM10002 and ZA250CM10003.

TABLE I. SDD MODELS USED IN RESARCH

Model	ZA250CM10002	ZA250CM10003
Capacity	250 GB	250 GB
No. drives	562	1090
Average age (months)	21.7	11.1
Operating days	204 287	276 281
Failed drives	2	8
Annular failure rate	0.36 %	1.06 %

The Backblaze records S.M.A.R.T. attributes for each HDD and SSD every day and publishes this data quarterly in form of the open data set [5]. S.M.A.R.T. is an abbreviation for Self-Monitoring, Analysis, and Reporting Technology and was developed to report on various indicators related to the reliability of HDD drives to predict failures. This technology is also applied to SSDs, but due to differences in technology, different parameters are available. The monitored parameters of the SSD include, operating time and number of power-on cycles, temperature, the degree of wear of the cells, the total amount of data read and written to SSD. The SMART parameters available for monitoring vary from manufacturer to manufacturer, so in addition to the above, there may be parameters that are specific to a particular disk model.

The Backblaze data set of SMART parameters is published quarterly in the form of csv files for each calendar day. Each line of the csv file contains a set SMART parameter for a specific model of SSD or HDD and disk serial number. This data was imported into the SQL database using SQLite3 after which the several SMART parameters for the target disk model were extracted using the appropriate SQL query.

The first parameter, SMART 241, is the total amount of data written by the host in GB and it represents the load under which SSD is subjected during its lifetime. In the first part of the research, the influence of host writing on the number of erasure cycles for both drive models is being analyzed. SMART 173 parameter contains an average, minimum and maximum count of erased SSD blocks for that particular day. Using this parameter, it is possible to predict the remaining life of the disk where the manufacturer specifies a lifetime of at least 1500 erase cycles. Obtained results for SSD model ZA250CM10002, are shown in Fig 3. as gray lines for every drive model, where the mean value for the number of erase cycles for all drives is shown by the blue solid line, while blue dashed lines represent one standard deviation from the mean value.

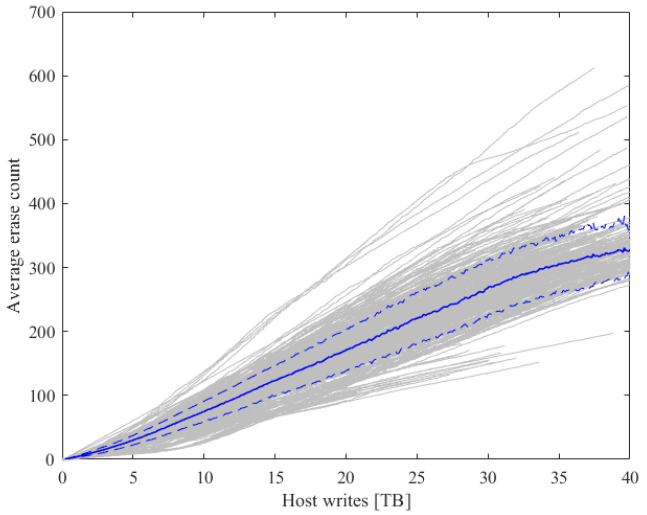

Fig. 3. Average erase count for ZA250CM10002 SSD

Results show that the average erase count is equal to 300, after 40 TB of data are written to SSD, this drive model has a wear level of around 20%. Given that these drives on average operate for almost two years, their useful life will be almost ten years. Obtained results for SSD model ZA250CM10003, are shown in Fig 4. and this drive has a much higher number of erased cycles from the previous dive model, 600 erased cycles after 30 TB of written data. This reveals that the wear level for drive model ZA250CM10003 is equal to 40% and since these drives on average operate for one year, their useful life will be just two and a half years, four times shorter than for previous model ZA250CM10002.

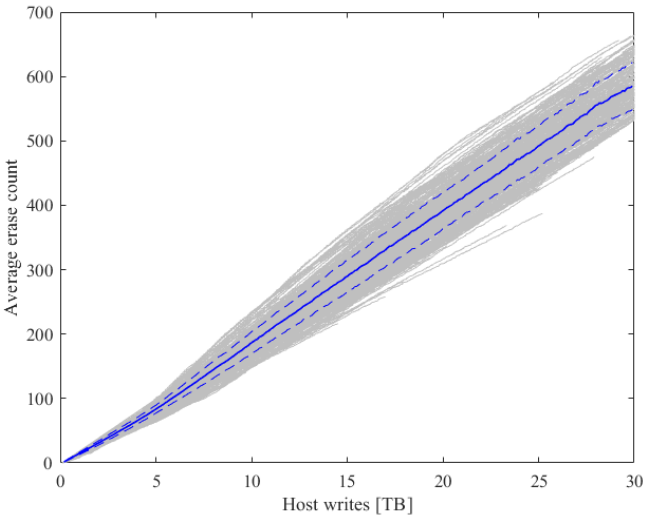

Fig. 4. Average erase count for ZA250CM10003 SSD

The SMART 233 parameter represents the total amount of data written to disk in GB. The ratio between this parameter and total host writes represents the WA parameter as shown by Eq. 1.

$$WA = \frac{SMART\ 233}{\text{SMART } 241} \tag{1}$$

In the second part of research influence of host writes on WA is analyzed for both drive models. Results are shown in Fig 7. and Fig 8. as gray lines for every drive model, where the mean value for the number of erase cycles for all drives is shown by the blue solid line, while blue dashed lines represent one standard deviation from the mean value. In case of the ZA250CM10002 model, result shown in Fig 5. suggest that this drive might use compression since some of the drives have WA factor lower than one. Later WA factor increases to a stable value which is less than two.

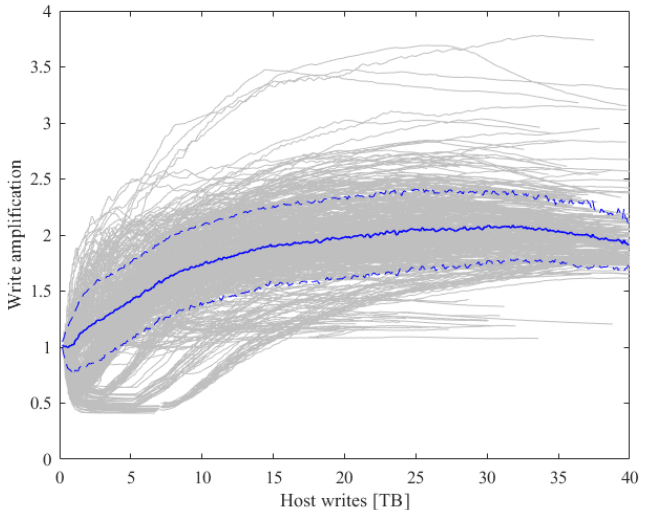

Fig. 5. Write amplification for ZA250CM10002 SSD

In case of the ZA250CM10003 model, result shown in Fig 6. suggest that this drive does not use compression like previous drive model and WA factor increases to a stable value which is around two and a half.

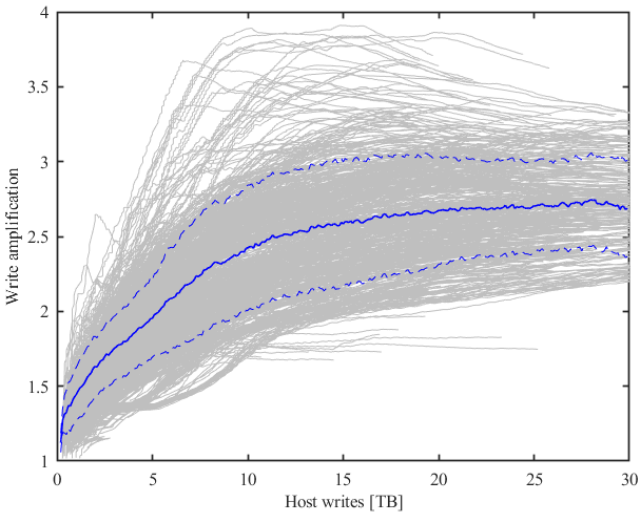

Fig. 6. Write amplification for ZA250CM10003 SSD

IV. CONCLUSION

In this research, we performed the analysis of the write amplification on the SSD drive operating in the Backblaze data center. The result showed that for two similar SSD models from the same manufacturer there is a significant difference in write amplification which suggest that the latter SSD model will probably reach the end of its service life much sooner than the older model. The main conclusion is that the drive ZA250CM10002 might use compression since some of the drives have WA lower than one at the beginning of their lifetime. Further research will be focused on the analysis of data which will be published by the Backblaze in following year to check these claims.

ACKNOWLEDGMENT

The research in this paper is part of the project 451-03-68/2022-14/200132 funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] D. Reinsel, J. Gantz, J. Rydning, "Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big", IDC White Paper sponsored by Seagate, 2017, Framingham, MA, USA

[2] A. Klein,The SSD Edition: 2021 Drive Stats, https://www.backblaze.com/blog/ssd-edition-2021-drive-stats-review/, accessed 10th April 2022.

[3] Y. Cai, S. Ghose, E. Haratsch, Y. Luo, O. Mutlu, Errors in Flash-Memory-Based Solid-State Drives: Analysis, Mitigation, and Recovery, DOI: https://doi.org/10.48550/arXiv.1711.11427

[4] M. K. Jibbe, B Chan, Analysis of SSD health and Prediction of SSD life, Storage Developer Conference, September 21, 2016, Santa Clara, USA

[5] F. Chen, D. Koufaty, X. Zhang, Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives, SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.

[6] X. Hu, E. Eleftheriou, R. Haas, I. Iliadis, R. Pletka, Write Amplification Analysis in Flash-Based Solid State Drives, SYSTOR '09: Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, Haifa, ISrael, May 2009, pp, 1–9, https://doi.org/10.1145/1534530.1534544

[7] BackBlaze SMART data set, https://www.backblaze.com/b2/hard-drive-test-data.html, accessed 10th April 2022.