White Paper

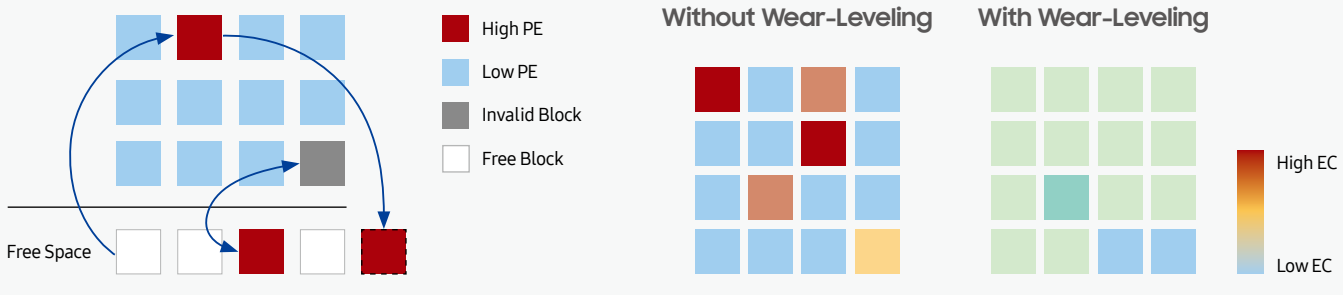
Over-Provisioning Benefits for Samsung Data Center SSDs

Over-provisioning is a function that provides additional capacity specifically for data to be erased from an SSD, without interrupting system performance. The dedicated over-provisioning space may be adjusted to the user's preference, delivering benefits that include faster speed and longer SSD life.
This white paper provides in-depth information about over-provisioning, as well as instructions on how to adjust the over-provisioning space, and considerations to be made before doing so.

Background

Unlike Hard Disk Drives (HDDs), Solid State Drives (SSDs) store electrons on NAND cells when writing data. With NAND flash, the stored data cannot be overwritten when new data is stored or erased. Since the writing and erasing operations (Program/Erase) of an SSD are carried out in different units, referred to as pages and blocks respectively, multiple cycles of Program and Erase are inevitable in writing and managing data. As more of such cycles are repeated, some electrons get trapped between cells and over time, these cells reach the end of their lifetime and encounter durability issues. Such a phenomenon is called the wear-out of cells and is responsible for the physical limits of the lifetime of NAND.
Therefore, proper management of NAND is crucial in extending the lifetime of the SSD. When data is repeatedly written in a certain area, the corresponding cells quickly wear out, so such repeated writing to the same cells should be prevented. Wear-leveling, a function that prevents repeated writing operations to a certain region, enables cells to be utilized evenly by swapping the blocks exposed to a high number of P/E cycles with free blocks, allowing the user to use the SSD longer under given conditions.

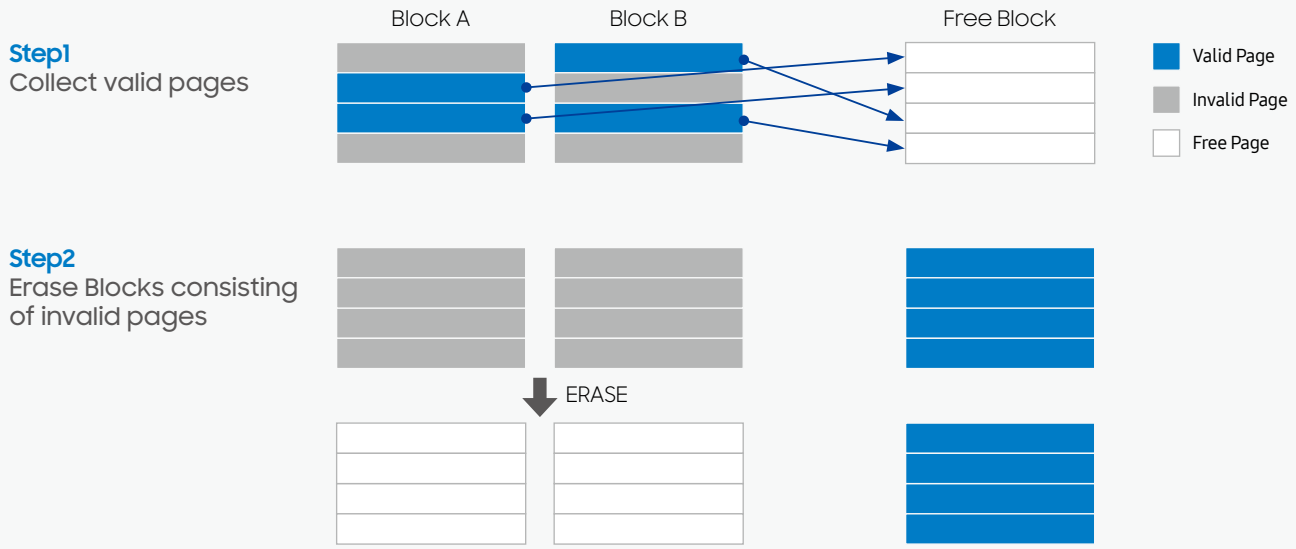
Figure1: Wear Leveling



SAMSUNG

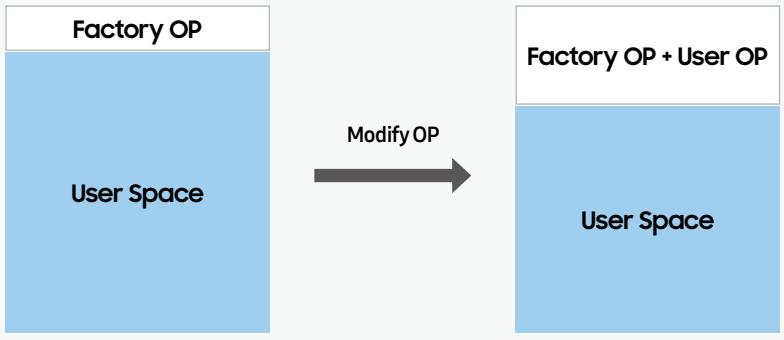
Since overwriting is impossible with NAND flash, existing data must first be erased in order to write new data to that cell, which slows down overall write performance of the SSD. Generally, it takes longer to erase data than to write it because, as mentioned previously, write operations are carried out in pages while erase operations are executed in blocks. To alleviate this decrease in write performance, a process called garbage collection (GC) is implemented to create free blocks within the SSD. This technology secures free blocks by collecting valid pages into a single location and erasing the blocks consisting of invalid pages. However, this too may sometimes result in slower performance in the unexpected case that garbage collection interferes with the host write. Therefore, free space in the SSD is required to allow the firmware (FW) feature to run smoothly. This process in which extra space is allocated is called over-provisioning (OP).

Figure 2: Garbage Collection



What is OP (over-provisioning)?

As previously stated, over-provisioning refers to a function that secures extra space to allow for efficient use of the SSD by allocating a certain amount of the SSD's NAND flash to an over-provisioning space. This space can only be accessed by the SSD's controller and not by the host. Consisting of free blocks only, the OP region assists in efficient delivery of free blocks when wear-leveling or garbage collection is in progress and contributes to improved performance and lifetime of the SSD. Typically, Samsung DC SSDs are set to provide 6.7% of capacity for OP by default, but the user can manually adjust the size of the space if he or she requires additional OP depending on the user environment.



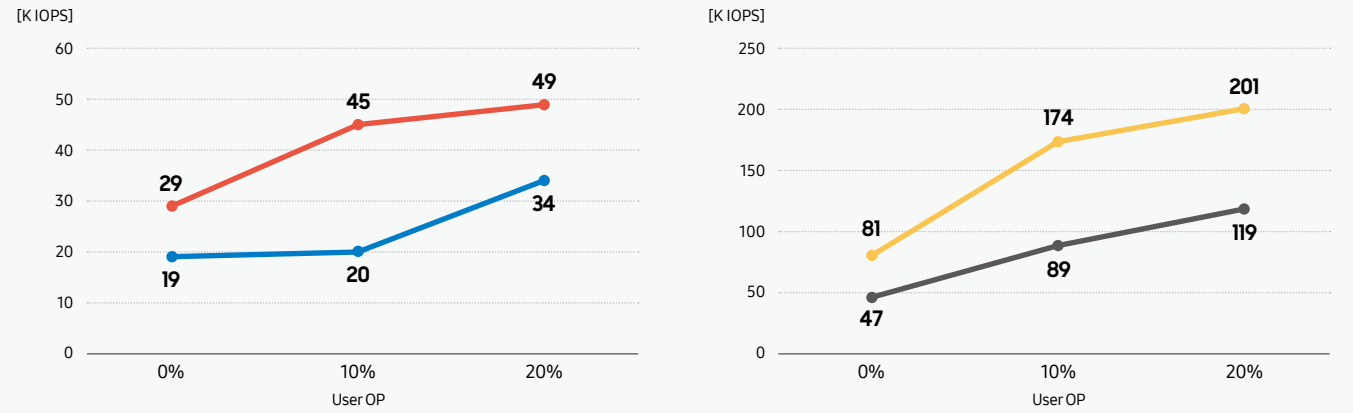
How do I calculate the OP ratio?

OP Ratio Formula: OP (%) = ((Physical Capacity – User Capacity) / User Capacity) * 100
Ex) When 120 GB of a 128 GB SSD is used as the user capacity while 8 GB is assigned to the OP, the OP (%) is ((128 – 120) / 120) * 100 = 6.7 %.

What are the advantages of increasing OP?

Although there is no difference between the sequential and random write performance for fresh-out-of-the-box (FOB) NAND, the random write does not perform as well as the sequential write once data has been written over the entire space of the NAND. Random writes, smaller in size than sequential writes, mix valid and invalid pages within blocks, which causes frequent GC and results in decreased performance. If the OP is increased, more free space that is inaccessible by the host can be secured, and the resulting efficiency of GC contributes to improved performance. The sustained performance is improved in the same manner.

Figure 4. Random Write Performance (QD32) per User OP for SATA / NVMe DC SSD



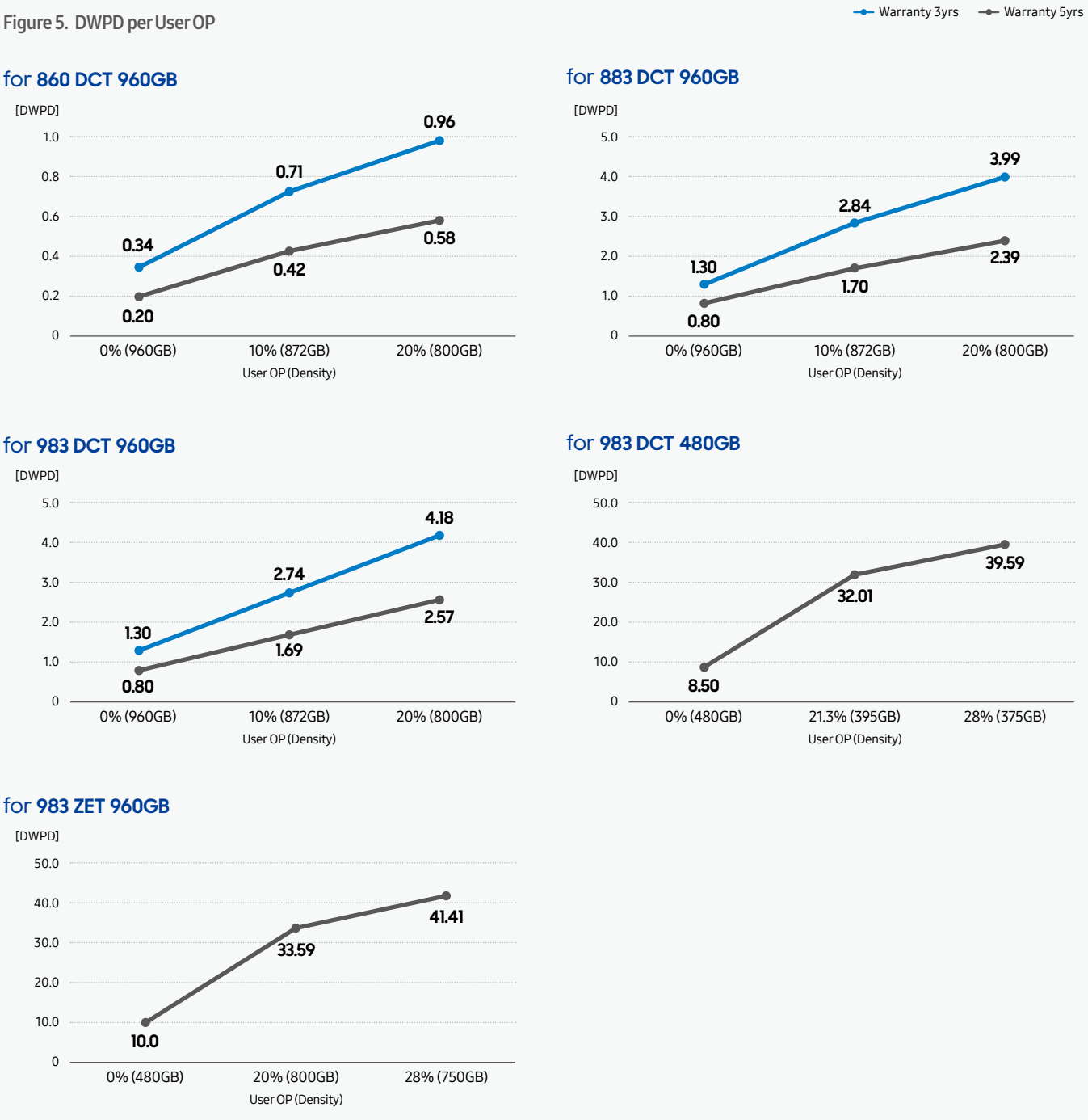
From the aspect of the lifetime of the product, internal operations such as GC cause the number of NAND writes to become greater than that of the host writes, and this also results in an increase in WAF (Write Amplification Factor), defined as the ratio of host writes to NAND writes. An increase in the WAF value indicates that unexpected NAND use is growing, and the lifetime of the product may be shortened before reaching the total byte written (TBW). A sufficient OP space decreases the NAND usage by improving the efficiency of internal NAND operations. It has the advantage of increasing the daily workload per day (DWPD) usable during the warranty period.

Example for Samsung DCT Series

The graphs below show the estimated DWPD for each warranty by allotting additional user OP for Samsung's Data Center products (860/883/983 DCT, 983 ZET). The values in the graphs were calculated using the formula below and is a calculated value of each SSD, not a guaranteed value. The DWPD rises in accordance with the increase in the OP rate. Users can find the numbers needed for the calculations below through S.M.A.R.T. attributes, which are explained in more detail in the next section "Estimating life-time of an SSD using S.M.A.R.T. attributes."

$$WAF = \frac{\text{Physical Write Amount}}{\text{Host Write Amount}} \qquad DWPD = \frac{\text{NAND PE Cycle} * \text{Raw Density}}{\text{Logical Density} * 365 * \text{Warranty Years} * WAF}$$

Figure 5. DWPD per User OP



SAMSUNG

How do I estimate the lifetime of an SSD with S.M.A.R.T. attributes?

- Looking at the S.M.A.R.T. attributes table allow user to understand the wear of their SSD given a particular workload and time period. Alternatively, these attributes allow users to extrapolate the lifetime of their SSD. Please note that the below attributes are only available with SATA SSDs.

- ID 241 and ID 251 indicate the write amount of the host and NAND, respectively, and users can calculate the WAF of their SSD with these values.

- ID 177 indicates the number of wear-leveling operations and can also be interpreted as the overall average program/erase cycle. This value, along with the WAF value, allows the user to find out the DWPD.

- ID 247* represents the time in seconds that the SSD has been in operation since the workload timer was started. (Users can start/stop said timer at their discretion or let it run continuously. It is controlled through their SSD software tools).
  * Please note that ID 247 is expressed in minutes for 860 DCT.

- ID 246 shows the share of I/O operations that were read commands since the workload timer (ID 247) was started and is expressed as a percentage. (Conversely, the share of write I/O operations can be determined by subtracting the given smart attribute reading from 100).

- ID 245 measures the wear of the SSD given the workload (ID 246) and the period of time over which these workloads have been sustained (ID 247). It is displayed as a per mille reading of the total wear of the SSD over its useful lifetime (i.e. a reading of 1000 would mean that the SSD has been worn out over the given time & usage pattern).

- Example applying to 883 DCT
  A user has witnessed that the usage pattern of his SSD has recently decreased from 80% to 70% read I/O operations, and he would like to understand what impact this change has on the lifetime of his SSD. He has decided to run a test for 1 week. At the end of his test run, the S.M.A.R.T. attributes read as follows:
  ID 245: 4, ID 246: 70, and ID 247: 604,800 (7 days x 24 hours x 60 minutes x 60 seconds)
  To find the estimated end of life given the above readings, the user would need to work out the following calculations:
  First, the user wants to understand how many more cycles could be run under the given test scenario before the SSD wears out completely. Therefore, the calculation yields: 1000 / 4 = 250 cycles.
  Second, the user then multiplies this number by the duration of the test run to find the total expected lifetime of the SSD in seconds. This calculation yields 250 x 604,800 = 151,200,000 seconds.
  Finally, given the relatively abstract nature of large numbers expressed in seconds, the user then wants to express the calculated lifetime in years, months or weeks. If the user chooses to express the lifetime in years, he needs to make the following calculation: 151,200,000 / (365 x 24 x 60 x 60) = 4.79 years.

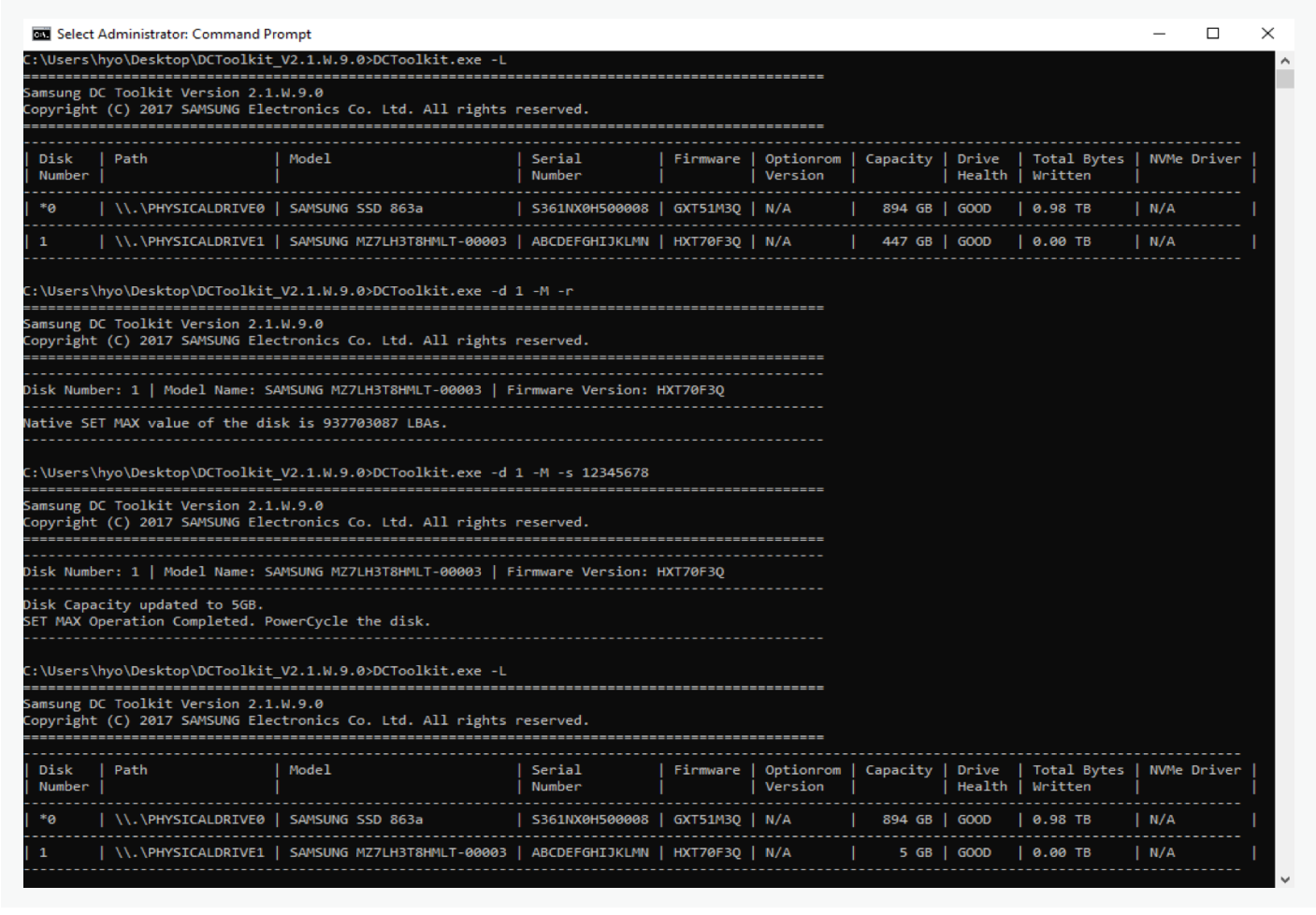
ID	Attribute name	Status Flag	Threshold (%)
5	Reallocated Sector Count	110011	10
9	Power-on Hours	110010	-
12	Power-on Count	110010	-
177	Wear Leveling Count	010011	5
179	Used Reserved Block Count (total)	010011	10
180	Unused Reserved Block Count (total)	010011	10
181	Program Fail Count (total)	110010	10
182	Erase Fail Count (total)	110010	10
183	Runtime Bad Count (total)	010011	10
184	End to End Error data path Error count	110011	97
187	Uncorrectable Error Count	110010	-
190	Airflow Temperature	110010	-
194	Temperature	100010	-
195	ECC Error Rate	011010	-
197	Pending Sector Count	110010	-
199	CRC Error Count	111110	-
202	SSD Mode Status	110011	10
235	POR Recovery Count	010010	-
241	Total LBAs Written	110010	-
242	Total LBAs Read	110010	-
243	SATA Downshift Control	110010	-
244	Thermal Throttle Status	110010	-
245	Timed Workload Media Wear	110010	-
246	Timed Workload Host Read / Write Ratio	110010	-
247	Timed Workload Timer	110010	-
251	NAND Writes	110010	-

How do I adjust OP?

Users can use either the DC Toolkit or Linux HDParm Disk Management to adjust the available space for OP. This adjustment is limited to the space that is not in use only. If the user wishes to increase available space for OP, then he or she must clear up some space that is already in use.

How to adjust the OP setting with DC Toolkit

Step	Description	CMD
1	Identify the device connected to the system.	DCToolkit.exe -L
2	Check the MAX ADDRESS setting available range.	DCToolkit.exe -d 1 -M -r
3	Set MAX ADDRESS value.	DCToolkit.exe -d 1 -M -s 12345678
4	Confirm set value.	DCToolkit.exe -L



Conclusion

Although increasing the OP has the advantage of improving both the performance and lifetime of the user's SSD, it also decreases the available space to the host. This means that rather than excessively increasing it, an appropriate amount of space for the user's workload must be considered when adjusting the OP. Users can experience the most benefit from OP if the preferences are set up when the SSD is FOB. Therefore, it is recommended that the user's SSD product is FOB when the OP setting is set up.

For more information about the Samsung SSD, visit **www.samsungssd.com.**

Samsung Electronics Co., Ltd.
129 Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677, Korea **www.samsung.com** 2019-03

SAMSUNG