WHITE PAPER
Memory and Storage
NAND-based Solid State Drives

Over-Provisioning NAND-Based Intel® SSDs for Better Endurance

How over-provisioning improves NAND-based Intel® SSDs endurance and performance in real-world workloads.

Authors

Zhdan Bybin
Senior Application Engineer

Mohammed Khandaker
Technical Marketing Engineer

Monika Sane
Design Engineer

Table of Contents:

Executive Summary

Over-provisioning increases SSD endurance by allowing extra space for the flash controller to manage incoming data. Over-provisioning improves wear-leveling and random write performance, and reduces the write amplification factor (WAF), thereby improving the endurance of NAND-based SSD.

Background

Writing data to an SSD is significantly different than writing data to a hard disk drive (HDD). Writing to a spinning HDD is performed by magnetizing and demagnetizing sectors on a thin layer of metal mounted on a circular platter, and information is written in binary format (value 1 or 0). Overwriting an HDD is simply a matter of changing the magnetization value, allowing the data to be directly overwritten.

However, overwriting NAND-based SSDs is a much more complicated process. The underlying technology media, NAND flash, is primarily an array of floating gate transistors where electrons are trapped and stored in a floating gate. Applying an electromagnetic field causes the trapped electrons to tunnel through the insulator into the substrate.

The presence or absence of electrons on the floating gate determines value 1 or 0 in the case of the single level cell (SLC) NAND, and the number of electrons determines the voltage level in multiple level cell (MLC, TLC, etc.) NAND. Due to the physical and electrical functioning properties of the floating gate, existing data must be erased before new data can be written to it. The process to write data to SSD media is called program (P), and the process to erase data from it is called erase (E), together called the P/E cycle of the NAND flash. Every P/E cycle leads to very slight wear of the media. Moreover, there is a finite, predefined number of P/E cycles, after which the media becomes unusable.

Another important difference is how the data is organized and partitioned logically on the media. Flash memory is divided into blocks, each block contains pages, and each page is a collection of memory cells. Writing data to an SSD happens in pages, however erasing data happens in blocks. Thus, in order to overwrite a block containing valid and invalid (discarded by the host) pages, the valid pages must be temporarily moved elsewhere on the media, so that the entire block can be erased, then overwritten. This temporary data movement creates the undesirable phenomenon called write amplification (WA). It is undesirable because the actual amount of data written to the SSD media is larger than the amount of data that the host intended to write, thereby leading to further media wear.

To minimize wear and increase the lifetime of the SSD media, there are various algorithms and techniques, such as wear-leveling, garbage collection, etc. Over-provisioning is one such technique available to users of NAND-based SSDs, and is the focus of this document. Over-provisioning an SSD means to make more of the SSD's capacity available to the controller than what was initially designed in. This technique increases SSD endurance by allowing extra buffer space for the flash controller to manage incoming data—it improves wear-leveling and random write performance, and ultimately decreases write amplification factor (WAF).

In contrast, Intel® Optane™ SSDs don't use garbage collection or wear-leveling techniques because of the entirely different endurance mechanism and superior underlying nature of the Intel® 3D XPoint™ media. In environments where very high endurance is required, Intel recommends using Intel® Optane™ technology-based SSDs such as Intel® Optane™ SSD DC P4800X Series, or Intel® Optane™ SSD 900P Series, with their exceptional endurance rating of up to 60 drive writes per day (DWPD).

Introduction to Over-Provisioning

As previously discussed, SSD endurance is the ability to withstand the repeated writing of data, and the ability to retain that data for a period of time (data retention). NAND endurance is measured in terms of terabytes written (TBW) or DWPD and is dependent on a number of factors: the maximum media P/E cycles specification, capacity, workload, and firmware (FW) techniques that are employed to lower the WAF. Under demanding enterprise workloads, NAND-based SSDs can wear out quicker due to higher WAF which is defined as the ratio of NAND writes/host writes. A minimal WAF - as close to 1.0 as possible - is desirable to minimize wear and increase SSD lifetime. In some cases, WAF can be < 1 if the SSD controller has compression mechanisms built in.

$$\text{WAF} = \frac{\textit{NAND WRITES}}{\textit{HOST WRITES}}$$

WAF is different for each workload. One of the more demanding workloads is the one described in the Joint Electron Device Engineering Council (JEDEC) endurance spec. The endurance workload can be obtained at www.jedec.org

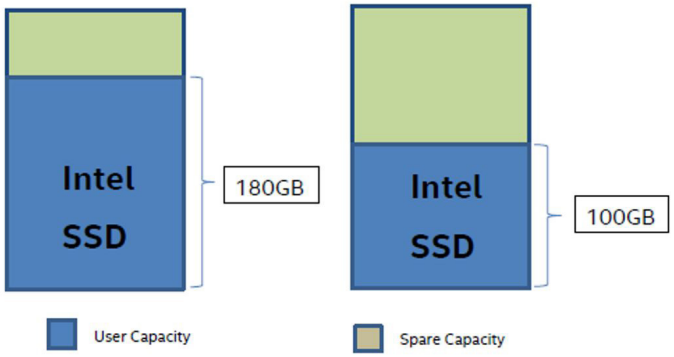
Intel measures and publishes its data center SSD endurance in accordance to the JEDEC specification (JESD 218A, JESD219) as well as for sequential write workload. However, because this is not an industry requirement, other SSD vendors may report endurance numbers measured under the 100% 4KB random write workload. Generally this workload has a WAF that is ~10% lower than a JESD219 workload.

$$\text{Endurance (Peta Bytes Written, PBW)} = \frac{\textit{Raw Drive Capacity} * \textit{NAND PE Cycles}}{\textit{WAF}}$$

$$\text{Endurance (DWPD for 5 years)} = \frac{\textit{Endurance (PBW)}}{\textit{Rated drive density} * 5 * 365}$$

To provide an SSD controller the spare capacity needed to move the data around when programming or erasing partially invalid pages or blocks, each SSD has "factory over-provisioned" area. This area is not addressable by the host/user and may vary in size depending on SSD model and capacity. Effective size of this area can influence the WAF for a workload that uses 100% logical block address (LBA) span of the drive.

Sacrificing user-addressable capacity by manually increasing the effective SSD's spare area allocation will result in endurance gains due to WAF decrease, as well as improvements in random write performance and quality of service (QoS) due to fewer SSD controller housekeeping activities. This method of over-provisioning is similar to the HDD concept called "short stroking" the drive.



Methods of Over-Provisioning

Over-provisioning should be performed on an SSD in a completely clean state. This can be an SSD that is fresh out-of-the-box or an SSD on which a secure erase has been performed.

The three most common methods of over-provisioning an SSD are as follows:

1. Limiting the logical volume capacity during partitioning in OS (user will see full capacity in Disk Manager or fdisk).
2. Limiting the Maximum LBA on the drive level, so that in OS, it will appear as a lower-capacity drive.
3. Limiting an application to use only a certain LBA range.

Please note, method 3 above will not work for the scenario in which the filesystem is deployed on full LBA range.

For method 2 above, while working with Intel® SSDs for the Data Center, customers can use either Intel or 3rd-party tools. Following are the tools that work with native over-provisioning:

a. The easiest tool to use for over-provisioning is Intel® SSD Data Center Tool (Intel® SSD DCT), which supports both SATA and NVMe* Intel® SSDs, in Linux* and Windows* environments - https://www.intel.com/content/www/us/en/support/articles/000006289/memory-and-storage.html

   Intel SSD DCT User Guide can be found here: https://www.intel.com/content/www/us/en/support/memory-and-storage/000020016.html?wapkw=data+center+tool+user+guide

The command sequence for over-provisioning:

 # sudo isdct show -intelssd (To get index of the drive)

 # sudo isdct delete -intelssd 0 (Note: ATA security needs to be NOT frozen and NOT in the locked state, please refer to https://www.intel.com/content/www/us/en/support/articles/000006094/memory-and-storage.html).

 # sudo isdct set -intelssd X MaximumLBA=(xGB\x%\LBA\'native')

(example: isdct set -intelssd 0 MaximumLBA=80%)

Power cycle the drive or reboot the system.

```
[root@localhost ~]# isdct show -intelssd

- Intel SSD DC S4500 Series PHYS73120349240AGN -

Bootloader : Property not found
DevicePath : /dev/sg0
DeviceStatus : Healthy
Firmware : SCV10111
FirmwareUpdateAvailable : The selected Intel SSD contains current firmware as of
 this tool release.
Index : 0
ModelNumber : INTEL SSDSC2KB240G7
ProductFamily : Intel SSD DC S4500 Series
SerialNumber : PHYS73120349240AGN

[root@localhost ~]# isdct set -intelssd 0 maximumlba=200GB
Set MaximumLBA successful. Please power cycle the device.
[root@localhost ~]#
```

The way we recognize an over-provisioned drive is illustrated in picture below (notice how MaximumLBA < NativeMaxLBA) :

 # sudo isdct show –all –intelssd 0

```
MaximumLBA : 390625000
MediumPriorityWeightArbitration : Device does not support this command set.
ModelNumber : INTEL SSDSC2KB240G7
NVMePowerState : Device does not support this command set.
NativeMaxLBA : 468862127
OEM : Generic
PLITestTimeInterval : 10080 minutes
PhySpeed : 6.0 Gbps
PhysicalSectorSize : 4096 bytes
PhysicalSize : 200000000512
PowerGovernorAveragePower : 4000 milliwatts
```

An over-provisioned SSD can be reverted to its native capacity as shown in image below. Also, it is best practice to put the SSD in standby mode before power cycling it. This can be done as follows:

 # sudo isdct set –intelssd 0 maximumlba=native

```
[root@localhost ~]# isdct set -intelssd 0 maximumlba=native
Set MaximumLBA successful. Please power cycle the device.
[root@localhost ~]# isdct start -intelssd 0 -standby

- StandbyImmediate PHYS73120349240AGN -

Status : Completed successfully.
```

b. HDParm* (latest version) 9.49 and above - 3rd party tool, Linux* only, only SATA drives

Just as when using Intel SSD DCT, secure erase must be performed on the SSD before over-provisioning it.

 # sudo hdparm --user-master u --security-set-pass 123 /dev/sdX

 # sudo hdparm --security-erase 123 /dev/sdX

The command for over-provisioning would be:

 # sudo hdparm -N /dev/sdX (To find the maximum sector count)

 # sudo hdparm -NpXXXXXXXXX -yes-i-know-what-i-am-doing /dev/sdX (This enables host protected area and sets the number of visible sectors to the count appearing immediately after "-Np")

c. nvme-cli – open source tool developed by an Intel engineer, Linux only, only NVMe drives

Endurance Calculations for NAND-Based Intel® SSDs

Intel data center SSDs have timed workload SMART indicators (E2h, E3h, E4h) that allow users to easily calculate the lifetime of the SSD under any given real-world workload using actual NAND wear statistics. The E2h attribute measures the wear endured by the SSD during the timed workload; E3h tracks the workload's read/write ratio; and E4h reports the number of minutes spent during the workload. These attributes must be reset prior to applying the characteristic workload, by issuing the SMART Execute Immediate ATA command (for SATA SSDs) or NVMe Vendor Unique Set Features D5h command (for NVMe SSDs). Intel recommends applying a full characteristic cycle of the expected workload, i.e., 8-hours, 24-hours, 1 week etc., while the minimum requirement is a 60 minute workload with enough write pressure to make E2h increase its counter by 1, which is equivalent to ~0.001% overall media wear.

If Intel SSD DCT is used, the tool will do all the calculations on the user's behalf.

The commands are as follows:

```
# sudo isdct set –intelssd 0 EnduranceAnalyzer =reset
```

(to reset the attributes prior to the test)

Once reset, raw values for E2h/E3h/E4h will show '65535' and will stay the same until the representative workload is applied for at least 60 minutes.

```
[root@localhost fio]# isdct set -intelssd 0 enduranceanalyzer=reset
Set EnduranceAnalyzer successful. Completed successfully.
[root@localhost fio]# isdct show -smart E2 -intelssd 0

- SMART Attributes PHYS731203492 40AGN -


- E2 -

Action : Pass
Description : Timed Workload - Media Wear
ID : E2
Normalized : 100
Raw : 65535
Status : 50
Threshold : 0
Worst : 100
```

After the test is done, Intel SSD DCT can provide the estimated life calculation using the following command:

```
# sudo isdct show -all –intelssd 0
```

Without Intel tools, the same calculations can be made using E2h/E3h/E4h 'raw' values. Here is a real-world application example after testing a 240GB Intel® SSD DC S4500:

Attribute	Attribute Name	Value @ test start	Value @ test end
E1	Host Writes	14308	26173
E2	Timed Workload Media Wear	65535	41
E3	Timed Workload Host Read/Write Ratio	65535	65
E4	Timed Workload Timer	65535	8624

Based on this data, we can calculate following:

Host writes = E1end – E1start = 11865 units * 32MB per unit = 370 GB

Time spent during test = 8624 min / 60 / 60 = 6 days

Workload read/write ratio = 65%/35% R/W

Media wear = 0.040 %

Estimated drive's life remaining = 6 days * 100 % / 0.04 % = 15000 days = 41.1 years

Flexible Capacity and Endurance Examples for Intel® SSD DC S4500 Series and Intel® SSD DC S4600 Series.

The following chart shows the increase in endurance to over-provisioning % (OP%) for the Intel SSD DC S4500 Series, and user capacity with respect to different OP%.

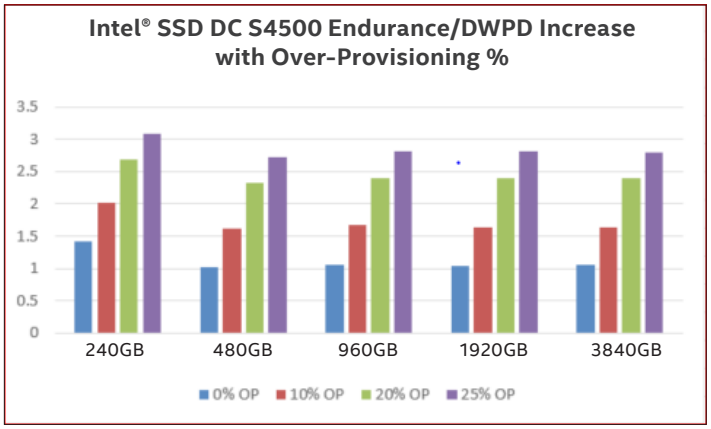


Chart 1. Intel® SSD DC S4500 Series Endurance (DWPD) increase with different over-provisioning levels

The following tables provide values of estimated calculations for popular endurance levels, i.e., how to get from 1 DWPD to 3 DWPD, or from 3 DWPD to 5 DWPD, or for popular over-provisioning (OP) and capacity levels, i.e. 10%, 20% or 400GB, etc. It also shows different endurance levels if sequential workload is dominant compared to JEDEC workloads or 4K random write (RW) workloads.

Flexible Capacity and Endurance Calculations for Intel® SSD S4500 Series:

	Size (GB)	No OP JEDEC Endurance		No OP Seq Write Endurance		Size	10% OP 4K RW Endurance	
S4500		PBW	DWPD (5 yrs)	PBW	DWPD (5 yrs)	10% OP	PBW	DWPD (5 yrs)
2.5"	**240**	0.62	1.42	1.64	3.75	**216**	0.79	2.01
	480	0.90	1.03	3	3.42	**432**	1.28	1.63
	960	1.86	1.06	5.94	3.39	**864**	2.63	1.67
	1920	3.27	1.06	11.62	3.31	**1728**	5.18	1.64
	3840	7.64	1.09	22.53	3.21	**3456**	10.30	1.63

Size	2DWPD 4K RW Endurance		Size	20% OP 4K RW Endurance		Size	3DWPD 4K RW Endurance	
	PBW	DWPD	20% OP	PBW	DWPD		PBW	DWPD
215	0.8	2.03	**200**	0.89	2.45	180	1.01	3.08
403.25	1.5	2	**400**	1.52	2.08	345	1.89	2.99
816	3	2.01	**800**	3.14	2.15	700	3.84	3
1632	5.93	1.99	**1600**	6.26	2.14	1400	7.65	2.99
3264	12.12	2.03	**3000**	13.95	2.55	2800	15.25	2.98

Flexible Capacity and Endurance Calculations for Intel® SSD S4600 Series:

	Size (GB)	No OP JEDEC Endurance		No OP Seq Write Endurance		Size	10% OP 4K RW Endurance	
S4600		PBW	DWPD	PBW	DWPD	10%OP	10%OP	DWPD
2.5"	**240**	1.4	3.19	2.19	5.01	216	1.53	3.88
	480	2.95	3.36	4.37	4.98	432	3.23	4.09
	960	5.25	3	8.11	4.63	864	5.81	3.68
	1920	10.84	3.09	16.16	4.61	1728	11.78	3.73

Size	5DWPD OP 4K RW Endurance		Size	10DWPD OP 4K RW Endurance	
	PBW	DWPD (5 yrs)		PBW	DWPD (5 yrs)
184	1.68	**5.01**	**105**	1.92	**10.01**
380	3.49	**5.03**	**220**	4.04	**10.05**
725	6.6	**4.99**	**420**	7.72	**10.06**
1465	13.35	**4.99**	**860**	15.71	**10**

Flexible Capacity and Endurance Calculations for Intel® SSD P4500 Series:

SKU	Size (TB)	No OP JEDEC Endurance		Sequential Write Endurance		Size	10% OP 4K RW Endurance	
P4500		PBW	DWPD	PBW	DWPD	10% OP	PBW	DWPD
U.2 15mm AIC/U.2 AIC	1	1.38	0.75	5.06	2.75	0.9	2.17	1.32
	2	1.89	0.5	9.65	2.60	1.8	3.6	1.09
	4	4.84	0.65	19.76	2.70	3.6	8.18	1.24
	8	7.08	0.45	38.61	2.60	7.2	14.44	1.1
U.2 7mm	500GB	0.67	0.73	2.64	2.89	450	1.08	1.31
	1	1.85	1.01	5.52	3.02	0.9	2.69	1.64
	2	2.29	0.63	10.31	2.82	1.8	4.09	1.24
	4	4.99	0.68	20.87	2.86	3.6	8.6	1.31

SKU	Size	20% OP 4K RW Endurance		Size	3DWPD 4K RW Endurance	
P4500	20% OP	PBW	DWPD	TB	PBW	DWPD
U.2 15mm	0.8	2.94	2.01	0.685	3.74	**2.99**
	1.6	5.24	1.79	1.33	7.22	**2.97**
	AIC / U.2	11.47	1.96	2.7	15.1	**3.06**
	AIC	21.08	1.8	5.3	29.15	**3.01**
U.2 7mm	400	1.47	2.01	340	1.89	**3.04**
	0.8	3.47	2.38	0.73	3.98	**2.99**
	1.6	5.76	1.97	1.36	7.45	**3**
	3.2	11.65	1.99	2.75	14.88	**2.96**

Flexible Capacity and Endurance Calculations for Intel® SSD P4510 Series:

SKU	Size (TB)	No OP JEDEC Endurance		Sequential Write Endurance		Size	10% OP 4K RW Endurance	
P4510		PBW	DWPD	PBW	DWPD	10% OP	PBW	DWPD
U.2 15mm	1	1.92	1.05	5.62	3.00	0.9	2.78	1.69
	2	2.61	0.7	10.90	2.90	1.8	4.43	1.35
	4	6.3	0.85	21.84	2.80	3.6	10.02	1.52
	8	13.88	0.9	44.25	3.00	7.2	21.3	1.62
M.2 110mm	1	0.98	0.54	5.13	2.81	0.9	1.91	1.16
	2	1.95	0.53	10.26	2.81	1.8	3.79	1.15

SKU	Size	20% OP 4K RW Endurance		Size	3DWPD 4K RW Endurance	
P4510	20% OP	PBW	DWPD	GB	PBW	DWPD
U.2 15mm	0.8	3.56	2.44	735	4.02	**3**
	1.6	6.13	2.1	1400	7.63	**2.98**
	3.2	13.29	2.27	2870	15.72	**3**
	6.4	27.74	2.37	5840	31.86	**2.99**
M.2 110mm	0.8	2.78	1.9	680	3.69	**2.97**
	1.6	5.52	1.89	1350	7.4	**3**

As mentioned previously, over-provisioning will also positively affect random write performance of a NAND-based SSD if 100% LBA span is used, or if filesystem is applied to the whole addressable capacity of the drive. The smaller the default factory over-provisioned area is, the higher the impact will be. The chart below shows the impact of OP% on random write performance using a 240GB Intel® SSD DC S4600 as an example.

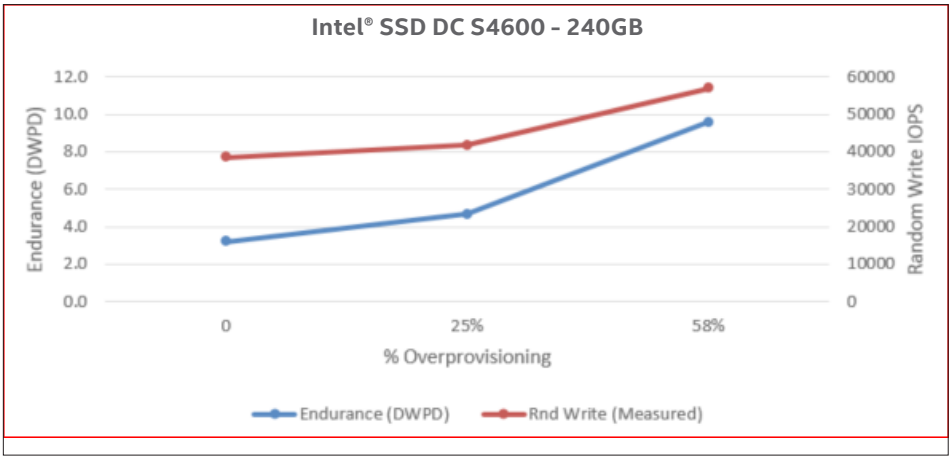


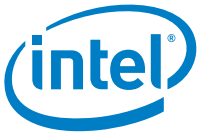
Chart 2. Intel® SSD DC S4500 Series Random Write Performance

System Configuration for all performance testing: Intel® Xeon® CPU E5-2699 v4 @ 2.20GHz on Intel® S2600WT motherboard, Intel® C612 Chipset, BIOS Version SE5C6 10.86B.01.01.0019.101220160604 32GB DDR4, FIO version 2.18, CentOS 7.0, Kernel 4.8.6 (DAS patch)

Conclusion

As seen in the preceeding formulas and tables, while sacrificing user-addressable capacity, over-provisioning provides a positive effect on NAND-based SSD endurance, write amplification factor, and random write performance. In general, over-provisioning allows flexibility in an SSD's endurance and capacity where the user can go from a 1 DWPD-rated SSD to 3 DWPD, or from 3 DWPD to 5, or even up to 10 DWPD.

If higher endurance levels are required – for example, for hot data tier, or caching, or DRAM displacement – Intel® Optane™ SSDs can be used. Due to the different media nature, Intel® Optane™ SSDs do not gain benefits with over-provisioning; they already have very high endurance, with up to 60 DWPD rating.



For more information, visit intel.com/ssd

337671-002US