# Understanding Linux Storage I/O Access in the Age of SSDs

SSDs bring unique benefits to enterprise storage. Their high speeds, low power and low latency drive both standard and emerging applications toward new performance thresholds.

SSDs have also introduced a new concept into the storage market: storage devices that wear as they are written (versus HDDs, which wear as data is accessed and written). The market saw initial user trepidation with SSDs, partly because users didn’t know their I/O patterns and had difficulty estimating them. As a result, many storage architects deployed SSDs with endurance ratings far higher than their workloads required.

As SSD endurance needs have changed (trending toward lower endurance), proper matching has become imperative to control costs and ensure reliability — especially in the era of quad-level cell (QLC) NAND SSDs that are read-centric with lower write endurance.

This paper explains how a tool built into Linux (iostat)[1] and another distributed by Micron (Storage Executive software) can help you better understand your workload and application storage I/O profiles. Learn how to:

- Use iostat to characterize a storage workload.
- Measure SSD-specific wear using Micron’s Storage Executive.
- Match the right SSD type and endurance ratings to the right applications and their workloads.

[1] In this paper, “Linux” refers to a Linux distribution with the iostat tool built in.

# Background

Measuring storage I/O profiles of different applications and workloads enables you to make purchase decisions based on your workload needs. This has become more important with the growth of enterprise SSDs use.

> NAND media in SSDs must be erased before it can be rewritten. This two-step process is called a program/erase (P/E) cycle. Different NAND types support different P/E cycle counts, giving different SSDs different amounts of write endurance.

## SSD Endurance Trends

When SSD adoption was just starting (around 2007), the idea of a storage device that wears when written was new. 10, 20 or more drive writes per day (DWPD) was normal. As a result, many system designers initially overestimated the amount of wear their applications applied to SSDs to ensure a safety margin.

However, trends now show SSD endurance measured in DWPD is rapidly decreasing. This suggests a better understanding of workload read/write profiles. Figure 1 shows how DWPD requirements have decreased over time.

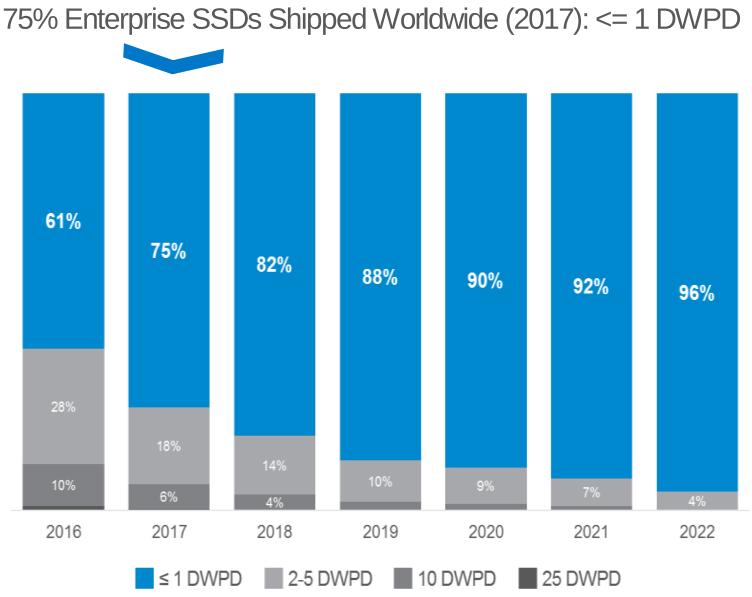

***Figure 1: SSD Endurance Trends***
(Source: Analyst consensus, Forward Insights Datacenter, May 2018)

## SSD Wear

As SSDs are written and rewritten, their wear state is communicated to the host system. SATA SSDs, for example, communicate their degree of wear through Self-Monitoring, Analysis and Reporting Technology (SMART).

It is important to note that when an SSD reaches its wear threshold, it migrates into a read-only state where the data on the SSD can still be read but can’t be overwritten, nor can new data be added.

# Estimating Storage I/O

To determine if an SSD is suitable for your application/workload, you must characterize the I/O profile of the workload. This characterization is typically a four-step process, as shown in Figure 2. The process combines SSD-specific and application-specific data to estimate the amount of data written to the SSD (as a function of time) and the workload’s read/write ratio.

Many operating systems (OSs) support integrated tools to help measure the read ratio of data I/O on storage devices. When configured to monitor specific values, these tools can keep a running log of storage I/O transactions.

Micron supplies SSD-level monitoring software that helps characterize the amount of data written to storage through standard SMART reporting.

The combination of OS and Micron tools can help users have a more complete understanding how applications use storage and align this to a specific SSD through a more complete understanding of the workload’s storage IO profile.

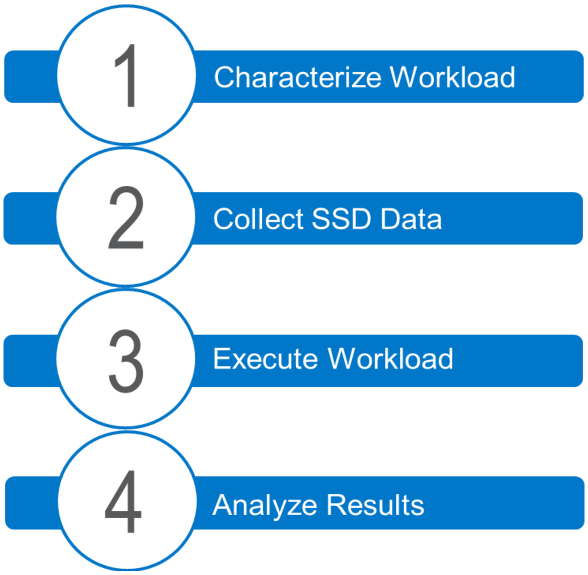

*Figure 2: Workload Profiling Steps*

## 1. Characterize Workload: Estimate Read/Write Ratio With iostat

Most Linux distributions include iostat, which is a tool that can be used to monitor storage I/O in real time. If your distribution does not include iostat, you can install sysstat (the parent package of iostat) using the following command:

```
sudo apt install sysstat
```

To monitor a specific device, run the following:

```
iostat -p sdX (where X = device ID sda, sdb, etc.)
```

**Note:** Tech Republic provides an introduction to iostat for storage, and the Unix & Linux StackExchange community hosts a conversation on using iostat.

**Important:** When characterizing a workload, ensure that the workload runs long enough to reach steady state (about 2 hours is common, but it may vary based on workload specifics).

Figure 3 shows an example of iostat output, which reports on all storage devices in the system. In this example, the SSD is **/dev/sdb** (other devices are included only for clarity):

```
[root@SSD_node]# iostat /dev/sdb -xmcN -t 5

Linux 3.10.0-957.el7.x86_64 (localhost.localdomain)          02/27/2019           _x86_64_  (4 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.01    0.00    0.01    0.00    0.00   99.98

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sdb               0.00     0.00   32.09    0.94     0.20     0.01    12.93     0.00    0.11    0.11    0.07   0.04   0.12
```

*Figure 3: iostat Output*

The value of **avgrq-sz** is key to understanding the test workload's average block size. IOPS can then be used to estimate the read/write ratio using rMB/s and wMB/s. In the example above, avgrq-sz = 12.92 x 512 bytes, or about 6.5KB. Using rMB/s and wMB/s to estimate % read and % write gives a ratio of .20 to 0.01 (about 95% read and 5% write, which is a very read-heavy workload). The value differs for other workloads.

Estimating read/write ratios can help quantify a workload as read-centric, write-centric or mixed-use. This, in turn, can help you select an SSD with endurance characteristics well-matched to the workload. While estimating workload read/write ratios using iostat data can help target a specific SSD type, this data shows the I/O patterns being sent to the storage device but does not factor in I/O concatenation or bifurcation that may take place at different layers of the storage I/O stack. Therefore, this data may not translate directly into the I/O pattern seen by the SSD.

## 2. Collect SSD Data: Capture Initial SMART Data

Micron offers a convenient software tool, called Storage Executive, to manage Micron SSDs and collect data from them (including wear indications expressed via SMART data in the SSD). Storage Executive provides broad functionality, including the ability to do the following:

- View all drives installed in a system and see current drive status, capacity, temperature, firmware version and driver information.
- View SMART attributes.
- Update firmware.
- Remove all data from a drive by performing a drive sanitize or physical security ID (PSID) revert operation.
- Improve drive performance and endurance by allocating overprovisioned capacity.
- Adjust drive endurance with Micron's Flex Capacity feature (supported drives only).
- Perform a drive self-test.

This paper shows how Micron's Storage Executive can be used to monitor the amount of data written to an SSD under test by collecting that SSD's SMART data before and after the workload test. Visit Micron's Storage Executive Software page for details on obtaining and using the tool.

**Note:** SMART wear data is cumulative; it does not reset (or return to a prior value). Therefore, SMART data must be collected before and after running a workload test. The difference between the cumulative host write sector count (ID 246) before and after the test indicates the amount of wear applied during the test.

**Note:** Third-party tools can be used to collect SMART data correctly; however, they sometimes incorrectly identify the SSD's SMART attributes (this does not affect measured results). When using a third-party tool, check the Micron SSD data sheet to identify the SMART attribute name associated with the amount of data written.

Table 2 below shows an example SMART output from Storage Executive for a typical Micron enterprise SSD taken before starting the test workload. SMART ID 246 (in bold) shows the cumulative amount of data written to this SSD in 512-byte increments (sectors).

| SMART Data (Before Workload) | | | |
|---|---|---|---|
| Device Name : /dev/sdb | | | |
| ID | Attribute Name | Attribute Data | |
| 1 | Raw Read Error Rate | 0 | Errors/Page |
| 5 | Retired NAND Blocks | 0 | NAND Blocks |
| 9 | Power On Hours Count | 1896 | Hours |
| 12 | Power Cycle Count | 6 | Cycles |
| 170 | Reserved block count | 0 | Blocks |
| 171 | Program Fail Count | 0 | NAND Page Program Failures |
| 172 | Erase Fail Count | 0 | NAND Block Erase Failures |
| 173 | Average Block-Erase Count | 58 | Erases |
| 174 | Unexpected Power Loss Count | 0 | Unexpected Power Loss events |
| 180 | Unused reserved block count | 10099 | Blocks |
| 183 | SATA Interface Downshift | 0 | Downshifts |
| 184 | Error Correction Count | 0 | Correction Events |
| 187 | Reported Uncorrectable Errors | 0 | ECC Correction Failures |
| 188 | Command Timeouts | 8 | Outstanding Commands Since Last Reset |
| 194 | Enclosure Temperature | 21 | Current Temperature (C) |
| | | 32 | Highest Lifetime Temperature (C) |
| 195 | Cumulative Corrected ECC | 0 | Corrected ECC |
| 196 | Reallocation Event Count | 0 | Events |
| 197 | Current Pending Sector Count | 0 | 512 Byte Sectors |
| 198 | SMART Off-line Scan Uncorrectable Errors | 0 | Errors |
| 199 | Ultra-DMA CRC Error Count | 0 | Errors |
| 202 | Percentage Lifetime Used | 0 | % Lifetime Used |
| 206 | Write Error Rate | 0 | Program Fails/MB |
| 210 | RAIN Successful Recovery Page Count | 0 | TUs successfully recovered by RAIN |
| **246** | **Cumulative Host Write Sector Count** | **260744152135** | **512 Byte Sectors** |
| 247 | Host Program Page Count | 8222331979 | NAND Page |
| 248 | FTL Program Page Count | 480116578 | NAND Page |

```
SMART attributes are retrieved successfully
CMD_STATUS    : Success
STATUS_CODE   : 0
Copyright (C) 2018 Micron Technology, Inc.
```

*Table 2: Example Starting SMART Values*

To determine the total amount written to this SSD during any workload test, do the following:

1. Record the starting value of ID 246 before starting the workload (the above example).
2. Run the complete workload test (record the time to complete the test).
3. Record the ending value of ID 246 after the workload completes (see section 4. Analyze Results).

The difference between the values of ID 246 in step 3 and step 1 equals the total number of 512-byte sectors written to the SSD. (It may be easier to express this value in KB or MB through arithmetic conversion for subsequent analysis.) Note that these calculations are relative to a single SSD; hence, they may differ from values noted earlier (this is expected). Record the SMART ID 246 value before continuing to the next section.

## 3. Execute Workload

Workload execution is application-specific. Ensure you capture the starting SMART ID 246 value and that iostat is available. Record the total duration of the workload when execution completes.

## 4. Analyze Results

After the test workload completes, use Storage Executive to capture SMART data. Table 3 shows an example output.

| SMART Data (After Workload) | | | |
|---|---|---|---|
| Device Name : /dev/sdb | | | |
| ID | Attribute Name | Attribute Data | |
| 1 | Raw Read Error Rate | 0 | Errors/Page |
| 5 | Retired NAND Blocks | 0 | NAND Blocks |
| 9 | Power On Hours Count | 1896 | Hours |
| 12 | Power Cycle Count | 6 | Cycles |
| 170 | Reserved block count | 0 | Blocks |
| 171 | Program Fail Count | 0 | NAND Page Program Failures |
| 172 | Erase Fail Count | 0 | NAND Block Erase Failures |
| 173 | Average Block-Erase Count | 58 | Erases |
| 174 | Unexpected Power Loss Count | 0 | Unexpected Power Loss events |
| 180 | Unused reserved block count | 10099 | Blocks |
| 183 | SATA Interface Downshift | 0 | Downshifts |
| 184 | Error Correction Count | 0 | Correction Events |
| 187 | Reported Uncorrectable Errors | 0 | ECC Correction Failures |
| 188 | Command Timeouts | 8 | Outstanding Commands Since Last Reset |
| 194 | Enclosure Temperature | 20 | Current Temperature (C) |
| | | 32 | Highest Lifetime Temperature (C) |
| 195 | Cumulative Corrected ECC | 0 | Corrected ECC |
| 196 | Reallocation Event Count | 0 | Events |
| 197 | Current Pending Sector Count | 0 | 512 Byte Sectors |
| 198 | SMART Off-line Scan Uncorrectable Errors | 0 | Errors |
| 199 | Ultra-DMA CRC Error Count | 0 | Errors |
| 202 | Percentage Lifetime Used | 0 | % Lifetime Used |
| 206 | Write Error Rate | 0 | Program Fails/MB |

| | | | |
|---|---|---|---|
| 210 | RAIN Successful Recovery Page Count | 0 | TUs successfully recovered by RAIN |
| **246** | **Cumulative Host Write Sector Count** | **260747179063** | **512 Byte Sectors** |
| 247 | Host Program Page Count | 8222641789 | NAND Page |
| 248 | FTL Program Page Count | 480139697 | NAND Page |

```
SMART attributes are retrieved successfully
CMD_STATUS   : Success
STATUS_CODE  : 0
Copyright (C) 2018 Micron Technology, Inc.
```

***Table 3: Example Ending SMART Values***

## Example SMART Data Written to SSD (as seen by the SSD)

You can monitor the exact amount of data written to the SSD using its wear reporting mechanism. Note that the example shown in Table 4 uses SMART, but other SSDs may use different reporting methods.

When SSDs report the amount of data written to them, you can determine the amount of data written by capturing the SSD's SMART data at the start and end of the test. This example uses SMART ID 246, which is the number of 512-byte sectors the host wrote.

We subtracted ID 246's starting value from its ending value and converted the result (512-byte sectors) into megabytes (MB). The result is the total MB written to the monitored storage device during the workload test. Combining this data with the test runtime yields the test workload's data write rate.

| SMART ID 246 | |
|---|---:|
| Ending Value | 260,747,179,063 |
| Starting Value | 260,744,152,135 |
| #512 Byte Sectors Written | 3,029,928 |
| SSD Write Rate (MB/s) | ~5 |

***Table 4: SMART Calculations***

## Example Workload Analysis

You can combine the SSD choice targeting data from iostat with measured SMART data from an example drive, workload and total test run time for better insight on SSD choice. This example analysis uses a hypothetical SSD with a warranted TBW of 3500TB for five years.

**Important:** This is an example only. Actual results may differ. Estimating workload-SSD match should be done using the actual SSD being considered and the workload applied.

### SMART Data

After targeting an SSD type (in this example, iostat data suggests targeting read-centric SSDs), you can estimate whether the rated endurance of the read-centric SSD being considered is enough to support a five-year replacement schedule.

Suppose you run the test workload continuously on the example SSD and use a five-minute test interval and collect the SMART data seen in Table 4. This SMART data indicates that over the course of a five-minute test, the workload wrote about 1478MB to the SSD (a write rate of about 5 MB/s).

Suppose the SSD you are considering offers an endurance (expressed as Total Bytes Written [TBW]) of 3500TB. Assuming the workload's write rate is constant and that this five-minute sample reflects the workload I/O profile over time, we can extrapolate the total amount of data that would be written to the SSD during its five-year warranty period and compare that to the SSD’s rated endurance of 3500TB.

```
5 MB/sec x 5 years = ~741TB of data written over 5 years
```

Because the example SSD is rated for 3500TB for five years, the measured data suggests our application would write far less that the SSD’s five-year TBW (SMART data suggests this application would write about 741TB over five years).

# Benefits

Understanding a workload’s read/write ratio can help you choose the right SSDs to benefit the right workload. For example, a read-centric workload excels with read-centric SSDs and does not benefit from the additional endurance of higher-write endurance, higher-cost SSDs. Combining the read/write ratio and measured data written over time helps validate if an SSD’s rated endurance is sufficient.

Profiling storage I/O with measured data is the best way to ensure an optimal match between workload and SSD. While a litany of conventional wisdom describes “typical” or ”historical” application and workload I/O profiles, many of these are configuration-dependent.

Matching appropriate applications and workloads to SSDs has become more important with the introduction of QLC SSDs. These SSDs are designed for read-centric workloads (and provide better longevity when writing large I/O sizes). When SSDs first introduced the concept of storage devices that wear into the storage market, this concept was new; however, readily available tools now enable you to more precisely understand storage I/O needs.

# Conclusion

Built-in and Micron-supplied tools combined with workload execution timing can help characterize application and workload storage I/O profiles. This, in turn, enhances understanding of data placement and a more precise match between SSD and workload.

micron.com