A MICRON **TECHNICAL MARKETING BRIEF**

Over-Provisioning the Micron® 1100 SSD for Data Center Applications

Of the many factors that affect datacenter system design, cost is almost always a primary concern. We often assume that high performance systems and applications come with higher cost. There may be other applications where cost is the primary consideration. Until recently, cost-focused designs often were forced to used legacy hard disk drives (HDD). However, as the cost per gigabyte of solid state drives (SSDs) has dramatically decreased, entry-level SSDs have become quite cost-competitive (per GB stored) compared with HDDs, especially when we consider Enterprise-class 10K and 15K HDDs. SSDs can also offer exceptional IOPS performance and very low and consistent latency. With these market and technological changes, we are seeing SSD adoption growing in the datacenter – extending to very affordable, entry-level Enterprise SSDs as well as SSDs intended for client applications.

Micron’s 1100 Client SATA SSD, built with cost-effective 3D three-level cell (TLC) NAND and its category-leading endurance, is an attractive choice for many datacenter applications.

The Micron 1100 Client SATA SSD

Built-in encryption engines perform at full interface speed, without using CPU bandwidth. Encrypted SSDs run at the same speed as their non-encrypted counterparts.

Over-provisioning as described in this paper can help enhance the performance and reliability of the 1100 in such applications.

There are design and specification differences between client-oriented SSDs like the 1100 SSD and those more deliberately targeted at the data center such as Micron’s 5100 SATA SSD. These differences can result in implementation challenges when considering Client SSDS for datacenter use. However, careful consideration of the targeted workload and some
simple changes to the Client SSD’s configuration prior to installation can help mitigate these challenges.

This technical brief describes the performance attributes of client SSDs and how they can be easily reconfigured to better suit some data center applications. We will describe the practice of over-provisioning (OP) and how we can use OP to modify the Micron 1100 SSD to better enabling data center applications. We will show how to do this using the command line interface (CLI) included with Micron’s Storage Executive software.

Design Considerations Not Related to Performance

When deciding on a client SSD for a data center application, there are several design attributes all system architects should consider.

Important: The values shown in the tables and throughout this document are estimates for storage system designers’ reference and are not be intended as product specifications. Referring to the 1100 data sheets, consider that specified performance numbers are for expected client workloads, not for the data center. Variables in the host system: processor speed, memory speed, amount of memory, etc. – all will affect the performance measured at the storage device.

Unexpected Power-Loss Protection

SSDs designed for the data center or the enterprise typically have very robust power loss protection. This comes in the form of optimized firmware along with robust backup circuits for power loss protection. These circuits and code protect stored data against possible corruption (due to unexpected power loss). They also enable the drive to ensure that all the data in the SSD's volatile memory is written to non-volatile NAND FLASH memory.

Performance and price are major considerations when choosing to place a client SSD in a data center application. However, other factors, such as Power Loss Protection, Data Path Protection, and media Endurance are also important, and should be considered during the design-in phase.

This protection also allows the drive to quickly boot on the next power-on cycle after an unexpected power loss. However, this comprehensive combination of firmware and hardware protection adds cost to Enterprise-class SSDs.

Client SSDs, SSDs designed for desktop and notebook computing, are designed with lower-cost power-loss protection. On client SSDs, data which has already been committed to NAND is fully protected from power loss events just as on enterprise-class SSDs. However, any data in the SSD's volatile memory may not be protected.

The amount of data in the Micron 1100's volatile memory is typically small (on the order of 1MB or less). From the Client SSD user's perspective, losing this amount data may be a familiar event to the computer user. It may be experienced as a loss of the last couple of sentences typed in a document when power is lost. Modern office efficiency applications often manage events like this by regularly journaling file changes in real time, allowing the user to recover unsaved changes.

This design characteristic of the client SSD means that the data center architect should consider power source protection, such as UPS as well as data redundancy practices.

Please see the Micron White Paper, *How Micron SSDs Handle Unexpected Power Loss*, for more details on these levels of power loss protection for client and enterprise SSDs.

Data Path Protection

During the normal course of operations, data moves to and from the interface through assorted electronic components to storage locations along what is known as the data path. At various points along this data path, and especially in components such as the DRAM, rare, naturally occurring events can cause disturbances which can change a "1" signal to a "0", or vice versa, seemingly spontaneously. Although rare, an event like this may have negative consequences, so detecting and compensating for such events is very important.

Competing interests (cost versus intended use, design versus actual deployment and a host of others) change the level of protection for the data path for client and enterprise SSDs. Fortunately, with regard to the Micron 1100 SSD, Micron's engineers have chosen to deliver the same levels of data path protection for this client SSD as delivered for concurrently delivered for Micron's data center SSDs, such as the Micron 5100. With a combination of memory-path error correction code (MPECC), firmware and error correction hardware, the 1100 SSD is capable of detecting 2 bits in error, and correcting 1 per I/O transfer, throughout the data path.

We discuss design consideration for data path protection in greater detail in the Micron Technical Marketing Brief: *A Comparison of Client and Enterprise SSD Data Path Protection*.

Endurance

Client SSDs are designed for cost-effectiveness in environments where they are unlikely to be writing more than 20GB to 50GB per day and will usually not be under workload on a continuous, 24-hour, 7-day a week basis. Therefore, the NAND components for these SSDs may be of lower endurance than the NAND chosen for enterprise-class SSDs. If we put a Client SSD under the same workload as an Enterprise SSD, the Client SSD would have a far shorter product lifespan.

Because of this, it is strongly recommended that these Client SSDs be designed into systems which are expected to deliver workloads highly weighted towards reads rather than writes. A 90:10 or 95:5 read-to-write ratio is recommended, even with the added OP conditions discussed here.

Performance Considerations

In examining data sheets, it's clear that Client SATA SSDs have outstanding MB/s and IOPS write performance, with sequential write (128kB transfers) regularly exceeding 500 MB/s, and random write performance (4kB or 8kB transfers) approaching 100,000 IOPS. However, because of the expected workload, this specified performance is measured when the Client SSD is in the Fresh out of Box (FOB) performance state. In desktop applications, this is appropriate, because the SSD will rarely experience a long, sustained workload, and so will almost always remain at or near this FOB state. For more information about SSD performance states, please see this paper on micron.com.

Enterprise SSDs may list somewhat lower performance numbers on data sheets, but not because they're actually slower. Rather, this is due to performance in the enterprise measured at a "steady state." This is a performance state where the SSD experiences a little performance change as a consistent workload is applied. Steady-state is defined by the Storage Networking Industry Association (SNIA) in their Performance Test Specification, Enterprise.

When comparing client and enterprise SSDs, it is critical to understand that client SSD performance is specified in the "Fresh-out-of-Box" performance state, while enterprise and data center SSDs are specified in "Steady State."

We will discuss and show steady state performance data under various conditions here, but for a more detailed discussion on the performance states of SSDs, see Micron's technical marketing briefs, *Differences in Personal vs. Enterprise SSD Performance*, and *Best Practices for SSD Performance Measurement*.

What is Over-Provisioning?

Over-provisioning (OP) on a data storage device refers to storage capacity that is not addressable by the operating system. The device may use this extra space to optimize internal processes. All NAND FLASH-based SSDs contain some level of OP. Importantly, OP will allow an SSD to maintain a certain level of steady state write performance. As a rule of thumb, less OP means lower steady state write performance, while more OP enables higher steady state write performance. However, as the amount of OP increases, it may be that points of diminishing returns can be found (beyond which more OP may not further improve write performance, or the improvement may not be economical).

Over-Provisioning, or OP, is any unused space on an SSD, which can be used to optimize background operations. Formally, the OP space should be un-addressable by the host computer, but any unused space will be used by the SSD for this purpose.

SSDs designed for client computing typically have minimal OP because of the light write performance demands of desktop computing and to efficiently manage cost for consumer-level applications. Additionally, in client applications, it is not unusual for a computer user to operate the drive at something less than 100% of available capacity filled.

In this case, the unused space can be used by the drive to optimize its internal operations; any unused space can be considered as OP even if it is addressable by the operating system.

Enterprise SSDs are different in that they tend to have more built-in OP to improve steady state performance when the addressable user space is full or nearly full. This higher level of OP native to Enterprise SSDs is the main reason for their improved steady state write performance. The implication is that reserving normally addressable NAND storage space on a client SSD can allow a less expensive drive to improve its write performance, even approaching that of an Enterprise SSD.

The 1100 SSD can be adjusted in the field to increase its steady state write performance by providing some additional OP. We will discuss how to do this and show estimates for steady-state performance improvement under synthetic workloads.

OP and Steady State Write Performance

SSD performance changes as they are written. It is important to note that this is a phenomenon for writes, and that reads suffer far less variability over time. Also, this change in performance is far more dramatic for small-block, random transfers than it is for large-block, sequential transfers.

In describing the performance of write operations, we define the key performance states for write workloads as follows:

- **Fresh-Out-of-Box State**: The performance state when a drive is as-shipped from the factory, or after SECURITY ERASE or SANITIZE BLOCK ERASE commands are used to purge user data. *This is the performance state in which client SSDs are specified.*
- **Transition State**: The performance state where the SSD is transitioning between FOB and steady state. This state can be quite variable in performance characteristics.
- **Steady State**: The state where data throughput speeds reach a point with less than +/- 10% variability over a designated time interval. The time to reach steady state varies with the drive being tested and the nature of the workload. *This is the performance state in which data center and enterprise SSDS are specified.*

For purposes of this discussion, we will consider the specific example of a 1TB-class 1100 SSD whose native capacity is 1024GB. This SSD’s performance will be evaluated at its native capacity (1024GB, with 0% OP in addition to the native OP), and at reduced capacities of 1000GB (2.3% OP) and 800GB (21.87% OP). These capacity points are intentionally chosen to illustrate the effects of “a little more” OP can do versus “a lot more” OP.

Although not discussed here, the arguments would be similar for other 1100 SSD native capacities, including the new 2TB product with OP selected at similar percentages. Again, these are estimates, not specifications.

Dynamic Write Acceleration for Client SSDs

The Micron 1100 SSD is designed with a feature called Dynamic Write Acceleration (DWA). DWA is a proprietary method of enhancing the FOB performance of Micron’s TLC 3D NAND by using portions of the NAND array in Single Level Cell (SLC) mode. DWA dynamically apportions parts of the NAND array in SLC mode to write new data from the host. This can be viewed as an internal write buffer, varying in both size and physical location within the array depending on current conditions of the drive and the environment.

DWA allows the drive to deliver outstanding SLC-like FOB performance under many client workloads while maintaining the capacity and cost-effectiveness of TLC NAND. However, DWA does exhibit some higher level of performance variability under some workloads, particularly in the Transition State, as we’ll describe.

Performance Under 4k Random Write Workload

A constant 4k random write workload is not one for which the 1100 SSD is ideally suited. However, this rugged workload is useful to illustrate the performance states of the 1100 SSD and to exhibit the performance effects of additional OP. For all of the data presented in this paper, we presume that Native Command Queueing is enabled with a fully utilized command queue (queue depth 32).

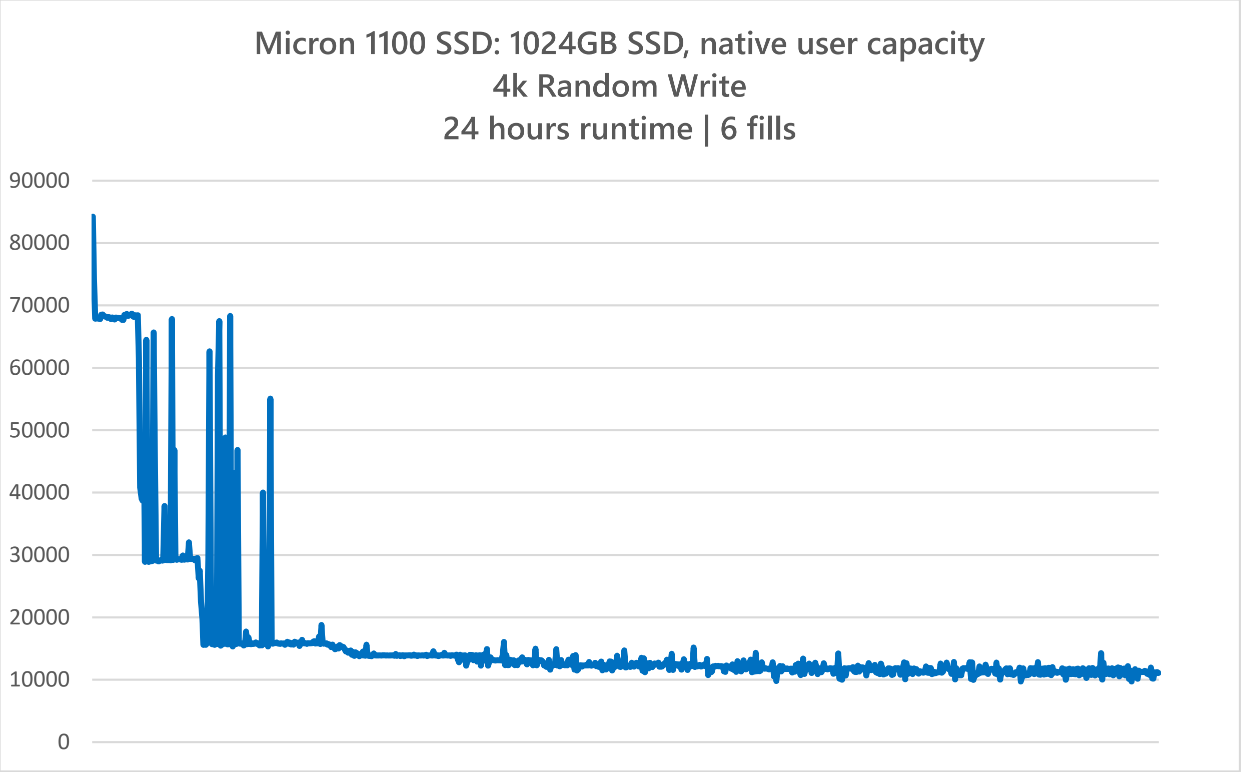


Figure 1: Write Saturation: 4K Random Write IOPS on 1024GB 1100 SSD at Full Capacity

It is very important to note that the 1100 SSD, in its normal client environment, is designed to run with the write cache enabled. Data storage system designers may be tempted to disable write cache in order protect against data loss during unexpected power loss events. However, *disabling write cache for this model may result in drastically reduced performance and endurance.*

Figure 1 shows a 4K random write saturation plot (measured over a period of roughly 22 hours, with total transfers equivalent to about 6 times the capacity of the SSD). The FOB performance is at the left side of the plot, measuring over 80,000 IOPS. IOPS drops through a range of high variability between 70,000 and 18,000 IOPs in the Transition state. This region of high variability is expected as it shows the effect of Micron's DWA – the dynamic management of the SLC cache. Once through the Transition state, the 1100 SSD settles into a Steady State performance level at approximately 11,000 IOPS.

This is a rather drastic decline in performance, but is expected in a Client-class SSD.

Now, let's look at the effect of an OP which sets the user capacity down to 1000GB, with 24GB, or an additional 2.3%, of space set aside. This is shown in Figure 2.

This seemingly small amount of extra space can have a measurable effect on performance. Now, the Steady State portion of the saturation curve is flatter and more consistent. This flat portion of the curve sits at roughly 13,000 IOPs, compared to 11,000 for the native capacity – a small, but notable improvement.

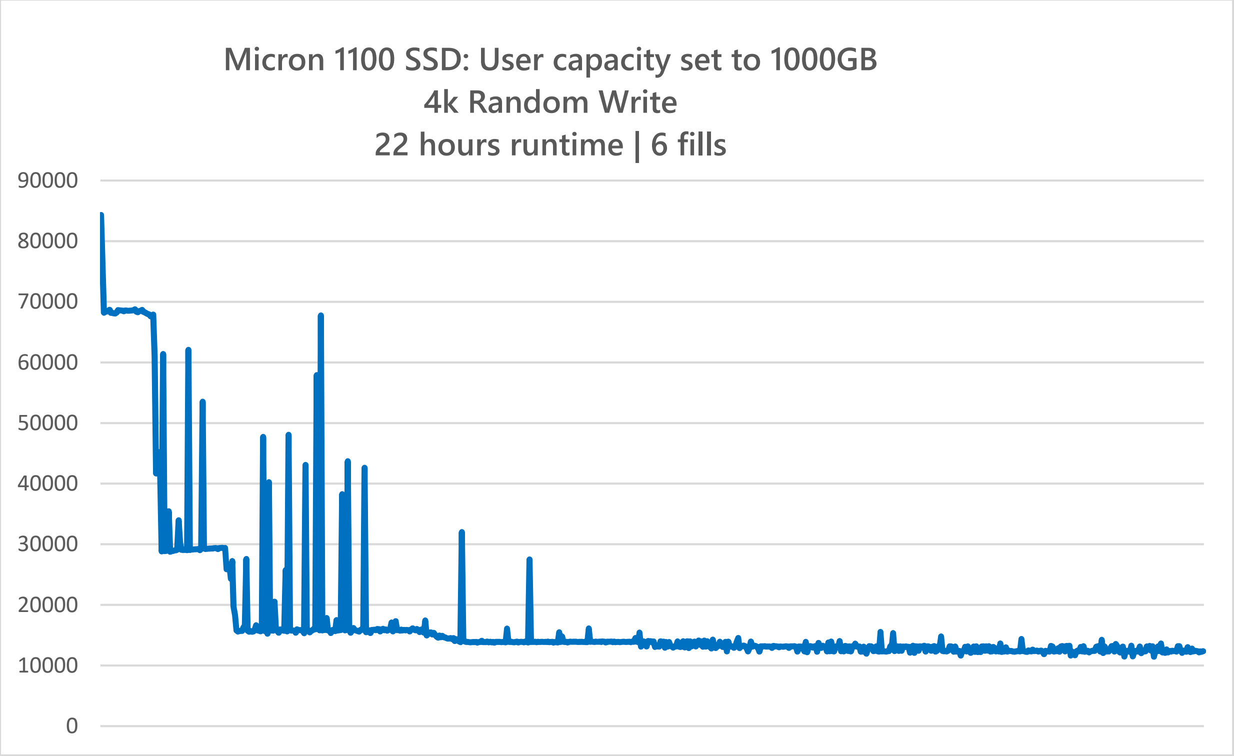


Figure 2: Write Saturation: 4K Random Write IOPS on 1024GB 1100 SSD at Full Capacity

Next, we'll take a larger increase in OP and reduce the usable capacity to 800GB.

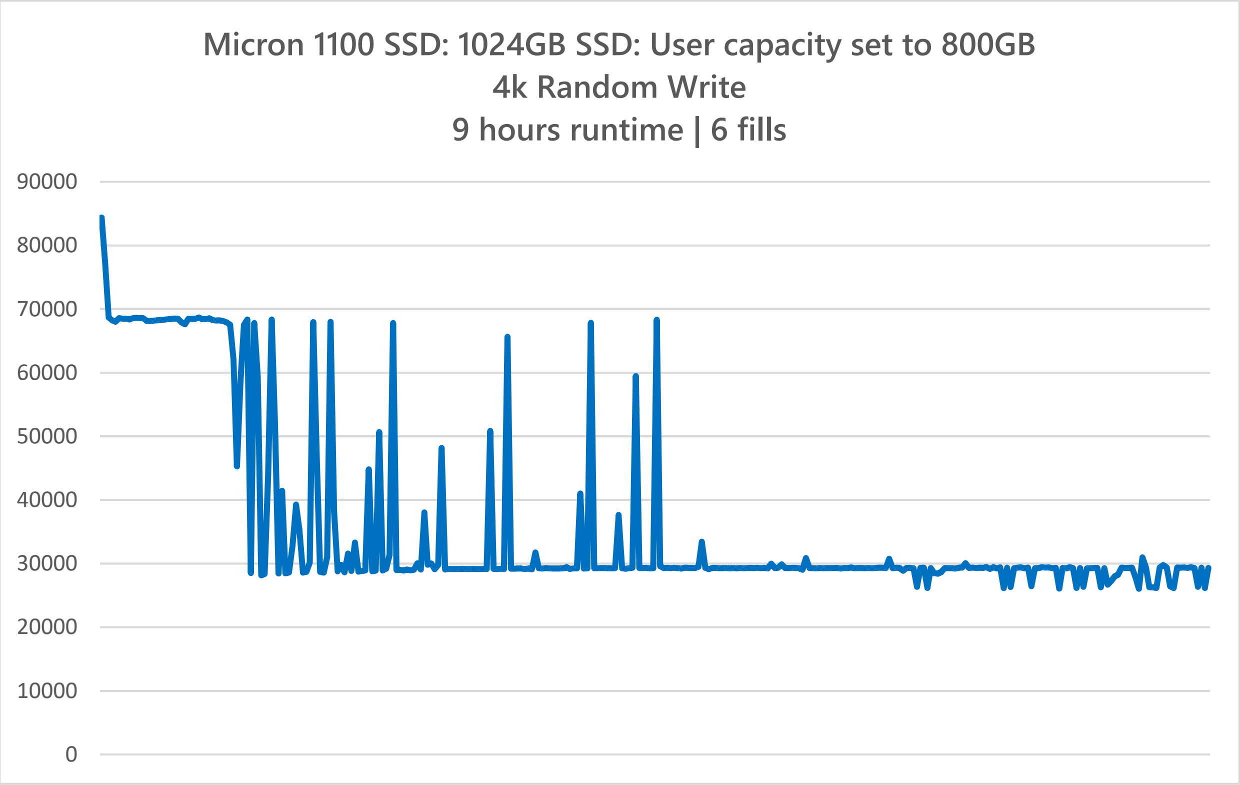


Figure 3: Write Saturation: 4K Random Write IOPS on 1024GB 1100 SSD at Full Capacity

Figure 3 shows the write saturation plot for this condition, and shows a more marked improvement.

With the 21.9% additional OP, the steady state part of the curve is very flat and consistent. The drive is showing a steady state performance at approximately 29,000 IOPs*, nearly three times the steady state performance of the drive at its native capacity*, although requiring the sacrifice of some addressable capacity.

OP can improve Write Amplification by allowing garbage collection to work more efficiently. Other factors that can affect "WAF" are the randomness of transfers (sequential is better), the size of transfers (larger is better), and the alignment of writes ("4k-alignment" is best).

OP and Mixed Read/Write Performance

Having considered performance in 100% write environments, it's critical to understand that such environments are quite rare in real deployments, and not recommended for a client-class drive like the 1100 SSD. For an 1100 (and other similar drives), reads should be much more prevalent than writes.

Consider the example of a content delivery network (CDN) application. Content delivery services may periodically store (write) new video or music content, for example, after which the system spends the rest of a month experiencing almost exclusively read transactions from customers out on the internet.

So, moving from the technical example of a 100% write theoretical workload to a 95% read workload is a good representation of an application in which the 1100 SSD may excel.

At any percentage of write workload, even as low as 5%, the drive can eventually be put into steady state, but the transition state may last longer. As with a 100% write workload, the steady state throughput can be increased by providing some level of OP. However, at these low write rates a large OP may not be necessary.

OP and Typical Write Endurance

Since NAND FLASH blocks can only be written after they are erased, flash-based SSDs may need to rearrange stored data to provide erasable (and then writable) cells. We call the process of rearranging data and erasing space to prepare for new writes, "garbage collection." Garbage collection will cause additional program operations, and will increase the total amount of writes beyond that required only by write commands from the host computer.

The amount of increase is described by an SSD attribute called Write Amplification Factor (WAF), given by the sum of host writes and background writes, divided by the number of host writes. These numbers may be measured in bytes written or in NAND blocks written, with the result being a ratio greater than 1.0.

For example, if background operations cause an additional 75% write workload beyond host write requests, the WAF will be 1.75.

More detailed instructions for calculating WAF are available in Micron's Technical Note, *SMART Attribute: Calculating the Write Amplification Factor*.

With increased over-provisioning, internal operations in the drive become more efficient and therefore produce a lower WAF. Importantly, WAF grows very rapidly when the percentage of capacity used exceeds 90% for client SSDs with very little native OP. Keeping available capacity on any client SSD at *a minimum* of 10% is a good rule of thumb for any workload including both the data center applications in this discussion and desktop applications.

Write amplification is the direct cause of decrease in performance at steady state, as background writes steal bandwidth from input/output operations with the host computer. The implication of this, along with the data

Decreased performance when used capacity exceeds 90% is not unique to SSD. In fact, the Windows operating system will caution the user when used space exceeds 90%, encouraging efficient use of storage space, regardless of the device type.

presented here regarding steady state performance is that setting aside 20% or so of space for OP will result in lower WAF and will therefore improve performance and long-term reliability because each cell in the NAND FLASH array receives less overall writes.

Micron SSD data sheets specify "endurance" as part of the warrantable performance of the drive. Endurance may be defined as the total number of bytes written to the drive over the course of its service life. For client SSDs like the Micron 1100 SSDs, this total-bytes written (TBW) value will typically be measured in tens or hundreds of terabytes (10^{12} bytes). Although product warranties are bounded by calendar time, write endurance, viewed in isolation, is largely independent of time. So, a drive which only has 20GB of data written to it per day will last very much longer in calendar time than a similar drive which experiences 200GB of writes per day.

Performing OP Using Micron's Storage Executive CLI

Configuring over-provisioning for an SSD consists of two steps:

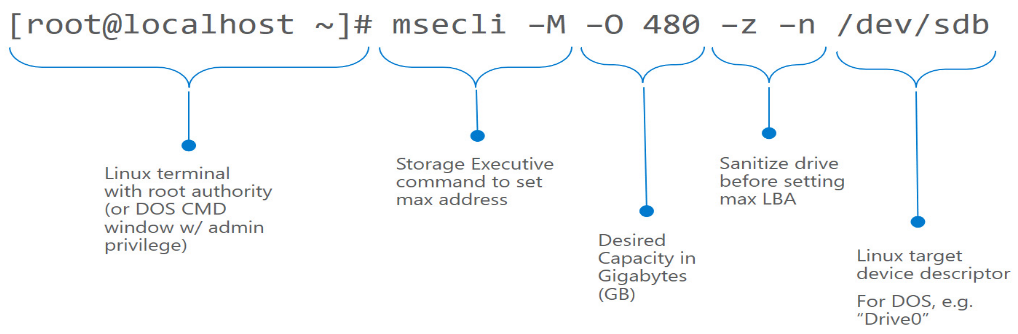
1. **Purge**: Remove all user data on the drive by issuing a SECURITY ERASE or SANITIZE BLOCK ERASE command, resetting the drive to FOB state.
2. **Set a new max address**: Consult Table 1 below, selected max address configurations, and see the IDEMA specification for LBA counts that are generalizable to any capacity point.

Micron's Storage Executive software provides a method for performing each of the steps above for SSDs across Micron's product portfolio. A single command allows Storage Executive to perform the same function on Micron's most current products (supporting the Accessible Max Address Configuration, AMAC, protocol) and on Micron's legacy products (supporting the Device Configuration Overlay – DCO – protocol). It is possible to use third-party tools for this function. However, Storage Executive makes this protocol determination automatically.

The structure of the command in the command line interface (CLI) in Storage Executive is shown in Figure 4. MSECLI stands for Micron Storage Executive Command Line Interface, shown below in Linux, with information for Windows CMD in parenthesis.

The operation to set the max address may only be performed once per drive power cycle. Any attempts to set a max address more than once during the same power cycle will be aborted by the drive.

To change a configuration on a drive which has already been over-provisioned, simply set a new max capacity as necessary. The drive may be returned to the native OP configuration by using the max addresses for 100% native capacity.



```
[root@localhost ~]# msecli -M -O 480 -z -n /dev/sdb
```

Linux terminal with root authority (or DOS CMD window w/ admin privilege)
Storage Executive command to set max address
Desired Capacity in Gigabytes (GB)
Sanitize drive before setting max LBA
Linux target device descriptor
For DOS, e.g. "Drive0"

Figure 4: OP Example: 512GB SSD Set to 480GB Accessible Capacity Using Linux

Note that the command argument which determines the desired user capacity is -O, with the upper case letter O. Using the command argument -o (lower case o) will allow the user to define the user capacity with regard to "max LBA," the maximum LBA (logical block address) number for a desired capacity. This is a more complicated method, but is acceptable, given the correct LBA number.

Table 1 shows the proper setting for the maximum LBA for several common capacity settings. Any arbitrary capacity can be selected using the LBA calculations available from the International Disk Drive Equipment and Materials Association (IDEMA) at idema.org (link to PDF).

Desired Capacity in GB	IDEMA LBA Count	IDEMA Max LBA	IDEMA Max LBA (hex)	User Available Bytes
120	234,441,648	**234,441,647**	DF94BAF	120,034,123,776
128	250,069,680	**250,069,679**	EE7C2AF	128,035,676,160
240	468,862,128	**468,862,127**	1BF244AF	240,057,409,536
256	500,118,192	**500,118,191**	1DCF32AF	256,060,514,304
480	937,703,088	**937,703,087**	37E436AF	480,103,981,056
512	1,000,215,216	**1,000,215,215**	3B9E12AF	512,110,190,592
960	1,875,385,008	**1,875,385,007**	6FC81AAF	960,197,124,096
1500	2,930,277,168	**2,930,277,167**	AEA87B2F	1,500,301,910,016
2000	3,907,029,168	**3,907,029,167**	E8E088AF	2,000,398,934,016
2048	4,000,797,360	**4,000,797,359**	EE7752AF	2,048,408,248,320

Table 1: Select IDEMA Max LBA Settings for Popular Capacities (SATA – 512 Bytes Per Sector)

Please note that the Micron 1100 SSD supports 512-byte sectors. Other Micron products, particularly SAS and PCIe/NVMe SSDs, support 4096-byte sectors, or other configurations. The GB-to-LBA calculation for so-called "4k sector" SSDs will be different, and Table 1 is not applicable to these SSDs. Therefore, it is easiest to simply define the OP setting using the -O argument with the targeted user capacity.

Achieving OP Through Other Means

Implementing OP through the Storage Executive CLI tool is the most robust and permanent way for users to implement OP. However, other methods may be used if they are implemented correctly and make sense from an application perspective.

Because OP is essentially spare capacity, any process that manages the amount of user data on the drive could potentially be manipulated to produce OP.

A purge (Security erase, or Sanitize) should be performed prior to attempting any of the methods described below. Note that a purge will irretrievably delete any existing data on the SSD.

OS Free Space: OP may be dynamically provided by maintaining free space from an operating system perspective. Any method by which the specified percentage of available space on the drive or array is maintained will work effectively. However, some discipline must be enforced on the user so this can have variable results in the long term.

RAID Free Space: In a hardware RAID environment, OP may be achieved with configuration settings to leave capacity on every drive, un-assigned to any volume, virtual drive, or LUN. One drawback to this method is that a system administrator may alter this configuration and reduce or eliminate the reserved OP.

File System Partitions: Setting data partitions to a capacity less than the native capacity of the drive can be a very effective means of providing OP. In fact, this is the method that Storage Executive uses in its GUI implementation. However, this method does still require some discipline on the user or system administrator to leave any unpartitioned space completely unused from an operation system level. Any tampering with nominally unpartitioned space could lead to performance degradations which are difficult to troubleshoot.

Conclusion

SSDs targeted for the client market such as the Micron 1100 tend to have minimal amounts of native OP. These SSDs can deliver outstanding performance in typical desktop and notebook applications, but this minimal native OP limits their write performance and endurance characteristics in typical data center applications. However, providing additional over-provisioning to the 1100 SSD can enable their use in some data center applications, especially those that are read-intensive.

To use the 1100 SSD in the data center, it is important to understand the characteristics of the targeted workload with regard to the write/read balance, the amount of data written per unit time, and the behavior of the 1100 SSD under the specific workload given any particular OP setting.

This paper has described approximate behavior of the Micron 1100 SSD at particular OP settings, under a synthetic 4k random write workload, to be used as a guide for further study by the system designer. Although additional OP can improve steady state performance and NAND FLASH endurance, this should not be construed to change any warranty provisions regarding the lifetime of the product.

micron.com/ssd

All information is provided on as "AS IS" basis without warranties of any kind, including any implied warranties, warranties of merchantability or warranties of fitness for a particular purpose.

No hardware, software or system can provide absolute security and protection of data under all conditions. Micron assumes no liability for lost, stolen or corrupted data arising from the use of any Micron product. Products are warranted only to meet Micron's production data sheet specifications. Products, programs and specifications are subject to change without notice.