A guide to understanding differences in performance and use cases

Client solid state drives (SSDs) — those designed primarily for personal computer storage — are suitable in some, but not all, data center applications. Data center SSDs, on the other hand, are designed from the ground up for data center use.[1]

When considering the use of a client SSD in a data center application, it is imperative to understand the input/output operations per second (IOPS) performance and design differences between the two.

This technical brief discusses some of these differences.

Different SSDs for different applications

SSD designers optimize performance and cost based on intended use. Directly comparing SSDs designed for different uses (when examining data sheets, for example) can be difficult. It is like comparing different products intended for fundamentally different uses.

We can make more informed decisions when we understand some of the performance implications of using a client SSD in an application for which it was not designed.

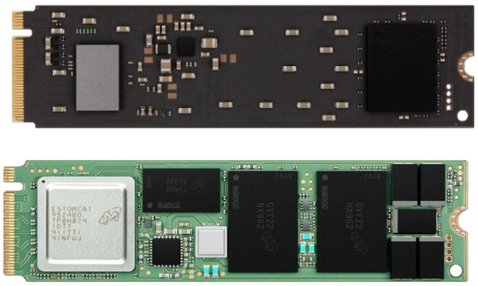

Figure 1: Client and data center SSDs may look similar[2]

Focus Areas

The technical brief highlights some of the differences between SSDs designed for client use and data center use. It is designed to help SSD users make informed choices about which type of SSD they deploy for which application.

Performance evolution

It is generally known that SSD performance changes with time. As the SSD migrates from fresh-out-of-box state to steady state, its write performance evolves. The performance evolution patterns may differ between client and data center SSDs.

This paper offers an example of this evolution difference and describes some of the background reasons this difference occurs.

Over-provisioning

Where there is more physical NAND capacity on the SSD than there is advertised storage capacity, that SSD is over provisioned.

Over provisioning helps optimize some required background tasks like garbage collection as well as write performance.

Different over provisioning levels and the resultant effect on an example workload is described.

Power loss protection

Power loss protection (PLP) is designed to protect data being written to – and already written to – the SSD from sudden power loss – including data that has successfully be written to the NAND.

Client and data center SSD PLP is escribed including which portions of the client and data center paths are protected, as well as why client and data center SSDs offer different PLP implementations.

1. Statement refers to intended design use and does not reflect actual suitability for either SSD type in any use case.
2. Representative examples only. Figures may not be to scale.

Consider an IOPS performance comparison between a client SSD (optimized for personal storage such as mobile computing) and an SSD optimized for data center use (such as highly active real-time databases). Because data center SSDs are designed for demanding workloads like this (and client SSDs are not), we expect the data center SSDs to excel (while the client SSDs may not).

A common test illustrating this point is a 4KB random 100% write workload over an extended period.

Figure 2 shows how the performance of each SSD type changes with time. FOB is "fresh out of box," meaning the SSD has experienced little to no data written to it. Steady state is the performance state where performance changes little with time.[3]

Each SSD's IOPS are shown on the vertical axis while time is shown on the horizontal axis.

Although the exact shape of these curves may change with different SSDs and workloads, all SSDs undergo this performance change. With this example workload, the data center SSD shows higher steady state performance. Steady state write performance is an important factor for data center customers.

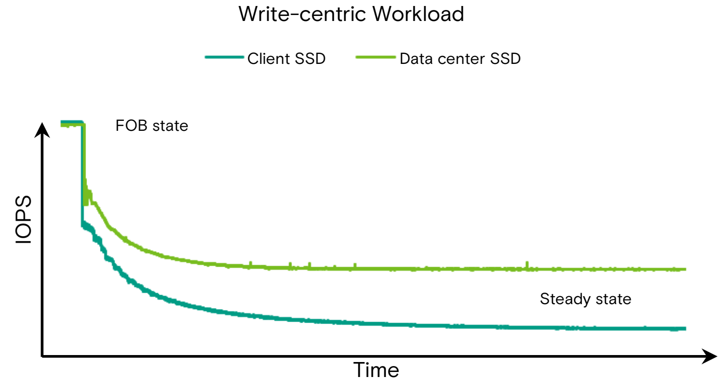

Figure 2: SSD performance over time[4]

It is important to note that the comparison in Figure 2 is only one aspect of drive performance. It is not a complete representation for all applications, uses, or standard benchmarks. It illustrates that good performance is relative to the target application and use.

Factors affecting write performance: Understanding over-provisioning

Over-provisioning is additional media space on an SSD that does not contain user data. Every SSD has some level of over-provisioning.[5] Figure 1 shows the 4K random write performance of a client and a data center SSD over time. The data center SSD has considerably more over-provisioning. That additional media space plays a critical role in steady state random write performance.

This section explains why.

Introduction to garbage collection

When NAND (the media used in the SSDs discussed here) has been written, the media must be erased before it can be rewritten. This is different from hard disk drives (HDDs). HDDs use "write in place" media. If the HDD media already contains data, we can overwrite the data in a single step. NAND takes two steps (erase and write).

NAND is organized by pages (the smallest portion that can be written) and blocks (the smallest portion that can be erased). Blocks contain many pages (the exact number depends on the NAND design). When we want to erase a NAND page so we can write new data to it, we cannot erase just that page — we have to erase an entire block. If the block has some data we want to keep, we have to move that data by writing it somewhere else on our SSD before we erase the block.

A process known as garbage collection accomplishes this in two steps. The first step identifies the data we want to keep and moves it to a free location on the SSD. Once complete, the second step erases the block to produce pages to which we can write new data.

3. See https://www.snia.org/sites/default/files/technical-work/pts/release/SNIA-SSS-PTS-2.0.1.pdf for additional details on SSD performance states.
4. Representative example. May not be indicative of all SSDs of either type.
5. For more information on over provisioning, see https://www.snia.org/sites/default/files/SSSI/NVMe_SAS_SATA_Endurance_White_Paper.pdf

The example in Figure 3 helps illustrate garbage collection on a hypothetical client SSD. This example contains 256 NAND pages, shown as squares (real SSDs have far more NAND pages), and each column of cells represents a block.

The green squares represent pages with data we want to keep. The black squares are pages that are ready to receive new data. The blue squares are pages with data that we need to keep, but that we also need to move to be able erase the block without losing (erasing) any of this data.

This example client SSD uses minimal over provisioning.

In this example, the SSD must move the data in the blue cells before it erases the block (column). Note that there are few areas into which the data can be moved (black cells). This is due to limited over-provisioning in this example.

Figure 4 shows a similar example but with an SSD that has typical data center over provisioning level. As before, this SSD must first copy the data we want to keep into new pages so it can erase the block.

In the data center example, the SSD has far more choices where to move blue squares before erasing the block (over-provisioning effectively enables more black squares). This enables better optimization, making garbage collection more efficient.

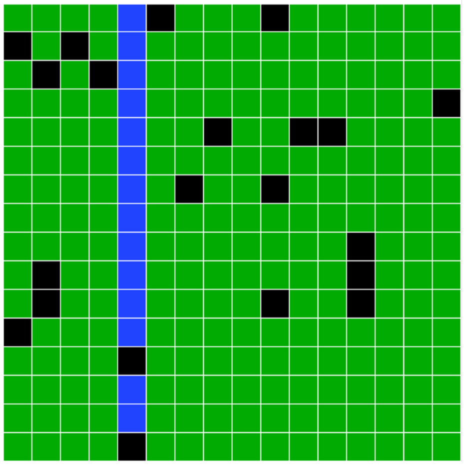

Figure 3: Garbage collection example with typical client SSD over-provisioning

Over-provisioning and random workloads

SSD over-provisioning is calculated as a ratio and expressed as a percentage – we can see the effect over-provisioning has on write IOPS performance when we adjust over-provisioning on the same data center SSD, applying the same random workload iteratively.

Figure 5 shows how different over-provisioning levels can affect IOPS performance. In the example, we performed the same test on the same data center SSD containing the same firmware installed in the same system. We only varied the level of over-provisioning (OP).

For these tests:

$$\frac{\text{Total amount of NAND on the SSD}}{\text{Total amount of NAND available for data storage}}$$

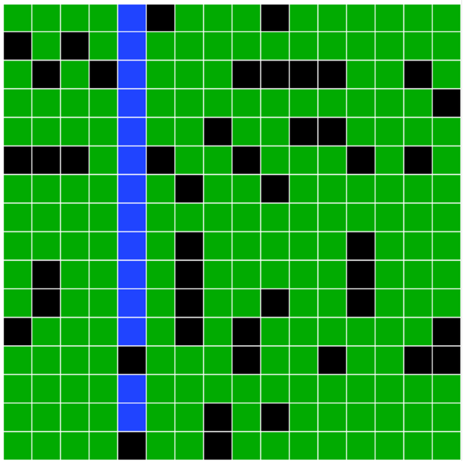

Figure 4: Garbage collection example with typical data center SSD over-provisioning

- We restored the SSD to FOB before we started each test and applied a small transfer, random, mixed IO workload.
- We started with the default capacity (blue) and then increased the over-provisioning using Micron's Flex Capacity feature to +17% (over default) and then +50% (over default).

Figure 5 shows the test results:

- Additional over-provisioning increases the IOPS performance at steady state.
- It does not affect IOPS performance at FOB.

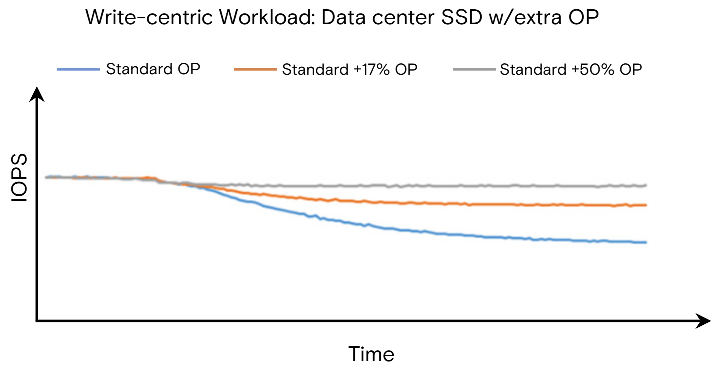

Figure 5: Effects of additional OP in a data center SSD[6]

6. Representative example. May not be indicative of all SSDs of either type. Source: Micron SSD applications engineering.

The values may change based on the SSD and workload evaluated. The relative results and overall principle remain the same: Increasing over-provisioning (even on a data center SSD) can improve IOPS performance for workloads with a significant write component (mixed I/O). Here is why: As the write amplification[7] decreases, the random steady state performance improves. This is because of the improvements in garbage collection efficiency, as discussed in the previous section.

Over-provisioning and sequential workloads

Sequential workload IOPS performance is affected far less by changing over-provisioning levels compared to random workloads. This is because sequential workloads place the data in a more orderly manner as they write it.

Figure 6 illustrates this process. Using the same hypothetical example SSD, Figure 6 shows an example of data placed by a sequential workload. Because the data is more orderly (compared to random workload placement), garbage collection does not happen as frequently.

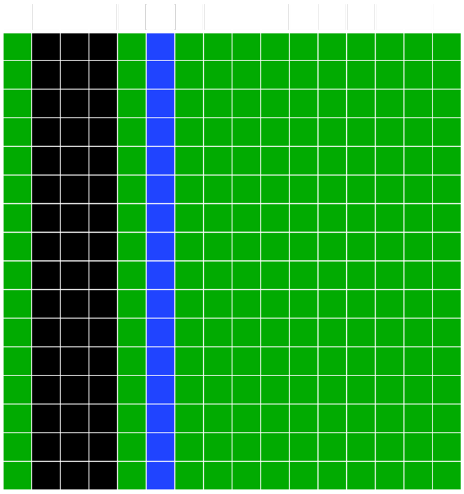

Figure 6: Garbage collection on a sequential workload

Both client and data center SSDs typically show good sequential workload performance.

Write buffering and steady state performance

Traditionally, write buffering has been used to increase instantaneous, or burst, I/O performance. Incoming write traffic is buffered into fast storage (usually DRAM) and then migrated to slower, long-term storage (NAND). Because buffers are typically limited in size, they are not a major factor in steady state performance. Once the buffer fills, it brings no benefit (to absorb an incoming write, we must drain data from the buffer into the NAND).

For client and data center SSDs, the write buffer may improve steady state, random IOPS performance. This is because SSDs extensively use parallelism to improve IOPS performance. If we can increase parallelism, we increase IOPS performance.

One method for increasing parallelism is write accumulation. Write accumulation is a process by which several smaller write operations are combined into a larger write operation across multiple physical NAND die.

This process optimizes write operations: It enables the greatest amount of data to be written with the least amount of media busy time. To take advantage of write accumulation, the SSD must have some form of write buffer in which to accumulate write commands.

Although client and data center SSDs can use this technique, the exact implementation may differ. Micron data center SSDs have stored energy to write all the data in a write accumulation buffer to NAND should the SSD lose power (due to sudden removal, for example). Without a power protection mechanism, this sudden power-loss may result in data risk.

Typical client SSDs do not have this capability. This is because in conventional personal storage applications such as personal computing, this difference is inconsequential. (The SSD cannot be removed without powering the system down. If it is, the operating system also halts because it, too, is stored on the SSD.) One may disable the write buffer on some client SSDs, but performance may be reduced.

Power-loss protection[8]

Client and data center SSDs both use nonvolatile NAND memory for long-term data storage. Different types of NAND store a different number of bits in each cell. For example, triple-level cell (TLC) stores 3 bits per cell while quad-level cell (QLC) stores 4 bits per cell. The more bits per cell, the higher the NAND (and SSD) potential density.

TLC and QLC NAND have some characteristics: these devices using these NAND types can be vulnerable to data loss in the event of an unexpected power loss for the SSD.

7. Write amplification is the extra writes into the flash storage due to background processes. For more information see: https://www.snia.org/education/online-dictionary/term/write-amplification
8. PLP examples are described. Different SSDs may implement PLP differently.

Client and data center SSDs may have various levels of power-loss protection (PLP). Client SSDs protect data at rest. Data center SSDs protect data at rest and data in flight. "Data at rest" is data that has been successfully written to the storage media. "Data in flight" refers to data that has been sent to and acknowledged by the SSD (but may not yet be committed to the media, such as data temporarily buffered in volatile memory) or any write that is in progress but not yet complete.

Client SSD PLP — Data at rest

For many client SSDs, data at rest protection is usually sufficient. Figure 7 shows typical client SSD PLP for a DRAM-less design. Client PLP only protects data at rest (data that has already been written to the NAND), shown in the portion of the SSD surrounded by green. Data in flight is not protected against sudden power loss.

Figure 8 shows typical client SSD PLP for an SSD with DRAM. As with DRAM-less SSDs, PLP only protects data at rest (data that has already been written to the NAND), shown in the portion of the SSD surrounded by green. Again, data in flight is not protected against sudden power loss.

In both types of client SSDs, the SSD controller SRAM is not protected against PLP.

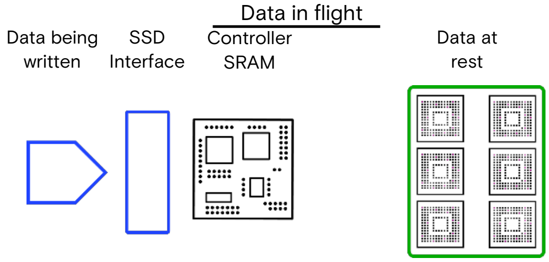

Figure 7: PLP Client SSD with no DRAM (DRAM-less)

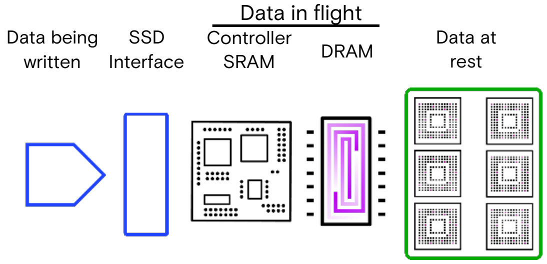

Figure 8: PLP Client SSD with DRAM

Data Center SSD PLP — Data at rest and data in flight

Figure 9 shows an example of data center PLP which extends from the NAND (as in client PLP), through the DRAM buffer, and to the SSD controller's SRAM. This PLP protects committed writes not yet stored in nonvolatile memory, as well as writes to nonvolatile memory already in process and in the controller's SRAM.

Data Center SSDs have extended PLP because data loss in the data center is more critical than in client computing. Client devices are typically single user, so while data loss protection is important, it affects only one user. Modern desktop applications are often able to compensate for this small risk by journaling the user's activity so that unsaved changes can be recovered in the event of an unexpected power loss.

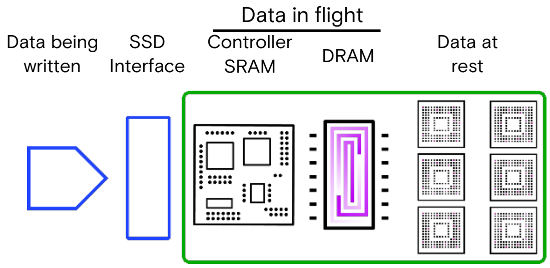

Figure 9: PLP Data center SSD

On the other hand, data center SSDs are often installed in platforms supporting hundreds of users and mission-critical systems. Data loss here potentially affects hundreds of users or more and can have greater consequence. With data center SSDs, is it essential to protect data at rest, like in client SSDs, but also data in flight. Any writes in progress must be completed, and any data buffered in volatile memory must be committed to the NAND device and protected.

Summary

Many factors affect SSD performance in a given application. How the application accesses the SSD (randomly or sequentially) can influence SSD IOPS performance, as can the basic design of the SSD itself. It is important for system designers to understand some of the key differences between client and data center SSDs to ensure an optimal fit for their use models.