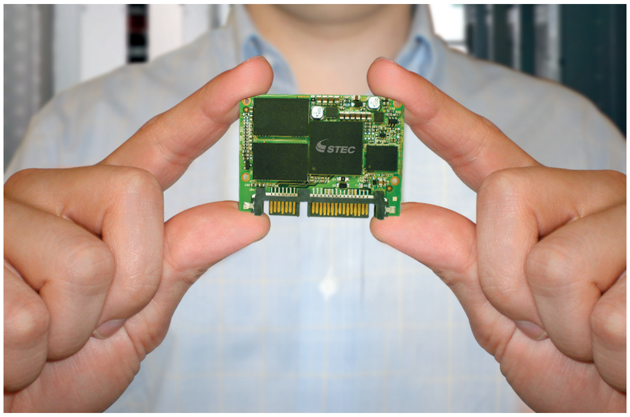
Benchmarking Enterprise SSDs



When properly structured, benchmark tests enable IT professionals to compare solid-state drives (SSDs) under test with conventional hard disk drives (HDDs) and other SSDs.

STEC®
The SSD Company™

Contents

Executive Summary

When properly structured, benchmark tests enable IT professionals to compare solid-state drives (SSDs) under test with conventional hard disk drives (HDDs) and other SSDs. To be truly useful, SSD benchmarks should show the performance of the SSD under the kinds of heavy workloads that are typically encountered in enterprise applications.

In addition, the benchmark should evaluate performance not only under optimum conditions, but over the entire life of the drive. Some SSD benchmark scores that appear very fast can be misleading because they do not reflect long-term performance. To be of real value, SSD benchmark tests need to measure the stability of consistent performance over time. This requires preconditioning the SSD under test, accomplished by writing random data patterns to completely fill all NAND blocks and engage the drive's wear-leveling and flash management routines. Properly managing data flow and internal NAND will make the benchmark a more useful gauge of SSD performance under real-world conditions.

Well-designed benchmark tests provide a level playing field and help simplify the evaluation of SSDs. The purpose of this paper is not to provide actual benchmark data, but to present best practices for benchmark tests that reflect SSD performance in the enterprise environment.

Introduction

Traditional benchmark tests used to evaluate HDD performance cannot accurately measure SSD performance and endurance. This is the case because an HDD is characterized by rotational latency and seek times. Existing benchmark scripts are designed to measure performance over a short time interval. By comparison, SSDs have no spin or seek functions, so the access results obtained using HDD benchmark scripts typically show very high IO rates and low latency. SSDs that are tested with these benchmarking scripts will deliver wildly optimistic performance scores that do not accurately reflect how the drive will perform under enterprise conditions.

Enterprise-class SSD capabilities

While some SSDs are intended for consumer applications that involve fast large-block reads, other SSDs are designed for enterprise workloads. Unlike consumer SSDs, enterprise solutions must support large numbers of simultaneous users running different types of traffic independently of each other. This usage pattern leads to random patterns of data traffic. The controller of an enterprise-class SSD is designed to support multi-threaded access, potentially involving hundreds of simultaneous data streams between the host and the SSD. Enterprise-class SSDs must perform extremely well, even for small-block transfers of varying sizes, and simultaneous reads and writes.

How well the enterprise-class SSD controller handles simultaneous flash management and host data transfers differentiates it from a consumer-grade drive. Controllers in enterprise SSDs are designed to maintain consistent performance behavior while transferring data, regardless of the amount of flash capacity in use, and irrespective of the volume of traffic being generated to the drive at any point in time. Wear-leveling operations and background media error correction algorithms are designed so that data transfer performance to the host is unchanged while these operations run in the background. An enterprise-class SSD is designed to handle these heavy workloads 24-7-365 for five years or more.

SSDs require special benchmarks

Benchmarks are useful for testing SSDs if they provide data that shows how the device will perform in the real-world environment. But a benchmark is not so useful if the test does not accurately represent the environment within which the drive will be deployed. Current benchmark tests are optimized for HDDs, and very good at identifying HDD results. HDD benchmarks are focused on identifying rotational latency and seek times associated with rotating media, as well as the movement of read/write heads across the surface of the disks.

These benchmarks are effective at demonstrating some of the relative strengths of SSDs as compared to HDDs, such as sustained bandwidth and maximum-read input/output operations per second (IOPS). Unfortunately, many benchmarks that were originally constructed to evaluate HDDs are run over a short time interval, and the results do not provide a measure of long-term SSD performance. Compared to HDDs, SSDs are more challenged by managing host data flow and internal NAND than by accessing the data.

For this reason, SSD benchmarks should map to the following actual SSD performance parameters:

- Mixed reads and writes that are characteristic of actual enterprise workloads, rather than 100 percent read or 100 percent write patterns
- Constant IO to the devices at reasonable queue depths that reflect realistic-use environments
- Sufficient volumes of write data to fill the open NAND blocks, so wear-leveling and flash management routines are engaged
- Running constant random write-only tests to determine devices' capabilities under worst-case loads
- Writing random data patterns to defeat any zero fill that the drives may use

Limitations of existing benchmarks

Validation tests need to evaluate more than just IOPS

Typical original equipment manufacturer (OEM) qualification models are designed to test and validate SSDs for reliability, functionality and performance. It is important to note that SSD performance is not only about sheer numbers of IOPS, but about steady performance over time and consistent latency. Most testing at this level is designed to ensure that the devices will not have any anomalies in steady-state performance over heavy and sustained use. But a faster device is not always the optimal device for demanding enterprise applications, especially if there are certain times or intervals when the performance or latency will be affected by the drive.

Benchmarks should measure more than just sequential throughput

Many OEM benchmarks focus on sequential throughput only. These benchmarks, which are widely available online, provide data about how fast a drive can write a stream of data, measured in sequential read or write megabytes per second (MB/s). Many vendors optimize their device profiles to score well on these tests. The problem with this approach is that in the enterprise space, very few storage-use cases are actually sequential in nature.

Even inherently sequential operations from a single application will not necessarily result in sequential access in a disk array or a multi-user enterprise environment. To understand why this is true, take the example of three to five servers simultaneously accessing their own sequence-based files on a storage system. Look closely at the storage system, down at the device level, and you will see that IO is suddenly very random in nature as the multiple data streams are split and the devices locate and retrieve data from different locations on the drives.

Validation testing should be based on real-world applications
An interesting aspect of benchmarking is that the benchmark can be tricked into doing exactly what the user wants it to do, leading to a skewed result. Benchmarks are interesting tools, but in reality, the best way to measure expected performance is by running storage devices with the actual applications that the drive will need to support in the real-world enterprise environment.

An enterprise email application is an excellent example. In the email application, reads and writes are not consistent and constant, but unpredictable in their frequency, duty cycle and block sizes. The best way to measure the capabilities of a device in this use case is to run it in your systems and directly measure the performance on the device. If this method is difficult to perform with a production system, you should obtain IO validation tools that simulate the application environment. Microsoft Exchange has the Jetstress and LoadGen tools for IO workload testing, and Oracle has a tool called ORION. All of these tools closely approximate actual enterprise software applications.

Need to measure steady-state performance at realistic queue depths
When you ask data center managers how they measure storage workloads on their drives, they will tell you that their focus is often on keeping outstanding IO requests, or "queue depths", as small as possible – usually under 16 and often as low as 8, or even 4. Benchmarks with queue depths of 32, 64 or greater may provide a higher number of IOPS, but the right question to ask is how well the device will perform in a real-world enterprise data center when it is running at a normal queue depth.

Mixed workloads can dramatically affect performance
Another measure that is often overlooked in benchmark testing is the mixed workload that includes both reads and writes. Most SSDs perform well at 100 percent write and even better with 100 percent read operations. It is important to recognize what happens when you set up a device to perform 70 percent reads, or operate a 50/50 read/write ratio. The performance of many drives can slow down by 50 percent or more in these mixed workload environments.

Parameters for benchmarking enterprise-class SSDs

As a leading provider of enterprise-class SSDs, STEC follows the guiding principle that to evaluate real-world SSD performance, it is necessary to test worst-case usage scenarios. Benchmarking involves these key concepts:

The baseline: 100 percent random writes over time
Measuring the performance of the drive as it is being written with random data invokes the internal management schemes of the SSD controller to wear-level and manages the flash, showing how the SSD's behavior can change as it is being filled. This benchmark is designed to expose potential weak spots as quickly as possible, and can be used to precondition the drive.

A 100 percent random-write IO will initially run quickly if the drive is empty at the start of the test. When all the raw blocks in the device have been written, the SSD will start to engage its wear-leveling and flash management algorithms, and their impact can be measured and evaluated.

As you can see from Figure A, the actual IOPS under 100 percent write operations are achieving from ~2,700 IOPS to as low as 170 IOPS.

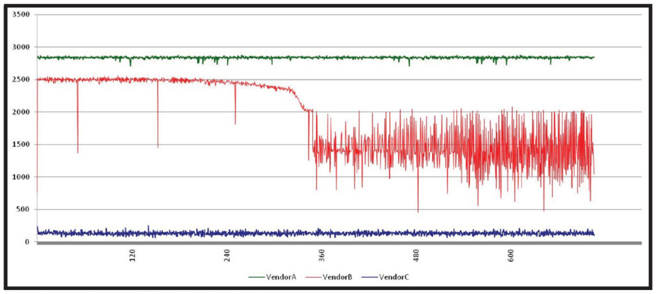


Figure A

This example shows how a working SSD can lose performance as it moves into wear leveling or other internal tasks for managing the operations of the SSD.

Preconditioning the SSD under test
Preconditioning involves filling the SSD with random write data past the raw capacity of the NAND flash memory. This is done to engage the device's wear-leveling and error-handling algorithms. An advantage of this technique is that it accounts for the over-provisioning found in most SSD designs.

An empty SSD will run much faster than an SSD filled with data. Preconditioning the SSD with randomly written data patterns enables the device to return accurate "steady-state" performance results that can be sustained over the useful life of the device. You should note that sequential write data is insufficient to force the device to engage the block allocation and wear-leveling algorithms that impact performance.

Mixed read and write operations
IO operations involving mixed reads and writes are a weakness observed in many SSD designs. It is typical to benchmark the absolute performance of an SSD when it is performing 100 percent reads or 100 percent writes, simply because these characteristics are relatively easy to measure. What the data sheets typically do not mention is that when simultaneously mixing both reads and writes, most SSDs actually slow down under load due to write prioritization and the need to service the write operations, which take time to complete. STEC enterprise-class SSDs are architected to address these issues.

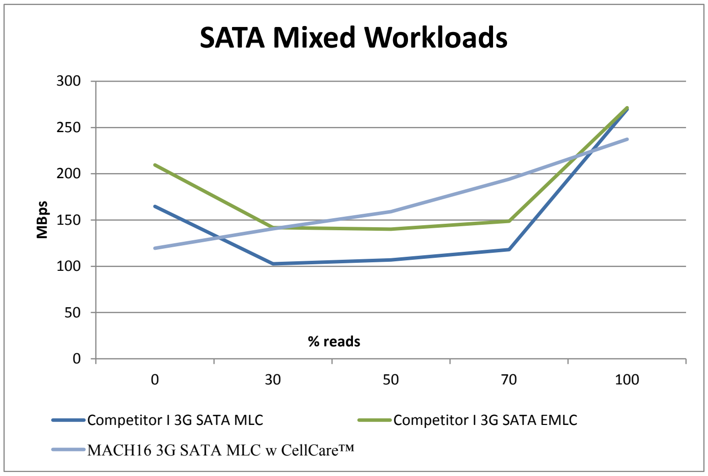


Figure B

As shown in Figure B, many designs show a "bathtub curve" effect, showing highest performance at both pure read and pure write operations, and lowest performance with a 50/50 ratio of writes and reads. At STEC, we recommend running a range of different write/read combinations to measure their impact. We use read percentages of 100, 80, 60, 40 and 20, and 100 percent writes.

Aligned vs. nonaligned IO
IO alignment can have a significant impact on SSD performance, even when the drive is fully preconditioned with random write data. Aligned IO for an SSD enhances efficiency of the device for managing NAND writes, and can boost SSD endurance by reducing the number of read-modify-write operations that can cause extra writes to occur in the background on the SSD.

Today's NAND flash memory architectures use 4K- or 8K- page write data. Aligning the IO to the page size, or a multiple of 4K or 8K, enables the SSD to maintain efficient write management without the need to constantly invoke schemes of read-modify-write methods needed to store data across two flash pages. However, this may be out of the control of the SSD integrator, depending on the operating system and file system/partition format. As a result, the benchmark should test both aligned and non-aligned capabilities, as shown in Figure C.

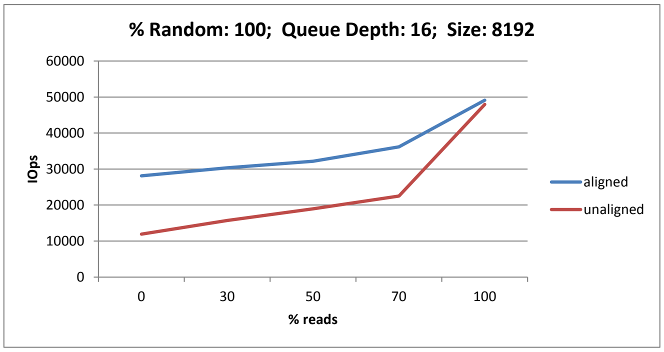


Figure C

Varying IO block sizes and queue depths
This benchmark is similar to tests run on HDDs. It is important to have all the preconditioning and environment decisions on alignment done before these tests are performed in order to see the actual steady-state performance an SSD device can achieve.

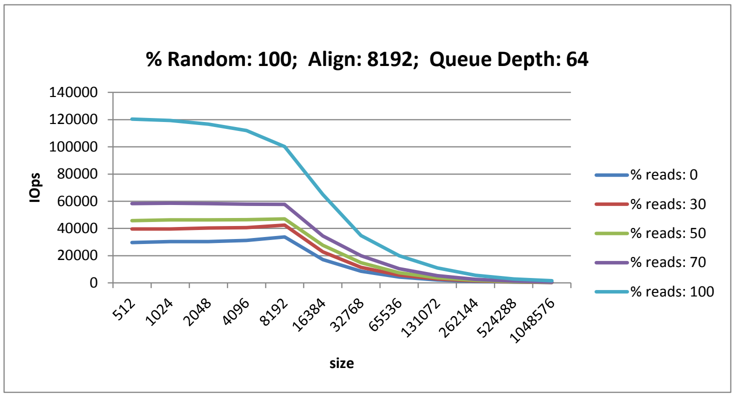


Figure D

The STEC data shown in Figure D shows consistent performance at all block sizes across the range of read/write ratios without the "bathtub curve" in mixed operations that can affect other SSD designs.

Queue depth is important in storage devices. Efficiencies can be gained from increasing queue depth to SSDs because this allows for more efficient handling of write operations, and may also help reduce write amplification that impacts the endurance of the SSD.

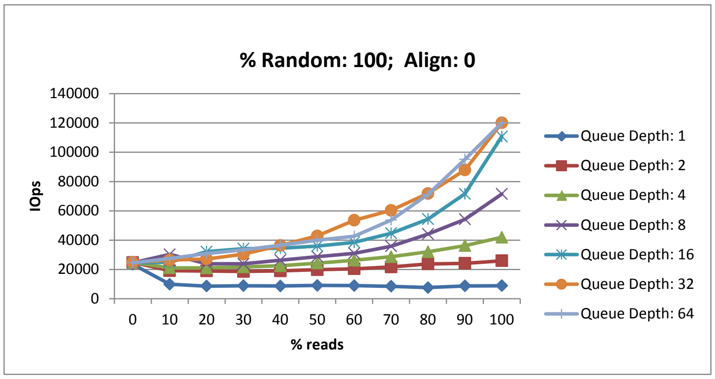


Figure E

Figure E shows device performance at different queue depths, and the performance that can be expected from an SSD depending on the use case, queue depth and read/write ratio. This can also help the system integrator adjust the system parameters to manage queue depth settings to an optimum level for the device being integrated.

Latency

Latency is a measure of how fast a device can respond to a given command such as reading or writing data. Even if the commands are not constant, the time it takes to respond to a command is still important. This is the time it takes for the application to get or write its data. The longer it waits for data, the slower the application may be. Reducing latency is the single most effective way to speed up application performance. SSDs should not only be measured for average latency. The best SSDs have a very tight latency distribution. Latency should be measured under random small-block (8K, 4K) workloads.

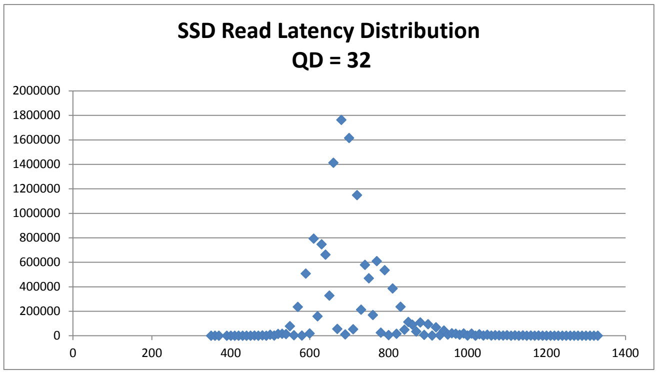


Figure F

Conclusion

The SSD with the fastest raw read or write results using existing HDD benchmark tools may not be the drive with the best performance in a working enterprise environment. SSD performance depends heavily on the workload and the drive's actual time in use. Actual performance can vary significantly from OEM specifications and best-case benchmarks.

Discovering how an SSD will actually perform in a real-world environment involves testing the drive in an environment that duplicates the prevalent enterprise workloads where the device will be deployed under worst-case conditions.

STEC benchmark recommendations

First: Precondition

- Avoid running benchmarks on empty drives.
- Precondition SSDs before testing with block-level data.
- Use random data patterns of non-zero content when writing
- Fill the drive to at least 2x the user capacity to include filling the overprovisioning space inside the SSD.
- Sequential preconditioning is different from random preconditioning. Select the correct preconditioning for the testing planned. If in doubt, use random preconditioning.

Next: While Testing

- Make sure you have preconditioned the drive.
- Run a range of tests that are consistent across devices tested. It is a good idea to run on the same hardware for baseline testing.
- In addition to 100 percent read or write operations, always run a set of mixed read and write tests.
- Consider wether another preconditioning run is necessary between tests (example: switching mostly random tests to purely sequential tests).
- Run tests over extended periods of time, allowing the drive to fill up and engage wear leveling and other algorithms, and other internal impacts on performance. A test that only runs for 30 seconds may not show real performance of the device. Consider a test of over two minutes for each profile.

Finally: Real-World

- Note that benchmarks only tell the story of what you will see if you run the benchmark. Realistic results or expectations of how a device will run in a real-world application are not generally possible from benchmark tests. They are only good for head-to-head comparisons.
- If your testing plan is intended to identify a device for use in your data center, it is important where possible to run tests in real-world environments using a file system and real data.

Following these recommendations will help you avoid unpleasant surprises, and ensure that the SSDs you select are capable of meeting the challenges of today's enterprise usage models over the anticipated life of your applications.

Appendix:

There are many tests available to the user depending on their operating system and intended testing needs:

Pure Benchmarks

• Iometer	Windows and Linux
• FIO	Linux
• IOzone	Linux, BSD, POSIX, Windows

User Environment Testing

• Email:

0 Jetstress	Windows
0 LoadGen	Windows

• Database:

0 ORION	Windows and Linux (Emulates Oracle databases)	
0 Super Smack	Linux	MySQL or PostgreSQL
0 sql-bench	Linux	MySQL

There are also many other benchmark/environment tests for purchase, depending on requirements.

For more information on STEC products, solutions and technology, please visit www.stec-inc.com

facebook.com/userstecinc

twitter.com/stec_inc

youtube.com/user/stecincssd

+1.949.476.1180
3001 Daimler Street, Santa Ana, CA 92705