



SSD Performance – A Primer

An Introduction to Solid State Drive Performance, Evaluation and Test

August 2013

SSSI Member and Author:
Eden Kim, Calypso Systems, Inc.



The Solid State Storage Initiative

About SNIA

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization made up of some 400-member companies and 7,000 individuals spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies and educational services to empower organizations in the management of information. To this end, SNIA is uniquely committed to delivering standards, education and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at <http://www.snia.org>.

About SNIA SSSI and SNIA SSS Technical Working Groups

The SNIA Solid State Storage Technical Working Group, comprised of over 50 SSD Industry OEMs, SSD Controller manufacturers, test houses, vendors and suppliers, continues to study the test and measurement of the latest in SSD products and technologies.

Other industry groups are investigating SSD related storage phenomena and are developing new technical positions and specifications (see SNIA NVM Programming TWG for Solid State Storage as persistent memory, SNIA SSSI Understanding SSD Performance Project, SNIA SSSI WIOCP Program, and SNIA IOTTA Repository for IO Traces, Tools and Analysis). The reader is invited to investigate and download information from these entities at <http://www.snia.org/forums/ssi>.

The latest SSS TWG Performance Test Specifications, released and draft versions, are also available for download at <http://www.snia.org/publicreview>.

Solid State Storage Drives (SSDs) are replacing Hard Disk Drives (HDDs) in many Enterprise and Client applications due to their increased performance, smaller form factor, lower power consumption and variety of device interfaces. However, SSD performance will vary due to several factors, including:

- Software application and operating system used (workload and IO stack),
- Type of NAND flash (MLC, eMLC, SLC, TLC),
- SSD device architecture (number of dies, amount of over provisioning),
- SSD design optimization (Error Correction Codec, flash translation layer, firmware) and
- Hardware & test environment (motherboard, CPU, RAM, SSD interface, test software).

So how does one select an SSD? What are the criteria? What are the important performance metrics and what do they mean? What does the user need for his/her intended use? And finally, how does one evaluate and interpret SSD performance tests and results?

This primer addresses these and other questions so the reader can understand which SSD performance criteria are most important when evaluating and selecting an SSD.



Table of Contents

| | |
|---|----|
| Foreword | 4 |
| I. Introduction to Performance | 5 |
| 1. What do I need to know when selecting an SSD? | 5 |
| 2. What is SSD Performance? | 5 |
| 3. How does SSD Performance differ from HDD Performance? | 6 |
| 4. What are the important SSD Performance Metrics? | 7 |
| 5. What should I consider when choosing an SSD? | 7 |
| II. Comparing SSD Performance | 8 |
| 1. Client SSD Performance Considerations | 8 |
| 2. Enterprise SSD Performance Considerations | 9 |
| 3. The Dimensions of Performance | 10 |
| • Three Performance Metrics – IOPS, TP & LAT | 10 |
| • Access Patterns and Test Stimuli | 11 |
| • IO Stack | 12 |
| • Workloads | 13 |
| 4. A quick look at Performance Results | 14 |
| • Summary Performance Comparison – HDD, SSHD, SSD | 14 |
| • A brief look at Test Parameters and Set-up Conditions | 15 |
| • A closer look at IOPS | 16 |
| • A closer look at Throughput | 20 |
| • A closer look at Latency | 21 |
| • How Parameter Settings Affect Performance | 22 |
| III Conclusion | 25 |
| Special Thanks / About the Author | 25 |
| Appendix | |
| 1. Test Practices | 26 |
| 2. SNIA Report Format | 26 |
| 3. Drive Preparation - PURGE, Pre-conditioning & Steady State | 26 |



Foreword

I found this paper to be an excellent tutorial on the performance of Solid State Drives (SSDs). It covers this topic in a very easy-to-understand way, yet provides detailed technical information that the reader can either dig into for a better understanding, or simply skip without missing the main points. The concepts presented within this paper cover SSD design considerations, performance measurement techniques, and the hardware and software used to test the SSDs. The target audience for this paper covers a wide spectrum of people, including end users, IT managers, SSD marketing executives, test operators, and SSD designers and OEMs.

When Eden Kim (author) asked me to write this introduction, I started thinking about how all of this SSD performance work began. It is interesting to look back and see how we got to where we are today. I think I started things within SNIA by making a presentation to the SNIA Board of Directors in April 2008 on a new upcoming storage technology that was starting to show some promise. That technology was enterprise storage made out of NAND flash. Interest in this new technology started growing and ramped up very quickly. In July at SNIA's summer symposium the SNIA hosted 3 days of presentations and tutorials from companies involved in the development of non-volatile solid state storage. 151 individuals from 57 different companies responded to SNIA's "call for interest" in this new technology. One month later, in August, the SNIA Board approved the creation of the Solid State Storage Initiative (SSSI). Then in October, the Solid State Storage (SSS) Technical Work Group (TWG) was officially launched.

The SSSI was created to provide a rich venue for collaboration for developers and users of solid state storage. Vendor-neutral education and support for standards and interoperability were the primary goals for the group. The SSS TWG, the sister group to the SSSI, became responsible for the standards piece and immediately began working on a performance test specification. This work resulted in the release of SNIA's Solid State Storage Performance Test Specification (PTS) which was initially made available in June 2010. The reason I mention all of this is because I believe this paper to be a very concise summary of the work done by the SSSTWG from the time it was created to now.

Anyone interested in understanding how SSDs work, what kind of performance they can achieve, and how they compare with HDDs, including all developers and users, will benefit from reading this paper. Personally, there are 2 simple but important conclusions that I took away from this paper: First, the storage ecosystem needs standardized testing for SSDs in order to understand how different SSDs compare to each other so that intelligent decisions can be made in designing, implementing, purchasing, and using SSDs. Second, the SSSI Reference Test Platform (RTP) and the Performance Test Specification (PTS) – both developed within the SNIA - provide an industry baseline for benchmarking that should be widely adopted by the storage community to ensure consistent test results.

Happy Reading!

Phil Mills, IBM

SNIA Director and Secretary 2003-2010

SSSI Founder and Chairman 2008-2010



I. Introduction to Performance

I. What do I need to know when selecting an SSD?

There are many things to consider when selecting an SSD, including:

- Price
- Capacity
- Performance
- Power efficiency
- Data integrity
- Durability
- Reliability
- Form factor
- Connection type (device interface)

Naturally, these criteria will vary in importance depending on the SSD use case – client, enterprise, mobile, etc.

While one may debate which of the above are most important, this primer will focus on **SSD Performance** (and those factors that affect performance). Performance is often listed as a prime selection criteria, along with price and capacity, when selecting an SSD because the “amount of performance you can afford” directly affects the user experience and has a tremendous impact on overall computing value. One needs to balance that preference with hard realities of budgets – over-buying can often prove as disappointing as under-buying. It is critical that the SSD be the best “fit” for the planned usage model.

Issues related to price, capacity, data integrity, endurance, and others, while also important in the overall selection process, are outside the scope of this document.

2. What is SSD Performance?

“Performance” relates to how well the SSD functions when accessing, retrieving or saving data. For example, how fast a computer boots up is heavily dependent on how fast the boot drive (whether HDD, SSD or Solid State Hybrid Drive) can handle the file requests from the operating system (OS). How fast a software application loads and runs and how quickly files are accessed or stored also relate to the overall performance of the drive.

It is important to note, however, that not all aspects of the perceived user experience are due to the performance of the storage device. Overall performance of any system – notebook, desktop, workstation, server, or storage array may be highly dependent on other factors.

For example, suppose I want to retrieve a file using my corporate network. The speed at which I can get my data may be limited (or bottlenecked) by the ability of the hardware controlling the storage device, the transport mechanism between me and the system storing my data as well as a litany of other factors. This data retrieval speed may also be enhanced or limited by using other acceleration techniques like a faster memory cache in the file server.



Similarly, the ability of my CPU itself to process commands, or a limitation on input/output (IO) commands getting to the CPU or to the SSD could limit the perceived IO transaction rate of the storage device.

Finally, the time it takes to complete a task (or IO request) sent by the host may be limited by a number of factors that influence the round trip time of the request from the user space to the storage device and back.

3. How does SSD Performance differ from HDD Performance?

The performance of SSDs differs significantly from HDDs due to the unique attributes of NAND flash based SSD architecture. Unlike HDDs, SSDs have distinct performance states. As a result, SSD performance:

- 1. Changes over time;
- 2. Depends on the write history of the SSD; and
- 3. Depends on the type of stimulus being applied to the drive.

Performance measurements are highest when the SSD is brand new and has no write history (or is Fresh-out-of-Box, or FOB). SSD performance will settle over time to a lower, more stable range (known as Steady State – a specifically defined Performance Test Specification (PTS) region where performance response is relatively time invariant). The Steady State level of performance will depend on how much and what kind of data was previously written to the SSD.¹ And finally, the level of performance will depend on the type of IO request being made (random or sequential access, size of the blocks, the ratio of reads to writes, and how the IO access blocks are placed on the storage media – also known as block alignment).

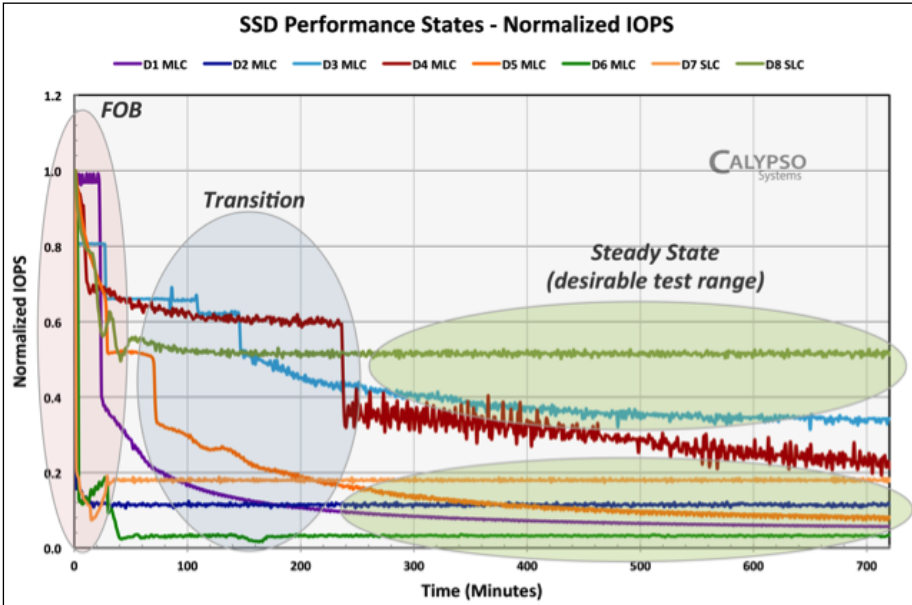


Figure 1. SSD Performance States

¹ Steady State (SS) is a specifically defined performance range from which test measurements are taken. The SSD is first prepared using a prescribed Pre-conditioning methodology. The SS window is then calculated pursuant to the test specifications. Note: there can be more than one Steady State depending on the test and the defined Steady State test criteria.



Figure 1 to the left shows how performance changes over time for several different SSDs. Each line presents a different SSD with continuous random 4KiB² block size 100% Writes over 6 hours. Each drive shows normalized IOPS performance starting in the **FOB** state, settling through a Transition period and ending at a **Steady State**.

It is important to note that the same test (such as IOPS) applied to a given SSD can yield different results depending on the *hardware/software environment*, the *test methodology* and the *test parameter settings themselves*. This is extremely important to remember when examining SSD performance specifications and when comparing performance between different SSDs.

For these reasons, and to address market demands for standardized testing, the SNIA Solid State Storage Technical Working Group has developed, validated and published a detailed Solid State Storage Performance Test Specification (PTS). The PTS prescribes a specific SSD test methodology, suite of performance tests, standard data disclosure and formatting, as well as a recommended Reference Test Platform (RTP) for comparative SSD performance test of Client and Enterprise SSDs. Note too that the PTS is flexible – enabling the test operator to examine how SSDs may behave when used outside of their intended usage models.

4. What are the important SSD Performance Metrics?

HDD and SSD performance is most often described in terms of three basic metrics: Input Output Operations per Second (**IOPS**), **Throughput** (usually expressed in Megabytes per second or MB/s) and **Response Time** (or Latency and typically expressed in milliseconds or microseconds). A metric is the measure of performance against a given parameter – for example, speed expressed in miles per hour or height in inches.

IOPS refers to the IO operation transfer rate of the device or **the number of transactions** that can occur in a given unit of time (in this case seconds). The IO transaction rate is measured in IOPS.

Throughput – abbreviated as “TP” and often expressed as **Bandwidth** – refers to the **rate of data transfer** or, in this case, the amount of data that is transferring to or from the SSD or HDD. Throughput is measured in MB/sec (and an oft used metaphor is the “water through a fire hose or straw” analogy).

Response Time (or Latency) – abbreviated as “LAT” when shortened from Latency – refers to the time it takes for a command generated by the host to go to the storage device and return, i.e. the **round trip time for an IO** request. Response time is measured in milliseconds or microseconds and is often reported as an Average (AVE) or Maximum (MAX) Response Time.

5. What should I consider when choosing an SSD?

Having identified the three fundamental SSD performance metrics (IOPS, TP and LAT), one should choose an SSD that provides the best overall price / performance for the intended application(s), and fulfills the form factor, reliability, power and endurance requirements for the intended use of the SSD.

² One KiB = 1024 bytes. One KB = 1000 bytes. SSD Industry convention uses base 2 (e.g. KiB, MiB, GiB) to describe storage capacities (e.g. 4 KiB block size) and base 10 (e.g. KB, MB, GB) to describe transfer rates (e.g. total GB written or MB/sec).



The user should carefully investigate:

1. The SSD's intended **use** (Enterprise, Client, Mobile, etc.)
2. The SSD response to **specific workloads** (access pattern, use case and metrics)
3. **How the SSD is tested** (test methodology, test platform and test parameter settings)

In order to accomplish these objectives, it is necessary to understand **what factors affect SSD performance** and to know **how to fairly evaluate and compare SSD performance**.

II. Comparing SSD Performance

Now that we have identified the basic SSD performance metrics (IOPS, TP and LAT), what do the advertised measurements mean? What is the relationship between IOPS, TP and Latency? What are good, medium or bad performance ranges? How do I compare the measurements listed for different SSDs? Which measurements are more important? Do I need faster IOPS or faster Response Times?

The answer to these types of questions is invariably: *it depends*. The relevant parameters are ultimately determined by the use case anticipated for the SSD – i.e. what type of user applications will be run and how much and what kind of workload will be applied to the SSD in its intended use.

I. Client SSD Performance Considerations

Client use cases typically involve a single user running various Client software applications over the course of the day, i.e. the workload for the SSD is not 24 hours a day, 7 days a week. It is based mostly on software applications and operating system. Client SSD demand intensity is low relative to Enterprise class SSDs which have multiple users or concurrent workloads that continuously access the SSD.

For Client SSDs, IOPS are often listed as “up to” a certain number of IOPS, at “sustained” IOPS or perhaps (rarely) at “Steady State” IOPS. What is the difference between these claims and what do the different performance states mean? How many IOPS are good? How many IOPS are enough?

One should always try to understand the performance state in which any measurements are taken. Each performance state is highly dependent on a variety of factors that can have a significant impact on performance. FOB and “up to x IOPS” statements are transient levels of performance that may never be seen once the user installs and begins to use the SSD.³ Claims of sustained IOPS can also be problematic (and misleading) if the sustained level is not precisely explained.

Which metric is more important? In absolute terms, a higher value for IOPS and Bandwidth (MB/sec) is better (more) while a lower value for Response Times or Latency is better (faster). How many more IOPS will make a difference in my use case? Are the additional IOPS worth the additional price? What if IOPS are better but the Response Time is worse when comparing SSD A to SSD B? Which is more important: higher IOPS or lower Response Time; higher IOPS or higher Bandwidth? And at what cost?

³ FOB occurs only after a device PURGE and usually only lasts 2 SSD “capacity fills” – i.e. 200GB for a 100GB SSD. A 100GB SSD used as a boot drive will use up its FOB state in the first 200GB written, including software installations and use. A SSD used as a secondary drive may provide a period of enhanced FOB performance, but again limited to the 2 times user capacity immediately following a device PURGE. Note also that a device PURGE will delete all user data on the device being purged.



For many Client use cases, IOPS and Bandwidth may be more important than Response Times (so long as the response times are not excessively slow). This is because the typical Client user would not typically notice a single IO taking a long time (unless the OS or software application is waiting for a single specific response).

However, choosing an SSD based solely on the highest IOPS or Throughput rate can have a point of diminishing returns. The user's computer may not be able to utilize IOPS rates past a certain point. The Client use case may consist of low end-user SSD demand ("read the SSD, wait while the user processes the data on the screen, maybe then write a little to the SSD"). Also, response times may begin to rise as the IOPS rate saturates (IOPS increase past the point where they can be efficiently processed). Similarly, higher Bandwidth, while always good to have, may be advertised based on a block size or read write (RW) mix that does not match the workload that the user's computer generates or that the SSD sees.

Thus, it is critical to understand what and how the advertised metrics are obtained and to know what kind of workloads will be generated in the intended use case. Are the advertised metrics accurate and the test methodology reasonably disclosed? Are the advertised metrics relevant to my intended use (e.g. video streaming, gaming, internet surfing, email, office suite, desktop publishing, graphics, video editing, etc.)?

2. Enterprise SSD Performance Considerations


The Enterprise market requires SSDs with higher overall performance. Enterprise SSDs are generally measured at Steady State under a full workload for seven days a week for 24 hours per day operation. The constant usage requires a substantially different design, fewer errors, higher general performance and greater data integrity.

Enterprise SSDs are often aggregated as RAID, Tiered, Direct Attached Storage (DAS), Network Attached Storage (NAS) and Storage Attached Network (SAN) for more complex storage solutions. SSDs are also packaged as higher performance solid state storage in PCIe connected devices – both as traditional SSD-like devices as well as more advanced, higher speed products (such as Persistent Memory, Memory cache and Direct Memory). In all cases, the traditional metrics of IOPS, TP and LAT still apply.

Enterprise SSD use cases also tend to be more homogenous than Client with SSDs tuned for the access patterns associated with specific applications such as On-line Transaction Processing (OLTP), Virtualized Machines (VM), Virtual Desktop Infrastructure (VDI), Video-on-Demand (VOD) edge servers, Tiering, data center, database and others.

This allows the Enterprise user to deploy SSDs with high performance in the areas that relate to the intended Enterprise use case. For example, OLTP environments may focus on small block random workloads while VOD and video streaming may focus on large block sequential workloads. Some SSD products may be designed for read-intensive applications such as web hosting, cloud computing, meta-data search acceleration and data center virtualization while other products may focus on write-intensive applications such as datacenter logging or snapshots.

While IOPS are an important metric in the Enterprise, it is often the management of response times that is paramount. It is often said "IOPS are easy, latencies are hard." Competing Enterprise SSDs may offer substantially the same number of IOPS, but may differ significantly in response times and latencies. Thus, Enterprise focus on response times can differ from Client.



For example, while Client SSD use cases may largely be concerned with average response times, the Enterprise use cases are often more interested in maximum response times and the frequency and distribution of those response times.

In a Client application, it may be acceptable to have many response time spikes (instances of very high response times for a number of IO operations) since the result may merely be an annoying delay to the user (hourglass or spinning gear for a while).

In the Enterprise, the SSD must often meet a Quality of Service (QoS) level - the requirement that a given application complete all requested processes within a specified time limit (and effectively disallowing an application from accepting or processing a given percentage of requests whose response times exceed a fixed threshold). This means there is often a maximum response time ceiling below which 99.99% or 99.999% of the IOs must occur ("4 nines or 5 nines" of Confidence).

In addition to the treatment of response times, the selection of NAND flash type can affect performance if throttling is employed. Different types of NAND flash (SLC, MLC, eMLC or TLC) have different endurance ratings – but still must meet service life requirements of the Enterprise SSD. To ensure the SSD meets the warranted duration, manufacturers may throttle the write rate (to preserve the NAND flash life). This throttling, in turn, limits the number of program-erase (P/E) cycles written to the NAND.

It should be noted that throttling is not limited to Enterprise SSDs. Client SSDs may also throttle IOPS to save power. By limiting the number of NAND dies allowed to be used simultaneously, significant power savings can be achieved. This can help Client SSDs meet stringent power budget specifications for mobile and consumer SSD applications.

Regardless of the type of SSD – Enterprise or Client – the evaluation of performance metrics and measurements will depend on the use case, the intended workload and the operating (or test) variables and parameters that affect the SSD performance.

3. The Dimensions of Performance

Three Performance Metrics – IOPS, Throughput (TP) & Latency (LAT)

Each metric represents a different dimension of performance. Evaluation of SSD performance should consider all three dimensions: IO transaction rate (IOPS), sustained data transfer rate (TP) and the latency (LAT) or response times (RT) of the IOs. The optimal values for each metric, and the relationship between them (relative importance of each) is a function of the anticipated use case workloads.

It is important to evaluate all three dimensions of performance because the user will want to know:

- How many IO operations can be completed (IOPS)?
- How much data can be transferred (Throughput in MB/s)?
- How much delay there is for a given IO operation (Response Time/Latency)?

For example, when I boot my system (on a laptop or Enterprise virtual desktop), how fast will the OS load the small block read and write operations (IOPS)? The department manager may look at this as “how



quickly are my workers getting productive?” How fast will my system load large graphic files for editing or stream a training video from my storage device (Throughput)? How long will I have to wait to find/load/save a specific file (Response Time/Latency)?

Access Patterns and Test Workloads

An access pattern is the type of storage and retrieval operations to and from a storage device. Access patterns are described in three main components:

- Random/Sequential – the random or sequential nature of the data address requests
- Block Size – the data transfer lengths
- Read/Write ratio – the mix of read and write operations

Any particular workload or test stimulus is approximated by some combination of access patterns. That is, an access pattern is one component of a synthesized equivalent IO workload. For example, “RND 4KiB 65:35 R/W” describes an access pattern consisting of: a sequence of IO commands, each one 4KiB long (block size), to random locations on the storage device, in the proportion of 65% Reads to 35% Writes.

Random/Sequential

A read (or write) operation is sequential when its starting storage location (Logical Block Address or LBA) follows directly after the preceding operation. Each new IO begins where the last one ended. In contrast, an IO operation access is considered to be random when its starting LBA is not contiguous to the ending LBA of the preceding IO operation.

Sequential operations can be faster than random operations. When HDD recording heads are on a single track, sequential IOs can be faster than random IOs. In NAND flash SSDs, sequential IOs can be faster than random IOs when SSD mapping tables are less fragmented.

NAND SSDs use a virtual mapping scheme whereby LBAs are mapped to Physical Block Addresses (PBAs). SSD virtual mapping is used for several reasons. For example, wear leveling algorithms distribute newly recorded data to new cell locations to promote even wear on the memory cells and thus improve the memory cell life or endurance. Hence, the SSD must keep track of the LBA – PBA associations.

The SSD controller keeps track of the LBA to PBA association by using a mapping table. Depending on its design, the SSD mapping table can become highly fragmented (and slower) if comprised of many random access entries whereas the mapping table can be more efficient (faster) when mapping a more organized table of fewer sequential access entries.

Block Sizes

Block sizes (BSs) can be in varying sizes from small (e.g. 0.5 KiB, 4KiB, 8KiB) to large (e.g. 128KiB to 1024KiB or more). Blocks are more efficiently stored in an SSD when they are aligned with the NAND flash memory cell boundaries so that, for example, a 4KiB block will fit exactly in a 4KiB NAND page size. All things being equal, more small block IOs can be accessed in a given period of time than large block IOs – although the amount of data could be the same (128 IOs of 4KiB data transfer length vs 4 IOs of 128KiB data transfer length).

The smallest granularity that can be written to NAND flash depends on the design of the underlying NAND flash. NAND flash writes to a single page and erases a block of many pages (hence “page write, block erase”). NAND typically has either a 4KiB or 8KiB page size. Blocks that are aligned with the page boundaries will be more efficiently stored and faster to retrieve than misaligned blocks.

Larger sequential block sizes can favor a greater Throughput while smaller sequential block sizes can favor a higher IOPS rate.

Read/Write Ratio

Read operations on SSDs are generally faster than write operations. Due to the fact that a NAND memory location cannot be overwritten in a single IO operation (as HDDs can overwrite a single LBA), a NAND flash write operation can take several steps performed by the SSD controller. The number of write steps depends on how full the drive is and whether the SSD controller must first “erase” the target cell (or even re-locate some data by performing a more time costly read/modify/write operation) before writing the new data. NAND read operations, on the other hand, can be completed in fewer steps.

In SSDs, an access pattern comprised of small block sequential read operations will yield higher IOPS while an access pattern comprised of large block sequential reads will yield higher Throughput. Conversely, access patterns comprised of small block random writes will yield lower IOPS and lower Throughput.

So far, we have talked about two metrics: IOPS and TP. Now let us turn to our third metric: Response Time (also called Latency). Response Time measures how long a particular IO transaction takes to complete (which is affected by everything that touches the IO as it traverses the “IO Stack”). Knowing how an IO is affected by the IO Stack is important to understand when evaluating Response Time measurements.

IO Stack

The performance we measure for a particular access pattern can be very different depending on where in the system we do the measurement. **Device level** (or Block IO) tests typically measure access patterns as close to storage hardware as possible (desirable for SSD drive testing), whereas **file system level** tests are more often used to test the software application in user space (which measures the overall system level performance of the storage device).

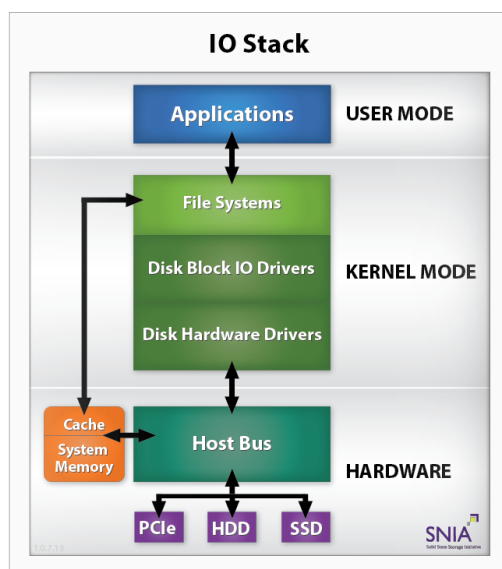


Figure 2. IO Stack

Access patterns generated by software applications must traverse the **IO Stack** (Figure 2) - to get from the user space to the SSD and back again.

This means that the file system and various drivers will affect the IOs as they pass them up or down the IO Stack.

Examples of how IOs are affected by the IO Stack include: **coalescing** small IO data transfers into fewer larger IO data transfers, **splitting** large sequential IO data transfers into multiple concurrent random IO data transfers and using the faster file system cache while deferring IO commits to the SSD.



This is important because while a given application may, for example, predominantly generate small block random IO traffic in user space, the IO Stack could change these IO requests as they get passed from user space through the IO Stack to the SSD and back.

SSD manufacturers, test houses and developers test SSD device performance at the Block IO Level. The Block IO level is where SSD performance can be most fairly compared without the effects of the IO stack. Performance benchmark testing should also use Block IO level test on identical test platforms in order to compare performance between different SSDs.

Workloads

What are workloads? In very general terms, a workload can be described by the access patterns measured during an observation period (e.g. 10 minutes of random 4KiB 100% Writes). Having determined that the key performance metrics (IOPS, TP and LAT) are described in terms of their access patterns, the key questions are:

1. What access patterns are *generated by my application*, and
2. What access patterns are *applied to my SSD*.

Workload IOs at the user level can be very different from those seen at the device (e.g. SSD) level because the IO stack can alter or change the nature of the access patterns as the IOs traverse the stack.⁴

User level workloads are complex, multiple stream access patterns that can change over time. A system can start with a largely small block random read/write (RW) boot workload, shift into a predominantly sequential read (R) workload as software loads, and then generate a mixed stream workload as multiple programs interact – all while the operating system software generates its own mix of access pattern activities in the background.

While user workloads are complex mixtures of various access patterns, there are *commonly accepted* access patterns used by the industry for certain applications.⁵


Some commonly accepted workloads are described as a *monotonic* access pattern - such as OLTP (random 8KiB 65:35 RW) and VOD (sequential 128KiB 90:10 RW) - while other workloads are described as a *composite* access pattern – the combination of two or more access patterns into a composite workload stimulus.

Once the user understands the nature of the application workload, the SSD can be tested for those specific access patterns. Additionally, the test operator can optimize the test stimuli to more closely emulate the target workload by adjusting various test set-up conditions and operating parameters.⁶

⁴ Empirical performance metrics collected by the SNIA SSSI Workload IO Capture Program (WIOCP) program have shown that there can be over 40,000 IOPS seen at the file system level while only 1,700 IOPS are observed at the SSD. This may be due, in part, to the file system using the faster file system memory cache for many IOs and delaying the commit of those IOs to the SSD - thereby reducing the effective IOPS rates seen at the SSD.

⁵ These application workloads have been characterized at the file system level. IO trace capture and other tools are used to capture the IO stream to analyze and understand the mix of access patterns generated by application workloads. It is important to understand that these commonly accepted synthetic models do not exactly represent any given real world workload, but instead are simplifications that are sufficiently close so as to provide a good comparative base.

⁶ For example, synthetic test access patterns need to add the dimensions of time and demand intensity in order to emulate the application workload. The timing of the data flow (and idle times) and the total OIO (or Outstanding IO Demand Intensity) need to be defined in order to better test the SSD and emulate the target workload. (Note: the term "synthetic" as used here refers to the ability to reliably create a known and repeatable test workload. It does not mean to artificially create an unrealistic workload.)



The SNIA PTS presents discrete access pattern measurements under various test conditions for each of the main metrics. This enables users to associate the discrete measurements with the components of the commonly accepted user workloads, such as the ones mentioned above.

In addition to commonly accepted workload access patterns, much industry work is ongoing to empirically ascertain the nature of user workloads in more detail. For example, the SNIA SSSI Workload IO Capture Program (WIOCP) is an open industry effort to collect user IO statistics of workload access patterns to help in the modeling of more accurate synthetic workload access patterns.

Users can observe and compare SSD performance by using the standardized device level tests and default settings required by the PTS. Users can also optimize (and disclose) optional PTS test settings to observe the effects of different access patterns and parameters. The PTS methodology can additionally be used to model new access patterns as they become published and accepted as representative of different user workloads. Examples of standard and modified PTS test parameters and their effects on performance are discussed throughout this document.

4. A quick look at Performance Results

Summary Performance Comparison - HDD, SSHD, SSD

How does one use the PTS to evaluate and compare SSD performance? First, data should be collected by using the SNIA PTS on an SSSI Reference Test Platform (RTP) to get standardized and normalized results. The user should then select the key measurements from each of the three metrics (IOPS, TP and LAT) to compare the overall performance.

There are commonly accepted block size and RW mixes that are used to evaluate SSD performance. While users may have an interest in other specific access patterns, industry convention uses device level “corner case” (or boundary) testing to evaluate SSD performance (see Appendix – Test Practices).

While IOPS, TP and LAT have long been used for the testing of HDDs, the SNIA PTS methodology has added solid state storage specific processes to accommodate the uniqueness of SSD performance. Specifically, the PTS defines a standardized Pre-conditioning and Steady State determination to ensure the SSD is properly prepared as well as SSD specific tests that address SSD specific performance behaviors.

Figure 3 to the right shows summary performance data for HDD, Solid State Hybrid Drive (SSHD), mSATA, SATA, SAS and PCIe SSDs. In this chart, there are four measurements for IOPS, two for TP and two for Response Times. This provides an overview of overall performance. Of course, different access patterns can be selected for summary view.



| Summary Performance Data – HDD, SSHD, SSD | | | | | | | | | | |
|---|------------------------------|-----------------|-------------------------|-------------------|-----------------|--------------------|-------------------------------|---------------------|----------------------------------|--|
| Class | Type | FOB IOPS | IOPS (higher is better) | | | | Throughput (larger is better) | | Response Time (faster is better) | |
| Storage Device | Form Factor, Capacity, Cache | RND 4KiB 100% W | RND 4KiB 100% W | RND 4KiB 65:35 RW | RND 4KiB 100% R | SEQ 1024KiB 100% W | SEQ 1024KiB 100% R | RND 4KiB 100% W AVE | RND 4KiB 100% W MAX | |
| HDD & SSHD | | | | | | | | | | |
| 7,200 RPM SATA Hybrid R30-4 | 2.5" SATA 500 GB WCD | 125 | 147 | 150 | 135 | 97 MB/s | 99 MB | 15.55 msec | 44.84 msec | |
| 15,000 RPM SAS HDD IN-1117 | 2.5" SAS 80 GB WCD | 350 | 340 | 398 | 401 | 84 MB/s | 90 MB/s | 5.39 msec | 97.28 msec | |
| Client SSDs | | | | | | | | | | |
| mSATA SSD R32-336 | mSATA 32 GB WCD | 18,000 | 838 | 1,318 | 52,793 | 79 MB/s | 529 MB/s | 1.39 msec | 75.57 msec | |
| SATA3 SSD IN8-1025 | SATA3 256GB WCD | 56,986 | 3,147 | 3,779 | 29,876 | 240 MB/s | 400 MB/s | 0.51 msec | 1,218.45 msec | |
| SATA3 SSD R30-5148 | SATA3 256GB WCE | 60,090 | 60,302 | 41,045 | 40,686 | 249 MB/s | 386 MB/s | 0.35 msec | 17.83 msec | |
| Enterprise SSDs | | | | | | | | | | |
| Enterprise SAS SSD R1-2288 | SAS 400GB WCD | 61,929 | 24,848 | 29,863 | 53,942 | 393 MB/s | 496 MB/s | 0.05 msec | 19.60 msec | |
| Server PCIe SSD IN1-1727 | PCIe 320GB WCD | 133,560 | 73,008 | 53,797 | 54,327 | 663 MB/s | 772 MB/s | 0.05 msec | 12.60 msec | |
| Server PCIe SSD IN24-1349 | PCIe 700GB WCD | 417,469 | 202,929 | 411,390 | 684,284 | 1,343 MB/s | 2,053 MB/s | 0.03 msec | 0.58 msec | |

Figure 3. Summary Performance Data

All of the above measurements were taken by Calypso Systems, Inc. on the SSSI RTP/CTS test platform pursuant to PTS 1.1.


Test results for a 15,000 RPM SAS HDD and a 7,200 RPM SSHD are shown along with test results for various SSDs. The HDD and SSHD drives show lower, unchanging performance with no difference between FOB and Steady State. For the SSDs, the FOB column shows a higher level of performance followed by lower level of performance during Steady State measurements.

For SSDs, FOB IOPS shows the test starting point as well as the peak IOPS value that SSD advertisers may claim. The IOPS, TP and LAT measurements show Steady State performance after the SSD has been pre-conditioned and written to Steady State.

General performance increases from mSATA through SATA, SAS and ending with the highest performance PCIe SSD listed. Note the MAX Response Time spike for the write cache disabled (WCD) SATA3 SSD. Note also the improved random 4KiB IOPS values when write cache is enabled (WCE) for the next SSD.

A brief look at Test Parameters & Set-up Conditions

Test Parameters are extremely important in determining SSD performance. Test set-up parameters include **PURGE**, write cache enable/disable setting (**WCE/WCD**), Pre-Conditioning Active Range (**PC AR**), Test Active Range (**Test AR**), Workload Independent Pre-Conditioning (**WIPC**), Workload Dependent Pre-Conditioning (**WDPC**), Steady State (**SS**) determination, and Stimulus **Segmentation** (or Banding).



Each of these set-up conditions can profoundly affect SSD performance measurement. For example, for Client SSDs, enabling write cache and limiting the AR can produce very high IOPS values (see SATA3 WCE SSD in Figure 3). WCE allows the SSD to use the much faster memory cache. Also, limiting the Pre-conditioning AR effectively creates over provisioned flash which can allow more controller optimization. On the other hand, write cache disabled with a full AR will result in lower IOPS. For Enterprise SSDs, the demand intensity settings (Outstanding IOs) can have a similarly large effect on response times. Performance test results should always be examined in the context of these key test set-up and parameter settings.

A closer look at IOPS

FOB

FOB is the state immediately following a device PURGE. In this state, there is no write history on the device and all the memory cells are free and available for writes. The SSD will show its peak IOPS performance during the FOB state and then quickly settle through a Transition State until a Steady State is reached. FOB is a useful metric because:

- FOB represents the starting point for tests that begin with a PURGE
- FOB provides a comparison for “up to x IOPS” marketing claims

WSAT

The Write Saturation test (WSAT) is a type of IOPS test (but should not be confused with the Steady State IOPS test described in the next section). The default PTS settings for WSAT are to apply random 4KiB 100% Writes after a device PURGE. As a monotonic access pattern, WSAT is easily applied using a variety of test software stimulus generators (both free and commercial). WSAT tests can also be used to investigate a wide range of performance behaviors by changing (and disclosing) parameters such as access pattern (e.g. using sequential 128KiB Writes) or by running the test for a period of Time or Total GB Written (TGBW).

The WSAT plot is useful for a variety of reasons:

- WSAT quickly shows how SSD performance evolves over time
- WSAT is useful to observe the effects of changes in variable and parameter settings
- WSAT can be run for a period of Time, an amount of Total GB Written or to Steady State Determination (pursuant to the PTS Steady State 5 Round Determination – see PTS)
- Different SSDs can be easily compared to a monotonic single block size access pattern

Anatomy of a WSAT Curve

WSAT plots share common characteristics: an initial peak FOB state; a Transition State – often evidence by a characteristic “write cliff” – leading to a stable and level Steady State. The write cliff will often appear at a recognizable multiple of the SSD’s stated total user capacity. While the WSAT curve may vary between drives, drives of a similar architecture/firmware may show similar shaped WSAT curves (or signatures). See Figure 4 to the right.

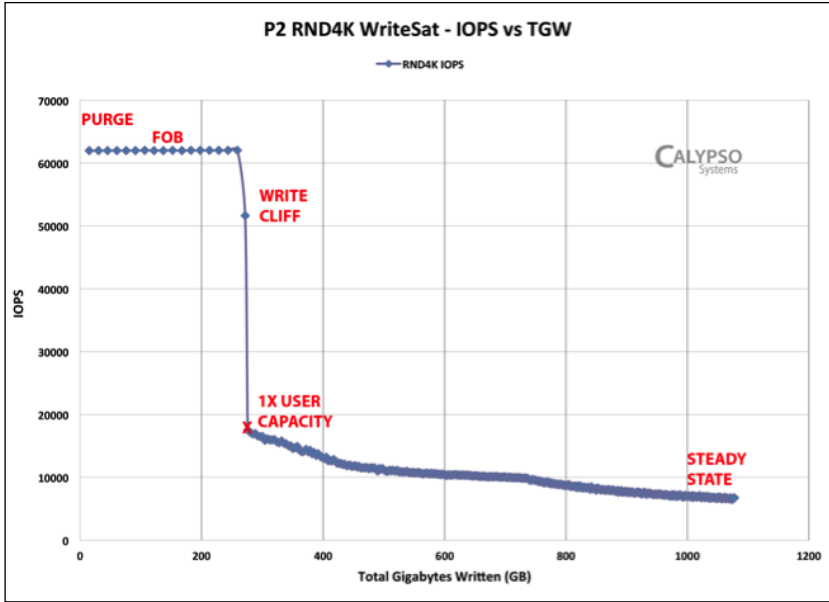


Figure 4. WSAT – IOPS v TGBW 6 Hr / WCE OIO=32

WSAT: Comparing drives

WSAT is also useful for comparing SSD performance. The chart in Figure 5 below shows various MLC (Client) SSDs on a WSAT IOPS v Time plot. Each SSD shows an initial peak FOB state followed by transition to Steady State. IOPS values on the Y axis are normalized – i.e. each SSD shows its maximum IOPS as 1.0 IOPS.

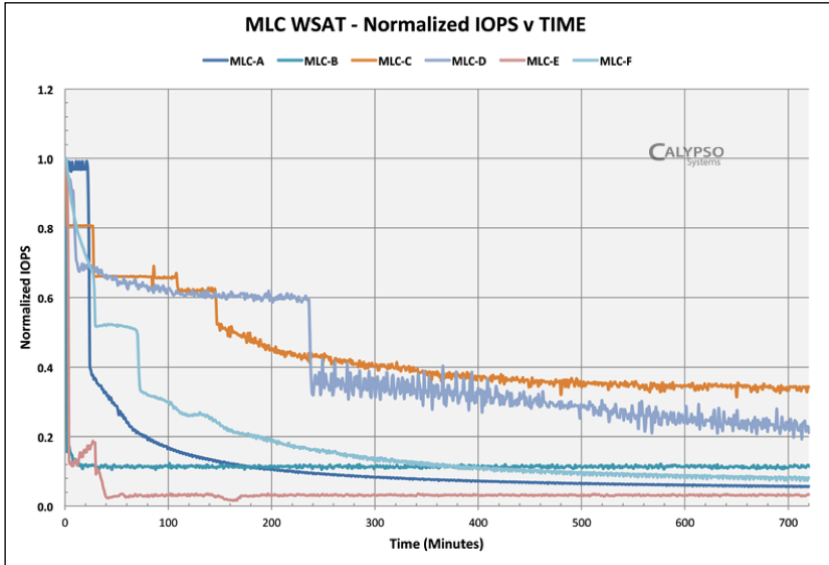



Figure 5. WSAT – IOPS v Time - Comparing SSDs



In the above chart, MLC-F, after an initial FOB peak, write cliff and plateau (at 60% of maximum IOPS), shows a second write cliff and ensuing continuous downward slope. Remember that under the PTS, WSAT Steady State (SS) can be determined in three ways: 1.) by using the 5 Round Steady State Determination formula; 2.) by Time (default is 24 hours); or 3.) by Capacity (default amount is four times the stated user capacity). SNIA report headers list Steady State criteria and conditions.

However, the main objective of a WSAT plot is to view performance evolution over a given test period, not to necessarily achieve an IOPS Steady State. The Steady State IOPS test (see next section) is a more appropriate test for determining stable Steady State IOPS measurements.

For example, one could apply the WSAT 5 Round Steady State Determination (5 points separated by 30 minutes of random 4KiB Writes – see PTS WSAT Steady State) early on the post FOB plateau and possibly meet the SS criteria. However, observing when and how the random 4KiB W performance settles (over 720 minutes in this case), and how many performance levels are observed, is the main focus of WSAT – i.e. seeing all of the random 4KiB Write behaviors over the specified time period.

Each SSD may have a unique WSAT curve, or signature. Some curves show a higher FOB level and shorter plateau while others show a lower FOB level but a longer plateau. Other SSD curves range from showing a step-function of succeeding levels, showing a gradual slope with few steps, to showing a dramatic write cliff followed by a long level period. WSAT can also be used to compare other performance variables.

While WSAT curves show behavior evolution, the reader is cautioned not to make overall SSD quality judgments based on the shape of these curves. Rather, the reader should examine WSAT together with the other Steady State tests of IOPS, Throughput and Latency.

Steady State IOPS

PTS IOPS measures different block sizes (BS) and RW mixes at Steady State. A higher IOPS is generally better. The IOPS test creates a 56-element matrix of 8 BS × 7 RW mixes in a single chart. This allows the user to quickly select the Steady State IOPS BS / RW mixes that relate to the user's workload of interest.

The table below shows the standard PTS IOPS Tabular Data chart in a SNIA report format. This report header contains all of the required disclosure information about the test set-up and test parameters associated with the IOPS test. See the Appendix for a more detailed discussion of SNIA report headers.



| Test Run Date: | | 03/02/2013 03:30 PM | | Report Run Date: | | 05/13/2013 05:48 PM | | |
|--|---|---------------------|-----------------------------------|-------------------|---------------------|---------------------|-----------------|-----------|
| IOPS Test (REQUIRED) - Report Page | | | | | | | | |
| SNIA SSS TWG | Solid State Storage Performance Test Spec (PTS) | | IOPS - Block Size x RW Mix Matrix | | | | Rev. | PTS-C 1.1 |
| Vendor: | | ABC Co. | SSD Model: | ABC Co. MLC-A 250 | | TEST SPONSOR | CALYPSO Systems | |
| Page | | | | | | | | 4 of 6 |
| Test Platform | | Device Under Test | | Set Up Parameters | | Test Parameters | | |
| Ref Test Platform | Calypso RTP 2.0 | Mfgr | ABC Co. | Data Pattern | RND | Data Pattern | RND | |
| Motherboard | Intel 5520HC | Model No. | MLC-A | AR | 100% | AR & Amount | 100% | |
| CPU | Intel XEON 5580W | S/N | 123 456 | AR Segments | N/A | Test Stimulus 1 | IOPS Loop | |
| Memory | 8 GB PCI600 DDR2 | Firmware ver | FFFF | Pre Condition 1 | SEQ 128K W | RW Mix | Outer Loop | |
| Operating System | CentOS 6.3 | Capacity | 250 GB | TOIO - TC/QD | TC 2/ QD 16 | Block Sizes | Inner Loop | |
| Test SW | CTS 6.5 1.13.8 | Interface | SATA 6Gb/s | Duration | Twice User Capacity | TOIO - TC/QD | TC 4/QD 16 | |
| Test SW Info | 1.10.7/1.9.16 | NAND Type | MLC | Pre Condition 2 | IOPS Loop | Steady State | 1 - 5 | |
| Test ID No. | R30-5146 | PCIe NVM | N/A | TOIO - TC/QD | TC 4/ QD 16 | Test Stimulus 2 | N/A | |
| HBA | LSI 9212-4e4i | Purge Method | Security Erase | SS Rounds | 1 - 5 | TOIO - TC/QD | N/A | |
| PCIe | Gen 2 x 8 | Write Cache | WCE | Note | - | Steady State | N/A | |
| Client IOPS - ALL RW Mix & BS - Tabular Data | | | | | | | | |
| Block Size (KIB) | Read / Write Mix % | | | | | | | |
| | 0/100 | 5/95 | 35/65 | 50/50 | 65/35 | 95/5 | 100/0 | |
| 0.5 | 49,053.4 | 32,137.8 | 21,205.1 | 21,452.2 | 22,868.5 | 45,001.1 | 112,880.9 | |
| 4 | 62,079.1 | 27,433.1 | 18,450.9 | 18,515.8 | 19,760.5 | 37,734.1 | 70,630.4 | |
| 8 | 32,683.6 | 16,954.9 | 11,394.2 | 11,547.9 | 11,968.9 | 22,078.8 | 45,403.6 | |
| 16 | 16,306.5 | 10,776.8 | 7,430.6 | 7,536.9 | 7,756.2 | 12,587.9 | 26,747.9 | |
| 32 | 8,137.5 | 6,903.3 | 4,821.4 | 4,894.0 | 5,156.5 | 7,500.5 | 15,215.7 | |
| 64 | 4,070.8 | 4,097.6 | 2,980.1 | 3,044.3 | 3,218.5 | 4,650.9 | 8,169.2 | |
| 128 | 2,034.1 | 2,113.2 | 1,830.5 | 1,912.9 | 2,034.1 | 2,827.4 | 4,224.2 | |
| 1024 | 253.4 | 263.6 | 293.6 | 317.5 | 352.9 | 474.9 | 540.6 | |

Figure 6. IOPS Tabular Data

Block sizes are listed by row with RW mixes by column. The first column represents 100% Writes (0/100) while the far right column is 100% Reads (100/0). Note that the random 4KiB 100% W, random 8KiB 65:35 RW and the random 4KiB 100% R are highlighted in yellow to illustrate commonly viewed BS/RW mixes.

Random 4KiB 100% R and 100% W are commonly listed accesses for several reasons including:

- Random 4KiB RW is a large component of many workloads
- Random 4KiB is a corner case access used for small block random tests (e.g. WSAT)

There is interest in the random 4KiB / 8KiB 65:35 RW mix because:

- Random 4KiB & 8KiB 65:35 RW mix is seen in many operating systems and applications
- Many Enterprise and Client workloads show a high percentage of accesses with a 65:35 RW mix
- It is a convenient reference for a blend of a block size's R and W values

In other words, by extracting these 3 IOPS values, one can quickly see the SSD IOPS performance at small block random 100% R and 100% W as well as commonly observed blended 65:35 RW mix workloads.

General trends can also be seen by examining a single row of BS (e.g. random 4KiB row) or by viewing a single RW mix column (e.g. 100% W). This can indicate possible SSD optimization for random 4KiB Writes, for example, as this is a popular benchmark BS/RW mix. The progression of IOPS values in the 100% W column can also indicate potential optimization for a single block size (such as random 4KiB seen above).

A closer look at Throughput (TP)

Steady State Throughput

TP measures large block sequential 100% R and 100% W. A higher value is generally better. The Client PTS-C measures a block size of 1024KiB while the Enterprise PTS-E measures both sequential 128KiB and sequential 1024KiB block sizes at 100% R and 100% W. See Figure 7 below.

Throughput (TP) is measured at Steady State (SS) by applying the PTS Pre-conditioning methodology and 5 Round Steady State Determination formula. The relevant test platform, test set-up, test conditions and parameter values are listed as well as the SS determination (SS in Rounds 1-5 for this test).

| Test Run Date: | | 03/02/2013 03:30 PM | | Report Run Date: | | 05/13/2013 05:48 PM | |
|--|-----------------------------|---------------------|-------------------|-----------------------|---------------------|---------------------|------------|
| Throughput Test (REQUIRED) - Report Page | | | | | | | |
| SNIA | Solid State Storage | | | TP - SEQ 1024 KIB R/W | | Rev. | PTS-C 1.1 |
| SSS TWG | Performance Test Spec (PTS) | | | | | Page | 4 of 5 |
| Vendor: | ABC Co. | SSD Model: | ABC Co. MLC-A 256 | | TEST SPONSOR | CALYPSO Systems | |
| Test Platform | | Device Under Test | | Set Up Parameters | | Test Parameters | |
| Ref Test Platform | Calypso RTP 2.0 | Mfgr | ABC Co. | Data Pattern | RND | Data Pattern | RND |
| Motherboard | Intel 5520HC | Model No. | MLC-A | AR | 100% | AR & Amount | 100% |
| CPU | Intel XEON 5580W | S/N | 123456 | AR Segments | N/A | Test Stimulus 1 | TP Loop |
| Memory | 8 GB PC1600 DDR2 | Firmware ver | ABCDEF | Pre Condition 1 | SEQ 128K W | RW Mix | Outer Loop |
| Operating System | CentOS 6.3 | Capacity | 256 GB | TOIO - TC/QD | TC 1/QD 32 | Block Sizes | Inner Loop |
| Test SW | CTS 6.5 1.13.8 | Interface | SATA 6Gb/s | Duration | Twice User Capacity | TOIO - TC/QD | TC 1/QD 32 |
| Test SW Info | 1.10.7/1.9.16 | NAND Type | MLC | Pre Condition 2 | TP Loop | Steady State | 1 - 5 |
| Test ID No. | R30-5156 | PCIe NVH | N/A | TOIO - TC/QD | TC 1/QD 32 | Test Stimulus 2 | N/A |
| HBA | LSI 9212-4e4i | Purge Method | Security Erase | SS Rounds | 1 - 5 | TOIO - TC/QD | N/A |
| PCIe | Gen 2 x 8 | Write Cache | WCE | Note | - | Steady State | N/A |
| Client Throughput - ALL RW Mix & BS - Tabular Data | | | | | | | |
| Block Size (KIB) | | Read / Write Mix % | | | | | |
| 1024 | | 241.5 | 532.6 | | | | |

Figure 7. Throughput Tabular Data

Note that the SSD interface is via a SAS 6Gb/s HBA (LSI 9212-4e4i) installed on an Intel 5520HC motherboard. This SAS 6Gb/s HBA interface allows the measured bandwidth (241 MB/s W, 532 MB/s R) to exceed the 3Gb/s bandwidth limit of native motherboard IC IOHR SATA2 register interface.

Figures 8 and 9 below show Steady State Convergence for both sequential 1024KiB 100% R and 100% W. Note: In this example, the Steady State Convergence occurred in the first five Rounds and is thus the same as the Steady State Window (i.e. Steady State Convergence can take longer than five Rounds)

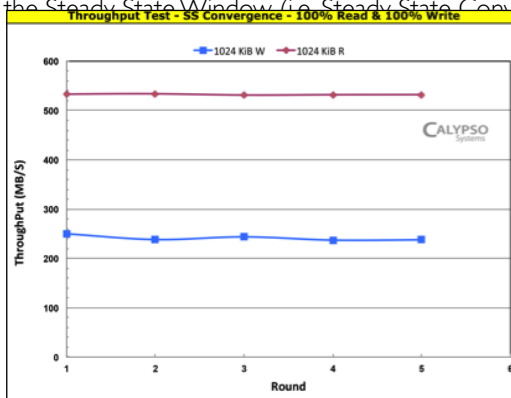


Figure 8. Throughput SS Convergence

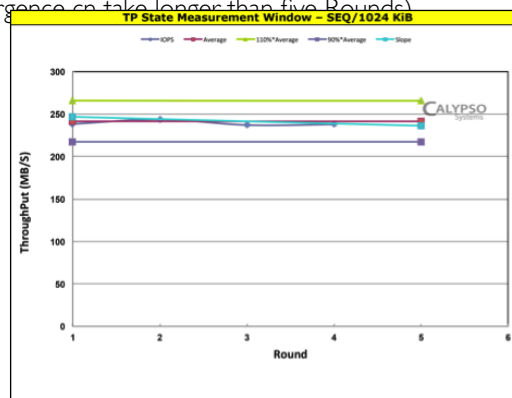
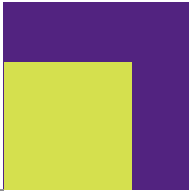


Figure 9. Throughput SS Window



Reads and Writes

A closer look at Latency (LAT)

SS Latency

The Latency (LAT) test measures Response Times of a single IO. Lower Response Times are generally better. LAT reports both AVE and MAX Response Times for 3 BS and 3 RW mixes. See Figure 10 below.

| Test Run Date: | | 05/19/2013 02:24 PM | | Report Run Date: | | 05/19/2013 04:52 PM | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|--------------------|-----------------------------|----------------|-------------------------------|---------------------|------------------------|------------|----------------------|--|--|--|------------------|--------------------|--|--|-------|-------|-------|-----|------|------|------|---|------|------|------|---|------|------|------|----------------------|--|--|--|------------------|--------------------|--|--|-------|-------|-------|-----|-------|------|------|---|-------|------|------|---|-------|------|------|
| LATENCY Test (REQUIRED) - Report Page | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SNIA | | Solid State Storage | | LATENCY - Response Time OIO=1 | | Rev. PTS-E 1.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| SSS TWG | | Performance Test Spec (PTS) | | | | Page 4 of 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Vendor: | | ABC Co. | | SSD Model: | | ABC Co. MLC-A 256 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | TEST SPONSOR | | CALYPSO Systems | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Test Platform | | Device Under Test | | Set Up Parameters | | Test Parameters | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Ref Test Platform | Calypso RTP 2.0 | Mfgr | ABC Co. | Data Pattern | RND | Data Pattern | RND | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Motherboard | Intel S520HC | Model No. | MLC-A | AR | 100% | AR & Amount | 100% | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CPU | Intel XEON 5580W | S/N | 123456 | AR Segments | N/A | Test Stimulus 1 | LAT Loop | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Memory | 8 GB PC1600 DDR2 | Firmware ver | ABCDEF | Pre Condition 1 | SEQ 128K W | RW Mix | Outer Loop | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Operating System | CentOS 6.3 | Capacity | 256 GB | TOIO - TC/QD | TC 1/QD 1 | Block Sizes | Inner Loop | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Test SW | CTS 6.5 1.13.8 | Interface | SATA 6Gb/s | Duration | Twice User Capacity | TOIO - TC/QD | TC 1/QD 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Test SW Info | 1.10.9/1.9.16 | NAND Type | MLC | Pre Condition 2 | LAT Loop | Steady State | 1-5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Test ID No. | R30-5314 | PCIe NVM | N/A | TOIO - TC/QD | TC 1/QD 1 | Histogram | N/A | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| HBA | LSI 9212-4e4i | Purge Method | Security Erase | SS Rounds | 1-5 | TOIO - TC/QD | N/A | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| PCIe | Gen 2 x 8 | Write Cache | WCE | Note | - | Note | - | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Average and Maximum Response Time - ALL RW Mix & BS - Tabular Data | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"> <thead> <tr> <th colspan="4">Average Latency (ms)</th> </tr> <tr> <th rowspan="2">Block Size (KIB)</th> <th colspan="3">Read / Write Mix %</th> </tr> <tr> <th>0/100</th> <th>65/35</th> <th>100/0</th> </tr> </thead> <tbody> <tr> <td>0.5</td> <td>0.20</td> <td>0.24</td> <td>0.13</td> </tr> <tr> <td>4</td> <td>0.19</td> <td>0.24</td> <td>0.14</td> </tr> <tr> <td>8</td> <td>0.29</td> <td>0.42</td> <td>0.19</td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th colspan="4">Maximum Latency (ms)</th> </tr> <tr> <th rowspan="2">Block Size (KIB)</th> <th colspan="3">Read / Write Mix %</th> </tr> <tr> <th>0/100</th> <th>65/35</th> <th>100/0</th> </tr> </thead> <tbody> <tr> <td>0.5</td> <td>38.92</td> <td>9.31</td> <td>0.79</td> </tr> <tr> <td>4</td> <td>19.10</td> <td>9.37</td> <td>0.79</td> </tr> <tr> <td>8</td> <td>34.43</td> <td>9.38</td> <td>6.25</td> </tr> </tbody> </table> | | | | | | | | Average Latency (ms) | | | | Block Size (KIB) | Read / Write Mix % | | | 0/100 | 65/35 | 100/0 | 0.5 | 0.20 | 0.24 | 0.13 | 4 | 0.19 | 0.24 | 0.14 | 8 | 0.29 | 0.42 | 0.19 | Maximum Latency (ms) | | | | Block Size (KIB) | Read / Write Mix % | | | 0/100 | 65/35 | 100/0 | 0.5 | 38.92 | 9.31 | 0.79 | 4 | 19.10 | 9.37 | 0.79 | 8 | 34.43 | 9.38 | 6.25 |
| Average Latency (ms) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Block Size (KIB) | Read / Write Mix % | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | 0/100 | 65/35 | 100/0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.5 | 0.20 | 0.24 | 0.13 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 0.19 | 0.24 | 0.14 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 0.29 | 0.42 | 0.19 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Maximum Latency (ms) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Block Size (KIB) | Read / Write Mix % | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | 0/100 | 65/35 | 100/0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.5 | 38.92 | 9.31 | 0.79 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 19.10 | 9.37 | 0.79 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 34.43 | 9.38 | 6.25 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 10. Average & Maximum Latency Tabular Data

Latency AVE and MAX Response Times are also reported in 3D Bar plots for all BS/RW mixes. Each value represents the AVE or MAX of Response Times recorded over each one-minute test period for the BS/RW mix. As expected, R latencies are much lower (faster) than W latencies. Figure 11 shows AVE Response Times in the microseconds compared to Figure 12 showing MAX Response Times in milliseconds.

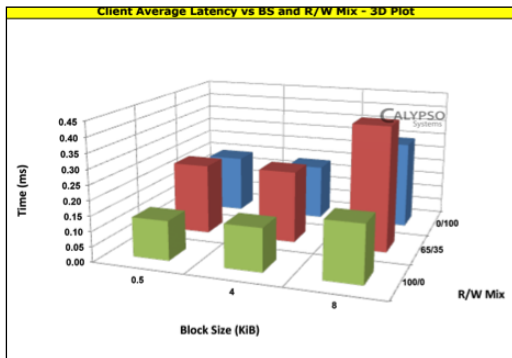


Figure 11. Average Latency

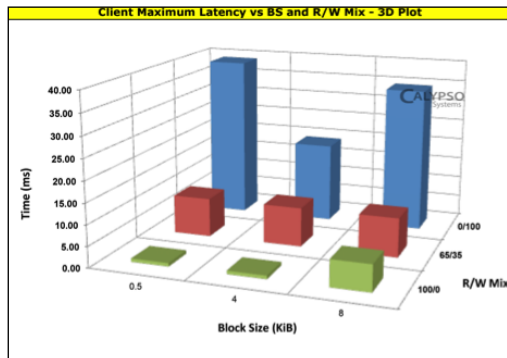


Figure 12. Maximum Latency

Response Time (RT) Histograms

Response Time histograms show the frequency and distribution of response times over the test period. A higher clustering of faster times (to the left) with fewer slower times (to the right) is generally better.

AVE Latency only shows the “average” of the response times recorded. MAX Latency only shows the slowest response time measured during the test period. Histograms provide a full picture of where all of the IO response times occur during the test period. Figure 13 is a histogram of response times. “Response Time Counts” are on the Y-axis and “Time Bins” in milliseconds (mS on the chart) are on the X-axis.

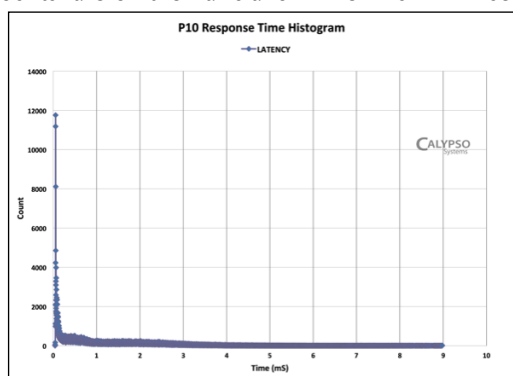


Figure 13. Latency Histogram

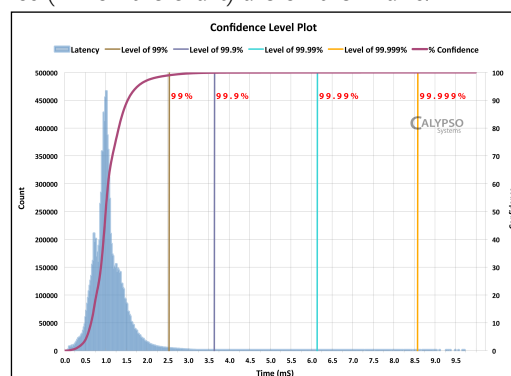


Figure 14. Confidence Plot

Confidence Plots

Confidence plots show the percentage of all response times that occur at a given time threshold – at what time value will, for example, 99.99% of the IOs occur. A higher percentage at a faster time is better.

Confidence plots can be used as a Quality of Service (QoS) metric to show how many IOs could be dropped if they occur slower than a given time value. The “number of nines” is shorthand for indicating QoS - 99.99% or 99.999% is “4 nines” or “5 nines” of Confidence. Figure 14 shows the 99%-99.999% points when that percent of all the IOs completed – i.e. “2 nines to 5 nines” of Confidence.

How Parameter Settings Affect Performance

Required PTS test parameters

The PTS has required test settings for both Client and Enterprise tests. This is done to emulate specific workload environments for the tests and to address known behaviors of NAND flash SSDs.

For example, the Pre-conditioning (PC) and Steady State determination methodology was developed to address the fact that SSD performance changes over time. However, the PC and SS set-up parameters are set differently in the PTS-C and the PTS-E to reflect the differences in workload environment for Client and Enterprise tests.

Establishing required parameter settings for PTS-C and PTS-E tests ensures that:

- Tests are relevant to the intended workload,
- SSD performance measurements can be repeated, and



- Results can be compared between SSDs.

Optional PTS Test Parameters

It is also important to know how parameter settings can affect performance. Optional test parameters can be changed (and disclosed) to investigate other dimensions of performance behaviors.

For example, SSDs are very sensitive to over provisioning (reserving LBAs from usage). LBAs reserved from Pre-conditioning Active Range (PC AR) and Test Active Range (Test AR) can effectively become over provisioned NAND space. This can improve SSD performance by making more NAND flash available to the controller. This especially helps small block RND W behavior. An example of over provisioned SSD performance is provided later.

Data Pattern (DP)

Data pattern refers to the degree of randomness of the data that comprise the access pattern. Most testing uses a random data pattern as opposed to a non-random (or repeating/compressible) data pattern.

Do not confuse the data pattern (nature of the binary data that is transferred) with the access pattern (nature of the block addresses that are used to access the data transferred). Some SSDs take advantage of non-random data patterns by using data compression and de-duplication algorithms to write less data to the SSD, effectively increasing performance.

Figures 15 and 16 below show the effects of using a random and a non-random data pattern. A highly compressible 100MB binary file from a database application was used for the non-random data pattern. The non-random data pattern is marked DP=File. The DP=File IOPS level (red line) is higher and writes more data in a shorter time than the DP=RND (blue line).

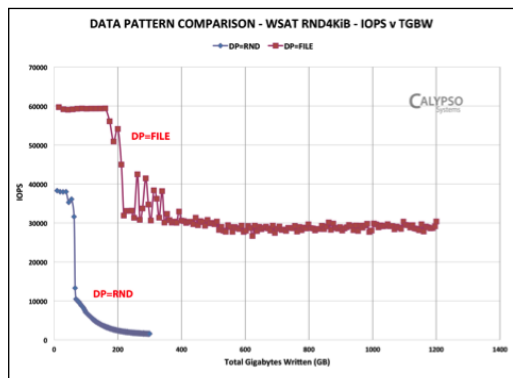


Figure 15. RND v Non RND DP IOPS v TGBW

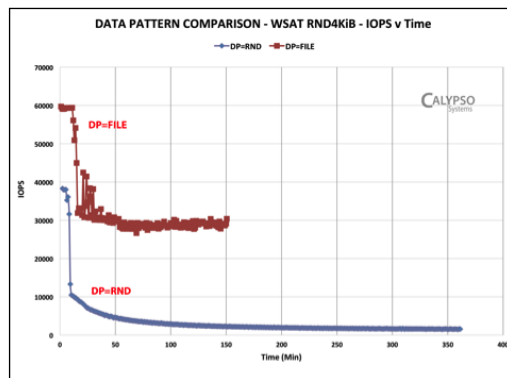


Figure 16. RND v Non RND DP IOPS v Time

Write Cache Setting: WCE v WCD

Write cache settings – write cache enabled or disabled (WCE and WCD respectively) - can be changed and disclosed. The required setting for the Client PTS-C is WCE; the required setting for Enterprise PTS-E is WCD. Figure 17 following shows the effects of write cache by comparing two WSAT runs: one with WCE and one with WCD. In the WCD data series, the initial FOB peak level and ensuing IOPS level is much lower than in WCE. This particular SSD is a Client 256GB SATA drive designed for WCE setting.

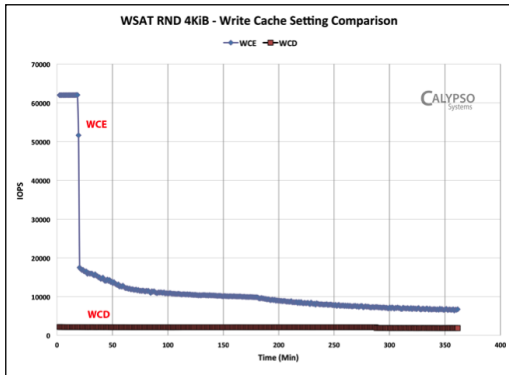


Figure 17. Write Cache Setting

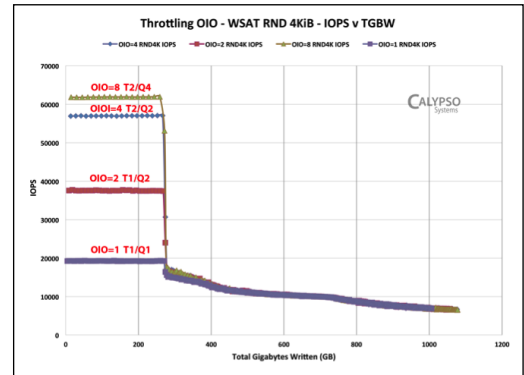


Figure 18. Varying OIO

Varying Outstanding IO (OIO): Demand Intensity

OIO = Thread Count x Queue Depth (TC x QD). Limiting TC and QD shows how the availability of IO stimulus, or the demand intensity, can affect IOPS levels. For the SSD in Figure 18 above, the optimal OIO=8. The OIO was reduced on successive test runs to view the effect of OIO throttling. Note: saturating OIO can also result in less than the maximum possible IOPS and an increase in response times.

Over Provisioning (OP): Limited Pre-conditioning (PC AR) & Test Active Ranges (Test AR)

Over provisioning (OP) occurs when user LBAs are reserved from both Pre-conditioning AR and Test AR. LBAs reserved from PC AR can create OP space and improve performance.

In Figure 19 below, AR=25% means that 25% of the LBAs are used for PC and Test AR while 75% of the LBAs are not used, resulting in an OP ratio of 3:1. AR=100% means all the LBAs are available for PC and Test AR resulting in an OP=0. A higher OP (and lower AR) will often show higher IOPS values.

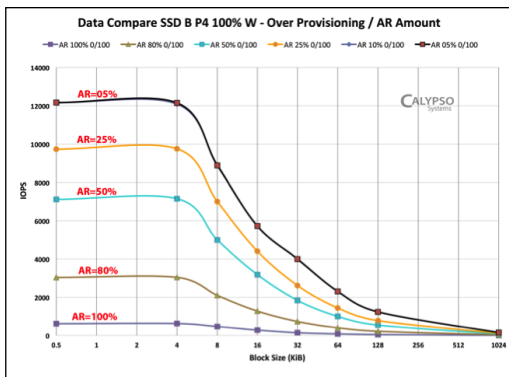


Figure 19. Over Provisioning

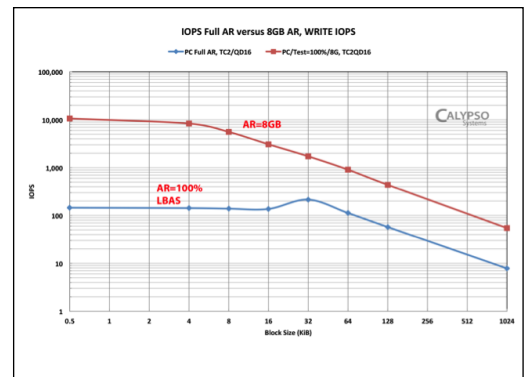


Figure 20. Varying Active Range

PTS-C v PTS-E Active Ranges: PC AR=80%/Test AR=8GiB v PC AR=100%/Test AR=100%

PTS-C allows use of a limited PC & Test AR. Pre-conditioning AR may be reduced to 80% to emulate the effect of TRIM. This limits the amount of OP the SSD (as seen above) to 20%. Figure 20 shows PTS-E PC & Test AR=100% and PTS-C with PC AR=80% and Test AR=8GiB (which has higher IOPS). Note: Marketing claims often inflate IOPS performance by limiting PC AR (thereby increasing OP) and Test AR to 8GB.



III. Conclusion

SSD product design and architectures continue to evolve. Advances in SSD NAND flash technologies, increasingly sophisticated SSD firmware, optimization of the IO software stack and more powerful hardware will lead to increased SSD performance levels into the hundreds of thousands to millions of IOs per second.

The use of the SNIA SSS Performance Test Specification and RTP can provide a means to accurately and reliably test these high performance SSDs and to create an industry wide baseline for SSD performance test.

The PTS test specification provides a proven methodology for testing SSD performance. The pending PTS-E 1.1 also adds new application based test workloads to help understand how the SSDs will perform under commonly accepted use cases. As additional work by industry groups provide more application specific access pattern information, increasingly sophisticated and accurate test workloads can be administered by the PTS.

The PTS provides a uniform and industry standard methodology for evaluating SSD performance.

Special Thanks

This white paper would not have been possible without the significant effort by many editors and contributors. I would like to express appreciation to all of those who helped create this primer and especially: Chuck Paridon *HP*, Doug Rollins *Micron*, Tom West *hyperI/O*, Paul von Behren *Intel*, Wayne Adams *SNIA Chair and EMC*, Phil Mills *IBM*, Paul Wassenberg *Marvell*, and Mike Peeler *Calypso*. A special thank you to Jennifer Coley, without whose constant feedback and questions, the explanations and descriptions would not be nearly as clear.

About the Author

Eden Kim is CEO of Calypso Systems, Inc., a solid state storage test and measurement company and the manufacturer of the Calypso RTP and CTS test software upon which the PTS and data contained herein were developed. The Calypso RTP/CTS was also used to develop, validate and run the PTS and is used in the SSSI Certified Test Lab for PTS test services. Calypso can be found at www.calypsotesters.com.

Eden is Chair of the SNIA Solid State Storage Technical Working Group which developed the PTS. Eden is also Chair of the SSSI PCIe SSD Committee, Chair of the SSSI TechDev Committee and a member of the SSSI Governing Board.

Eden is a graduate of the University of California.

Appendix

I. Test Practices

For comparative performance testing, it is critically important to use a **Reference Test Platform** and to **Isolate the SSD** under test. Select an *appropriately robust hardware system* that will generate an adequate workload to the SSD and will not bottleneck performance. Select an *OS that minimizes performance impact* on the test SSD and a *known test software* that adequately manages the test stimulus generation and measurement. Be sure to test as close to the *SSD device (Block IO)* level as possible.

Use the same test platform to normalize test measurements. Even using a slower hardware system with a modest test software tool, while not optimal, can generate valid comparative results if the *SAME* test environment is used to compare the SSDs. Ideally, the test operator should use an industry standard SSSI RTP.

2. SNIA Report Format

The SSS PTS prescribes a standardized report format for the reporting of PTS SSD performance data. SNIA reports provide a uniform way to present data and to disclose both required and optional test set-up and parameter settings related to the reported measurements. See Figure A-1 below.


| Test Run Date: | | 05/14/2013 02:11 PM | | Report Run Date: | | 05/25/2013 12:27 PM | |
|--|---|---------------------|----------------|------------------------|---|---------------------|---------------|
| Write Saturation Test (REQUIRED) - Report Page | | | | | | | |
| SNIA SSS TWG | Solid State Storage Performance Test Spec (PTS) | | | WSAT - RND 4KiB 100% W | | Rev. | PTS-C 1.1 |
| Vendor: | | ABC Co. | SSD Model: | MLC A 250 GB | | Page | 1 of 4 |
| | | | | TEST SPONSOR |  | | |
| Test Platform | | Device Under Test | | Set Up Parameters | | Test Parameters | |
| Ref Test Platform | Calypso RTP 2.0 | Mfgr | ABC Co. | Data Pattern | RND | Data Pattern | RND |
| Motherboard | Intel 5520HC | Model No. | MLC 250 GB | PC AR | 100% | AR & Amount | 100% |
| CPU | Intel 5580W | S/N | 123 456 | AR Segments | 1024 | Test Stimulus 1 | WDPC |
| Memory | 16GB ECC DDR3 | Firmware ver | FFFF | Pre Condition 1 | WDPC | TOIO - TC/QD | OIO 32-T2/Q16 |
| Operating System | CentOS 6.3 | Capacity | 250 GB | TOIO - TC/QD | OIO32-T2/Q16 | Steady State | 4X User Cap |
| Test SW | CTS 6.5 1.14.6 | Interface | SATA 6Gb/s | Duration | 24 hr | Time | 24 Hr |
| Test SW Info | 1.13.27/1.9.129-e16 | NAND Type | MLC | Pre Condition 2 | N/A | Test Stimulus 2 | N/A |
| Test ID No. | R30-5307 | PCIe NVM | N/A | TOIO - TC/QD | | TOIO - TC/QD | |
| HBA | LSI 9212-4e4i | Purge Method | Security Erase | SS Rounds | | Steady State | |
| PCIe | Gen 2 | Write Cache | WCE | Note | | Time | |

Figure A-1. SNIA WSAT Report Header

3. Drive Preparation: PURGE, Pre-conditioning and Steady State

PURGE - All performance tests should start with a Device PURGE. A PURGE is the use of an ATA SECURITY ERASE command, SCSI FORMAT UNIT command, or a proprietary command that puts the drive in a state “as if no writes have occurred.” PURGE provides a known and repeatable test starting point.

Pre-conditioning (PC) - One of the most important steps in performance testing is *Pre-conditioning the drive to a Steady State using (and disclosing) an industry standard methodology*. There are various PTS PC and SS methods. Regardless of the type of PC used, the selected PC should put the drive into a Steady State



and the PC and SS test methodology should be disclosed in the test reporting.

Test Loop Block Size Sequencing - PC for IOPS, TP & LAT applies a WDPC test loop sequence of BS & RW mixes until SS is reached. For example, the IOPS test loop runs one-minute test intervals for 56 different RW/BS mixes. There are 7 RW mixes and 8 Block Sizes comprising 56 one-minute test periods per loop. This test loop is applied after PURGE and WIPC until 5 consecutive one-minute Rounds for the tracking variable meet the SS Determination Criteria. See Figure A-2 below.

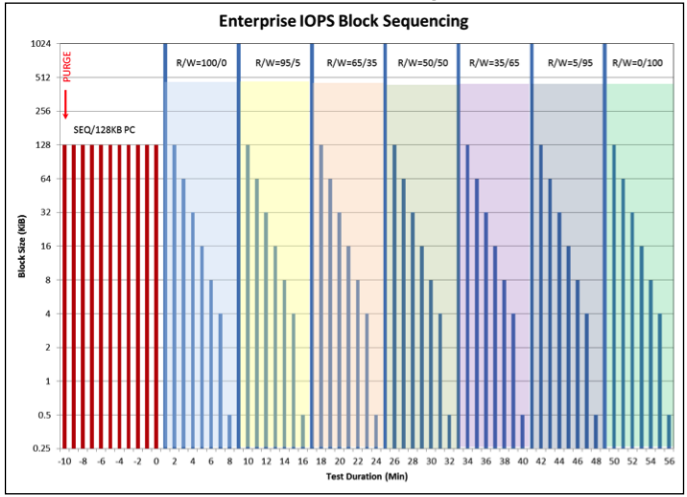


Figure A-2. IOPS Block Size Sequencing

SS Determination Criteria & Calculation - SS selects test results from a SS Window wherein 5 consecutive test round values for the tracking variable do not exceed a 20% data excursion nor a 10% slope ("20/10" criteria) when plotted to a least squares linear fit of the data points. Test results are only taken from the data in the SS Window. SS Window calculations are shown at the bottom of Figure A-3.

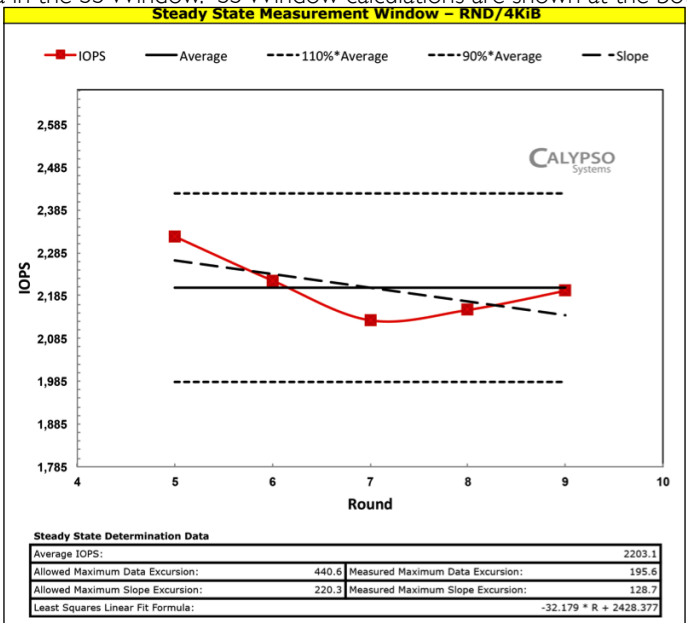


Figure A-3. Steady State Determination



Storage Networking Industry Association

425 Market Street, Suite 1020 • San Francisco, CA 94105 • Phone: 415-402-0006 • Fax: 415-402-0009 • www.snia.org