Burlywood

Igniting Data Center Storage Innovation™

WHITE PAPER

Why Your Concerns about Data Center SSDs are Justified: Can You Trust the Benchmarks?

By Tod Earhart
CTO and Founder

February 2023

Why Your Concerns about Data Center SSDs are Justified: Can You Trust the Benchmarks?

Tod Earhart, CTO and Founder, Burlywood, Inc. | February 2023

Contents

Why Your Concerns about Data Center SSDs are Justified: Can You Trust the Benchmarks?

Tod Earhart, CTO and Founder, Burlywood, Inc. | February 2023

ABSTRACT

Today, SSDs (Solid-State Drives) are mainstream, and systems and applications have advanced, demanding much more from them. These increasing demands have turned the tables where the HDD (Hard Disk Drive) infrastructure has become obsolete and a hindrance. The replacement technologies like PCIe (peripheral component interconnect express) and NVMe (nonvolatile memory express) have exposed major issues with SSDs as they approach their limits in performance and latency under complex workloads.

This paper continues the series and focuses on the process, tools, and measurements used to analyze the operational performance of SSDs during the evaluation and qualification process. We will show that the industry's standard tests and reported benchmarks can be extremely misleading as a tool to predict SSD performance in a real system.

EXECUTIVE SUMMARY

Essential points that are covered in this white paper:

- SSDs require internal operations for HDD emulation. The internal operations are design dependent and cause significant issues, including degraded performance, inconsistent latency, latency spikes and stalls, and early wear-out.
- SSD performance is extremely workload dependent. Internal operation activation and the intensity are driven by the workload characteristics and the SSD design.
- Workloads are complex, but only a broad understanding of the workload is needed to develop effective tests. It is best to view workloads directly, but they can be estimated if the proper tools are unavailable.
- All SSD metrics are workload dependent - including power consumption and endurance. An effective test must be representative of the actual workload to yield valid results.
- Tests must be run long enough to fully activate all internal operations and observe infrequent, high-latency events.
- The metrics to measure SSD operational performance are derived from HDDs. They need to be modified to accurately assess SSD performance.
- The standard benchmarks used today are HDD-based and ineffective when used to predict actual SSD performance. A new, SSD-centric benchmarking approach is needed.

SSD OPERATIONAL ISSUES

Commodity SSD designs are adequate for applications that only need a faster HDD. However, serious issues are exposed if an application requires the best performance and behavior an SSD can offer.
Some of the most common SSD operational issues include:

1. Inconsistent performance and latency, especially when drives approach capacity;
2. Intermittent latency spikes, sometimes exceeding multiple seconds;
3. Significant performance drop-off after weeks or months of usage;
4. Premature warranty wear-out based on rated endurance specifications;
5. Service life reductions that are workload related; and
6. Early device failures in production environments.

These issues result in severe, if not intolerable, impacts on systems and applications.

ROOT CAUSES

SSD emulation of an HDD using NAND flash media is complex. The properties and behavior of NAND flash storage media used in SSDs differ significantly from the magnetic disk media used in HDDs. SSDs require a sophisticated controller to perform many internal operations not required by HDDs. These operations manifest as second-order performance and latency impacts. These impacts become significant when complex workloads stress the SSD.

The design of an SSD's internal operations, storage media management processes, and activity scheduling are not standard and vary significantly across SSD vendors and models. The variations result in very different behaviors when the internal operations are fully activated.

When and how these operations are invoked is extremely dependent on the applied workload.

Internal Operation	Description
Garbage Collection	The process of freeing up space in the flash array by moving valid data out of used blocks to enable erasing and re-use of those blocks.
Wear Leveling	The process of moving data around within the array to achieve uniform wear caused by erasing. This maximizes drive life.
Bad Block Management	This process tests active blocks and retires them if they are defective or worn out, often requiring data movements
Array Usage and Queuing	The SSD consists of groups of data storage called planes. Planes support only one operation at a time (e.g., read, write, or erase). Delays occur when multiple operations are queued for the same plane.

Most internal operations involve data movement from one area of NAND to another. The amount of internal data movement is expressed as the Write Amplification Factor (WAF).

WAF = Total Data Written to the NAND Flash Array Total Data Written To the SSD

Most SSDs report their WAF since it is a measure of operational efficiency. The best possible WAF = 1, meaning no internal data movement, which is the case for HDDs. Typical SSD WAFs range from 2-5.

The WAF is both design and workload dependent. As the WAF increases, performance decreases, latency variability increases, average power increases, and drive life decreases.

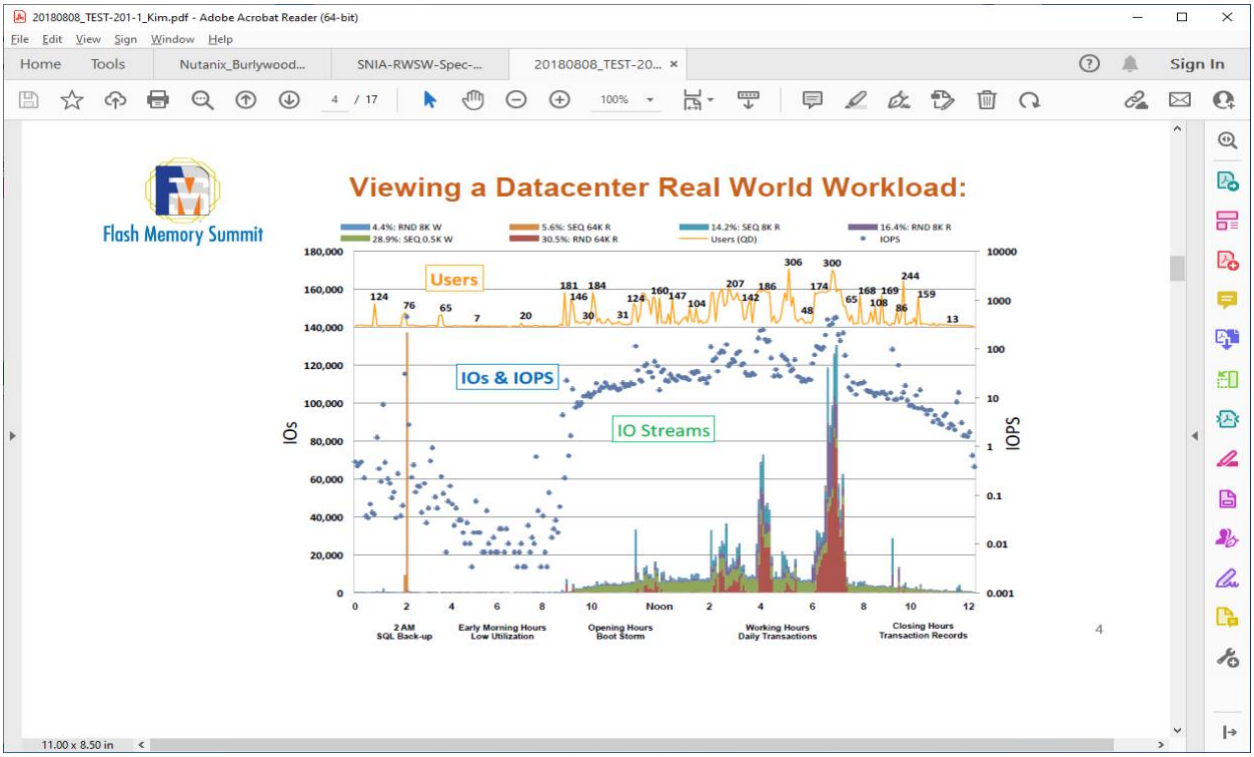
Key Learnings:

1. SSD operational performance varies significantly by vendor, model, and capacity;
2. SSD operational performance is highly dependent on the workload; and
3. The WAF (Write Amplification Factor) is an essential measure of SSD operational efficiency and is also workload dependent.

Understanding these facts is critically important when evaluating and comparing SSDs for your systems.

REAL WORKLOADS

The workload is the type, order, address, and size of the read and write commands presented to the SSD.



Source: Eden Kim, Calypso Systems, Inc., Flash Memory Summit Presentation, 2018.

The diagram shows that workloads can be extremely complex. The system software stack and hardware components above the SSD generate the workload.

A software stack has multiple layers, including the application layer that may be virtualized using VMs or containers, a hypervisor or scheduler, a base OS, a storage management layer, and connectivity to the SSDs. Each layer of the stack contributes to the workload, with the lowest layers having the most significant impact to the extent that they may mute or even hide the application layer characteristics.

There are no convenient tools to view or record the workload. Building this capability into systems has not been necessary because the workload has minimal impact on HDD performance. Knowing this, Burlywood has integrated real-time workload recording capability into its SSDs. This feature provides Burlywood and its clients invaluable information for understanding and optimizing system behavior and facilitating workload-aware™ SSD design.

Understanding your workload can seem daunting if you do not have the tools to view the workload at the SSD. Fortunately, it is not necessary to understand the workload in detail. It is only necessary to know the broad workload characteristics and how they activate the internal SSD operations previously mentioned in the "Root Causes" section.

The broad characteristics of a workload include:

1. Typical and most prevalent block sizes;
2. Traffic mix - read-to-write ratio both at critical, high load times and over the long term;
3. Command load – average and peak number of commands in flight at a time; and
4. Distribution of reads and writes across the address range.

To estimate your workload, start from the bottom of the software stack just above the SSD and work your way up. Consider the workload characteristics at each layer and how they may be affected by the layer below. If available, take advantage of software-based IO measurement tools as you work up the stack. It should not be necessary to go up too many layers to get a general sense of the workload.

Characteristic	Estimation Guidance
Block Size	Typically, a mix and is very application dependent. Most stacks tend towards larger block sizes for efficiency.
Read/Write Mix	Very application dependent. Usually, can be estimated based on the application and software stack
Command Load	Application, system, and configuration dependent. Command loading decreases as SSD performance increases and when more SSDs are used in parallel in a system
Distribution across address range	Almost every workload has a non-uniform address distribution (Hot / Cold mix)

Defining each of these characteristics creates a model of the workload. It is important to select realistic parameters to generate valid models.

More accurate workload models can be created with increased insight into the workload. A workload recording and analysis provides an accurate profile of the block sizes used, the read/write mix, command loading, and addresses used.

Key Learnings:

1. Workloads are generated by the entire software stack and can be complex;
2. It is difficult to view and measure the workload at the SSD;
3. A workload can be estimated using 4 broad characteristics; and
4. Directly recording and analyzing the workload generates the statistics needed to create an accurate model.

In order to yield valid results, tests used to evaluate and compare SSDs must simulate the workload.

WORKLOAD IMPACTS ON SSD INTERNAL OPERATIONS

Internal operations are activated by the workload characteristics.

SSD Operation	Activation and Intensity	Impacting Workload Characteristics
Garbage Collection	Starts after enough data has been written to fill the SSD once. The intensity increases with the volume of write traffic	Traffic Mix (Writes) Command Loading
Wear Leveling	High variation of the block erase counts across the NAND Flash array. It is driven by Hot/Cold mixed write traffic.	Distribution across address range
Bad Block Management	Always present. Increases with write activity and drive wear	Traffic Mix (Writes)
Array Usage and Queuing	Issuance of simultaneous commands to the NAND Flash array. Impact increases with the number of commands in flight, mixing reads and writes, and the intensity of internal operations.	Block Size Traffic Mix (Even) Command Loading

This is a complex system. The internal operations driven by the workload characteristics impact the host commands and each other. To observe the impacts of internal operations during testing a realistic workload model must be used.

Considering that the internal operations are design dependent and vary widely across SSD vendors, the internal operations must be appropriately activated to compare SSDs accurately.

Most internal operations are driven by writes and increase with write loading. When simulating a workload, it is important to use reasonable parameters or profiles for each workload characteristic to make the observed SSD behavior realistic.

Key Learnings:

1. The internal operations are driven by the specified broad workload characteristics;
2. The design and implementation of internal operations vary significantly across SSDs;
3. The interactions between the internal operations and host commands are complex and unpredictable;
4. An SSD must be tested under a realistic workload to predict its performance in a real system;
5. Realistic parameters or profiles must be used for every workload characteristic to yield realistic in-system SSD behavior.

Using a good workload model to observe valid SSD performance results is important.

SSD EVALUATION AND QUALIFICATION

The goal of evaluation is to compare SSDs against each other. The standard metrics and tests are inherited from the HDD legacy. Given the vast difference between SSD and HDD behavior, it is worth taking a closer look at the effectiveness of these tests and the relevance of the results.

HDD performance has little dependence on the controller design or the workload it is placed under. The overriding performance, latency, and power characteristics are driven by its physical properties such as rotational speed, head seek velocity, storage density, and mass.

Conversely, SSD performance is extremely dependent on the controller design and the workload presented.

Operational Performance Metrics

The operational performance metrics for SSDs include:

- Throughput in IOPS (input/output operations per second) or bandwidth;
- Latency - average, max, and tail (99%, 99.9%, …, 99.9999%);
- Power - Average and Peak;
- Endurance - DWPD (Drive Writes Per Day) or TBW (Terabytes Written); and
- Write Amplification Factor (WAF).

Excluding endurance and WAF, these metrics are the same as HDDs. Unlike HDDs, all these metrics are extremely workload dependent. Some additional explanations and interpretations are required to tune these metrics for SSDs.

Throughput: IOPS pertain only to a single block size, typically 512b or 4KB. SSDs internally operate on a wide range of block sizes, and real workloads have a mix of block sizes. Expressing average read and write bandwidth over a workload's mixed block size profile is more informative.

Latency: The most critical latency measurement is dependent on the application. Most applications are concerned with delays or disruptions caused by maximum or very long tail latencies.

Power: Power is workload dependent. Typical SSD data sheet specs report idle power and average power on a write-heavy benchmark test. For a given workload, these measurements may be misleading. For example, an SSD reporting good datasheet power specs may burn more power if it is inefficient (High WAF) and the workload has a high duty-cycle (little idle time).

Endurance: Endurance is also workload-dependent and needs to be reported against a workload. It is driven primarily by the Write Amplification on the anticipated workload and the NAND type.

WAF: The write amplification factor is an underlying factor for all other metrics. It is a valuable indication of the SSD's efficiency against a given workload. WAF reporting is standardized, and it can be queried in-system during operation.

The relative importance of these metrics varies with the system and application. They adequately describe the operation of an SSD, and a good system engineer can rank and assign weights to them.

Evaluation Test Effectiveness

SSDs must be tested in simulated workload environments to yield valid results. Suppose some of the workload characteristics in the test are unreasonable and/or they do not activate the internal operations in a way that the system does. In that case, the test results will be uninformative or even misleading.

The test time is also important. The SSD must be running at a steady state to avoid artificially positive results. The intensity of the internal operations ramps over time as the SSD is repeatedly filled, the flash array organization breaks down, and the age distribution of the stored valid data diverges. The metrics should be monitored during the test and recorded until they bottom out.

The interactions between the simulated workload and internal operations are complex and unpredictable. There are times when host commands will be serviced quickly and other times when many internal operations are stacked up and waiting on shared resources, resulting in significant latency events. The latency should be measured over an extended period to capture the outliers.

Standard Benchmarks

The standard benchmarks are another legacy inherited from the HDD infrastructure. These benchmarks are very effective for analyzing HDD performance and latencies. They are also effective for comparing an SSD against an HDD.

The standard HDD benchmarks are:

1. 4KB Random Read;
2. Sequential Read;
3. 4KB Random Write; and
4. Sequential Write.

SSDs are now mainstream, but these HDD-based benchmarks have remained the primary first-order performance measures touted in marketing literature. They are heavily relied on by engineers who evaluate SSDs for their systems, and they are very prominent in independent, third-party test suites.

Some important questions need to be answered:

- Are these still the right metrics?
- Are they providing helpful guidance about how an SSD will operate in a real system?
- Do they provide the correct information to compare SSDs effectively?

Recalling the previous sections about how the internal SSD operations are activated and the characteristics of real workloads, let's evaluate the benchmarks:

Test	Activated Operations	Workload Applicability
4K Random Read	Simple Command Queuing	Rare
Sequential Read	Simple Command Queuing	None - Rare
4KB Random Write	Garbage Collection Command Queueing	Rare
Sequential Write	Garbage Collection - Very Light Simple Command Queuing	None - Rare

These tests barely activate the internal operations and definitely do not stress them.
These tests have little resemblance to a real workload. At best, they may resemble small snapshots of a real workload, but the results may still be misleading if the internal operations are not properly activated.

An important feature missing from the standard benchmarks is a hot/cold write mix. Without it, wear leveling is never activated.

A strong argument can be made that these tests are ineffective at predicting how an SSD will operate in a real system, and the metrics are unsuitable for comparing SSDs.

Additional Testing Considerations

Most evaluation engineers understand that the basic benchmarks are insufficient to analyze an SSD's performance in a real system. However, they are limited by the test tools and suites available. Off-the-shelf test suites and existing tests that emulate specific application workloads need to be analyzed to ensure they activate the internal operations the same way as the actual system.

Very few SSD users have the bandwidth to define and create their own tests. If custom tests are created, the designer must have enough understanding of their workload and how to activate all of the internal SSD operations to ensure the test results are meaningful.

Given that only a broad understanding of a workload type is required for an effective test, a suite of tests representing general classes of workloads should be developed that yield more relevant results. A system designer could then choose the portion of the test suite that most closely resembles the target workload(s).

CONCLUSION

SSDs are mainstream and being pushed to their limits. A new level of sophistication is required to evaluate, measure, and compare SSDs to accurately predict their performance in real systems.

SSDs have inherited their test suites, benchmarks, and performance metric definitions from HDDs, but SSDs have much different behavior. The new approach must be SSD-centric and account for variations in SSD designs, workload impacts on SSD operation, and metric definitions that effectively predict SSD performance in a real system.

To design or select effective tests, it is important to understand the workload characteristics and how they activate the SSD's internal operations. The test stimulus must put the SSD in a realistic operational state to yield relevant performance, latency, power, and endurance metrics. If the test does not sufficiently exercise the SSD, the results will be invalid or even misleading.

Today's evaluation test tools and processes poorly predict behavior in real-life production systems. Burlywood has integrated real-time workload recording capability into its SSDs, enabling visibility, and understanding of all types of IO traffic. Analysis of the recording results in workload understanding, providing valuable information to guide more effective evaluation test design.

Burlywood uses workload recording and analysis data to optimize its SSD designs against real workloads. In contrast, other SSDs are optimized to the standard benchmarks.

Visit us online at www.burlywoodtech.com or contact us at innovate@burlywoodtech.com to explore how we can help you analyze your workloads and determine if your organization can benefit from our SSD technology.

1551 S. Sunset St., Suite E, Longmont, CO 80501 | + 1 408-320-5115 | www.BurlywoodTech.com