Introduction to Real World Workloads

A Primer

Abstract

Real World Workloads (RWWs) have a dramatic impact on Cloud, Datacenter and Enterprise storage performance because they are comprised of a constantly changing combination of IO Streams and Demand Intensity (or users). RWW IO Stream composition varies depending on where in the Software/Hardware Stack they are observed. The development of Real World Workload stimulus focuses on the capture, analysis and test of IO Streams at the file system or block IO level on physical and virtual servers.

RWWs allow Cloud, Datacenter and Enterprise Storage professionals, AI Machine Learning researchers, Distributed Edge Computing developers, Advanced Storage device designers and Storage & Server manufacturers to optimize and validate application and storage architectures. This allows users to decide how much, and what kind of, storage to design, purchase and deploy.

Eden Kim

Calypso Systems

SNIA Compute, Memory, and Storage Initiative

Table of Contents

List of Figures

Introduction

Storage performance is highly dependent on the workloads, or IO Streams, that are applied to storage. Real World Workloads (RWWs) are popular because they are comprised of IO Streams observed during real world application use. Real World Workloads are very different from synthetic workloads and affect performance in very different ways.

Storage that is optimized using synthetic workloads may not have the same performance as storage that is exposed to Real World Workloads. This is because Real World Workloads are comprised of constantly changing combinations of many IO Streams of varying Demand Intensity (or Queue Depth or users) whereas synthetic workloads apply a fixed number of IO Streams and Demand Intensity over a fixed duration.

Indeed, Real World Workloads are intended to emulate the IO Streams and Demand Intensity observed during real world use. On the other hand, synthetic workloads are intended to stress storage outside the range of normal operation.

Recent SNIA Performance Test Specifications (PTS) have enumerated standard practices for the capture, analysis and test using Real World Workloads. Calypso, the SNIA Compute, Storage & Memory Initiative (CSMI) and the SNIA Solid State Storage Technical Working Group (S3 TWG) periodically publish Reference Real World Workloads deemed to represent select use cases.

SNIA Reference Real World Workloads can be viewed on the CMSI site. SNIA Reference workloads can also be viewed, as well as replayed and downloaded, at www.testmyworkload.com (TMW site). The TMW site offers free IO Capture tools and workload visualization in addition to providing access to commercial IO Capture tools that have more advanced feature sets.

1. What are Real World Workloads?

Real World Workloads (RWWs) are the collection of IO (Input/Output) Streams observed between two points in the Software/Hardware (SW/HW) stack that occur during real world application use. Real World Workloads are typically observed between User space (applications) and Storage at the file system or block IO level.

What are IO Streams?

IO Streams are defined as specific Random or Sequential (RND/SEQ) accesses of a Read or Write (R/W) IOs of a given data transfer, or block, size. (e.g., RND 4K W, SEQ 128K R, RND 0.5K W, RND 128 byte R, SEQ 192K W, etc.).

Real World Workloads are Constantly Changing

Real World Workloads are combinations of many IO Streams (hundreds to thousands). These IO Stream combinations constantly change over time as they traverse the Software/Hardware (SW/HW) Stack and are exposed to abstractions by middleware, drivers and hardware components.

Applications, middleware and drivers can modify IO Stream composition. Changes in IO Streams also occur when IOs are exposed to different storage data paths (direct, remote, virtual, fabric, storage, media, etc.). Finally, Real World Workloads may also present block sizes of non-traditional size and varying User Demand Intensity when IO Streams are ultimately applied.

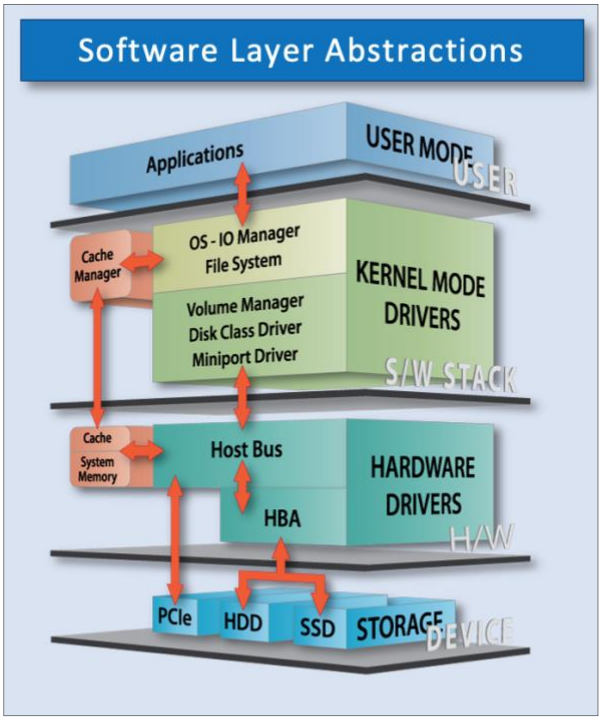

Figure 1 - SW/HW Stack. IO Streams merge, append, coalesce and split as they traverse the SW/HW Stack

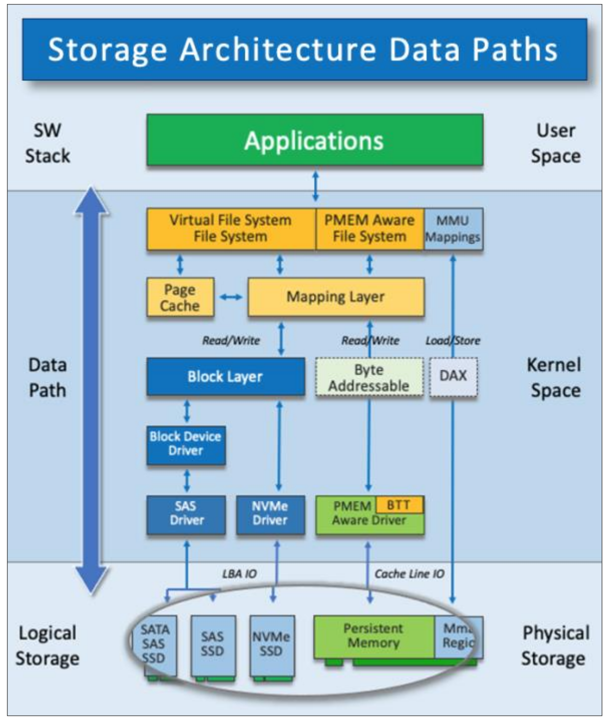

Figure 2 - Data Paths. Data Paths impact IO Streams as they are exposed to each layer of abstraction

"… *Your Performance Depends* …"

It is axiomatic that ***"… your performance will depend…"*** or ***"… your mileage may vary …"*** This is especially true for Real World Workloads. For applications, optimizations are often based on assumptions about the nature of incoming IOs without regard for what IO Streams actually occur during real world use. For example, virtualization, compression or encryption may assume the occurence of specific small or large block transfers, Random or Sequential accesses or Read or Write IOs when the actual performance will depend on what kind of IOs do, in fact, occur.

For storage and firmware design, factors such as R/W asymmetry, over provisioning, memory cache, channel design, table management, garbage collection and more can result in significantly different performance depending on what type of IO Stream and how much demand intensity is presented to storage.

What is the difference between Real World & Synthetic Workloads?

Real World Workloads are constantly changing combinations of IO Streams and Queue Depths (QD) observed between two or more points in the SW/HW stack. Real World Workloads are intended to emulate real world application and storage use by applying the IO Streams actually observed during real world application use.

Synthetic workloads are single, or very few, IO Streams that are applied over a fixed duration and Demand Intensity (or number of Users or QDs). Synthetic workloads, also referred to as "corner case benchmarks", are created by software workload generators and are typically used to stress applications and storage outside the range of normal operation.

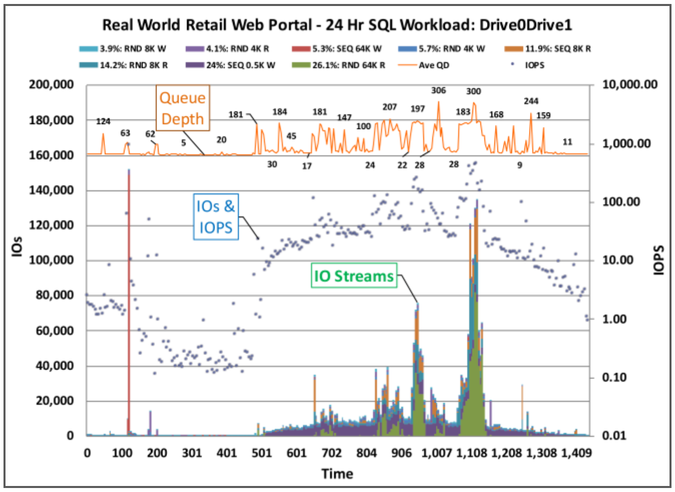

Figure 3 - Real World Workloads. Constantly changing combinations of IO Streams and Queue Depths

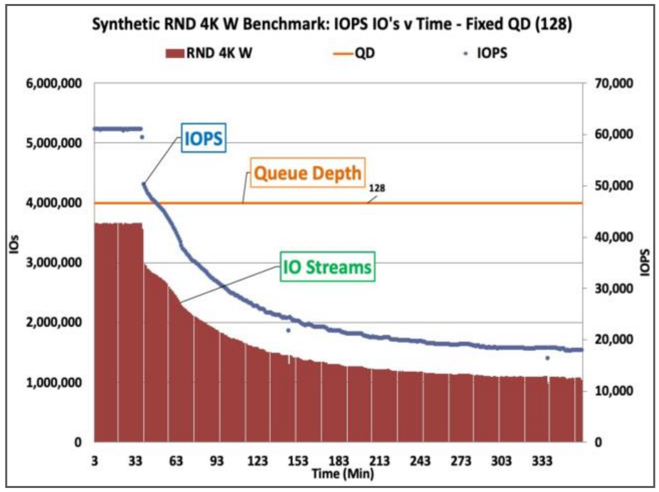

Figure 4 – Synthetic Workloads. Single, or few, IO Stream(s) at a fixed Demand Intensity and Duration

What is a Real World Workload IO Capture file?

A Real World Workload IO Capture file is a collection of IO Stream statistics observed at a specific level in the SW/HW Stack during actual application use. IO Captures are tabulations of IO statistics where no actual data content is captured. A full discussion of Real World Workload capture, analysis and test is presented in the SNIA Real World Storage Workload (RWSW) PTS for Datacenter Storage v1.0.7.

What is the difference between an IO Capture and IO Trace file?

IO Captures files aggregate IO Stream statistics over a series of defined “time-steps”. The use of time-steps allows the user to set IO Capture resolution to be coarse (hours, days, weeks) or fine (minutes, sec, mSec, uSec) and to reasonably control the IO Capture file size.

Because IO Captures are a series of time-steps, IO Captures can easily be converted into test script steps. Test scripts can be run for shorter or longer duration by adjusting the duration of each test step. Demand Intensity can also be set as desired for each test step.

IO Capture files tabulate IO workload metrics that are observed during the IO Capture. IO workload metrics, such as IO transfer size, Process ID and whether the IO is a Read or Write IO, are often provided by the OS kernel. Other specialized workload metrics, such as de-duplication, Random or Sequential access or data compressibility, may be determined by IO Capture tool algorithms.

Unlike IO Capture files, IO Trace files record a continuous stream of complex IO data which results in very large file sizes. IO Traces includes a wide variety of data (cpu statistics, error correction codecs, bus, driver and protocol commands, etc.) in addition to specific IO workload statistics. This continuous stream of complex data makes it difficult to clearly present IO Trace data.

IO Trace files can also be converted into IO Capture time-step files. However, the converted IO Trace file is limited to the IO workload metrics contained in the original IO Trace file. This means that converted IO Trace files may not include specialized IO workload metrics of interest.

IO Capture files are much smaller and more portable than IO Trace files. These time-step based files can used for real time monitoring and post processing of workloads.

2. How do I Capture a Real World Workload?

Free IO Capture tools are available at www.testmyworkload.com while commercial tools are available from Calypso Systems, Inc. IO Capture tools are Operating System (OS) specific and can be run as a Command Line Interface (CLI) or with a Graphical User Interface (GUI) Console.

A. IO Capture Tools

IO Captures tools are binary files that are installed on the target storage server. All logical storage devices recognized by the OS can be captured. IO Capture tools tabulate statistics on IO Stream workload metrics. Processing of IO statistics on the target server (where the IO Capture is being run) typically uses 1% or less of CPU Usage.

Figure 5 below shows the IOProfiler IO Capture tool and set-up parameters. “Duration” sets the time duration of the capture. “Time Resolution” sets the time-step resolution (as low as 1 uS). “Spatial Resolution” sets the LBA Range spatial resolution (as low as 0.01%). IO “Capture Level” sets the level for IO Captures (block, file system or both).

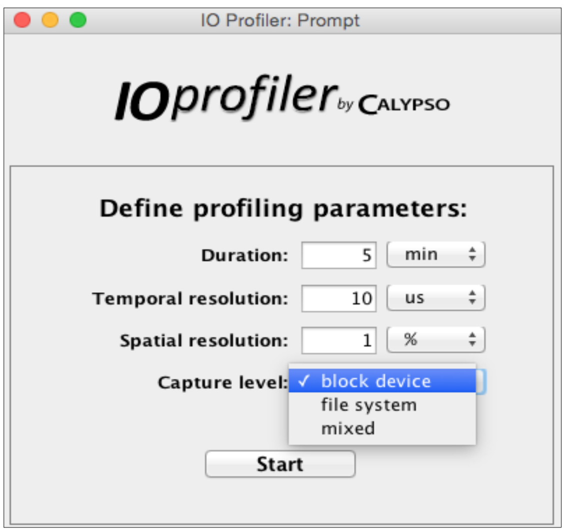

Figure 5 – IOProfiler IO Capture Tool. Set IO Capture Parameters before starting an IO Capture

Resolution Time-Step Duration	File Size 1 Minute IO Capture	File Size 20 Minute IO Capture	File Size 1 Hour IO Capture	File Size 24 Hour IO Capture
1 uSec	11 GB	213 GB	640 GB	15 TB
10 uSec	2 GB	34 GB	103 GB	2 TB
100 uSec	198 MB	4 GB	12 GB	288 GB
1 mSec	16 MB	324 MB	974 MB	24 GB
10 mSec	2 MB	34 MB	103 MB	2 GB
100 mSec	225 KB	4 MB	13 MB	312 MB
1 Sec	45 KB	900 KB	3 MB	72 MB
10 Sec	16 KB	320 KB	960 KB	23 MB
60 Sec	10 KB	200 KB	600 KB	14 MB
5 Min	N/A	69 KB	208 KB	5 MB

* Estimates Based on Single Drive Capture, Not Zipped, Block IO Level, No Compression/Deduplication

Figure 6 – IO Capture File Sizes. Estimate IO Capture file size based on Step Resolution & IO Capture Duration

B. IO Capture File Sizes

As previously discussed, IO Capture files are tabulations of statistics for IO traffic observed at the selected SW levels. IO Capture files do not contain actual data content. This allows for highly secure and portable IO Capture files.

Figure 6 above shows estimated size of IO Capture files based on Time-step Resolution and IO Capture Duration. Actual IO Capture file size depends on the number of drives captured, number of IO processes, number of metrics and total IO steps. For example, the SNIA Reference Retail Web Portal 24-hour capture has 288 five-minute steps for a total file size of 5MB zipped and 100MB raw.

A 9-hour DBRocks IO Trace from the SNIA IOTTA Repository is used to compare the file sizes of an IO Trace versus an IO Capture. The IO Capture file size for the 9-hour DBRocks trace, with 3,486 ten second steps, is 41MB zipped and 294MB raw. The IO Trace file size for the 9-hour DBRocks file is 23GB zipped and 234GB raw – a difference of three orders of magnitude larger for IO Trace files.

3. How do I View a Real World Workload?

Once captured or created, IO Capture files can be viewed for free at TestMyWorkload.com or fully curated with IOProfiler™ tools. IO Captures can be viewed as an excel or HTML file and can be replayed virtually using a Chrome browser. Static SNIA Reference Real World Workloads can be viewed on the CMSI site while active playback of SNIA Reference Real World Workloads are available on the TestMyWorkload site.

What do Real World Workloads Look Like?

Real World Workloads (RWWs) are dynamic workloads that are captured at the file system or block level for all logical storage recognized by the target server OS. Visualization of RWWs is primarily provided by IO Stream Maps and LBA Range Hits Maps.

A. IO Stream Maps

IO Stream Maps, Figure 7 below, show temporal (Time) locality of reference and are presented as Quantity of IOs (A) vs Time (B). IO Streams and workload metrics can be filtered by selecting the desired IO Stream or metric data series (C). IOs and IO Streams can also be selected to be viewed either as IOPS (by Frequency) or as MB/s (by Amount Transferred) (D).

IO Streams can be parsed and filtered by additional methods including setting the IO Stream Threshold (E) and Step Resolution (F), filtering the Process ID (G), or selecting the Cumulative Workload IO Streams (H) or logical Drive(s) (I). Any change in settings will refresh the related IO Stream Map (Figure 7) or LBA Range Hits Maps (Figure 8).

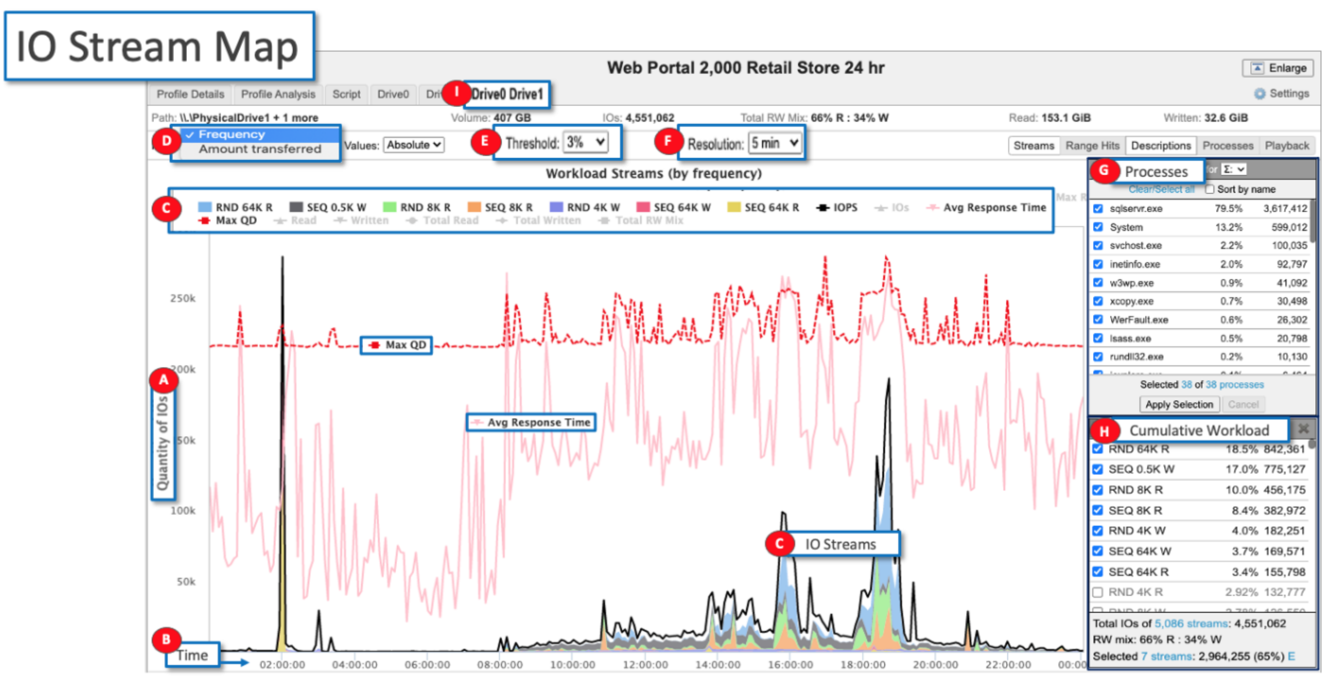

Figure 7 - IO Stream Map: Quantity of IOs v Time

B. LBA Range Hits Maps

LBA Range Hits Maps show IO hits by spatial locality of reference. LBA Range Hits Maps represent access pattern sequentiality and IO clusters.

Figure 8 shows individual LBA Range Hits (A) for Drive0 (B) in MB/s (transferred amount). The Y-axis (C) shows LBA Range 0 to 100% of capacity for Drive0 while the X-axis shows Time (D). LBA Range hit resolution can be set as low as 0.01% of the total LBA Range. Where multiple drives are shown, each drive presents LBA Range % in successive stacked ranges.

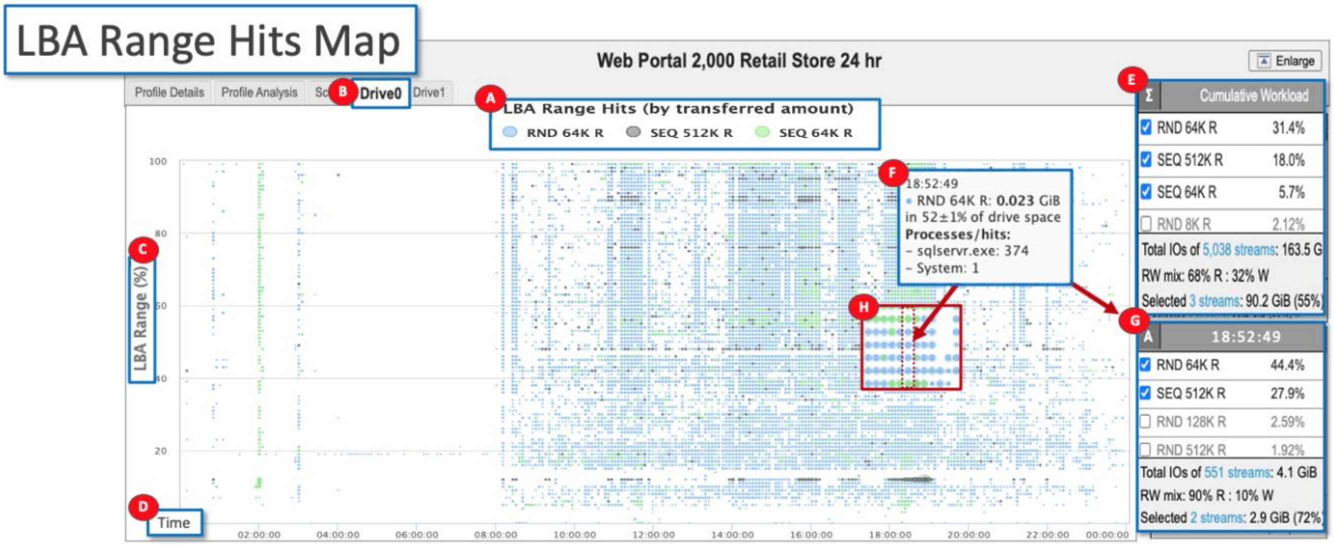

Figure 8 – LBA Range Hits Map

The Cumulative Workload Descriptions pane (E) shows the 3 most frequently occurring IO Streams the for the 24-hour Retail Web Portal IO Capture. There are 3 IO Streams in an overall RW mix of 68% Reads for 55%, or 90.2 GB of the total 163.5 GiB transferred.

The tooltip metrics box (F) for the individual Range Hit selected at Time point 18:52:49 show that RND 64K R IOs are comprised of 374 sqlsevr.exe IOs and 1 system IO for total Read amount of 0.023 GiB. Note that Range hit resolution for this IO Capture is set to 1% of the total LBA Range.

The Individual Workload pane (G) for all Range Hits at time 18:52:49 shows two IO Streams – RND 64K R (44.4%) and SEQ 512K R (27.9%) – occur 72% of the time. These two IO Streams represent 2 of the 551 IO Streams observed at 18:52:49 and represent a 90% RW mix.

LBA Range Hits Map time point 18:52:49 (H) shows the individual Range Hits for RND 64K R (blue dots) and SEQ 64K R (green dots).

C. Tracking IO Stream Occurrence to Logical Drive Locations

The stacked IO Stream Map and LBA Range Hits Map view aligns Time X-axes. This allows the reader to track IO Stream IO hits and the corresponding LBA Range hits.

Figure 9 24-hour Retail Web Portal stacked view shows IOs for Drive0Drive1 (A) in Amount Transferred (B). Note that Time points for 2:00:19 (D) and 2:15:21 (E) are highlighted with a red dotted line while successive stacked LBA Ranges are shown for Drive0 0-100% (F) and Drive1 101-200% (G).

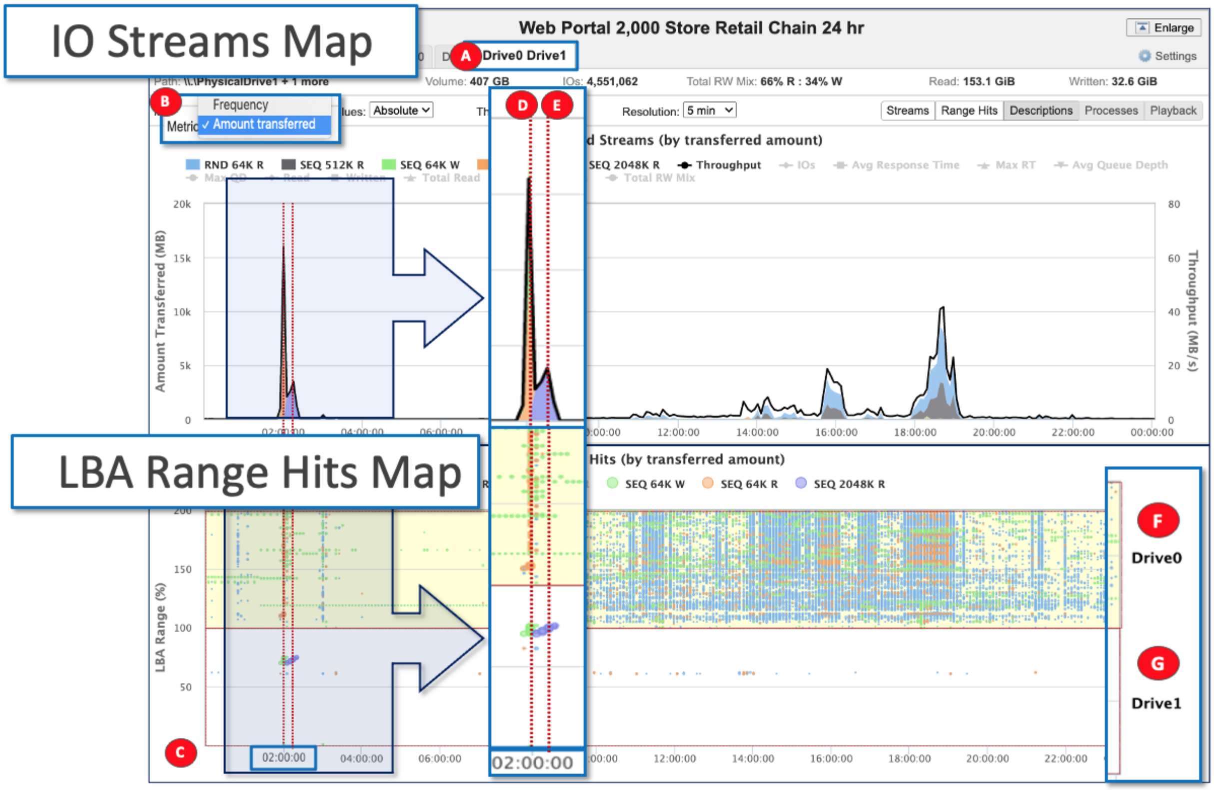

Figure 9 – Stacked view of IO Stream & LBA Range Hits Map

Figure 10 shows 2:00:19 am Back-up with SEQ 64K R & W on the IO Stream map (A1) writing SEQ 64K from Drive0 (A2) in to Drive1 (A3) in Figure 11. Figure 10 shows 2:15:21 am SEQ 2048 Reads on the IO Stream Map (B1) with all of the Reads coming from Drive1 (B2) in Figure 11.

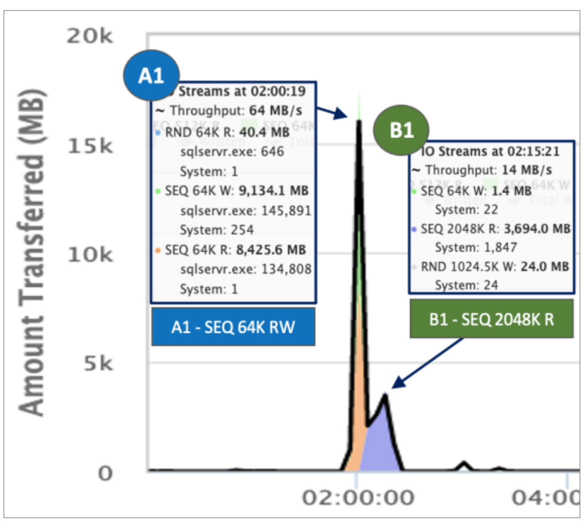

Figure 10 – IO Stream Map. Sqlserver SEQ 64K RW (A1). System SEQ 2048K R (B1).

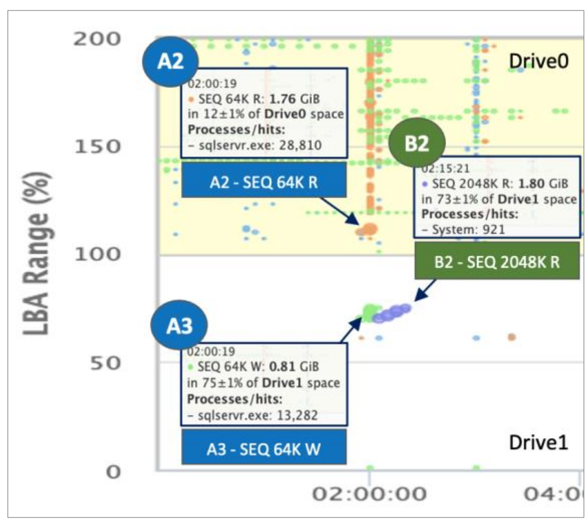

Figure 11 – LBA Hits Map. SEQ 64K R Drive0 (A2) to SEQ 64K W Drive1 (A3). SEQ 2048K Read on Drive1 (B2).

4. How do I Analyze Real World Workloads?

Real World Workloads can be analyzed by ***locality of reference*** and by ***IO Stream composition*** of the IO Capture(s). Because IO Captures can be taken at the file system and block level on any OS compliant host or target server, it is possible to analyze how and where workloads traverse the SW/HW stack as well as observe changes in IO Stream composition.

A. Locality of Reference

The IOProfiler (IPF) IO Capture applet allows you to ***compare file system vs block level*** IOs captured on different ***Storage Architectures*** (direct, NAS or fabric storage).

IO Captures are taken for all logical storage recognized by the target server OS. This includes Direct Attached Storage (DAS), logical units (LUNs), storage clusters, remote storage (e.g., RDMA), virtual storage (VM) and fabric storage (e.g., NVMe-oF). IO Capture workloads can be examined as they traverse the SW/HW stack from file system to block level storage, logical to device storage or from host to fabric/virtual storage.

1. File system vs Block level IOs

A two-drive 24-hour GPS Nav Portal IO Capture is shown in Figures 12 & 13 below.

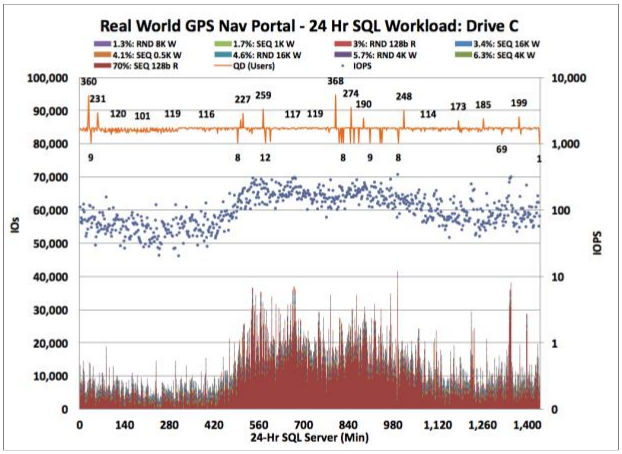

Figure 12 - GPS Nav Portal: Boot DriveC. 67:33 RW; 70% SEQ 128b R; Median QD 8, Max QD 368

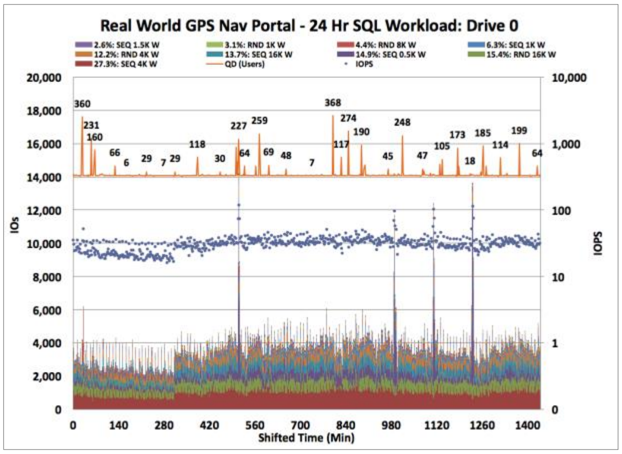

Figure 13 – GPS Nav Portal: Storage Drive0. 96% W; 61% RND/SEQ 4K/8K/16K; Median QD 8; Max QD 368

Boot DriveC file system IOs (Figure 12) are 67% R where 70% of the IOs are SEQ 128b Reads (R caches), 20% of the IOs are RND/SEQ 4K/8K/16K RW and where Median QD is 8 and Max QD is 368. Storage Drive0 Block IOs (Figure 13) are 91% W where 61% of the IOs are RND/SEQ 4K/8K/16K RW, where there are periodic SEQ 0.5K W spikes and where Median QD is 8 and Max QD is 368.

This is an example of application/user space IOs being transmuted as they traverse the SW/HW stack to storage. Here pre-fetching, pre-writes, look-ahead reads and other optimizations result in very different IO activity between the file system and storage level.

Other examples show how intervening software, middleware and driver layer abstractions change, modify, split, coalesce and fragment IO Streams. IO Captures can be run at the “top” (file system) and at the “bottom/edge” of the storage data path. Knowing the actual IO Stream composition at different SW/HW and storage levels helps to increase the efficacy of software optimizations.

2. Virtual Storage Cluster

IO Captures can be taken for storage cluster virtual drives by running IO Captures on the storage target server. IO Captures can be viewed for a single individual device or can be viewed as an Aggregate IO Stream Map and LBA Range Hits Map for multiple selected devices.

In Figure 14, the Profile Details tab (A) shows a Linux OS vmx (B) IO Capture for 20 logical storage devices: sda – sdu and the boot drive sd0 (C). Storage devices can be aggregated as desired by selecting the drives and clicking “Aggregate n Selected Drives” (D).

The IO Capture statistics on the Profile tab show that the capture duration was 30 minutes across 20 drives on a Linux Redhat 7.2 OS with a 10 sec temporal resolution and a 1% LBA Range spatial resolution. There is a total of 9.936 GB of storage across 20 volumes. There were 73,046 IOs with 140 PIDs. 4.3GiB Reads and 7.2GiB Writes were accessed for a total of 11.5GiB.

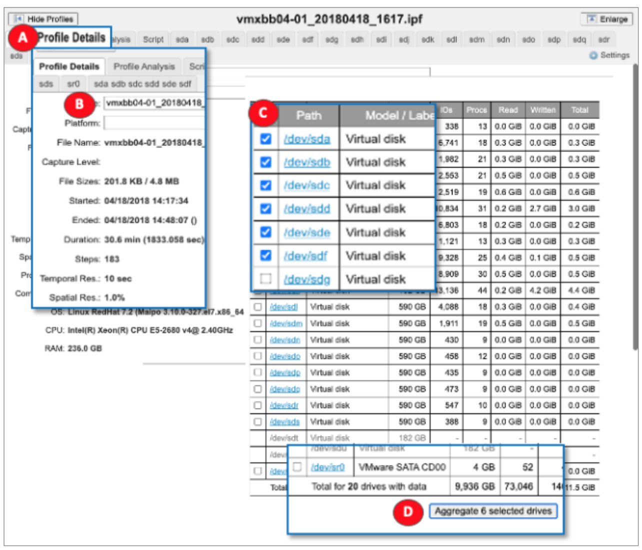

Figure 14 – Profile Details Tab. Profile Details shows IO Capture, administrative and storage volume information

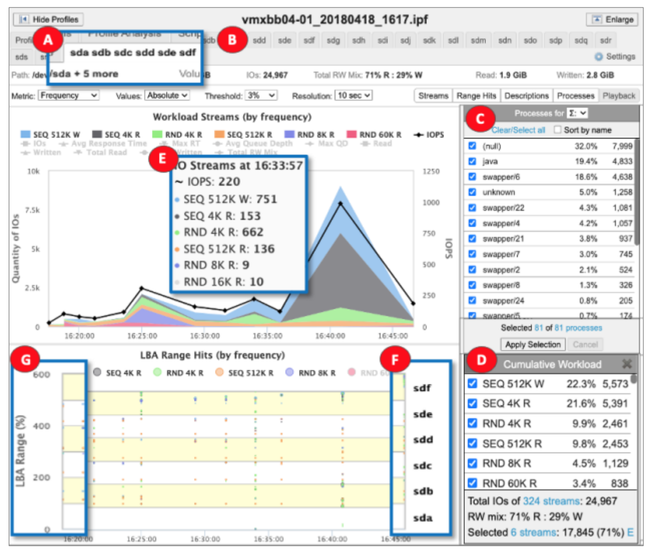

Figure 15 – Aggregate IO Stream & LBA Maps. See aggregate and individual storage IO and LBA Range Hits

Figure 15 above shows the aggregate IO Stream Map and LBA Range Hits Map for the 6 selected storage devices sda – sdf (A). Individual storage drive IO Stream and LBA Range Hits maps can be viewed by selecting the desired drive tab(s) (B).

The Process ID pane (C) and the Cumulative Workload pane (D) show PIDs and IO Streams for the selected drive(s) and can be used to filter/parse the IO Stream Map and LBA Range Hits Map. Note that the tooltip metrics (E) displays the time point IO Streams and metrics for the selected drive(s).

The LBA Range Hits Map shows each of the selected drives (F) across successive 100 unit LBA Ranges (G), i.e., the LBA Ranges for each drive are stacked on successive 100 unit ranges.

The 6 drive aggregate IO Stream and LBA Range Hits maps show the IO activity only for selected drives and do not include any other drive(s). Thus, IO and PIDs from the boot drive are not shown. Only IOs and PIDs associated with the Aggregate 6 Drives from the virtual storage cluster are shown.

Storage cluster LBA Range Hits Maps can help confirm IO traffic to specific storage tiers, hyper-converged storage or virtual storage drives.

3. RAID v Individual Device

RAID level vs storage device level IOs shows transmutation of IO Streams from software RAID LUN (md1) to individual storage devices (sdc-sde).

An IO Stream workload of SEQ 1024K W is applied to a software RAID 0 consisting of (3) 15K RPM SAS HDDs. The software RAID is set up as software RAID 0 (mdadm tool), Striped, N=3 (Whole Drives), Block Size (64K). IO Captures were taken at the software RAID 0 level (md1) and at the individual device (sdc, sdd and sde) level.

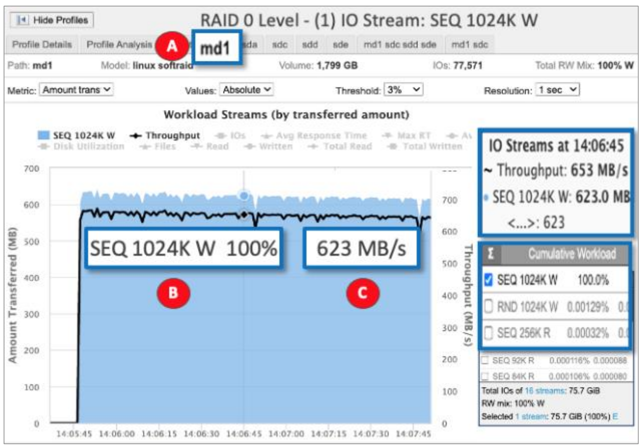

Figure 16 – IO Stream Map RAID md1. IO Stream Map showing SEQ 1204K Workload applied to RAID md1

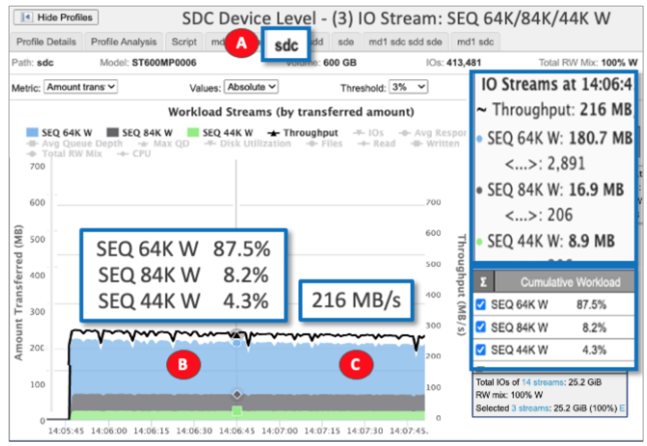

Figure 17 – IO Stream Map Device sdc. Transmuted IO Streams from RAID md1 to individual device sdc.

Figure 16 IO Stream Map RAID 0x3 md1 (A) shows a single SEQ 1024K W IO Stream (B) with a Throughput (TP) of 623 MB/s (C).

Figure 17 IO Stream Map Device sdc shows 3 IO Streams (87.5% SEQ 64K W, 8.2% SEQ 84K W and 4.3% SEQ 44K W) (B) for a total TP of 216 MB/s (C). Note sdc, sdd and sde IO Stream Maps are identical so sdd and sde IO Stream Maps are not shown above.

Here we see expected behavior where RAID 0x3 TP of 623 MB/s is approximately 3 times the individual device level TP of 216 MB/s. The SEQ 1024K W IOs applied to RAID 0 md1 are changed by the RAID software (mdadm RAID) and SW/HW Stack components (CentOS 7.0, software layers, device drivers, and HBA hardware) before they are applied to the individual device sdc, sdd & sde.

Other examples show IO Streams being changed by the SW/HW stack. In cases of virtualization, larger block RND and SEQ IOs fragment into smaller RND RWs. Encryption changes SEQ RW IOs to concurrent RND IOs. Deduplication eliminates recurring IOs of similar block size and LBA offset. TRIMs coalesce transfers into larger block sizes. Streaming media changes large block SEQ Reads into concurrent RND 4K/8K/16K RW IOs.

Note that the ultimate impact of SW/HW level optimizations and abstractions on IO Stream content is highly dependent on the OS, SW Stack, middleware optimizations and the storage architecture.

While workloads of similar type may share broad IO Stream characteristics, specific IO Stream content will depend on the OS, SW/HW Stack and storage architecture, will be influenced by concurrent running applications, will be impacted by the time of day and will be influenced by the total traffic (user demand) to the application.

B. IO Stream Composition

IO Captures are inherently complex and contain voluminous amounts of data. Indeed, one of the key features of IO Captures is the ability to curate workloads to more clearly present relevant data and metrics. IO Captures can be filtered and parsed to present relevant parameters, metrics and data for desired applications and workloads.

IO Stream Threshold and **Workload Descriptions** filter IO Captures by filtering the number of IO Streams displayed. **Process ID** (PID) and **Step Resolution** allows you to filter IO Stream Maps by the type of process associated with an IO or by adjusting the step granularity of the IO Stream Map.

1. Filter IO Stream Maps by IO Stream Threshold & Workload Descriptions

IO Stream Threshold (A) selects the number of IO Streams to be displayed by the percentage of total IO occurrence over the duration of the capture. For example, selecting 3% IO Stream threshold means that only those IO Streams that occur at least 3% or more of the time will be displayed.

Figure 18 shows IO Stream Threshold default setting of at 3% (A). IO Streams that meet this 3% threshold are shown in the Cumulative Workload Description pane (B). Here, the 7 most frequently observed IO streams that occur at least 3% of the time are shown.

IO Streams can be increased or decreased as desired by changing the Threshold setting or by selecting specific IO Stream(s) in the Cumulative Workload Description pane. For example, RND 8K W IO Stream at 2.78% occurrence (C) can be added by clicking the RND 8K W IO Stream. The IO Stream Map will refresh to present the new total IO Streams selected.

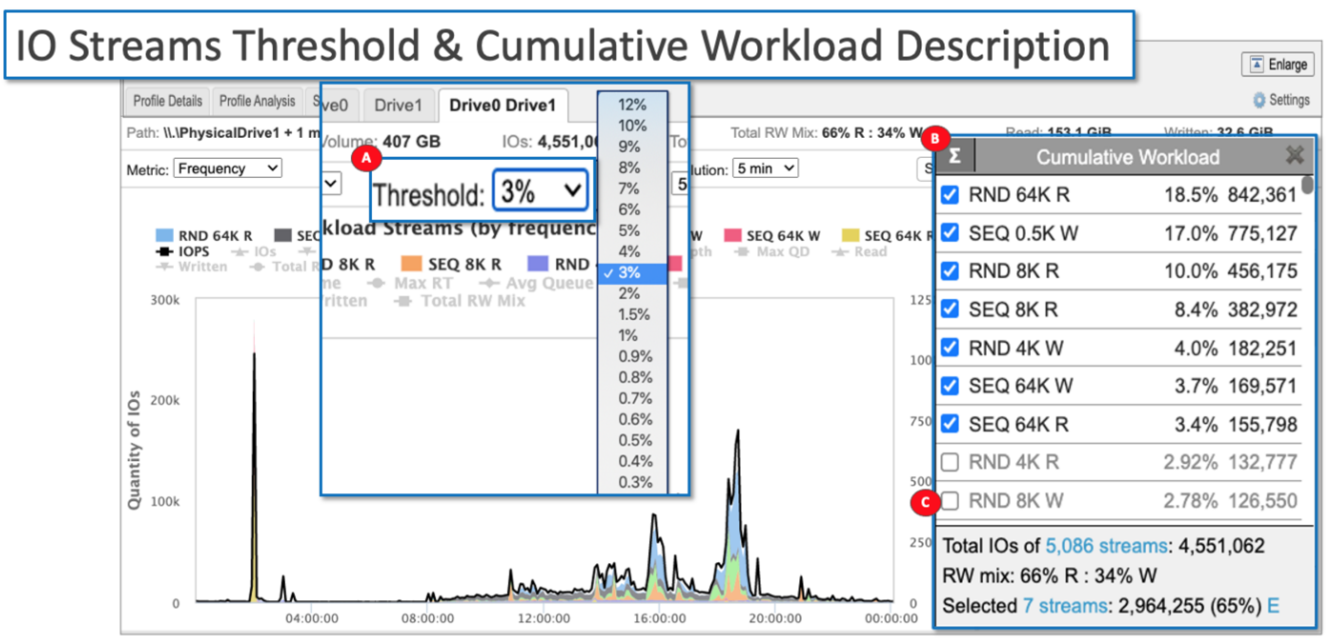

Figure 18 – Filter IO Stream Maps by IO Stream Threshold & Workload Descriptions

2. Filter Applications by Process IDs

Software applications can be filtered by Process ID. The Process ID (PID) pane shows the Processes that are associated with each IO (as assigned by the OS kernel). IO Captures of specific benchmark application software (A) can be filtered by selecting the desired PID IO Streams.

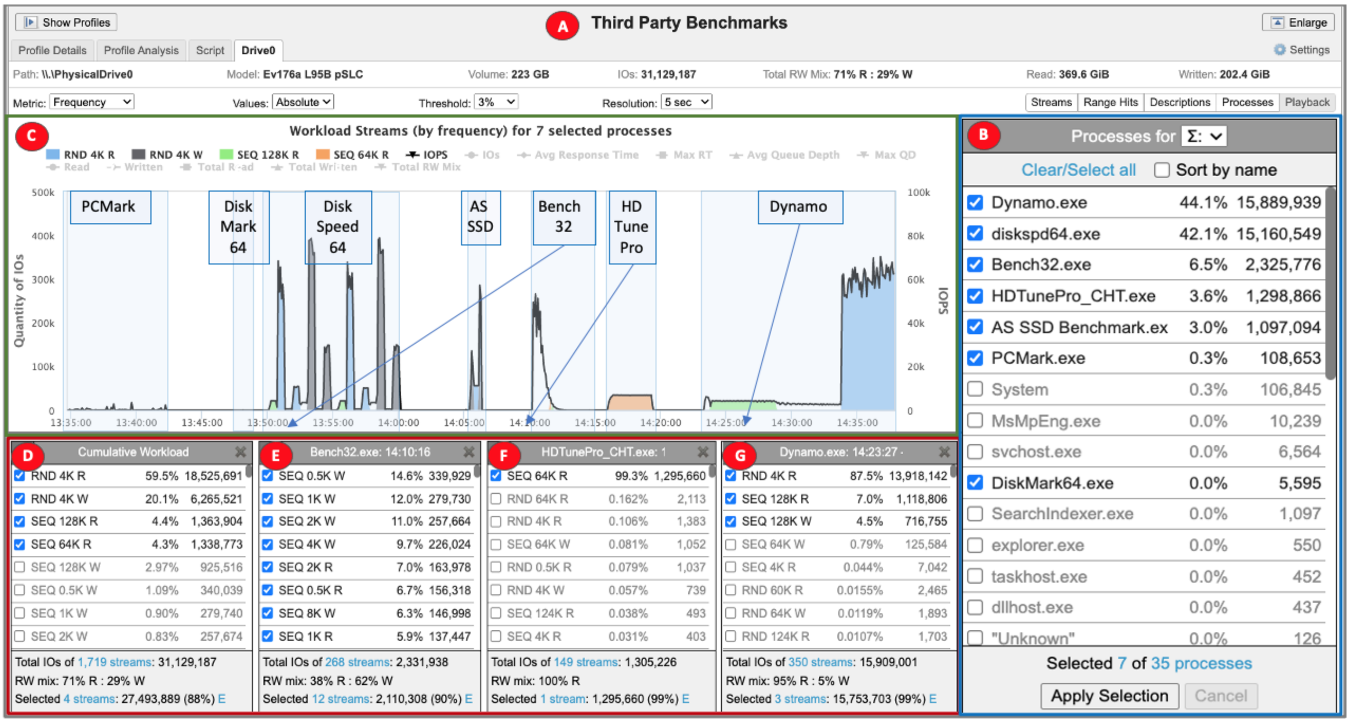

Figure 19 – Filter Applications by Process IDs

Figure 19 shows an IO Capture of seven third party benchmark applications (A). Selecting the desired Benchmarks in the PID Pane (B) filters the IO Streams by application in the IO Stream Map (C). The Cumulative Workload pane (D) shows all of the IO Streams in the IO Capture. The individual IO range panes show the IO Streams for Bench32 (E), HDTunePro (F) and Dynamo (G). Only IO Streams associated with the Benchmarks are shown (i.e., System and other PIDs are not presented).

In the PID pane (B), we see 7 of a total 35 Processes selected. Each benchmark tool is labelled on the IO Stream Map (C) with the individual Benchmark IO Ranges shown in panes E, F and G.

"Bench32" (E) at time point 14:10:16 has 12 IO Streams of SEQ RW in a 38:62 RW mix.

"HDTunePro" (F) at 14:19:1724 has 1,295,660 SEQ 64K R IOs at a 100% R RW mix.

"Dynamo" (G) at 14:23:27 has 13,918,142 RND 4K R, 1,118,806 SEQ 128K R and 716,755 SEQ 128K W IOs at a 95:05 RW mix.

Note that while each benchmark may show a large number of total IO Streams, the selected IO Streams represent 90 to 99% of the total IOs.

Filtering benchmark IOs by PID allows the user to see the IO Stream content of each benchmark tool instead of relying merely on arbitrary proprietary scaling scores (e.g., 6 out of 10) or assuming that the benchmark workload is indeed comprised of the IO Streams claimed by the vendor.

3. Step Resolution: Viewing Individual IOs to Long Duration Captures

Step Resolution sets the temporal resolution for presentation of the IO Stream Map. IO Stream Map Step Resolution granularity allows observation of individual IOs or can be used to present extremely large IO Capture datasets (e.g., 24-hour) in IO Stream Maps of smaller file size.

The minimum viewable Step Resolution is defined by the step resolution of the original IO Capture. For example, an IO Capture taken at 1 sec resolution cannot be viewed at less than 1sec granularity. However, the IO Capture Step Resolution may be set as desired to adjust the IO Stream Map for clarity. Fine grain resolution can be used to resolve single or few IO occurrences. Coarse grain resolution can be used to filter IO Stream maps for clarity and to limit IO Capture file sizes.

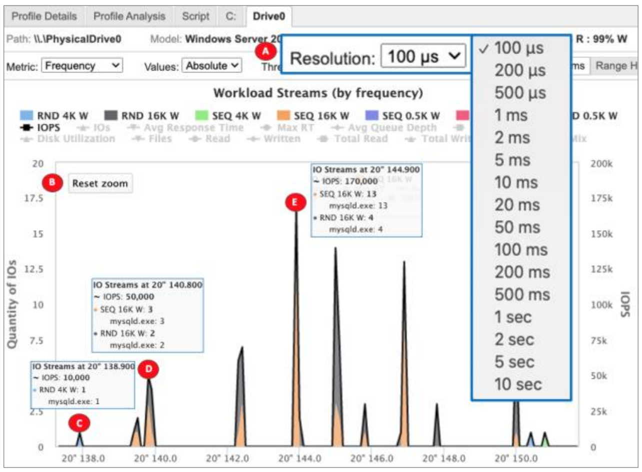

Figure 20 – 100uS Steps. Fine grain resolution allows observation of single or very few IO events/occurrences

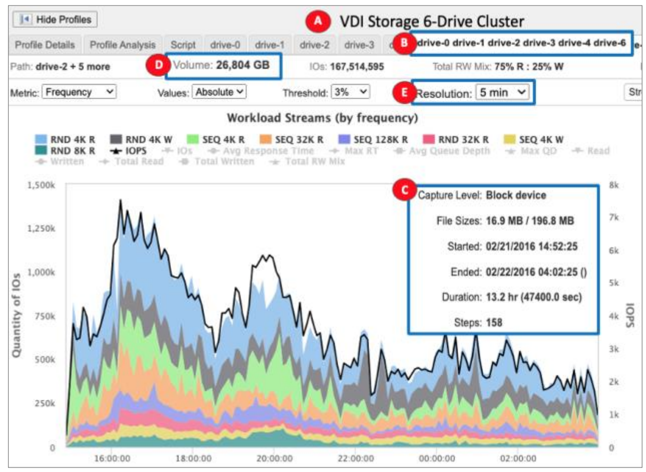

Figure 21 – 5 min Steps. Coarse grain resolution allows viewing of multiple/aggregate drive, long duration captures

In Figure 20, the original IO Capture Step Resolution was set to 100 uS (micro seconds) (A). At this resolution, it is possible to resolve single to several IOs. The zoom tool (B) shows IO events with a single IO (C), five IOs (D) and 17 IOs (E).

At Time point 20"138.000 (C) we see a single mysqlexe RND 4K W IO. At Time point 20" 140.800 (D) we see five mysqlexe IOs – three SEQ 16K W and two RND 16K W. At Time point 20"144.900 (E) we see seventeen mysqlexe IOs – thirteen SEQ 16K W and four RND 16K W.

In Figure 21, we see an IO Stream Map for a 13-hour VDI Storage 6-Drive Cluster (A). The aggregate 6-drive IO Stream Map (B) shows 26.8 GiB of original IO Trace data (D). However, with the original IO Capture Resolution set at 5 min (E), the IO Capture file size was only 16.9 MB zipped and 196.8 MB raw (C) – a reduction of 2.5 orders of magnitude in file size.

Fine grain Step Resolution and the Zoom feature can be used to show individual IO occurrences, IO clusters and PIDs associated with the target IOs. Using coarse grain IO Step Resolution in the original capture can allow the viewing of very long IO Capture – up to 24-hours or more – with an IO Capture size that is much smaller than the original IO Trace file.

IO Step Resolution supports examination of micro second IO bursts to viewing the broad stroke characteristics of very long duration IO Captures without the burden of extremely large IO Capture file sizes.

5. What are SNIA Reference Real World Workloads?

SNIA SSS TWG and the CMSI periodically select Reference Real World Workloads (RWWs) deemed to be representative of common use cases. SNIA Reference RWWs are posted on the CMSI website and TestMyWorkload websites.

While every RWW is unique, RWWs can be generally characterized by their overall RW mix, IO Stream content and Demand Intensity. While applications run on the same or similar OS, SW/HW Stack and storage architecture can share similar RWW characteristics, applications run on different OS, storage architectures and at different times can generate very different application and storage IOs.

A. SNIA Reference Real World Workloads: Table Comparison

Real-World Workload	RW Mix Normalized	IO Capture Level	Total IOs	Total IO Streams	9 Most Frequent IO Streams by % of IOs						Min QD	Max QD	Median QD
Retail Web Portal: **2-Drive, 24-hour**	65% R	Block IO	4.5 M	5,086	18.5% 10.0% 4.0% 3.4% 2.7 %	RND RND RND SEQ RND	64K R 8K R 4K W 64K R 8K W	17.0% 8.4% 3.7% 2.9%	SEQ SEQ SEQ RND	0.5K W 8K R 64K W 4K R	5	306	19
GPS Nav Portal: **1-Drive, 24-hour Storage Drive0**	100% W	Block IO	3.5 M	1,033	21.6% 11.7% 9.6% 3.4% 2.1%	SEQ SEQ RND RND SEQ	4K W 0.5K W 4K W 8K W 1.5K W	12.0% 10.7% 4.9% 2.4%	RND SEQ RND RND	16K W 16K W 8K W 2K W	6	368	8
GPS Nav Portal: **1-Drive, 24-hour Boot DriveC**	67% R	Block IO	11.5 M	7,169	60% 5.5% 4.9% 4.0% 3.5%	SEQ SEQ RND RND SEQ	128b R 4K W 4K W 16K W 0.5K W	2.9% 2.7% 1.4% 1.1%	SEQ RND SEQ RND	16K W 128b R 1K W 8K W	6	368	8
VDI Storage Cluster: **6-Drive, 12-hour**	75% W	Block IO	167 M	1,223	19.3% 9.1% 4.2% 3.3% 2.3%	RND SEQ SEQ SEQ SEQ	4K R 4K R 128K R 4K W 8K R	11.3% 8.2% 3.6% 3.3%	RND SEQ RND RND	4K W 32K R 32K R 8K R	64	1,024	128
SNIA CMSI Reference Workloads – Retail Web Portal, GPS Nav Portal and VDI Storage Cluster workloads can be viewed at www.testmyworkload.com													

Figure 22 – SNIA Reference Real World Workloads: Table Comparison

Figure 22 shows 4 SNIA Reference RWWs. Each workload shows overall RW mix, IO Capture level (file system or block), total IOs and IO Streams observed, most frequently occurring IO Streams, and the QD range of the workload. IO Stream Maps for each RWW show the PIDs for each RWW.

The Retail Web Portal workload is 65% R, has a high number of IO Streams (5,086), a wide range of non-fragmented traditional storage Block Sizes (0.5K to 64K) and a QD range of 5 to 306.

The GPS Nav Portal workload is shown both for boot DriveC and storage Drive0. Drive0 is 100% W, QD range of 6 to 368, 1,033 IO Streams and a range of smaller Block Sizes (1.5K to 16K). DriveC is 67% R, dominated by SEQ 128b Reads, has 7,169 IO Streams and a max QD of 368.

The VDI Storage 6-Drive Cluster Workload is 75% W, has 1,233 IO Streams, a range of Block Sizes associated with traditional disk storage (4K to 128K) and a QD range of 64 to 1,024.

B. GPS Navigation Portal: Boot DriveC & Storage Drive0

The SNIA GPS Navigation Portal has two single drive 24-hour reference workloads: DriveC boot drive (Figure 23) and Drive0 storage drive (Figure 24). DriveC is dominated by SEQ 128b Reads, is 67% R and has a QD range of 6 to 368. Drive0 has smaller block sizes, is 100% W, has a QD range from 6 to 368 and periodic SEQ 0.5K IO spikes.

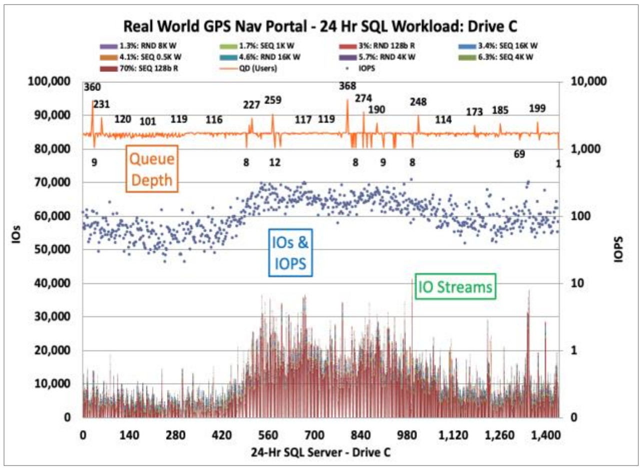

Figure 23 – GPS Nav Portal: DriveC Boot Drive

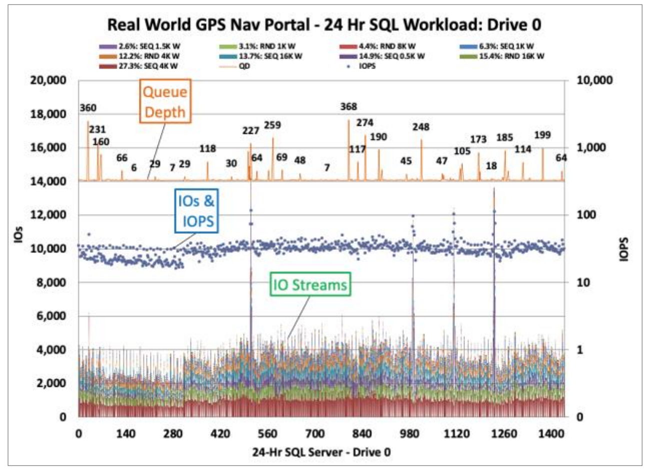

Figure 24 – GPS Nav Portal: Drive0 Storage Drive

C. Retail Web Portal: Drive0Drive1; VDI Storage: 6-Drive Cluster

The Retail Web Portal (Figure 25) is a 2-drive, 24-hour workload comprised of different retail events (such as morning boot storm, opening, daily & closing activities and 2 am back-up), a different IO Streams, QD from 5 – 306, block sizes from 0.5K to 64K, and a normalized RW mix of 65% R IOs.

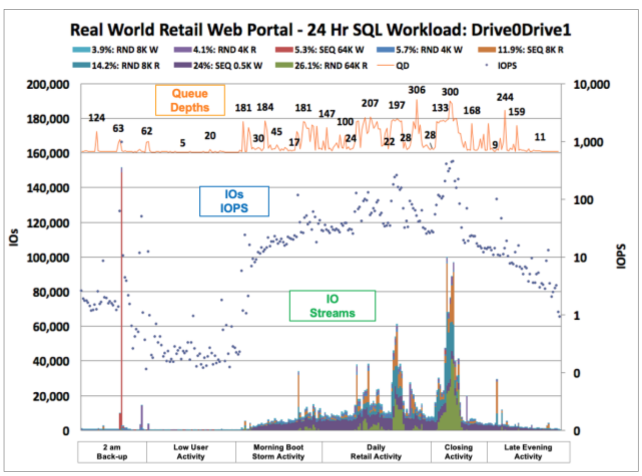

Figure 25 – Retail Web Portal: Drive0Drive1

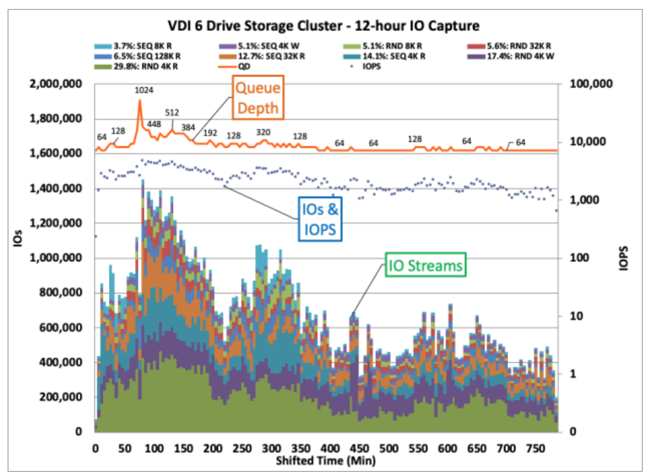

Figure 26 – VDI Storage 6-Drive Cluster

The VDI Storage 6-Drive Cluster (Figure 26) is a six-drive RAID 0 LUN, 12-hour workload comprised of non-fragmented storage block sizes, higher QDs (up to 1,024) and a RW mix of 75% W IOs. Note that 6-drive IO Streams are aggregated as a feature of CTS IOProfiler software.

6. Common Use Cases for Real World Workloads

A. Real World Workload Use Case Tree

Curated IO Captures (IO Profiles) captured at specific SW/HW levels helps optimize applications, validate middleware abstractions, and qualify/validate storage solutions.

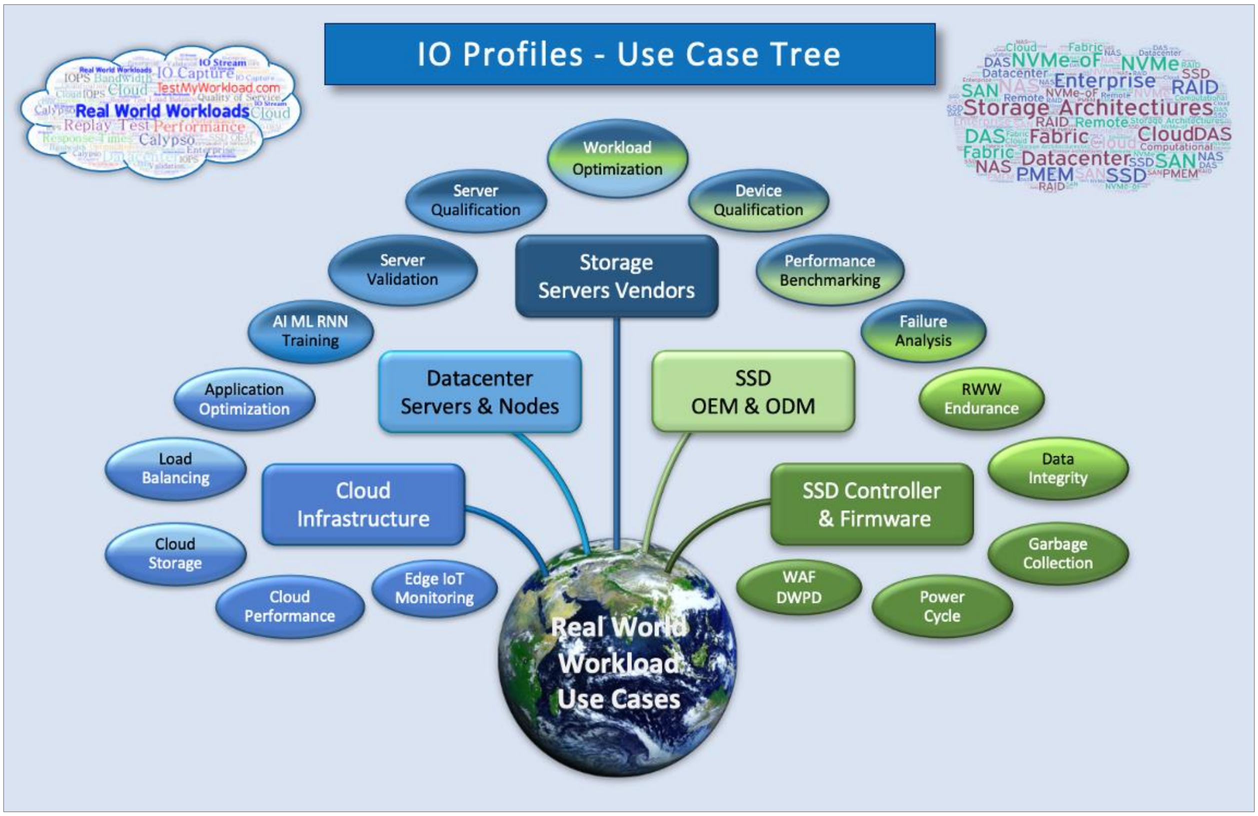

Figure 27 – Real World Workload Use Case Tree

IO Profiles have overlapping use cases for Cloud Infrastructure, Datacenter Servers & Nodes, Storage Server Vendors, Edge Computing and IOT designers, SSD OEMs and ODMs and SSD Controller and Firmware developers.

B. Using Real World Workloads to Test Applications & Storage

Real World Workloads can be used to test applications and storage. The IO Stream content of a workload capture can be used to create test stimuli for application and storage optimization, validation and qualification. IO Stream content can also be used for storage device performance benchmarking, endurance test and failure analysis.

IO Capture IO Stream maps can be made into test scripts by creating test IOs for each IO Capture step. In addition, the CTS IOProfiler software allows you to automatically generate RWW test scripts from the IO Stream Map GUI. These test scripts can be applied to direct, NAS or fabric storage at the file system or block level using the CTS workload generator. Auto generated test scripts have user accessible variables, pre-built test results plots and auto generated SNIA format reports.

7. RWW Tests – SNIA RWSW PTS for Datacenter Storage

The SNIA Real World Storage Workload (RWSW) PTS for Datacenter Storage (SNIA RWSW PTS) describes a methodology to analyze and curate workloads and sets forth standardized tests and reporting requirements for using Real World Workload (RWW) tests.

Types of Real World Workload Tests

The RWSW PTS describes four types of RWSW tests.

1. Self-Test
2. Replay Test
3. Thread Count/Queue Depth (TC/QD) Sweep Test
4. Individual Streams Test

A. Self-Test

Self-Test converts an IO Capture into a Self-Test “test result” that can be viewed as any other CTS or SNIA test. Self-Test is not an actual test but is used to present the performance of the original workload capture as if it were run as a CTS or SNIA test. The Self-Test generates performance plots of the target application and storage during the original IO capture. i.e., you can see and plot the native performance of the target server and storage while running applications during your IO Capture.

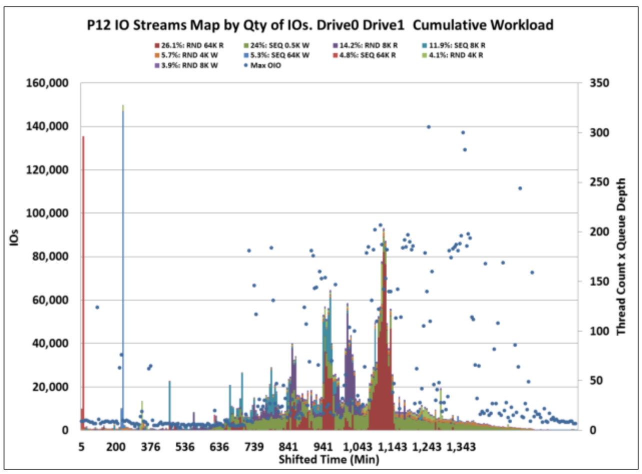

Figure 28 – Self-Test IO Stream Map & IOs. Sequence of IO Stream Combinations, IOs & QDs during IO Capture

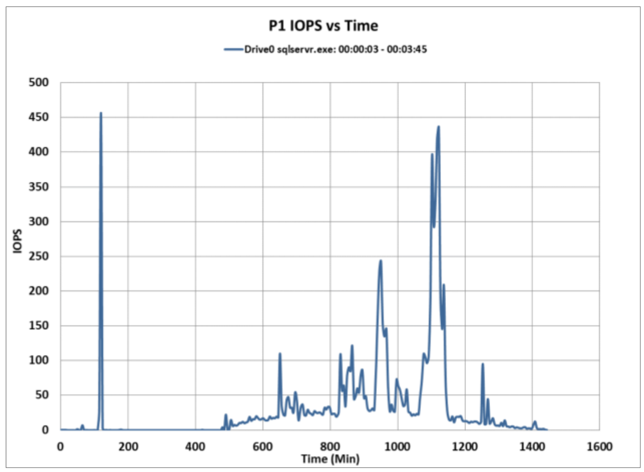

Figure 29 - Self-Test IOPS v Time. Cumulative IOPS while application runs on target server during an IO Capture

Figure 28 shows the IO Stream combinations observed during the Retail Web Portal IO Capture of Drive0Drive1. Each time-step shows the combination of IO Streams and QDs that occurred on the target server. Figure 29 shows these IO Streams as a cumulative IOPS vs Time plot.

B. Replay Test

The Replay Test creates a test script for the sequence of IO Stream combinations observed in the original IO Capture. The IO Stream Map is used to curate the IO Capture to present the desired metrics and IO Streams (see previous discussion on IO Stream Maps). Once the IO Stream Map is curated, the IO Steps are used to create a Replay Test script (manually or automatically by using CTS IOProfiler software).

Note that RWWs occur in a sequence of different IO Stream combinations for each step of the IO Capture. Because IO Stream combinations are selected and applied by time-step, the test script step duration can be adjusted to run exactly as they occurred or for a shorter or longer duration. In addition, parameters can be adjusted to increase or decrease the Demand Intensity of the Replay Test.

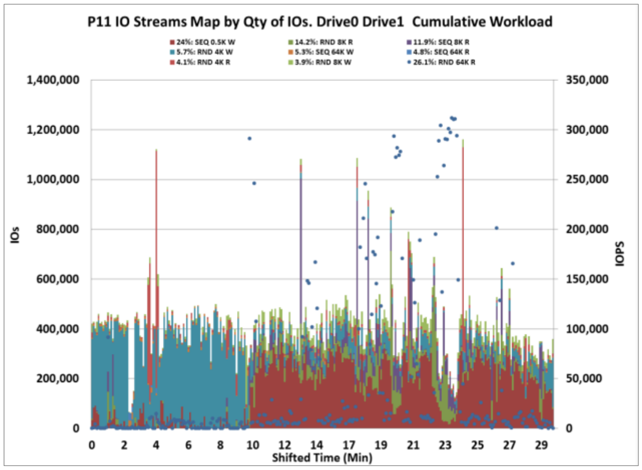

Figure 30 – IO Stream Map & IOs. Sequence of IO Stream Combinations, IOs & IOPS on NVMe SSD

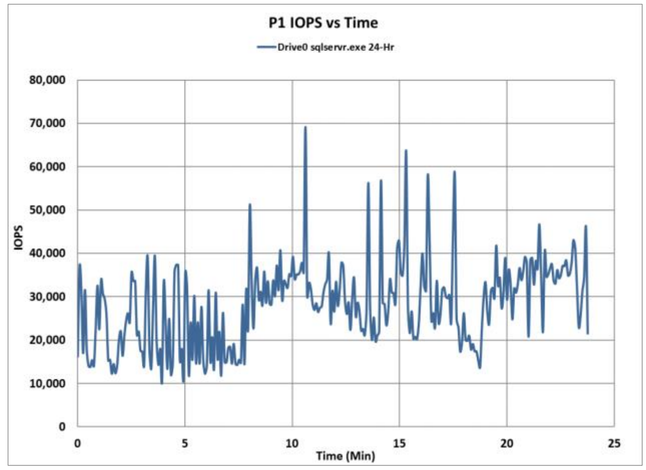

Figure 31 – Replay Test IOPS v Time. IOPS measured during a Replay test run on NVMe SSD

Figure 30 shows the Replay Test IO Stream Map & Quantity of IOs for Retail Web Portal Drive0Drive1. The Replay test applies the sequence of IO Stream combinations curated in the IO Stream Map to the target storage, in this case, an NVMe SSD, and reports IOPS v Time (Figure 31).

Note that the sequence and combinations of IO Streams are the same for the Self-Test (native application server) and the Replay Test (to NVMe SSD). Measured IOPS are much higher for the NVMe SSD than for the Self-Test native application server. This is due to throttling of Self-Test IOPS by the SW/HW stack, concurrently running applications and the native Demand intensity of the target server during the original IO Capture. For example, for the Retail Web Portal in Figure 25 on page 15, “early morning hours” show very low QD which results in correspondingly low IOs and IOPS for the Self-Test.

Figure 31 shows NVMe SSD Replay Test IOPS between 10,000 and 70,000 IOPS compared to Figure 29 native application server Self-Test IOPS between 50 and 450 IOPS.

C. Thread Count/Queue Depth (TC/QD) Sweep Test

The TC/QD Sweep Test evaluates RWW performance across a range of Demand Intensity (DI), i.e., measures performance saturation (in IOPS & ART) as the number of Users (or DI) increases. In this test, a fixed number of IO Streams is applied in each test step in the same percentage as they are observed in the original IO Capture. In this example, each test step consists of the nine IO Streams selected in the IO Stream Map. See Figure 34 Workload Streams Distribution below.

After a pre-conditioning, performance is measured over a range of Demand Intensity (or Thread Count x Queue Depth). Performance (in IOPS and Average Response Time (ART)) is then plotted as a function of increasing Demand Intensity, also known as the Demand Intensity Curve, or DI Curve.

Figure 32 shows a DI Curve for the GPS Nav Portal Drive0 workload. As DI increases, IOPS and ART increase until performance saturates. The “Max IOPS Point” shows the saturation point where IOPS are highest before there is a large increase in ART, also known as the DI Curve “knee”.

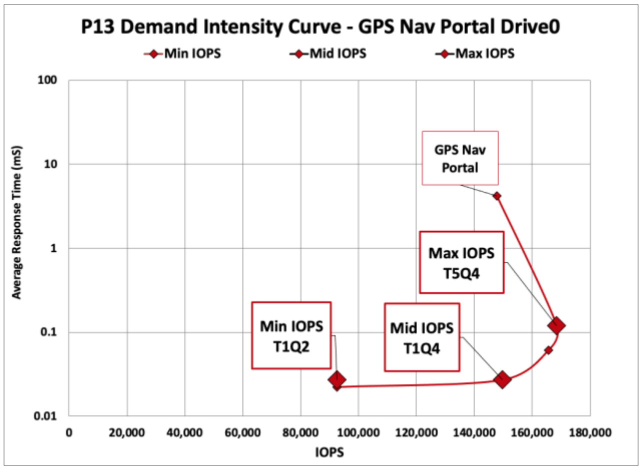

Figure 32 – DI OS Curve: GPS Nav Portal. Max IOPS shows the optimal “knee” before performance saturation

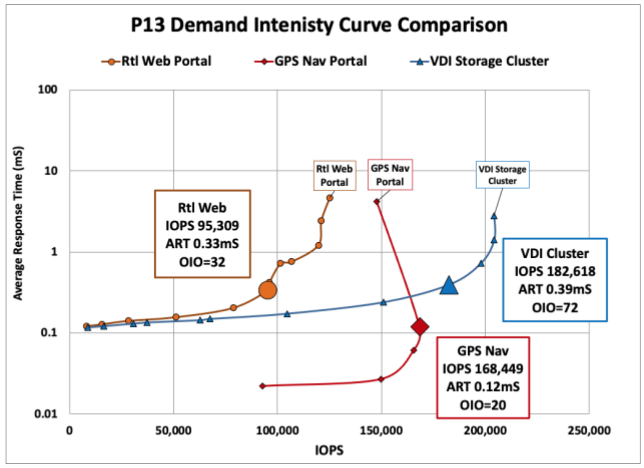

Figure 33 – DI Curve Comparison. “Better Performance” is at the “Lower Right Corner” of the DI OS Curve plot.

Figure 33 compares DI Curves for 3 RWWs run on a NVMe SSD: Retail Web Portal, GPS Nav Portal and the VDI 6-Drive Storage Cluster. “Better” performance is seen with Max IOPS points in the “lower right corner” of the DI Curve. Here, the VDI Cluster shows highest IOPS while the GPS Nav Portal shows the lowest Response Times. Note the Retail Web Portal shows the lowest performance.

D. Individual Streams Test

The Individual Streams Test measures the Steady State performance of individual IO Streams observed in a Real World Workload capture.

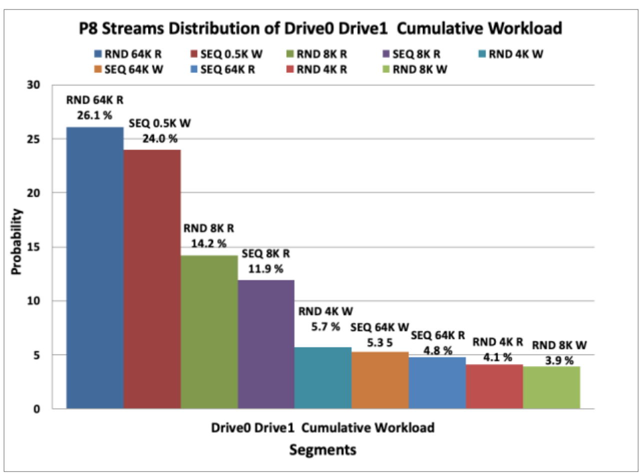

Figure 34 –Workload Streams Distribution. Individual IO Streams applied for each Steady State test step

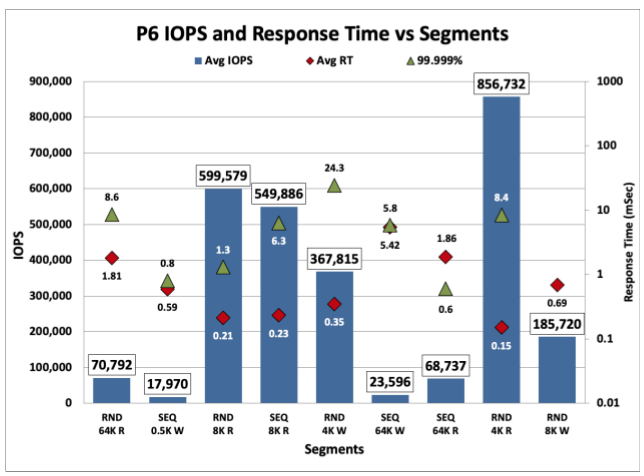

Figure 35 – Individual Streams Test: IOPS & RTs v Segments. Individual IO Streams run to Steady State

Figure 30 shows the 9 IO Stream Distribution for the Retail Web Portal RWW. Figure 31 shows individual IO Streams at Steady State and reports performance in IOPS, MB/s and Response Times.

8. HTML Playback files

Any test result displayed as an IO Stream Map (in the Calypso Test Software environment) can be saved as an HTML file. This applies to the Self-Test or any other Real World Workload test. The portable HTML file can be opened in any Chrome browser. The user can access various Tabs to view or playback of the IO Stream Map or to see pre built plots, logs, script and administrative information.

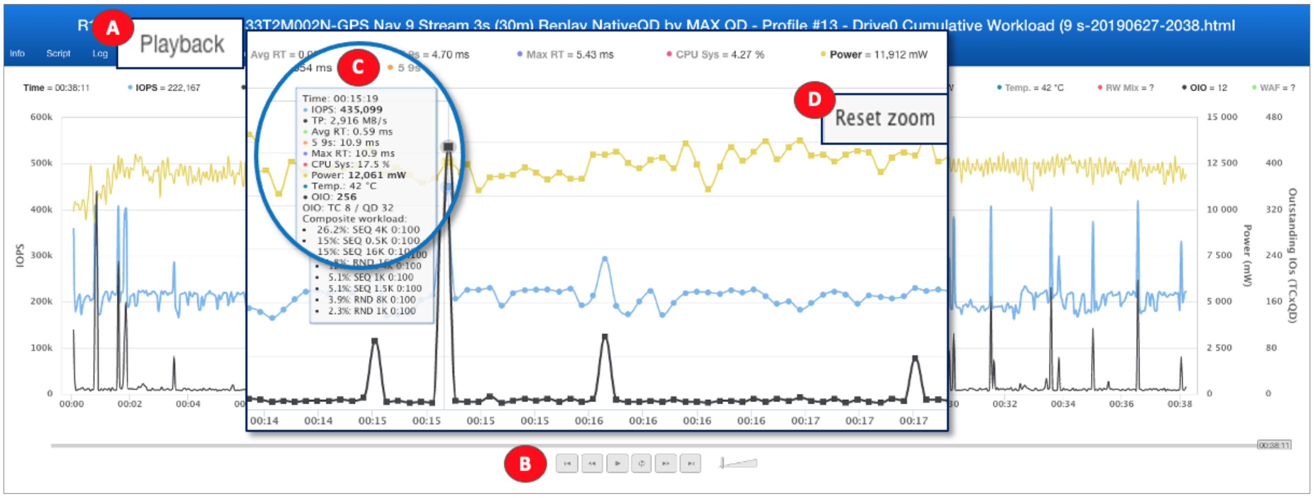

Figure 36 – HTML Results File: Playback of 24-hour Retail Web Portal Self-Test

In Figure 36, the Playback Tab (A) allows results to be replayed with the zoom, pause, loop and speed control (B). The Tooltip (C) highlights time point metrics on the IO Stream Map.

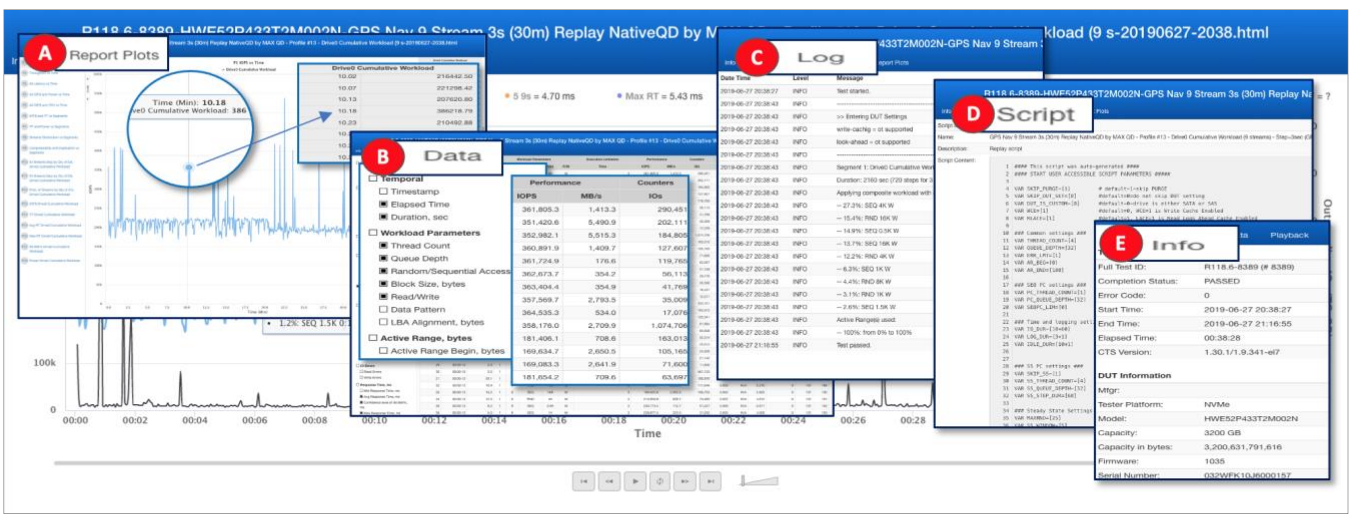

Figure 37 – HTML Results File: Results, Data, Script & Log Tabs

Figure 37 Tabs show Report Plots (A), Raw Data (B) by command line, Internal Test Logs (C), Script (D) with user accessible variables and the Info (E) with test administrative information.

The portable HTML file provides a secure audit trail with complete test results, data and reports. It can be emailed and viewed with only a Chrome browser without installation of, or access to, any other application or test software.

Conclusions

Real World Workloads are critically important for measuring application and storage performance. Real World Workloads used in conjunction with synthetic corner case benchmarks can provide a richer understanding of the use case environment and the design requirements of storage solutions.

For application optimization, knowing the IO Stream content ensures that middleware optimizations are designed based on valid assumptions of the incoming IO Streams and Demand Intensity. Indeed, an inaccurate understanding of Real World Workload content can lead to unintended consequences such as bottlenecks, response time spikes and unacceptable Quality of Service.

For the storage engineer, Real World Workloads can help illuminate the target application use case and inform the engineer of the operational Demand Intensity, the type and number of IO Streams and processes and help in the comparison of corner case benchmarks to Real World Workload performance.

Real World Workload IO Stream content can also help to evaluate real world Write Amplification and accurately predict Drive Writes Per Day for a given workload and application environment. This can help the storage designer understand what type of drive characteristics need to be provided for different use case customers.

For Cloud, Edge Compute, Datacenter and Enterprise storage engineers, Real World Workloads can be used to monitor Edge to Datacenter workload content, create validation replay scripts for suppliers, balance loads among servers and server farms, optimize storage server utilization, evaluate TRIM strategies, dis-aggregate workloads and storage architectures, analyze fabric/remote/virtual storage solutions, understand AI and RNN LSTM input and output levels and more.

For the end user, understanding Real World Workloads can help establish realistic storage qualification and validation testing, help to monitor load balancing among applications and servers, and determine how much and what kind of performance to buy.

For those interested in further understanding and using Real World Workloads, free tools are available at www.testmyworkload.com, reference workloads are posted at www.snia.org/forums/cmsi and commercial tools, including the IOProfiler, can be accessed by contacting info@calypsotesters.com.

About the Author

Eden Kim is the CEO of Calypso Systems, Inc. and chair of the SNIA Solid State Storage Technical Working Group. Eden is the primary author of the SNIA PTS specifications on SSD and Datacenter storage https://www.snia.org/tech_activities/work and author of many SNIA White Papers https://www.snia.org/forums/cmsi/knowledge/whitepapers.

Eden presents at industry trade association events in the US and China and speaks regularly at the SNIA PM Summit, SNIA SDC, Santa Clara Flash Memory Summit (FMS) and the China trade show series sponsored by DOIT in Beijing, Shanghai, Shenzhen, Wuhan and other locations. Mr. Kim is a graduate of the University of California.

Mr. Kim has published white papers which include discussions on SSD Test, Real World Workloads and Persistent Memory including: Datacenter Real World Storage Workloads, Optimizing NVMe-oF Performance with different Transports: Host Factors, SSD Performance: A Primer, Understanding SSD Performance Using the SNIA PTS Test Specification, The PTS User Guide, and the Introduction to Persistent Memory PTS White Paper.

Calypso Systems, Inc. is a supplier and manufacturer of advanced workload analysis, test and measurement software, hardware and test services and hosts the www.testmyworkload.com site. Calypso is also the supplier of the SNIA PTS Reference Test Platforms (RTP) for SSD and Datacenter testing. Calypso SSD RTPs are standard tools at SSD ODM & OEMs. Calypso IOProfiler software modules are fully functioned for the capture, analysis and test of RWWs for Cloud, Datacenter and Enterprise customers.

Questions concerning this white paper can be sent to edenkim@calypsotesters.com.

About the SNIA

The Storage Networking Industry Association (SNIA) is a not–for–profit global organization, made up of member companies spanning the global storage market. 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, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For more information, visit http://www.snia.org.

SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

Storage Networking Industry Association
5201 Great America Parkway, suite 320 • Santa Clara, CA 95054 • Phone: 719-694-1380 • Fax: 719-694-1389 • www.snia.org