SNIA™
Solid State Storage Initiative

www.snia-europe.org

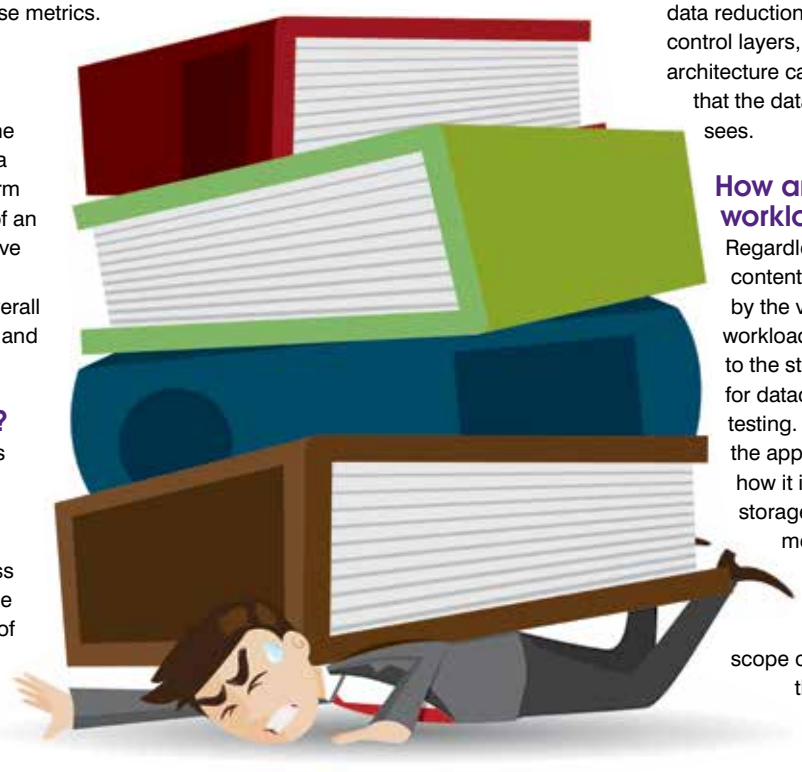
Understanding datacentre workload quality of service

By Eden Kim, Chair SNIA Solid State Storage Technical Work Group, Calypso Systems.

As solid state drives (SSDs) are deployed in datacentres in both hybrid HDD/SSD and all flash arrays (AFAs), it is becoming increasingly important to understand what metrics are relevant to assess SSD datacentre performance. While the traditional metrics of IO operations per second (IOPS), Bandwidth, and Response Times are commonly used, it is becoming more important to report and understand the 'Quality of Service' of those metrics. Response Time Confidence levels and an understanding of Demand Variation and Demand Intensity can help the IT administrator assess how a given SSD or array will perform relative to the requirements of an application workload or relative to a specific Response Time Ceiling thus helping in the overall system optimization, design, and deployment.

What are workloads?

Workload(s) are data streams generated by applications that are seen by the storage as a collection of access patterns. An individual access pattern is characterized by the spatial and temporal locality of the IO stream, Random or Sequential access, data transfer size, and Read/ Write mix. The workload is further described by the binary data content (or data pattern) of the transfer and its demand intensity (or number of threads and queues). The data pattern can be the result of a completely random transfer (i.e. random data pattern of 1's and 0's) or a workload that has some level of data reducibility (i.e. compressible or dedupable). The demand intensity is a result of the number of workers (or virtual machines) and jobs (requests) generated by the system and applications.



What are datacentre workloads?

Datacentre workloads, as seen by the storage, are the collection of data streams comprised of meta data and data content as managed through the various layers of the IO software stack. Caching layers, data reduction, data deduplication, storage control layers, storage pools, and back-up architecture can all affect the data streams that the datacentre storage ultimately sees.

How are datacentre workloads tested?

Regardless of the original data stream content and how it may be modified by the various software layers, the workload that is ultimately presented to the storage is what is important for datacentre storage performance testing. Of course, characterizing the application space workload and how it is ultimately presented to the storage is key. Various tools and methodologies are available to capture and replicate these workloads. A discussion of this topic is outside the scope of this article. Suffice it to say that whether the workload is a synthetic approximation of the application workload or a trace capture and playback,

the test operator ultimately has to apply the selected test workload to the storage and measure and analyze its performance.

Measuring workloads for performance analysis

Once the test workload has been determined, it is important to test the storage in a deterministic fashion to ensure that the actual storage is tested and that the key metrics are relevant to and for the test purposes. A valuable resource for general testing of SSD performance can be found in the SNIA Solid State Storage Performance Test Specification and more broadly discussed in accompanying white papers (see http://www.snia.org/sites/default/files/SNIASSSI.SSDPerformance-APrimer2013.pdf)

Industry standard test methodologies have been developed to ensure fair and accurate testing of SSDs both at the device and system level. Among the key points to remember are:

- Use a reference test platform with known hardware and software
- Precondition storage to a workload dependent steady state
- Set test parameter variables to match the intended application environment
- Use a robust stimulus generator and measurement tool with known attributes
- Report test results with disclosure of the test settings in a standardized format

Running a PTS DIRTH OLTP database test to compare datacentre performance

For this article, a database OLTP workload (RND 8KiB 65:35 RW mix – random, 8KiB transfer size, 65% read 35% write) was run in a PTS-E DIRTH test (Demand Intensity Response Time Histogram). The DIRTH test measures steady state performance of the test workload when running a test drive for varying Demand Intensity or Outstanding IOs (OIOs) expressed as Thread Count (TC) x Queue Depth (QD).

The drive is Purged, pre-conditioned with the test workload, and then run with varying TC and QD to determine the IOPS and Response Time (RT) saturation levels for different Demand Intensity points. Once the drive is profiled for OIOs and response times, the optimal TC/QD settings (highest IOPS at the lowest RTs) are used to create a Response Time histogram, which can then be used to determine Response Time Confidence levels and associated IOPS rates. See Figure 1: DIRTH Demand Variation Plot for OLTP workload.

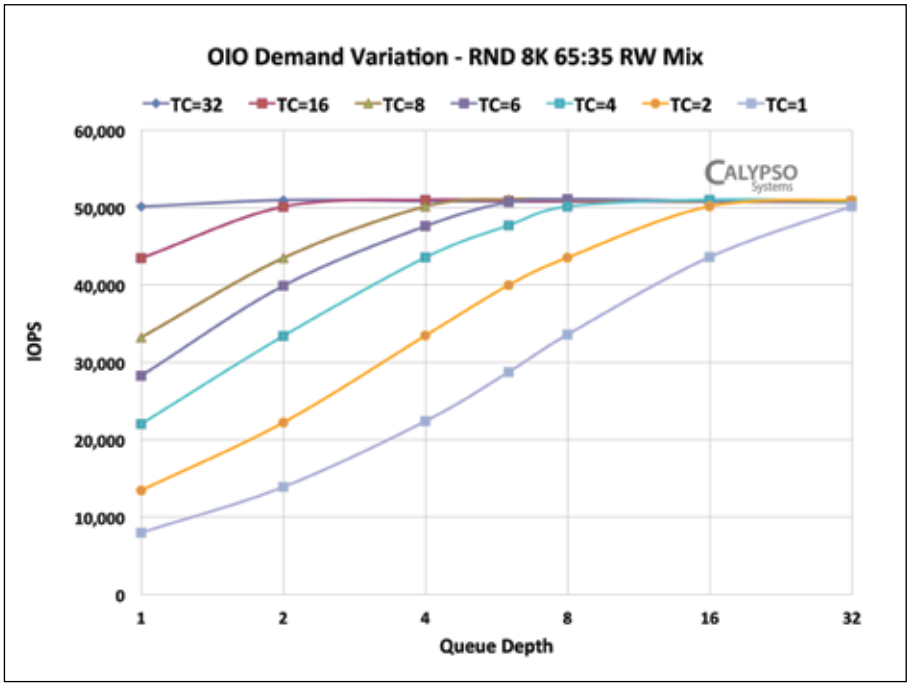


Figure 1: DIRTH Demand Variation Plot for OLTP workload

The DIRTH Demand Variation (DV) plot (Figure 1) shows how many OIOs (or TC x QD) are needed to generate the maximum IOPS. Here, it takes any OIO combination of 32 to achieve 50,000 IOPS – T1Q32, T2Q16, T4Q8, T8Q4, T16Q2 or T32Q1. While any OIO combination of 32 will result in 50,000 IOPS, the response times for each OIO will differ.

Average Response Time (ART) is measured for each TC/QD point to show where response time saturation occurs. The ART for T1Q1 is very low and increases for each OIO point until the maximum ART is seen for T32Q32. The ART saturation point is where the OIO response times increase. Here, the optimal OIO setting to achieve the maximum IOPS at the lowest ARTs is at T6Q8 or at an OIO/Demand Intensity of 48.

Demand Variation shows the operating range for the storage. Any OIO less than 48 will begin to starve the SSD and 'leave IOPS on the table' since there is not enough DI for the SSD to perform at its highest IOPS range. An OIO greater than 48 results in an increase in ARTs without a corresponding increase in IOPS. Thus, the optimal performance range for this storage is in an environment with an OIO of at least 48 – which can be viewed as 48 application threads/jobs (such as 48 virtual machines with a QD=1).

Response time confidence levels

What is a response time histogram?

A Response Time Histogram is a plot of the frequency and distribution of response times for every IO that occurs during the measurement period. In figure 2: Response Time Histogram, the x axis is time bins in mS while the Y axis is the IO count in log10 scale. During the histogram, every IO completion time is measured with every IO count cumulated in a corresponding time bin, not just the average and maximum response times.

Why confidence levels and why not use only ART and/or MRTs?

Average and Maximum Response Time (ART and MRT) are useful metrics but do not provide the quality levels that confidence levels provide. ARTs can misrepresent the

high range of an IO set while MRTs alone can misrepresent the range of all other IOs. An Average Response Time of, say, 5mS could, on its face be desirable. However if the Response Time ceiling is 6mS, the 5mS ART could be misleading: for example, when measuring ten IOs, if five IOs are 2mS and five IOs are 8mS, the ART is 5mS but there are five IOs (at 8mS) which exceed the 6mS RT ceiling.

Similarly, a single very high MRT could mask an otherwise low ART or merely represent a transient system response time spike due to Flash Translation Layer processes (such as garbage collection) that are independent of the specific IO command (i.e. such MRT spikes occur sporadically and are not caused solely by the immediate IO command).

What is response time quality of service?

Response Time Quality of Service (QoS) is a measure of the full span of response times by IO completion percentages. In other words, QoS shows at what time value a given percent of the IOs will complete. Thus, if a Response Time QoS level of 'five nines' is 20mS, then 99.999% of the IOs will complete in 20mS or less.

In Figure 2: Response Time Histogram, the ART, 2, 3, 4 and 5 nines confidence levels and the Response Time ceiling are shown as vertical colored bars. The cumulative confidence level is shown as the red line slope. (Additional information is provided in the plot header including: OIO expressed at TC/QD, IOPS and MRT.)

Tracking the 5 9's Response Time Confidence level shows at what completion time value 99.999% of the IOs will return - or when 99,999 out of 100,000 IOs will return. Conversely, 5 9's represents one dropped IO in 100,000 IOs relative to the stated 5 9's response time (in this case a 5 9's confidence time value equals 18.54 mS – or where one in 100,000 IOs returns in greater than 18.54mS).

What Is a response time ceiling & why is it used?

A Response Time ceiling is a time value threshold above which no IO response times will be accepted by the application. In other words, no IO response time can be greater than the stated RT ceiling. The RT ceiling is usually viewed with regard to a level of confidence such as the 5 9's

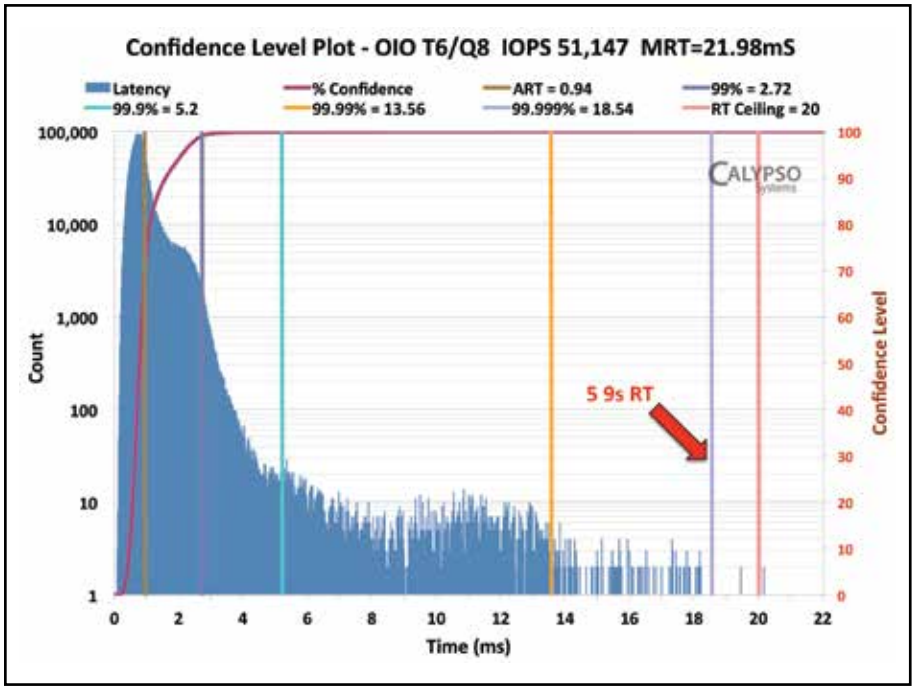


Figure 2: Response Time Histogram

confidence level. In Figure 2, the RT ceiling is set at 20mS (red vertical line). Examples of applications that will not accept any IOs after a certain amount of time, or IOs that exceed a given RT ceiling, include database applications wherein an individual request may be comprised of several, say ten, IOs.

All ten IOs must be returned within the stated RT ceiling in order for the request to be fulfilled. If one IO is late, then the total request is a failed response (unless other optimizations are in place). Note that storage that exceeds the RT ceiling can be 'tuned' to lower the 5 9's Response Time level or the system can be otherwise optimized to account for the given 5 9s response time level (by increasing cache levels and by other means).

The level of 'nines' can be set (such as 7, 8 or 9 nines) depending on the architecture and composition of the storage array or pool. The test operator should ensure that enough time is provided during the histogram to capture enough IOs to meet the designated confidence level (i.e. 9 nines equals 100,000,000 or more IOs).

Datacentre storage is beginning to utilize solid state storage as its primary, if not sole, storage media. The test and measurement of SSDs is well vetted at the SSD device level and the basic principles of SSD performance testing can be applied to AFAs. In addition to the traditional metrics of IOPS, Bandwidth, and Response Times, much can be gleaned from a closer observation of the 'Quality of Service' of those metrics.

Use of response time confidence levels and an awareness of varying demand intensity can provide a context for the range in which the storage pool can be expected to perform. Use of response time confidence levels and demand intensity/demand variation can provide the IT manager with the tools to understand both the native storage pool performance range for specific application workloads as well as provide valuable input for overall optimization of the storage pool and software/hardware stack.

About the SNIA Solid State Storage Initiative
The SNIA Solid State Storage Initiative (SSSI) fosters the growth and success of the market for solid state storage. SSSI educates markets about solid state storage, promotes and influences standards for solid state storage, and collaborates with other industry associations for success of solid state storage.

For more information on SNIA's Solid State Storage activities, visit www.snia.org/forums/sssi and get involved in the conversation at http://twitter.com/SNIASolidState