Technical Paper

Performance and Tuning Considerations for SAS® on the EMC XtremIO™ All-Flash Array

Release Information

Content Version: 1.0 March 2016.

Statement of Usage

This document is provided for informational purposes. This document may contain approaches, techniques and other information proprietary to SAS.

Contents

Introduction

This paper is a follow-up to our original EMC XtremIO ™ all-flash array (AFA) paper first distributed in October, 2014. Re-testing of the EMC XtremIO ™ scale-out flash arrays was conducted for a 4 'X-Brick' XtremIO ™ cluster, and updated 4.0 firmware. The previous EMC XtremIO ™ storage white paper was based on tests conducted on the same test host, varying only the storage subsystems for regression comparison. This paper presents testing on a new X-86 Host, 4 Node set, and as such will not be regressed against non-alike, previous test hosts.

This effort also includes a "flood test" of 4 simultaneous X-86 Nodes running a SAS Mixed Analytics workload, to determine scalability against the array, as well as uniformity of performance per node. This technical paper will outline performance test results performed by SAS, and general considerations for setup and tuning to maximize SAS Application performance with EMC XtremIO ™ arrays.

An overview of the flash testing will be discussed first, including the purpose of the testing, a detailed description of the actual test bed and workload, followed by a description of the test hardware. A report on test results will follow, accompanied by a list of tuning recommendations arising from the testing. This will be followed by a general conclusions and a list of practical recommendations for implementation with Foundation SAS®.

XtremIO™ Performance Testing

Performance testing was conducted with a 4 'X-Brick', EMC XtremIO ™ cluster, to garnish a relative measure of how well it performed with IO heavy workloads. Of particular interest was whether the XtremIO AFA would yield substantial benefits for SAS large-block, sequential IO patterns. In this section of the paper, we will describe the performance tests, the hardware used for testing and comparison, and the test results.

Test Bed Description

The test bed chosen for the flash testing was a mixed analytics SAS workload. This was a scaled workload of computation and IO oriented tests to measure concurrent, mixed job performance.

The actual workload chosen was composed of 19 individual SAS tests: 10 computation, 2 memory, and 7 IO intensive tests. Each test was composed of multiple steps, some relying on existing data stores, with others (primarily computation tests) relying on generated data. The tests were chosen as a matrix of long running and shorter-running tests (ranging in duration from approximately 5 minutes to 1 hour and 20 minutes. In some instances the same test (running against replicated data streams) was run concurrently, and/or back-to-back in a serial fashion, to achieve an average of *20 simultaneous streams of heavy IO, computation (fed by significant IO in many cases), and Memory stress. In all, to achieve the 20-concurrent test matrix, 77 tests were launched.

*Note – Previous test papers utilized a SAS Mixed Analytic 30 Simultaneous Test workload in a single SMP environment. This test effort utilizes a SAS Mixed Analytic 20 Simultaneous Test workload against each of 4 X-86 nodes; for single node comparison, as well as a 4 –node flood test (each running the 20 simultaneous workload). The 20 SAS Mixed Analytic 20 Simultaneous Test workload was chosen to better match the host CPU and Memory resources (see Hardware Description below) of the X-86 nodes. The 4-node flood test resulted in a total of 308 tests launched simultaneously against the storage.

Data and IO Throughput

The IO tests input an aggregate of approximately 300 Gigabytes of data, and the computation tests over 120 Gigabytes of data – for a single instance of each SAS Mixed Analytic 20 Simultaneous Test workload on each node. Much more data is generated as a result of test-step activity, and threaded kernel procedures such as SORT (e.g. SORT makes 3 copies of the incoming file to be sorted). As stated, some of the same tests run concurrently using different data, and some of the same tests are run back-to-back, to garnish a total average of 20 tests running concurrently. This raises the total IO throughput of the workload significantly.

As can be seen in the XtremIO Storage Monitor (Table 1 below), in its 1 hour and 20 minute span, the 4 simultaneous node workload quickly jumps to 13 GB/sec on the initial input READS for the tests. The test suite is highly active for about 40 minutes and then finishes two low-impact, long running “trail out jobs”. This is a good average “SAS Shop” throughput characteristic for a single-node instance that will simulate the load of an individual SAS COMPUTE node. This throughput is garnished from all three primary SAS file systems: SASDATA, SASWORK, and UTILLOC. Each node has its own dedicated SASDATA, SASWORK and UTILLOC file systems.

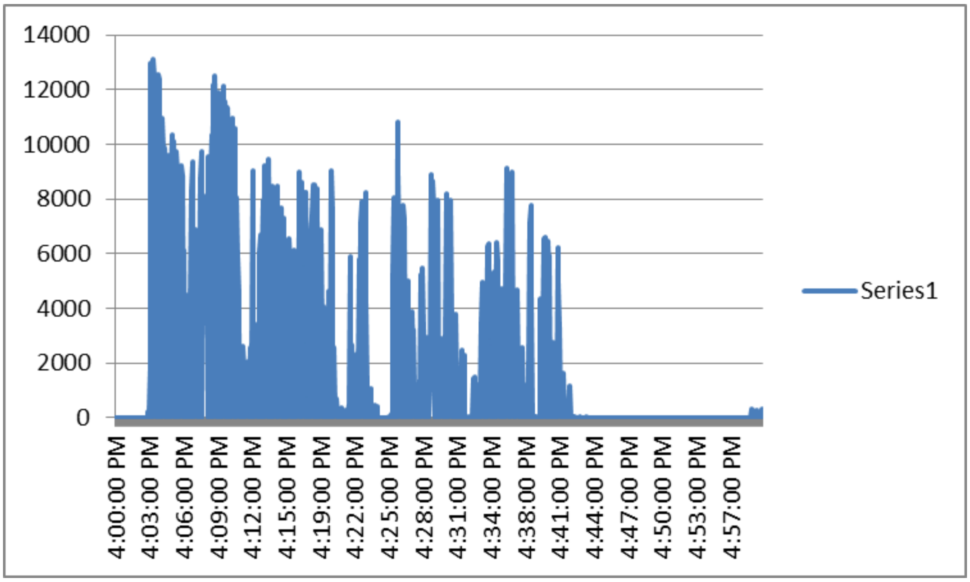


Table 1. XtremIO Throughput Monitor (Megabytes/sec) for 4-Nodes, SAS Mixed Analytic 20 Simultaneous Test Run

SAS File Systems Utilized

There are 3 primary file systems, all XFS, involved in the flash testing:

- SAS Permanent Data File Space - SASDATA
- SAS Working Data File Space – SASWORK
- SAS Utility Data File Space – UTILLOC

For this workload’s code set, data, result space, working and utility space the following space allocations were made:

- SASDATA – 3 Terabytes
- SASWORK – 3 Terabytes
- UTILLOC – 2 Terabytes

This gives you a general “size” of the application’s on-storage footprint. It is important to note that throughput, not capacity, is the key factor in configuring storage for SAS performance.

Hardware Description

This test bed was run against four host nodes utilizing the SAS Mixed Analytics 20 Simultaneous Test workload. The system host and storage configuration are specified below:

XtremIO Test Host Configuration

The four host server nodes information the testing was executed on is as follows:

Host: Lenovo x3560, RHEL 6.7 X86_64

Kernel: Linux 2.6.32-573.8.1.el6.x86_64

Memory: 256 GB

CPU: Genuine Intel Family 6, Model 63, Stepping 2, 2 Socket, 12 Cores/Socket, x86_64, 2501 MHz

Host Tuning:

The following udev rules were created:

- ACTION=="add|change", SUBSYSTEM=="block",ENV{ID_VENDOR}=="XtremIO",ATTR{queue/scheduler}="noop" ACTION=="add|change", SUBSYSTEM=="block",ENV{ID_VENDOR}=="XtremIO",ATTR{bdi/read_ahead_kb}="16384"

Multipath Settings:

The following Multipath settings were used:

- vendor XtremIO
- product XtremApp
- path_selector "queue-length 0"
- rr_min_io_rq 1
- path_grouping_policy multibus
- path_checker tur
- failback immediate
- fast_io_fail_tmo 15

Additional Settings Employed:

- custom tuned profile with the settings was implemented
- set_cpu_governor performance
- set_transparent_hugepages never
- /usr/libexec/tuned/pmqos-static.py cpu_dma_latency=1
- udevadm control reloadrules
- echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag
- blockdev --setra 16384 /dev/mapper/vg_xtremio_sasdata-lv_sasdata
- blockdev --setra 16384 /dev/mapper/vg_xtremio_saswork-lv_saswork

XtremIO Test Storage Configuration

XtremIO 4 “X-Brick” System:

- 25U in total, including the 2x Fibre Channel switches and 2x 1U InifiniBand switches (which act as the internal cluster interconnect of the XtremIO array controllers)
- An X-Brick is the fundamental building block of an XtremIO array. Each X-Brick is comprised of:
    - One 2U Disk Array Enclosure (DAE), containing:
        - 25x 400GB eMLC Hitachi Ultrastar SSDs
        - Two redundant power supply units (PSUs)
        - Two redundant SAS interconnect modules
        - One Battery Backup Unit
        - Two 1U Storage Controllers (redundant storage processors). Each Storage Controller includes:
        - Two redundant power supply units (PSUs)
        - Two 8Gb/s Fibre Channel (FC) ports
        - Two 10GbE iSCSI (SFP+) ports
        - Two 40Gb/s InfiniBand port
- Utilizing XIOS (XtremIO Operating System): 4.0.0 build 64
- 16 GB/sec potential throughput for the Cluster
- File System Type: XFS
- File systems configured with 16x 500g LUNs provisioned to each host used for 2 file systems: SASDATA and SASWORK (also containing UTILLOC).
- The LUNs were striped with LVM using a 64K stripe-size, formatted with XFS, and mounted with noatime/nodiratime options. This was matched by the SAS application using a 64K BUFSIZE.

Test Results

Single Host Node Test Result

The Mixed Analytic Workload was run in a quiet setting (no competing activity on server or storage) for the X-86 System utilizing EMC XtremIO^{TM} 4 X-Brick cluster on a single host node. Multiple runs were committed to standardize results.

The tuning options noted in the host sections above to LINUX operating systems for Red Hat Enterprise Linux 6.x. Work with your EMC representative for appropriate tuning mechanisms for any different OS, or the particular processors used in your system.

Table 2 below shows the performance of the EMC XtremIO TM X-Brick system. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 77 tests submitted. It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - X-86 w/EMC XtremIO 4 - X-Brick™	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	Memory Used in GB - Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	1.06	626	39	674	78

Table 2. Frequency Mean Value for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using EMC XtremIO TM All-Flash Arrays

The second column in Table 2 shows the ratio of total CPU time (User + System CPU) against the total Real time. Table 2 above shows the ratio of total CPU Time to Real time. If the ratio is less than 1, then the CPU is spending time waiting on resources, usually IO. The XtremIO system delivered a very good 1.06 ratio of Real Time to CPU. The question arises, "How can I get above a ratio of 1.0?" Because some SAS procedures are threaded, you can actually use more CPU Cycles than wall-clock, or Real time.

The third column shows the total elapsed run time in minutes, summed together from each of the jobs in the workload. It can be seen that the EMC XtremIO 4 X-Brick cluster coupled with the faster Intel processors on the Lenovo compute node executes the aggregate run time of the workload in approximately 626 minutes of total execution time.

It is important to note that while we aren't making direct comparisons of this newer generation, X-86 Node test to previous host testing; this newer, faster Intel processor set executed the workload in roughly half the time as the previous older generation host! The primary take-away from this test is that the EMC XtremIO 4-X-Brick cluster was able to easily provide enough throughput (with extremely consistent low latency) to fully exploit this host improvement! Its performance with this accelerated IO demand still maintained a very healthy 1.06 CPU/Real Time ratio!

Scaling from One Node to 4 Nodes on the EMC X-Brick System

For a fuller "flood test", the Mixed Analytic 20S Workload was run concurrently on 4 physically separate, but identical host nodes in a quiet setting (no competing activity on server or storage) for the X-86 System utilizing EMC XtremIO TM 4 X-Brick cluster. Multiple runs were committed to standardize results.

The tuning options noted in the single host node section above for Red Hat Enterprise Linux 6.x were utilized.

Table 3 below shows the performance of the 4 host node environments attached to the EMC XtremIO ™ 4 X-Brick cluster. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 77 tests submitted per node (308 in total). It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - X-86 w/EMC XtremIO 4 - X-Brick™	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	Memory Used in GB - Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	1.09	631	39	666	70
Node2	1.08	622	39	662	64
Node3	1.08	622	39	665	64
Node4	1.08	614	39	648	65

Table3. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using EMC XtremIO ™ All-Flash Arrays

The second column in Table 3 shows the ratio of total CPU time (User + System CPU) against the total Real time for each Node. If the ratio is less than 1, then the CPU is spending time waiting on resources, usually IO. The XtremIO system delivered a very good 1.06 ratio of Real Time to CPU.

The third column shows the total elapsed run time in minutes, summed together from each of the jobs in the workload. It can be seen that the EMC XtremIO 4 X-Brick cluster coupled with the faster Intel processors on the Lenovo compute node executes the aggregate run time of the workload in an average of 628 minutes per node, and 2,512 minutes of aggregate execution time for all 4 nodes.

Again, the newer, faster Intel processor set executed the workload on each node, in roughly half the time as the previous tested, older generation host! In addition to halving the execution time, the scale of 4 simultaneous test workloads was implemented versus the prior testing single workload. Again, the EMC XtremIO 4 X-Brick cluster was able to easily scale to meet this accelerated and scaled throughput demand, while providing a very healthy CPU/Real Time ratio per node!

The workload utilized was a mixed representation of what an average SAS shop may be executing at any given time. Due to workload differences your mileage may vary.

General Considerations

Utilizing the EMC XtremIO ™ array can deliver significant performance for an intensive SAS IO workload. It is very helpful to utilize the SAS tuning guides for your operating system host to optimize server-side performance with XtremIO™, and additional host tuning is performed as noted below.

General industry considerations when employing flash storage tend to recommend leaving overhead in the flash devices to accommodate garbage-collection activities and also focus on which workloads, if not all, to use flash for. Both points are discussed briefly:

- As per EMC, the XtremIO AFA will deliver consistent performance regardless of capacity consumption and because of this, does not require end-users to limit themselves to a fraction of the purchased flash capacity in order to enjoy the benefits of enterprise flash-based storage performance. This is the result of a number of architectural design choices employed by XtremIO – there are no system-level garbage collection activities to worry about (this activity is decentralized to the SSD controllers), all data-aware services (deduplication and compression) are always-on and happen in-line with no post-processing, and the XDP (XtremIO Data Protection) algorithm ensures minimal write activity and locking of the SSDs. More information available in the 'System Overview' section of : http://www.emc.com/collateral/white-papers/h11752-intro-to-XtremIO-array-wp.pdf
- In regards to what workloads should be moved to all-flash storage, the appropriate guidance will always be situation specific and you should consult your EMC and SAS specialists to assess individual suitability. It is quite true to say that the majority of flash storage platforms will provide higher throughput for read-intensive workloads than that can be seen in write-heavy equivalents. Because of this fact, most environments will tend to bias their initial flash adoption towards read-intensive activities – for example, a large repository that gets updated nightly, or weekly, and is queried and extracted from at a high level by users. The increase in available IOPS is only one potential benefit, AFA users must also consider how best to exploit the inherent reduction in command completion latencies seen with flash and how this may alleviate locking of threads at the processor level. Data-aware AFA (e.g. XtremIO) users must also identify what datasets may be best reduced using the array's deduplication and compression capabilities. In short, it could be said that any workload(s) which need more IOPS, lower response times, or increased storage density will benefit from the capabilities of a data-aware all-flash array, but as always, it will be up to the administration team to determine the worth of these potential benefits.

EMC XtremIO ™ all-flash arrays utilize in-line data reduction technologies including de-duplication and compression. These technologies yielded an average 2.9:1 reduction for de-duplication, and 3.8:1 reduction for compression, effectively reducing the amount of capacity needed on the XtremIO array by 2.9*3.8 = 11.02:1. This is a significant reduction. When you consider this reduction applied across all SAS file systems, including SASWORK and UTILLOC where data expansion is significant, this is extremely beneficial in capacity savings. Depending on a customer's specific SAS data characteristics, your experienced data reduction values will vary. Please work with your EMC engineers to determine your actual reduction across all 3 primary SAS file systems.

EMC and SAS Tuning Recommendations

Host Tuning

It is important to study, and use as an example, the Host and Multipath tuning listed in the Hardware section above. In addition pay close attention to LUN creation and arrangements, and LVM tuning. For more information configuring SAS workloads for REDHAT Systems please see:
http://support.sas.com/resources/papers/proceedings11/342794_OptimizingSASonRHEL6and7.pdf

In addition a SAS BUFSIZE option of 64K along with a RHEL Logical Volume Stripe size of 64K was utilized to achieve the best results from this XtremIO array. Testing with a lower BUFSIZE value may yield benefits on some IO tests, but may require host and application tuning changes.

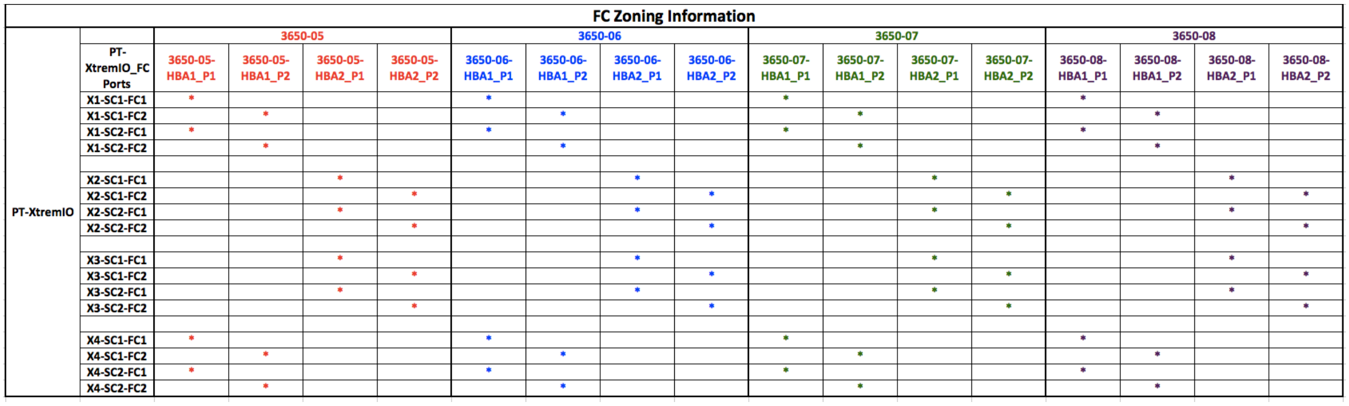
XtremIO Storage Tuning

XtremIO has no requirement to provide knobs for data placement such as creating RAID groups and/or creating back-end storage striped or concatenated LUN. Essentially, there is no storage tuning involved when it comes to XtremIO since any application on it has complete access to the fully distributed resources of the array. However, XtremIO is an active-active scale-out array. There remains a commonsensical approach to integrating SAN clients to the XtremIO AFA.

The storage array featured for this test bed is comprised of four X-bricks. Each X-brick has two storage controllers (SC). Each SC has two Fibre Channel ports and two iSCSI ports. The hosts depicted in this environment access the storage over FC. On the array, any or all volumes are generally accessible from any storage port; but to ensure balanced utilization of the full range of XtremIO resources available on the array, the host initiator ports (four per host) were zoned to the all storage ports in uniform (see Table 4 below). Per best practice on XtremIO, the maximum number of paths to the storage was limited to 16. EMC's Storage Array User Guide and Host Connectivity Guide are downloadable from support.emc.com:

(https://support.emc.com/docu62760_XtremIO-4.0.2-Storage-Array-User-Guide.pdf?language=en_US)

(https://support.emc.com/docu56210_XtremIO-2.2.x---4.0.2-Host-Configuration-Guide.pdf?language=en_US)



		FC Zoning Information															
		3650-05				3650-06				3650-07				3650-08			
	PT-XtremIO_FC Ports	3650-05-HBA1_P1	3650-05-HBA1_P2	3650-05-HBA2_P1	3650-05-HBA2_P2	3650-06-HBA1_P1	3650-06-HBA1_P2	3650-06-HBA2_P1	3650-06-HBA2_P2	3650-07-HBA1_P1	3650-07-HBA1_P2	3650-07-HBA2_P1	3650-07-HBA2_P2	3650-08-HBA1_P1	3650-08-HBA1_P2	3650-08-HBA2_P1	3650-08-HBA2_P2
PT-XtremIO	X1-SC1-FC1	*				*				*				*			
	X1-SC1-FC2		*				*				*				*		
	X1-SC2-FC1	*				*				*				*			
	X1-SC2-FC2		*				*				*				*		
	X2-SC1-FC1			*				*				*				*	
	X2-SC1-FC2				*				*				*				*
	X2-SC2-FC1			*				*				*				*	
	X2-SC2-FC2				*				*				*				*
	X3-SC1-FC1			*				*				*				*	
	X3-SC1-FC2				*				*				*				*
	X3-SC2-FC1			*				*				*				*	
	X3-SC2-FC2				*				*				*				*
	X4-SC1-FC1	*				*				*				*			
	X4-SC1-FC2		*				*				*				*		
	X4-SC2-FC1	*				*				*				*			
	X4-SC2-FC2		*				*				*				*		

Table 4 – Fibre Channel Zoning Configuration for XtremIO 4 – X-Brick Used in this test.

XtremIO Monitoring

XtremIO provides the ability to monitor every storage entity/component on the array either for front-end or back-end access. One particular metric merits mention here – the XtremIO AFA provides the ability to monitor and record the resource utilization of every XENV (XtremIO Environment) throughout the cluster. An XENV is composed of software-defined modules responsible for internal data path on the array. There are two CPU sockets per SC, and one distinct XENV runs on each socket. For example, X1_SC1_E1 pertains to the first XENV or socket on SC1, X-brick1. X1_SC1_E2 would be second XENV or socket on SC1, X-brick1. Table 5 below shows the utilization of the SAS workload against each XENV.

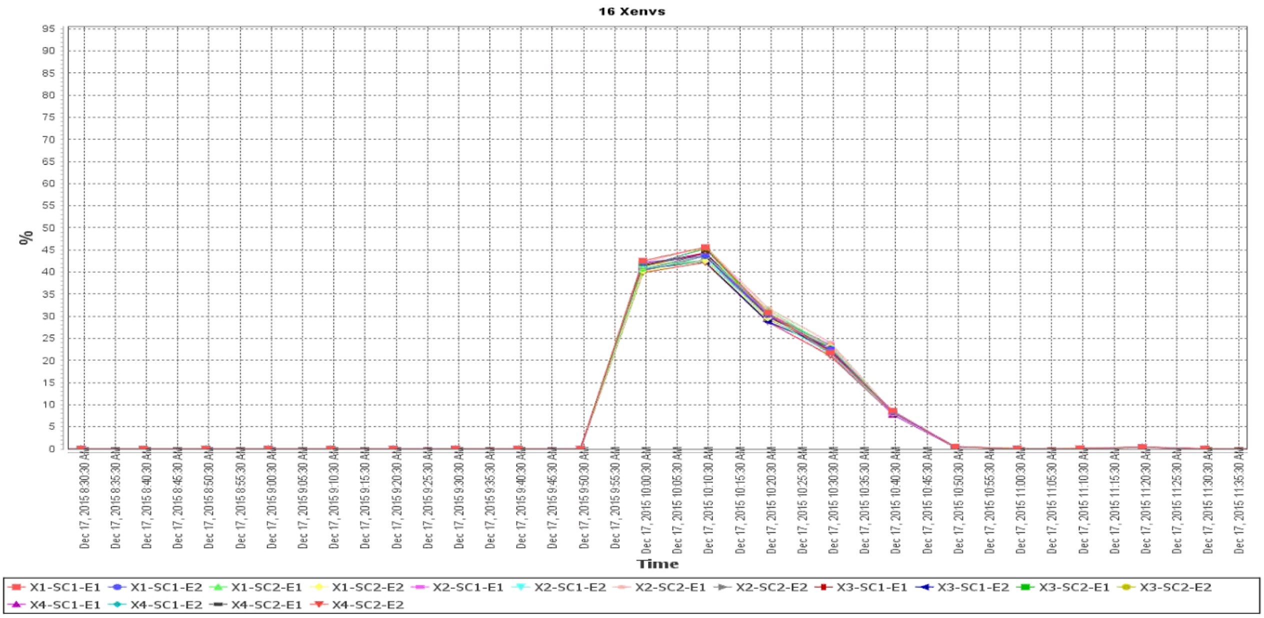


Table 5 – XtremIO Environment (XENV) Balanced Utilization.

The tightly knit grouping of the XENVs is a good indicator of a well-balanced storage array in terms of performance utilization. The trend showed peak at 45% and a declining pace right after. Clearly, the array was not saturated in this case, but in the event that it was, this is easily mitigated. XtremIO is designed to be fully scalable and since it uses a multi-controller scale-out architecture, it can therefore scale out linearly in terms of performance, capacity and connectivity through the addition of more X- Bricks. If expanded to eight for example, the new expected performance capacity is two times that of the existing tested 4-Xbrick cluster. Automatic data placement is part of XtremIO Data Protection implementation. You can read all about it from xtremio.com or https://www.emc.com/storage/xtremio/overview.htm.

Conclusion

The EMC XtremIO ™ all-flash array has been proven to be extremely beneficial for scaled SAS workloads when using newer, faster processing systems. In summary, the faster processor enables the compute layer to perform more operations per second, thus increasing the potential performance for the solution, but it is the consistently low response times of the underlying storage layer which allow this potential to be realized.

Operating the XtremIO array is designed to be as straightforward as possible, but to attain maximum performance, it is crucial to work with your EMC Storage engineer to plan, install, and tune the hosts for the environment.

The guidelines listed in this paper are beneficial and recommended. Your individual experience may require additional guidance by EMC and SAS Engineers depending on your host system, and workload characteristics.

Resources

SAS Papers on Performance Best Practices and Tuning Guides: http://support.sas.com/kb/42/197.html

Contact Information:

Josh Goldstein
EMC Corporation
2841 Mission College Blvd., 4th Floor
Santa Clara, CA 95054
+1(408) 625-7425
josh.goldstein@emc.com

Ted Basile
EMC Corporation
176 South Street
Hopkinton, MA 01748
+1(508) 435-1000
edward.basile@emc.com

Tony Brown
SAS Institute Inc.
15455 N. Dallas Parkway
Dallas, TX 75001
+1(469) 801-4755
tony.brown@sas.com

Margaret Crevar
SAS Institute Inc.
100 SAS Campus Dr
Cary NC 27513-8617
+1 (919) 531-7095
margaret.crevar@sas.com

To contact your local SAS office, please visit: sas.com/offices