Technical Paper

Performance and Tuning Considerations for SAS® using Veritas InfoScale™ Storage 7.0 on the EMC XtremIO™ All-Flash Array

Release Information

Content Version: 1.0 March 2016.

Contents

Introduction

This paper presents testing results and tuning guidelines for running Foundation SAS® on the Veritas InfoScale™ Storage 7.0 Cluster File System (VxCFS) with the EMC XtremIO™ scale-out flash arrays. Testing was conducted with an underlying EMC 4 'X-Brick' XtremIO™ cluster, and updated 4.0 firmware using a newer X86 Host, 4 node set. Regressions have not been made to previous EMC XtremIO™ scale-out flash arrays with dissimilar file systems. This effort consisted of a "flood test" of 4 simultaneous X-86 Nodes running a SAS Mixed Analytics workload, to determine scalability against the clustered file system and array, as well as uniformity of performance per node.

Veritas InfoScale™ Enterprise Storage not only provides a clustered file system across all the grid nodes, but also brings a complete storage management solution. It includes multi-pathing and volume management software, and works with the Veritas InfoScale™ Operations Manager (VIOM) which simplifies management and provides visibility of the full stack.

Veritas Dynamic Multi-Pathing software integrates seamlessly with leading storage arrays like EMC XtremIO™; including specific Array Support Libraries (ASL). This software facilitates communication between storage and server administrators through improved visibility of the storage network.

This paper will outline performance test results performed by SAS, and general considerations for setup and tuning to maximize SAS Application performance with InfoScale Storage™ 7.0 on EMC XtremIO™ arrays.

An overview of the 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 SAS® Foundation software.

Veritas InfoScale™ Storage 7.0 on XtremIO™ Performance Testing

Performance testing was conducted with InfoScale™ Storage 7.0 on a 4 'X-Brick', EMC XtremIO™ cluster, to establish a relative measure of how well it performed with IO heavy workloads. Of particular interest was how well the Veritas Cluster File System would perform 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 per node.

*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. 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 have an average of 20 tests running concurrently. This raises the total IO throughput of the workload significantly.

As can be seen in the Veritas Cluster Volume Manager (VxCVM) management utility, vxstat (Figure 1 below), in its 1 hour and 20 minute span, the 4 simultaneous node workload quickly exceeds 4 GB/sec on the initial input READS for the tests, then to 11 GB/sec for SASWORK WRITE activity. 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 obtained from all three primary SAS file systems: SASDATA, SASWORK, and UTILLOC.

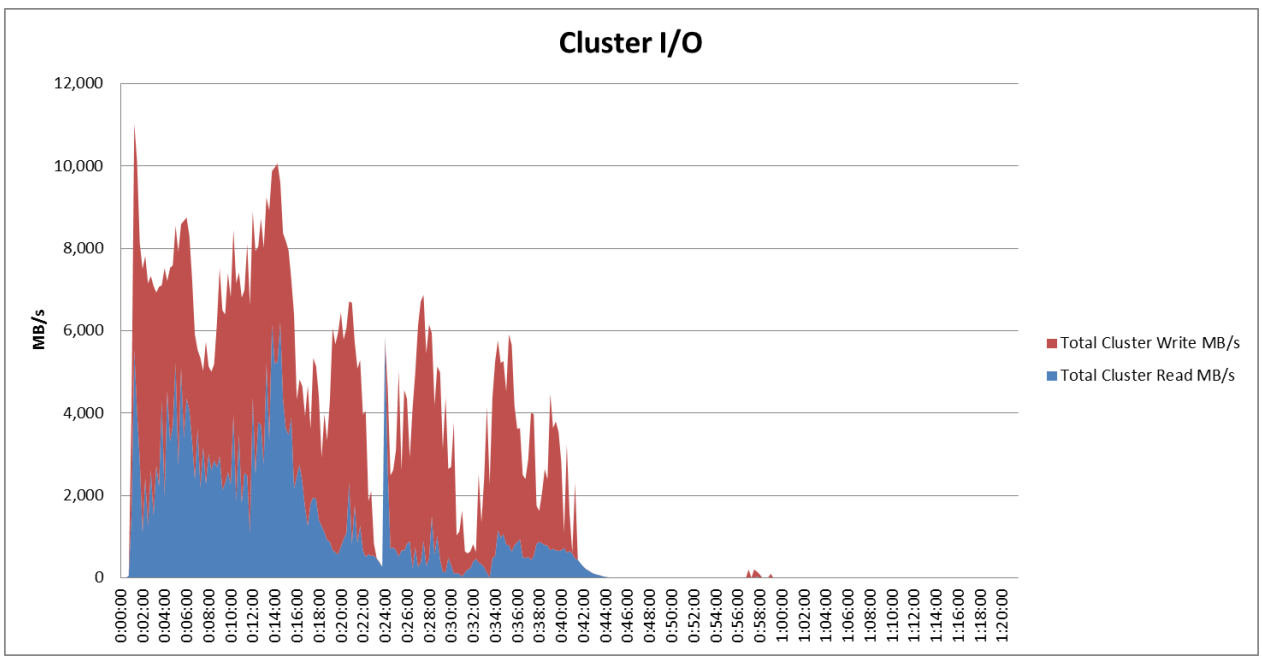

Figure 1. Vxstat Throughput Monitor (Megabytes/sec) for 4-Nodes, SAS Mixed Analytic 20 Simultaneous Test Run

SAS File Systems Utilized

There are 3 primary file systems, all Veritas Cluster File Systems, involved in the flash testing:

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

For this workload's code set, data, result file system, working and utility file system 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:

Test Hosts Configuration

The four host server nodes are described below:

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:

- Because Veritas InfoScale™ Storage 7.0 automatically configures Veritas Dynamic Multi-Pathing, and this includes Array Support Libraries for EMC XtremIOTM, there is no need to create any specific configuration as the paths to the storage array are automatically detected and configured.

- The InfoScale Operations Manager screenshot in Table 1 below shows paths automatically appear for each of the LUNs configured under InfoScale Storage™ 7.0.

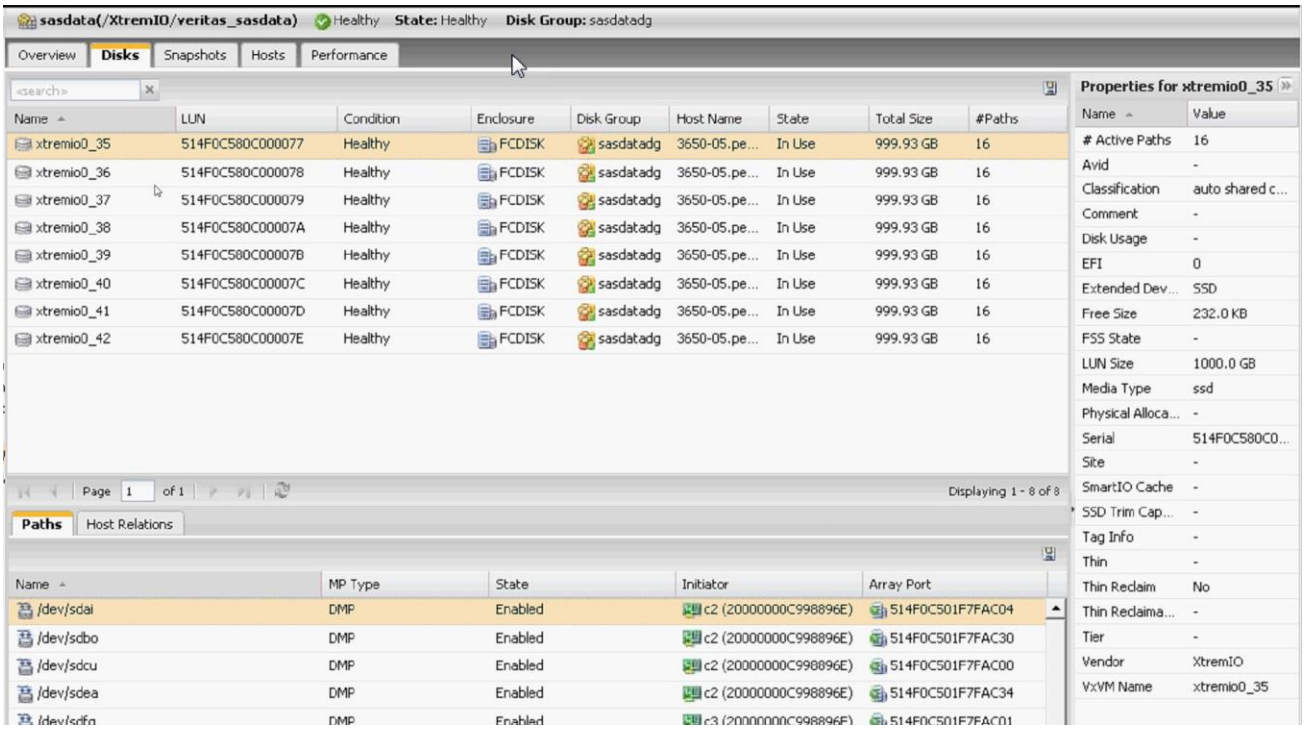

Table 1. InfoScale Operations Manager automated path detection.

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
- echo never > /sys/kernel/mm/redhat_transparent_hugepage/defrag

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: Veritas InfoScale™ Storage 7.0 Cluster File System (VxCFS)

- 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 Veritas Cluster Volume Manager using a 256K stripe-size, formatted with Veritas Cluster File System 7.0, and mounted with noatime/nodiratime options. File system block size was 8192. This was matched by the SAS application using a 256K BUFSIZE.

Test Results

The Mixed Analytic Workload was run in a quiet setting (no competing activity on server or storage) for the X-86 system utilizing InfoScale Storage™ 7.0 on EMC XtremIO™ 4 X-Brick cluster on a 4 host node set. Multiple runs were committed to standardize results.

The tuning options noted in the host sections above pertain to LINUX operating systems for Red Hat Enterprise Linux 6.x. Please note that since tuning is dependent on the OS and processor choices, you should work with your Veritas and EMC representatives to obtain appropriate tuning parameter values for your system.

Table 2 below shows the performance of the 4 host node environments running the SAS Mixed Analytic Workload with Veritas InfoScale™ Storage 7.0 on an EMC XtremIO™ X-Brick system. 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.

X-86 w/Veritas InfoScale™ on 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	1.0	718	685	96.83	39
Node2	1.0	715	684	97	39
Node3	1.02	701	682	95	39
Node4	1.0	727	693	100	39

Table 2. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using Veritas InfoScale™ 7.0 on EMC XtremIO™ All-Flash Arrays

The second column in Table 2 shows the ratio of total CPU time (User + System CPU) against the total Real time. If the ratio is less than 1, then the CPU is spending time waiting on resources, usually IO. Veritas InfoScale™ Storage7.0 on the XtremIO system delivered a good 1.0 ratio of Real Time to CPU. Node 3 experienced a ratio of slightly higher than one. 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 Veritas InfoScale*TM* Storage7.0 on 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 715 minutes per node, and 2,861 minutes of aggregate execution time for all 4 nodes.

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 57% of the time as the previous older generation host! The primary take-away from this test is that InfoScale Storage*TM* 7.0 on the EMC XtremIO*TM* 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 healthy 1.0 CPU/Real Time ratio. This is very good performance for a clustered file system, which typically has more processing overhead, and related latency than a local file system.

The workload utilized was a mixed representation of what an average SAS environment may be experiencing at any given time. Please note that in general the performance depends on the workload presented and will therefore vary from one environment to another.

General Considerations

Utilizing InfoScale™ 7.0 Cluster File System on 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 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 focus on which workloads 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.

Veritas, EMC, and SAS Tuning Recommendations

Host Tuning

IO Elevator tuning, and OS host tunable settings are very important for maximum performance. For more information on 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 256K coupled with a Cluster Volume Stripe size of 256K was utilized to achieve the best combined performance of the Cluster File System and storage array.

Veritas InfoScale™ 7.0 Tuning

When creating volumes that will be used to store SAS data, it is important to determine what the best layout for those volumes. Veritas Cluster Volume Manager offers the ability to tune to specific IO demands and make changes when those IO profiles change. Because SAS performs large sequential writes and reads, Veritas volumes can be created using a large stripe unit or block size, which means that more data will be write and read in parallel from each LUN. Adapting that parameter to the IO used by the application provides the best performance.

Testing was conducted with various block size transfers for READ and WRITE operations. While the underlying EMC XtremIO array optimally processes 64KB Blocks, the Veritas Cluster File System, in combination with SAS BUFSIZE parameter optimally performs with Cluster Volume stripe sizes of 256KB. Testing both the SAS BUFSIZE and Cluster Volume stripe size with 64KB, and 256KB yielded the best overall results with 256KB. There was slight additional latency experienced on the storage array moving using 256KB transfers instead of the array preferred 64KB transfers. This latency was outweighed by the improved Veritas Cluster File System performance at 256KB for large IOs. When utilizing storage arrays that offer optimal block performance that differs from 256KB, you may have to be experiment to achieve the best overall result for Veritas InfoScale™ 7.0.

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 3 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 3 – 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 4 below shows the utilization of the SAS workload against each XENV.

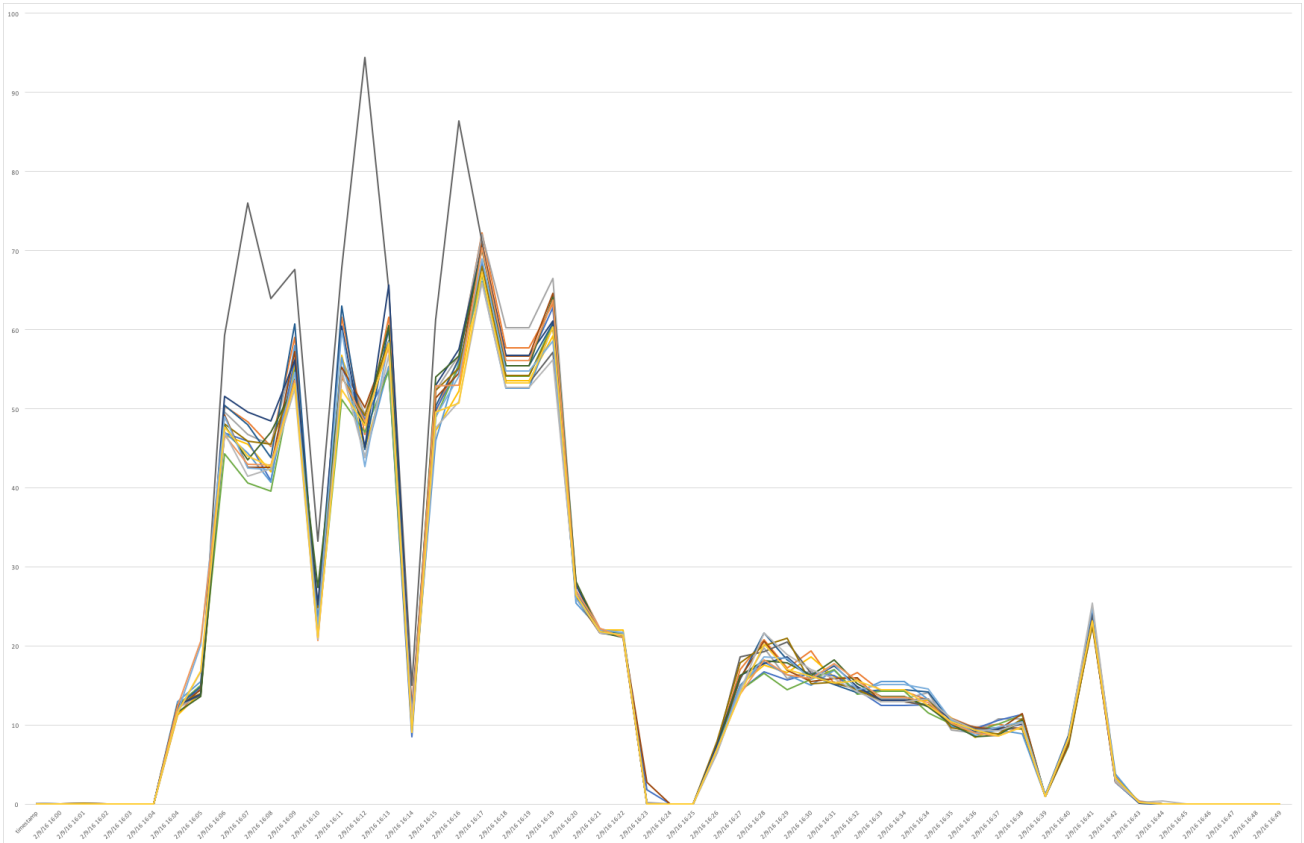

Table 4 – XtremIO Environment (XENV) Balanced Utilization.

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

Veritas InfoScale™ 7.0 on 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 performance of this clustered file system is very good. Coupled with the low latency underlying XtremIO™ storage, Veritas Cluster File System performs very well SAS workloads running on the faster processors in today's host servers.

To attain maximum performance for your site, it is crucial to work with your Veritas and EMC engineers 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 Veritas, 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:

Carlos Carrero
Veritas Technologies LLC
Paseo del Club Deportivo s/n. Building 13
Pozuelo de Alarcon, 28223 Madrid
SPAIN
+34 91 700 55 80
carlos.carrero@veritas.com

Shailesh Marathe
Veritas Technologies LLC
RMZ ICON, Baner Road,
Pune, MH 411045
IINDIA
+91 20 4075-4416
shailesh.marathe@veritas.com

Ed Menze
Veritas Technologies LLC
500 E Middlefield Rd
Mountain View CA 94043
+1 503-970-5127
ed.menze@veritas.com

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