Performance and Tuning Considerations for SAS® on the EMC XtremIO™ All-Flash Array
Release Information
Content Version: 1.0 October 2014.

Trademarks and Patents

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

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

EMC XtremIO™ all-flash storage offers high performance, and excellent scalability for SAS® workloads. EMC XtremIO™ scale-out flash arrays can be configured for large SAS workloads. 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 SAS®.

XtremIO™ Performance Testing

Performance testing was conducted with the EMC XtremIO™ all-flash array to garnish a relative measure of how well it performed with IO heavy workloads compared with high-performance spinning disk. Of particular interest was whether the XtremIO flash storage 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 53 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 30 simultaneous streams of heavy IO, computation (fed by significant IO in many cases), and Memory stress. In all, to achieve the 30-concurrent test matrix, 101 tests were launched.

Data and IO Throughput

The IO tests input an aggregate of approximately 300 Gigabytes of data, and the computation over 120 Gigabytes of data – for a single instance of each test. 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 30 tests running concurrently. This raises the total IO throughput of the workload significantly. In its 1 hour and 45 minute span, the workload quickly jumps to 900 MB/sec, climbs steadily to 2.0 GB/s, and achieves a peak of 4+ GB/s throughput before declining again. This is a good average “SAS Shop” throughput characteristic for a single-instance OS (e.g. non-grid). This throughput is from all three primary SAS file systems: SASDATA, SASWORK, and UTILLOC.
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. Fallow space is generally left on the file systems to facilitate write performance and avoid write issues due to garbage collection when running short on cell space. Per EMC it is not necessary to leave fallow space to alleviate garbage collection related latency. This array was not tested at a full mark for such.

Hardware Description

The host server information the testing was run performed on is as follows:

Host: HP DL980 G7  
OS: Linux version RHEL 6.5, Kernel Version 2.6.32-431.el6.x86_64 (mockbuild@x86-023.build.eng.bos.redhat.com) (gcc version 4.4.7 20120313 (Red Hat 4.4.7-4) (GCC)) #1 SMP Sun Nov 10 22:19:54 EST 2013  
Memory: 529,223,432 KB RAM,  
CPU: 64 x Intel(R) Xeon(R) CPU X7560 2.27GHz GenuineIntel, Model 46, CPU Family 6, Stepping 6, 2266 Mhz, 24576 KB Cache

Storage

Comparative performance testing was conducted between a high-performance spinning disk array, and the EMC XtremIO™ scale-out all-flash array. A comparison of the two storage types follows:

High-performance Spinning Disk Array Definition:

- Number and types of disks / controllers: 192 x 300GB 15K rpm Fibre Channel Drives and 4 controllers, consuming 84U all together
- RAID levels: Multiple RAID 5 sets (3 data/1 parity disk per set)
- File System Type: XFS
- File System/Logical Volume Arrangement: File Systems /SASDATA, SASWORK, /UTILLOC are placed across 1 Logical Volume utilizing all 192 spindles in the array
- Fibre Channel ports: 8 x 8 Gbps Adapters with ACTIVE/ACTIVE multi-pathing
- 8 GB/sec potential throughput for the array
The EMC XtremIO™ all-flash array Definition:

- 25 x 400 GB Hitachi Ultrastar™ Solid State Flash Drives in Each XtremIO X-Brick; total of 50 SSDs in the XtremIO cluster
- 2 "X-Brick" Cluster in a 13U Configuration
- 13U unit - 2 X-Bricks, each 5U plus two 1U Infiniband switches which act as the internal cluster interconnect of the XtremIO array controllers. 1U of space is left open for cable routing.
- For the entire 2 X-Brick Cluster - 8 x 8 Gbps Fiber Channel Adapters
- Utilizing Firmware 3.0 Build 42
- File System Type: XFS
- 8 GB/sec potential throughput for the Cluster

Test Results

The Mixed Analytic Workload was run in a quiet setting (no competing activity on server or storage) for both the high-performance spinning disk storage and the EMC XtremIO™ all-flash array. Multiple runs were committed to standardize results.

We worked with EMC storage engineers to tune the host and IO pacing (see General Considerations and Tuning Recommendations below). The EMC firmware level used in testing was 3.0, Update 42, the current (as of the date of publishing this document) generally available release. EMC XtremIO™ all-flash arrays utilize inline data reduction technologies including de-duplication and compression at Firmware 3.0, Update 42. 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 $2.9 \times 3.8 = 11.02:1$. 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 an extremely beneficial space savings. Depending on a customer’s specific SAS data characteristics, experienced data reduction ratios will vary.

The tuning options noted below apply to LINUX operating systems for Red Hat RHEL 6.5. 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 completely tuned environments comparing the EMC XtremIO™ system and the high-performance spinning disk system. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 101 tests submitted. It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

<table>
<thead>
<tr>
<th>Storage System</th>
<th>Elapsed Real Time in Minutes – Workload Aggregate</th>
<th>Memory Used in GB – Workload Aggregate</th>
<th>User CPU Time in Minutes – Workload Aggregate</th>
<th>System CPU Time in Minutes – Workload Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4
The first column in shows the total elapsed run time in minutes, summed from each of the jobs in the workload. It can be seen that the EMC XtremIO array reduced the aggregate run time of the workload by approximately 61 minutes, which is a big improvement.

Another way to review the results is to look at the ratio of total CPU time (User + System CPU) against the total Real time. Table 3 below 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 high-performance spinning storage array is a very large, expensive, and fast array. The Real time/CPU time ratio of the workload on the Spinning Storage array was 1.157, which is excellent.

The EMC XtremIO™ all-flash array achieved close to that performance at 1.14 with a 13 rack unit (RU) 2 X-brick Cluster, which is a relatively small XtremIO system (at the time of publishing XtremIO clusters scale to six X-Bricks in a cluster). Both of these numbers indicate the storage was keeping up with the CPU demand and provided excellent performance. The standard deviation and spread of the ratios for spinning storage is slightly higher than the XtremIO test, but the trade in the aggregate mean measure of Total CPU/Real Time for a reduced run time of 61 minutes is excellent.

The 1.141 Mean Ratio associated with the EMC XtremIO™ arrays indicates a good efficiency in getting IO to the CPU for service. This result was achieved using a default SAS 64K BUFSIZE to present to storage. Experiments with a 256K BUFSIZE yielded a better Mean CPU/Real Time Ratio, but at an increased aggregate run time of 1474 minutes (53 more minutes). For the IO intensive SAS Application set, this is an excellent outcome. 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.
In short our test results were very pleasing. It showed that the EMC XtremIO™ array, when tuned at install with recommended host-side parameters, and utilizing firmware level 3.0 Build 42, can provide significantly good performance for SAS workloads, in this case with a faster run time than a large, high throughput spinning array. The XtremIO™ array also reduced data effectively through inline de-duplication and compression while achieving this faster run time. 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 and Tuning Recommendations

General Notes

Utilizing the EMC XtremIO™ array can deliver significant performance for an intensive SAS IO workload. Work with your EMC and Host Systems Administrator on install tuning, and host CSTATE tuning (if your IT standards allow CSTATES to be altered) to maximize performance. Processor CSTATES recommendations are typically different for different processor types, and even models. They govern power states for the processor. By not allowing processors to enter a deep “sleep” state, they are more rapidly available to respond quickly to demand loads. Generally Keeping CSTATES at a 1 or 0 level works well, and for those processors supporting turbo-boost mode, it should be set to take advantage of that.

It is very helpful to utilize the SAS tuning guides for your operating system host to optimize server-side performance before XtremIO™ tuning, and additional host tuning is performed as noted below. See: http://support.sas.com/kb/42/197.html

Some general considerations for using flash storage include leaving overhead in the flash devices (although per EMC this is not required for the XtremIO arrays), and considering which workloads if not all, to use flash for.

**Reads vs. Writes.** Flash devices perform much better with Reads than Writes for large-block, sequential IO (SAS). If your scale of workload dictates that you can afford flash for all your file systems that is good. If not, you may wish to bias your flash usage to file systems and data that are read-intensive to get the maximum performance for the dollar. For example, if you have a large repository that gets updated nightly, or weekly, and is queried and extracted from at a high level by users, that may be where you wish to provision your flash storage.

EMC and SAS Tuning Recommendations

The following install tuning was performed on the host server during the installation of the XtremIO™ array. These are EMC and SAS recommended tuning steps.

- The following kernel boot options were used preventing CSTATES from going above 1:
Tuned Storage Profile – Enterprise Storage Profile

Firmware Level Utilized was 3.0 Build 42

The XtremIO X-Bricks consisted of 8 x 8 Gbps connections direct attached to HPDL980 utilizing native LINUX Multipathing.

Eight LUNs were created each for /saswork and /sasdata file systems. The file systems were striped across all LUNs with lvm using "lvmcreate --8", formatted with XFS.

The file systems were mounted with noatime, nodiratime

We created /etc/udev/rules.d/99-EMC-storage.rules containing:

```
#use noop scheduler for high-performance solid-state storage
ACTION=="add|change", SUBSYSTEM=="block", ENV{ID_VENDOR}=="EMC", ATTR{queue/scheduler}="noop"
#Reduce CPU Overhead due to entropy collection
ACTION=="add|change", SUBSYSTEM=="block", ENV{ID_VENDOR}=="EMC", ATTR{queue/add_random}="0"
#Schedule I/O on the core that initiated the process
ACTION=="add|change", SUBSYSTEM=="block", ENV{ID_VENDOR}=="EMC", ATTR{queue/rq_affinity}="2"
#after creating above udev files reboot or run "uddevadm control --reload-rules ; udevadm trigger"
```

We appended following lines to /etc/rc.local:

```
blockdev --setra 16384 /dev/mapper/vg_EMCE_saswork-lv_saswork
blockdev --setra 16384 /dev/mapper/vg_EMCE_sasdata-lv_sasdata
echo never > /sys/kernel/mm/redhat_transparent_hugepage/enabled
```

We added following device section to /etc/multipath.conf:

```
device {
  vendor "EMC"
  user_friendly_names yes
  path_grouping_policy multibus
  path_checker tur
  rr_min_io 1
  path_selector "round-robin 0"
  no_path_retry 0
  fast_io_fail_tmo 3
  dev_loss_tmo 30
}
```

We set fc hba queue depths to 128 in modprobe files:

```
#qlogic /etc/modprobe.d/qla2xxx.conf:
Options qla2xxx ql2xmaxdeth=128
#emulex /etc/modprobe.d/lpfc.conf:
Options lpfc lpfc_devloss_tmo=10
Options lpfc lpfc_hba_queue_depth=128
```

In addition a default SAS BUFSIZE option of 64K was utilized to achieve the best results from this XtremIO array, versus the typical 256KB block size from other storage types. Testing with a lower BUFSIZE value may yield benefits on some IO tests, but may require host and application tuning changes. The general SAS shop accepted default value for overall environment usage (mixed servers, mixed storage) has been 64K.

Conclusion

The EMC XtremIO™ all-flash array can be extremely beneficial for many SAS Workloads. Testing has shown it can significantly eliminate application IO latency, providing improved performance. It is crucial to work with your EMC Storage engineer to plan, install, and tune your host utilizing the XtremIO™ array to get maximum performance. The guidelines listed in this paper are beneficial and recommended. Your individual experience may require additional guidance by EMC depending on your host system, and workload characteristics.
Resources


Contact Information:

Name: Josh Goldstein
Enterprise: EMC
Address: 2841 Mission College Blvd., 4th Floor
City, State: Santa Clara, CA 95054
United States
Work Phone: +1(408-625-7425)
Email: josh.goldstein@emc.com

Name: Tony Brown
Enterprise: SAS Institute Inc.
Address: 15455 N. Dallas Parkway
City, State ZIP: Dallas, TX 75001
United States
Work Phone: +1(214) 977-3916
Fax: +1 (214) 977-3921
E-mail: tony.brown@sas.com

Name: Margaret Crevar
Enterprise: SAS Institute Inc.
Address: 100 SAS Campus Dr
Cary NC 27513-8617
United States
Work Phone: +1 (919) 531-7095
Fax: +1 919 677-4444
E-mail: margaret.crevar@sas.com
To contact your local SAS office, please visit: sas.com/offices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2014, SAS Institute Inc. All rights reserved.