Technical Paper

Performance and Tuning Considerations for SAS® on Red Hat™ using IBM Spectrum Scale™ on Elastic Storage Server GL4®

Release Information Content Version: 1.0 September 2017.

Introduction

This paper presents testing results and tuning guidelines for running SAS® Foundation on Red Hat™ using the IBM® Spectrum Scale™ for IBM Elastic Storage Server (ESS) GL4®. Testing was conducted with the ESS using an x86 four-node host set.

This effort consisted of a “flood test” against four 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.

This paper will outline performance test results conducted by SAS, and general considerations for setup and tuning to maximize SAS Application performance with Spectrum Scale 4.2.1 on an IBM Spectrum Scale and ESS GL4.

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

IBM Spectrum Scale on IBM ESS Performance Testing

Performance testing was conducted with Spectrum Scale 4.2.1 on an IBM Spectrum Scale/ESS system, to establish a relative measure of how well it performs with IO heavy workloads. There were several particular items of interest in this endeavor:

- Relative performance of the IBM Spectrum Scale/ESS GL4
- Performance of the IBM Spectrum Scale clustered file system with SAS Foundation workloads

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 intensive tests to measure concurrent, mixed job performance.

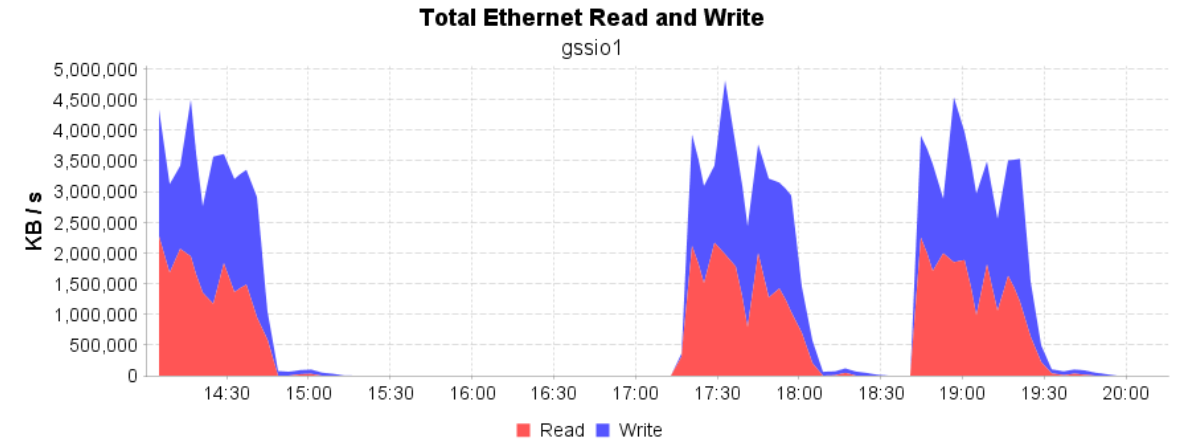
The actual workload chosen was composed of 19 individual SAS tests: 10 computation, two memory, and seven IO intensive tests. Each test was composed of multiple steps, some relying on existing data stores and others (primarily computation tests) relying on generated data. The tests were chosen as a matrix 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 approximate 20-concurrent test matrix, 77 tests were launched per node.

Data and IO Throughput

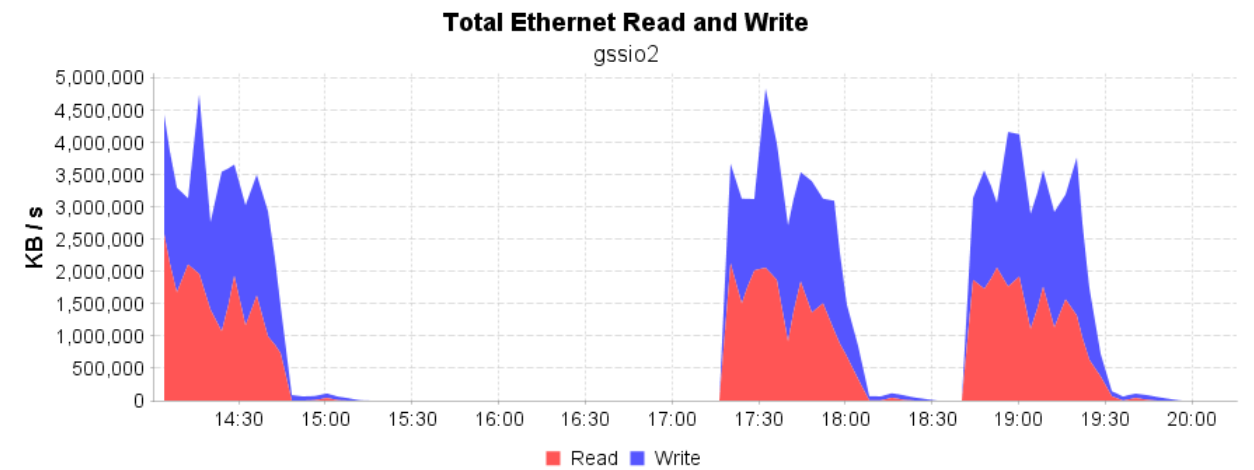
A single instance of the SAS Mixed Analytic 20 simultaneous test workload on each node inputs an aggregate of approximately 300 GB of data for the IO tests and approximately 120 GB of data for the computation tests. Much more data is generated as a result of the test-step activity, and threaded kernel procedures (for example, the SORT PROCEDURE can make copies of portions of the incoming file that are up to three times the size of the original). As stated, some of the same tests run concurrently, or back-to-back, or both, using different data. This results in an approximate average of 20 tests running concurrently and raises the total IO throughput of the workload significantly.

The Cluster IO Bandwidth chart from the IBM Spectrum Scale/ESS fabric infrastructure (Figure 1), shows that the four-

simultaneous-node workload runs for approximately 50 minutes. Several runs are shown in the chart to illustrate consistent results. The workload quickly exceeds 2 GB/sec per ESS fabric controller for initial input READS and 2+ GB/sec for initial SASWORK WRITES. The test suite is highly active for about 50 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 simulates the load of an individual SAS COMPUTE node. The throughput depicted is obtained from all three primary SAS file systems on all four nodes: SASDATA, SASWORK, and UTILLOC.



Series Name	Minimum	Average	Maximum
Read	1.900	551,341.662	4,065,288.700
Write	114.300	636,568.260	3,755,688.600



Series Name	Minimum	Average	Maximum
Read	1.800	558,022.487	4,366,759.100
Write	116.000	637,198.534	3,437,901.400

Figure 1. IBM Spectrum Scale/ESS READ/WRITE Bandwidth (MBps) via the 56 GbE cluster fabric for each of the ESS controllers

SAS File Systems Utilized

There are 3 primary 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, results file system, and working and utility file systems the following space allocations were made for the Spectrum Scale tests:

- SASDATA – 4 TB
- SASWORK – 4 TB
- UTILLOC – 4 TB

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

The system host and storage configuration are specified below:

Test Hosts Configuration

The four host server nodes:

Host: Lenovo x3650 M-5, RHEL 7.2
Kernel: Linux 3.10.0-327.36.3.el7.x86_64
Memory: 256 GB
CPU: Intel® Xeon® CPU E5-2680 v3 @ 2.50GHz
Host tuning: Host tuning was accomplished via a tuned profile script. Tuning aspects included CPU performance, huge page settings, virtual memory management settings, block device settings, etc. The script is attached in Appendix 1.

IBM ESS Test Storage Configuration

IBM ESS Technical Information

ESS Configuration

- Model: 5146-GL4
- Two IBM Power® System S822L as IO servers
- 256GbE (16 x 16 GB DRAM)
- An IBM Power System S821L server for xCat management server
- An IBM 7042-CR8 Rack-mounted Hardware Management Console (HMC)
- Storage interface: Three LSI 9206-16e Quad-port 6 Gbps SAS adapters (A3F2) per IO server
- IO networking: Three 2-port Dual 40GbE Mellanox ConnectX-3 adapter (EC3A) per IO server

- ALB bonding of three Mellanox adapter ports per ESS IO server
- Redundant Array of Independent Disks (RAID) controllers: PCIe IPR SAS Adapter. One IPR adapter per server for RAID 10 OS boot drive per server
- Switches:
    - One 1GbE switch with two VLANs providing two isolated subnets for service and management networks
    - IBM 8831-NF2 – 40GbE switch, Mellanox model SX1710
- Four IBM System Storage® DCS3700 JBOD 60-drive enclosures (1818-80E, 60 drive slots)
    - 58 x 4 of 2 TB 7.2K LN-SAS HDDs + two 400 GB SSDs
- 16 SAS cables

Software

- IBM Spectrum Scale (formerly IBM GPFS™) 4.2.1.1
- IBM ESS version 4.5.1
    - Red Hat 7.1
- MLNX-OS 3.3.6.1002

Network configuration

- IBM Switch Model: 8831-NF2 (Mellanox SX1710)
- Mellanox ConnectX-3 40GbE adapters IBM Feature Code # EC3A
- 36 Ports 40GbE / 56GbE Switch
- MLNX-OS Version 3.6.1002
- Global Pause Flow Control enabled
- TCP/IP only traffic

IBM Storage and Mellanox Network Description

The IBM storage and Mellanox network configuration used in the testing is described in the following two sections.

Storage

IBM Spectrum Scale and ESS combines the processor and IO capability of the IBM POWER8® architecture matched with the IBM System Storage assets. Together, they provide a platform for a multitier storage architecture enhanced with IBM Spectrum Scale to manage block, file, and object data in a shared file system environment. The capabilities of IBM ESS include: Proprietary device pool management, software RAID, large cache, and scalability. IBM ESS systems are delivered as an integrated package with the hardware/software stack validated. The system comes with IBM Spectrum Scale package pre-installed.

The **IBM ESS Model 5146-GL4** used in the test has four 60-drive just a bunch of disks (JBOD) enclosures. Each enclosure has 58 4.2 TB HDDs plus two 400 GB solid-state drives (SSDs) for a total of 240 drives. The predominant storage is near-line spinning disks with a raw capacity of about 246 TB. IBM ESS uses two IBM Power S822L storage servers and one IBM Power S821L management server. IBM Spectrum Scale is the storage cluster management software. Performance is more important for SAS workloads than capacity, thus seven high-speed 40Gb Ethernet ports dedicated to the ESS were connected to the switch.

Network

Typically, Ethernet is not the first choice in storage fabrics. Traditionally, the choice when running an analytics style workload, such as the SAS Mixed Analytics workload, has been Fiber Channel for block IO and possibly InfiniBand for file IO. Some (or possibly many) have experienced difficulties with getting Ethernet working correctly as a storage fabric. But the highly configurable IBM 8831-NF2/Mellanox SX-1710 switch has accomplished this task well. With 36 ports capable of running 40GbE and 56GbE and latencies as low as 220 nanoseconds, it is a perfect complement to the IBM ESS GL4 system with IBM Spectrum Scale.

The IBM Switch Model 8831-NF2 / Mellanox SX1710 is capable of operating as a 40GbE or as a 56GbE switch. In this test, the 20-test mixed analytics workload was run in both 40GbE and 56GbE modes.

The Mellanox Connect X-3 adapters in the configuration allow the fabric to be easily uplifted to a 56GbE modality by simply changing a software-only setting. This is a benefit of a software-defined converged infrastructure. Using 56GbE enabled the 20-test mixed analytics workload with four nodes to approach the disk subsystem throughput limits of IBM ESS GL4.

IBM ESS Linux IO Server Network Tuning

The following operating system network tunable parameters were changed from the default values for the Linux ESS IO Network Shared Disk (NSD) servers.

```
Ethernet MTU was changed to 9000 on the Linux client's interfaces, ESS IO servers, and for
each switch port.
ppc64_cpu --smt=2
ethtool -G enP4p1s0 rx 8192 tx 8192
ethtool -G enP9p1s0 rx 8192 tx 8192
ethtool -G enp1s0 rx 8192 tx 8192
mlnx_tune -r -c
ethtool -K enP9p1s0d1 tx-nocache-copy off
ethtool -K enP4p1s0d1 tx-nocache-copy off
ethtool -K enp1s0d1 tx-nocache-copy off
```

Note: All other ESS node network tunables were already pre-set/tuned as part of the ESS installation process.

IBM Spectrum Scale 4.2.1 Test-Specific Tuning

Multiple tuning parameters were tested during the runs of this test. The following settings were applied to the IBM Spectrum Scale 4.2.1 installation, in varying parameter combinations to determine optimal performance for this test:

Ethernet Fabric=40GbE, 56GbE (note that this 56GbE capability is a native feature of the Mellanox switch and not typical with other Ethernet switch vendors)
Blocksize=1MB, 4 MB, 8 MB, 16MB
prefetchPct=40
maxFilesToCache=50000
maxblocksize=16777216
maxMBpS=10000 (Linux client nodes)
maxMBpS=24000 (ESS nodes)
seqDiscardThreshhold=1073741824
workerThreads=1024 *autotune parameter
Pagepool=32GB, 64 GB, 128 GB (Linux client nodes) – Note the pagepool for the ESS nodes is set automatically to optimal values with ESS installation scripts and was not changed during test runs. Only the Linux SAS client's pagepool values were changed during testing.

IBM Spectrum Scale tuning was determined from previous SAS testing by the IBM ISV enablement team for Spectrum Scale Elastic Storage Server with SAS MA20 workloads at:

https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03541USEN&

You can also see performance results for similar Spectrum Scale tuning parameters within a previous record Spec SFS test can be found at the SpecSFS site:

https://www.spec.org/sfs2014/results/sfs2014.html

*The list above shows the Spectrum Scale cluster tunable settings that were manually changed to achieve the test results. Note that with this version of Spectrum Scale there is a tuning feature called *autotune* that allows us to manually change one parameter, `workerThreads`, which then automatically change several other related tunable settings for us. These automatically changed *autotune* parameters are not listed in this document but also contributed to the performance achieved.

IBM recommends a large page pool space for Spectrum Scale implementations with SAS workloads. The actual pagepool size depends upon the available system memory per client node in the Spectrum Scale cluster as well as the specific SAS workload requirements. For this workload, the test team had ample memory available beyond what the SAS application required and applied a large percentage of that memory per node to the cluster pagepool.

Test Results

The mixed analytic workload was run in a quiet setting (no competing activity on server or storage) for the x86 system utilizing Spectrum Scale 4.2.1 on an IBM ESS GL4 system. It was first run on a single host node, followed by a four-host node run. Multiple runs of each host node set were committed to standardize results. Multiple Spectrum Scale block sizes were tried as per the settings above, in combination with varying SAS BUFSIZE settings, host memory amount, and pagepool space size. The optimal settings for the fastest workload Real Time performance was:

- 56GbE Ethernet fabric
- 8 MB Spectrum Scale block size
- 256KB SAS BUFSIZE
- 128 GB pagepool space
- 256 GB Host RAM

The tuning options noted in the host sections above pertain to Linux operating systems for Red Hat® Enterprise Linux 7.2. Note that because tuning is dependent on the OS and processor choices, you should work with your Red Hat representatives to obtain appropriate tuning parameter values for your system.

Single Node Spectrum Scale 4.2.1 Results

Table 3 shows the performance of the single host node test environment running the SAS Mixed Analytics workload on Red Hat Linux with IBM Spectrum Scale 4.2.1 on an IBM ESS GL4 system. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 77 tests submitted on this single node. It also shows summed User CPU Time, and Summed System CPU Time in Minutes.

x-86 w/IBM Spectrum Scale on IBM ESS GL4	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	0.93	768	668	61

Table 3. Frequency mean values for CPU/Real Time ratio, total workload elapsed time, Memory, and User and System CPU Time performance using IBM Spectrum Scale 4.2.1on IBM ESS GL4, 8MB Spectrum Scale block size, 256K SAS BUFSIZE, 56 GbE fabric, 32 GB pagepool space. Note: All single node testing only utilized a 32 GB pagepool space.

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, the CPU is spending time waiting on resources (usually IO). IBM Spectrum Scale 4.2.1 on the IBM ESS system delivered an excellent 0.93 ratio of Real Time to CPU. The question, "How can I get above a ratio of 1.0?" arises because some SAS PROCEDURES are threaded, and 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 IBM Spectrum Scale 4.2.1 on the IBM ESS GL4, coupled with the fast Intel processors on the Lenovo compute node, executes the aggregate run time of the workload in an average of 768 minutes for the single-node test.

Testing was contained to 32 GB pagepool space sizing in the single-node runs. So, comparisons to large pagepool spaces are not available for single-node tests. Varying Spectrum Scale block sizes and SAS BUFSIZE showed the optimal result at 8 MB and 256 KB respectively, on a 56GbE fabric.

The primary takeaway from this test is that Spectrum Scale 4.2.1 on the IBM ESS GL4 was able to easily provide enough throughput (with extremely consistent low latency) to fully exploit this host environment. Its performance with this accelerated IO demand still maintained a healthy 0.93 CPU/Real Time ratio. This is an excellent performance for a clustered file system.

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

Four-Node Spectrum Scale™ 4.2.1 Results

Table 4 shows the performance of four host node environments simultaneously running the SAS Mixed Analytics workload with IBM Spectrum Scale 4.2.1 on an IBM ESS GL4 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 User CPU time, and summed System CPU time in minutes.

x86 w/IBM Spectrum Scale on IBM ESS GL4	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	0.89	800	659	56
Node2	0.97	809	759	54
Node3	0.97	764	684	52
Node4	0.96	781	699	54

Table 4. Frequency Mean Values for CPU/Real Time ratio, total workload elapsed time, Memory, and User & System CPU time performance using IBM Spectrum Scale 4.2.1 on IBM ESSGL4, with 8MB Spectrum Scale Block size, 256KB SAS BUFSIZE, 56 GbE fabric, 128 GB pagepool space

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). IBM Spectrum Scale 4.2.1 on the IBM ESS GL4 system delivered an excellent 0.89 to 0.97 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 IBM Spectrum Scale 4.2.1 on the IBM ESS GL4 system coupled with the fast Intel processors on the Lenovo compute node executes the aggregate run time of the workload in an average of 788 minutes per node, and 3,154 minutes of aggregate execution time for all four nodes.

Varying Spectrum Scale block sizes and SAS BUFSIZE showed the optimal result at 8 MB and 256 KB respectively, on a 56 GbE fabric, with a 128 GB pagepool space.

The primary takeaway from this test is that Spectrum Scale 4.2.1 on the IBM ESS GL4 system was able to easily provide enough throughput (with extremely consistent low latency) to fully exploit this host environment. Its performance with this accelerated IO demand still maintained a healthy 1.03 or better CPU/Real Time ratio. This is excellent performance for a clustered file system.

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

IBM ESS GL4 Array-Side Performance Indicators

As previously mentioned, the underlying IBM ESS storage is spinning JBOD with Spectrum Scale software providing the management and clustering capabilities. The IBM ESS capabilities that were used during testing include: Proprietary

device pool management, software RAID, large cache, and scalability through the Spectrum Scale cluster technology. No other advanced Spectrum Scale functions were used.

General Considerations

Utilizing the Spectrum Scale 4.2.1 clustered file system on IBM ESS GL4 can deliver excellent 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. Additional host tuning is performed as noted below.

IBM, 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 Red Hat systems, refer:
http://support.sas.com/resources/papers/proceedings11/342794_OptimizingSASonRHEL6and7.pdf

In addition, a SAS BUFSIZE option of 256 KB coupled with a Spectrum Scale file system block size of 8 MB was utilized to achieve the best combined performance of the parallel/clustered file system and Elastic Storage Server.

Spectrum Scale 4.2.1 Tuning

The Spectrum Scale tuning parameters used in this SAS MA20 test environment are generally recommended for the IBM ESS and other IBM FlashSystem products with SAS customer environments. In general, good performance results have also been achieved for other IBM FlashSystem products as well as non-flash storage products such as the Elastic Storage Server™/spinning disk Ethernet network attached storage with these same general Spectrum Scale tunable settings used in this test. Examples of this range from workloads such as the SAS MA20 workload, SAS MA30 workload, and non-SAS but similar types of workloads. Spectrum Scale is a mature and scalable clustering product that has been tested and proven with SAS workloads to have specific advantages with the use of cluster pagepool that rivals competing products.

IBM ESS GL4 Monitoring

For more advanced performance data gathering, reporting, and alerting the customer can purchase and install Spectrum Control™ software package.

Conclusion

IBM Spectrum Scale 4.2.1 on the IBM ESS GL4 system 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 excellent. The Spectrum Scale clustered file system performs admirably across a 40 GbE and a 56 GbE fabric for 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 Red Hat engineers to plan, install, and tune the hosts for the environment, as well as with IBM engineering guidance for IBM Spectrum Scale and IBM ESS. For additional information about IBM Spectrum Scale and IBM Elastic Scale Storage, contact your local IBM sales team or an IBM

Business Partner. For general questions, you may also contact 1-800-IBM-4YOU (1-800-426-4968) or **E-mail:** askibm@vnet.ibm.com, www.ibm.com/us-en/

The guidelines listed in this paper are both beneficial and recommended. Your individual experience may require additional guidance by IBM 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

IBM Papers on Spectrum Scale and IBM ESS:

IBM ESS on IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter search Spectrum Scale RAID

Introduction Guide to the IBM Elastic Storage Server: http://www.redbooks.ibm.com/redpapers/pdfs/redp5253.pdf

IBM Hyper-Scale in XIV Storage, REDP-5053:
http://www.redbooks.ibm.com/abstracts/redp5053.html

IBM Spectrum Scale: Big Data and Analytics Solution
http://www.redbooks.ibm.com/redpapers/pdfs/redp5397.pdf

Implementing IBM Spectrum Scale
http://www.redbooks.ibm.com/redpapers/pdfs/redp5254.pdf

A Deployment Guide for IBM Spectrum Scale Object
http://www.redbooks.ibm.com/redpapers/pdfs/redp5113.pdf

Contact Information:

Brian Porter
IBM
2889 W Ashton Boulevard
Lehi, UT 84043
Work Phone: +1(720) 430-7674
E-mail: bporter1@us.ibm.com

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

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

Appendix I

Tuned profile used for this testing follows.

```
# create /usr/lib/tuned/sas-performance/tuned.conf containing:
[cpu]
force_latency=1
governor=performance
energy_perf_bias=performance
min_perf_pct=100
[vm]
transparent_huge_pages=never
[sysctl]
kernel.sched_min_granularity_ns = 10000000
kernel.sched_wakeup_granularity_ns = 15000000
vm.dirty_ratio = 40
vm.dirty_background_ratio = 10
vm.swappiness=10


# select the sas-performance profile by running
tuned-adm profile sas-performance
```