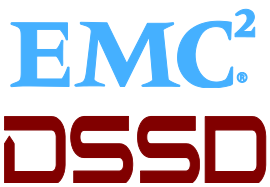
Technical Paper

Performance and Tuning Considerations for SAS® on the Intel® Xeon® E5 v3 Series Processors and EMC® DSSD™ D5™ Rack-Scale Flash Appliance



Release Information

Content Version: 1.0 April 2016.

Contents

Introduction

This paper represents test results for Foundation SAS® Workloads on the Intel® Xeon® E5 v3 Series Processors and EMC DSSD D5™ Rack-Scale Flash appliance. This paper presents test results on new X-86 Host servers. As such, results will not be regressed against non-alike, previous test hosts.

This effort also includes a scaled test bed of one, four, and eight simultaneous X-86 Nodes running a SAS Mixed Analytics workload, to determine scalability against the appliance, 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 Intel® Xeon® E5 v3 Series Processors and EMC DSSD D5™ Rack-Scale Flash Appliance.

An overview of the flash testing is 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 is followed by a general conclusions and a list of practical recommendations for implementation with Foundation SAS®.

Intel® Xeon® E5 v3 Series Processors and EMC DSSD D5™ Performance Testing

Performance testing was conducted with Intel® Xeon® E5 v3 Series Processors and EMC DSSD D5™ Rack-Scale Flash Appliance, to attain a relative measure of how well it performed with IO heavy workloads. Of particular interest was whether the DSSD D5 appliance would yield substantial benefits for SAS large-block, sequential IO patterns against the very fast Intel® Xeon® E5 v3 Series Processors. 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 test set on each 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 a single node, then four nodes simultaneously, and finally 8 nodes simultaneously (each node running the 20 simultaneous workload). The 20 SAS Mixed Analytic 20 Simultaneous Test workload was utilized to better match the host CPU and Memory resources (see Hardware Description below) of the X-86 nodes. The test sets resulted in the following numbers of overall simultaneous tests launched per test bed against the appliance:

- 1 Node – 77 Tests
- 4 Node – 308 Tests

- 8 Node – 616 Tests

Data and IO Throughput

The IO tests input an aggregate of approximately 300 Gigabytes of data per test set, and the computation tests over 120 Gigabytes of data per test set. Much more data was generated from test-step activity, and threaded kernel procedures such as SORT (e.g. SORT can make the equivalent of three 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 per set. This raises the total IO throughput of the workload significantly.

As can be seen in the following graph (Table 1 below), in its 40 minute span, the 8 simultaneous node workload quickly jumps to greater than 20 GB/sec on the initial IO throughput. The test suite is highly active for about 30 minutes and then finishes with "trail out jobs". This is a typical "SAS Shop" throughput characteristic for a single-node instance, and simulates the general load of an individual SAS COMPUTE node. This throughput was attained from all three primary SAS file systems being stored on the D5: SASDATA, SASWORK, and UTILLOC. Each node has its own dedicated SASDATA, SASWORK and UTILLOC file systems.

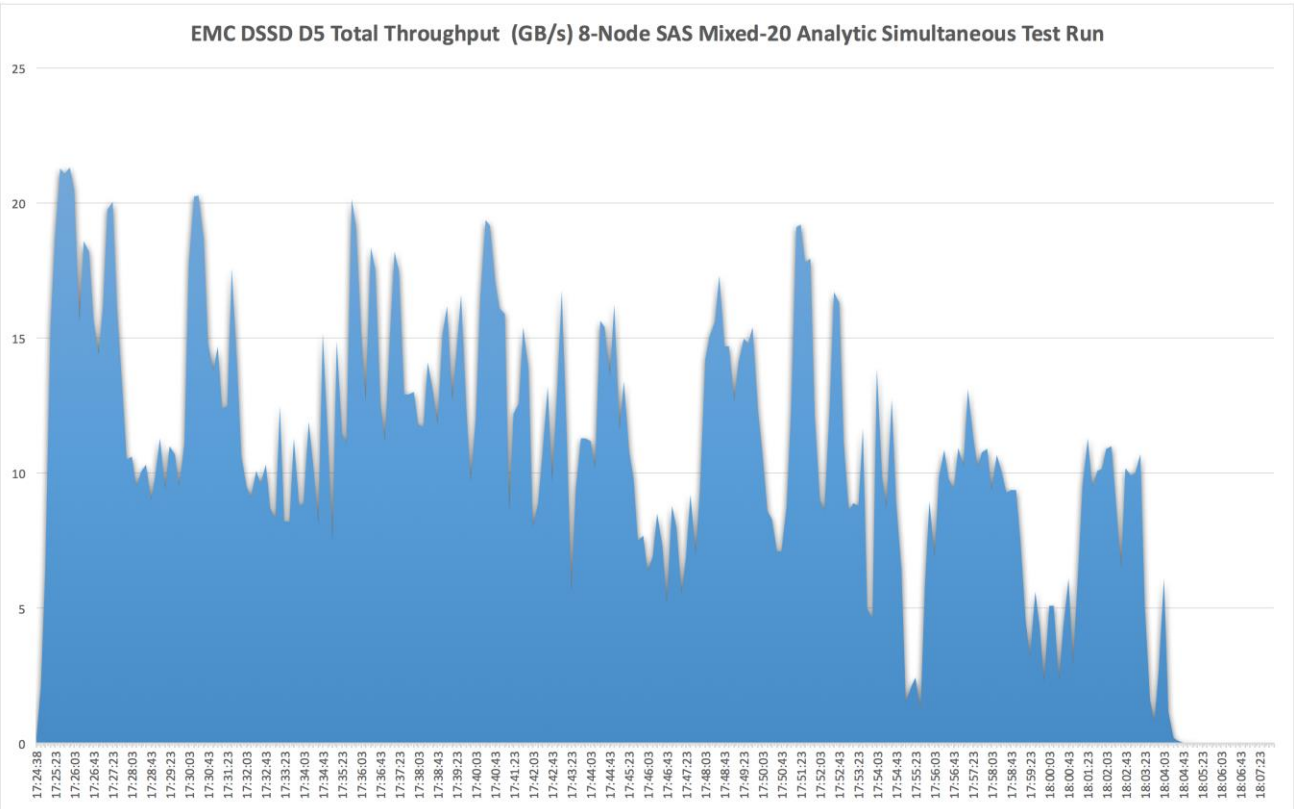


Table 1. EMC DSSD D5™ Rack-Scale Flash Appliance Test Throughput (Gigabytes/sec) for 8-Nodes, SAS Mixed Analytic 20 Simultaneous Test Run

SAS File Systems Utilized

There are three 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 the SAS Mixed Analytic 20 Simultaneous workload's code set, data, result space, the working and utility space allocations were as follows:

- SASDATA – 1 Terabyte
- SASWORK – 1 Terabyte
- UTILLOC – 1 Terabyte

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 one, four, and eight host nodes utilizing the SAS Mixed Analytics 20 Simultaneous Test workload. All nodes were identical in configuration. The system host and storage configuration are specified below:

Intel® Xeon® E5 v3 Series Processors Test Host Configuration

The host server nodes utilized consisted of:

Host: Dell® R730, RHEL 7.1, X86_64
Kernel: Linux 3.10.0-229.el7.x86_64
Memory: 384 GB
CPU: Genuine Intel® Xeon® CPU E5-2699 v3, Family 6 Model 63, Stepping 2, 2 Socket, 18 cores/socket, x86_64, 2.3GHz
HBA: 2 X DSSD PCIe Gen3 Client Cards (to connect the hosts to the DSSD D5)
Host Tuning:

- The bios power management settings on the hosts were set to "performance"
- /etc/modprobe/dssd.conf: options vpci vpci_sgl_enable=2 vpci_nvme_thread_policy=2

Additional Settings Employed:

- None

DSSD D5 Test Storage Configuration

DSSD D5**:**

- 5U in total, with no external switches
- 1 X DSSD D5 Rack-Scale Flash appliance. Comprised of:
  - 36 Flash Modules, each 4TBs
  - Two redundant Control Modules
  - Two redundant I/O modules, each with 48 x PCIe gen3 x 4 lane ports
  - Four redundant power supply units (PSUs)
  - Five redundant fan modules

    - Two 1Gb/s Ethernet management ports
    - 32 DSSD PCIe cables connecting the hosts to the D5 I/O Modules – each host has 4 cables from its client cards to the D5
- Utilizing Flood (DSSD D5 Operating System): 1.0
- 100 GB/sec potential throughput from the hosts to the storage
- File System Type: XFS
    - mkfs.xfs -f -l size=2037m,lazy-count=1 -i size=2048,align=1,attr=2,projid32bit=0
- Each host had three file systems, each based on a 1TB LUN from the D5: SASDATA, SASWORK, and UTILLOC. No LVM was used
- The SAS application used a 256K SAS 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 Intel® Xeon® E5 v3 Series Processors System utilizing EMC DSSD D5 on a single host node. Multiple runs were committed to standardize results.

Table 2 below shows the performance of the EMC DSSD D5 appliance. 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 - Intel® Xeon® E5 v3 Series Processors /EMC DSSD D5™	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	Memory Used in MB - Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	1.09	548	40124	757	75

Table 2. Frequency Mean Value for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using EMC DSSD D5™ Rack-Scale Flash appliance

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 DSSD D5 system delivered an excellent 1.09 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 DSSD D5™ Rack-Scale Flash Appliance coupled with the faster Intel processors on the Dell compute node executes the aggregate run time of the workload in approximately 548 minutes of total execution time.

It is important to note that while we aren't making direct comparisons of this Intel® Xeon® E5 v3 Series Processors Node test to previous host testing. This newer, faster Intel processor set executed the workload in less than half the time as the previous older generation host! The primary take-away from this test is that the EMC DSSD D5 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 1.09 CPU/Real Time ratio!

Scaling from One Node to Four Nodes on the EMC DSSD D5

For a fuller "flood test", the Mixed Analytic 20S Workload was run concurrently on four physically separate, but identical host nodes in a quiet setting (no competing activity on server or storage) for the Intel® Xeon® E5 v3 Series Processors System utilizing the EMC DSSD D5. Multiple runs were committed to standardize results.

Table 3 below shows the performance of the four host node environments attached to the EMC DSSD D5. 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 - Intel® Xeon® E5 v3 Series Processors w/EMC DSSD D5 ™	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	Memory Used in MB - Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	1.09	545	40124	568	75
Node2	1.10	535	40124	570	73
Node3	1.06	572	40124	579	76
Node4	1.09	582	40124	588	80

Table3. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using EMC DSSD D5™ Rack-Scale Flash appliance

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 DSSD D5 delivered an Excellent 1.09 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 DSSD D5 coupled with the faster Intel processors on the four compute node test bed executes the aggregate run time of the workload in an average of 558 minutes per node, and 2,234 minutes of aggregate execution time for all 4 nodes.

Again, the newer, faster Intel processor set executed the workload on each node, in less than half the time as the previous tested, older generation host. In addition to halving the execution time, the scale of four simultaneous test workloads was generated versus the prior testing single workload. Again, the EMC DSSD D5 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.

Scaling from One Node to Eight Nodes on the EMC DSSD D5

The Mixed Analytic 20S Workload was run concurrently on eight physically separate, but identical host nodes in a quiet setting (no competing activity on server or storage) for the Intel® Xeon® E5 v3 Series Processors System utilizing the EMC DSSD D5. Multiple runs were committed to standardize results.

Table 4 below shows the performance of the eight host node environments attached to the EMC DSSD D5. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 77 tests submitted per node (616 in total). It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - Intel® Xeon® E5 v3 Series Processors w/EMC DSSD D5 ™	Mean Value of CPU/Real-time - Ratio	Elapsed Real Time in Minutes – Workload Aggregate	Memory Used in MB - Workload Aggregate	User CPU Time in Minutes – Workload Aggregate	System CPU Time in Minutes - Workload Aggregate
Node1	1.02	611	40123	578	77
Node2	1.03	609	40123	576	75
Node3	1.02	606	40123	570	74
Node4	1.02	609	40123	571	74
Node5	1.02	610	40123	571	76
Node6	1.02	605	40123	567	74
Node7	1.03	599	40123	570	73
Node8	1.03	600	40123	573	72

Table4. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using EMC DSSD D5™ Rack-Scale Flash Appliance

The second column in Table 4 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 DSSD D5 delivered a very good 1.024 average 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 DSSD D5 coupled with the faster Intel processors on the eight compute node test bed executes the aggregate run time of the workload in an average of 606 minutes per node, and 4,849 minutes of aggregate execution time for all eight nodes.

In addition to halving the execution time, the scale of eight simultaneous test workloads attained excellent CPU/Real Time ratio with only a 9.2% increase in aggregate run time over the 4 node test. These are very dramatic results, with very low workload latency increase, despite doubling the workload. The low linearity of the scale results shows that the storage has a high level of available performance that can sustain large numbers of I/O requests from many systems simultaneously. That allows many hosts to run at nearly full I/O speed all while sharing the same storage.

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 DSSD D5™ Rack-Scale Flash appliance can deliver significant performance for an intensive SAS IO workload. The operating system specific tuning guide from SAS was not used, as DSSD D5 does not use standard data paths. No operating system tuning was performed.

DSSD D5 is the first of a new category of flash storage – Rack-Scale Flash. DSSD D5 is a completely new architecture designed for the most data-intensive applications, both traditional and next-generation, which require extreme levels of performance and the lowest possible latency.

DSSD D5 delivers ultra-dense, high-performance, highly available and very low latency shared flash storage for up to 48 redundant connected servers. D5 is connected to each node through PCIe Gen3 and leverages NVMe technology, to deliver the performance of PCI-attached flash. D5 is a standalone appliance that is dis-aggregated from compute, delivering the benefits of shared storage. The result is next-generation performance with latency as low as 100 microseconds, throughput as high as 100 GB/s, and IOPS of up to 10 million in a 5U system.

The appliance provides up to 36 flash modules with 144TB RAW (100TB usable) capacity. D5 is also has enterprise-class availability and serviceability features. These features include dual-ported client cards, dual active-active controllers, redundant components and industry-leading flash reliability and resiliency with Cubic RAID™, dynamic wear leveling, flash physics control and space-time garbage collection.

Data Reduction

D5 does not provide compression or deduplication, and therefore there were no data reduction services in use to decrease the I/O load to the storage.

Customers could use SAS compression to improve their I/O performance by reducing I/O requests to the storage.

EMC and SAS Tuning Recommendations

Host Tuning

A SAS BUFSIZE option of 256K was used for testing. Aside from this application side tuning, no other host side os tuning was required.

DSSD D5 Storage Tuning

The DSSD D5 comes pre-configured with Cubic RAID™ enabled, requiring no additional RAID setup or configuration. A single volume was created on the D5.

The DSSD software was installed on the hosts, and the hosts configured to access the volume (via the /etc/sysconfig/dssd-blkdev file).

Three x 1 TB objects x 8 hosts were created in the volume, which appeared as devices in /dev. Then mkfs.xfs and mount commands were executed to make the file systems available to the SAS application.

The objects were created with 4K "fragment" sizes. Fragment sizes on D5 objects can vary from 64B to 32KB when using the DSSD Flood API to write programs that directly use D5 objects. Unmodified applications can also make use of D5 by using the DSSD block device service. Because Linux only supports 512B and 4K storage device block sizes, the objects were created with 4K fragment size and the block service created block device entries in /dev to make these objects available to the XFS file system and then to SAS. Standard operations involve a large I/O request from SAS being split into 4K requests and sent to the block device. D5's block service uses all paths between the client and the D5 automatically, multi-pathing across the links. No logical volume service was needed to, for example, take multiple devices and stripe data across them because D5's Cubic RAID™ automatically stripes data across all of the flash modules (up to 36) in the EMC DSSD D5.

DSSD D5 Monitoring

DSSD D5 provides the Command Line Interface (CLI) and Browser User Interface (BUI) management and monitoring tools.

The CLI during the 8-node test for example showed read, write, and erase IOPs and read, write, and erase throughput. Because SAS was using a file system created 4K fragment size objects on the D5, these were 4K IOPs.

```
> monitor-pool
TIME                 R-IOPS W-IOPS E-IOPS   R-BW   W-BW   E-BW
2016-03-25 19:08:26  99.2K 182.9K   1.5K 406.2M   6.0G   6.4G
2016-03-25 19:08:36   1.1M 335.7K   2.3K   4.4G  11.0G   9.6G
2016-03-25 19:08:46 485.9K 508.6K   4.1K   2.0G  16.7G  17.2G
2016-03-25 19:08:56  23.1K 648.2K   5.0K  94.5M  21.2G  20.9G
2016-03-25 19:09:06  24.6K 640.4K   4.9K 100.6M  21.0G  20.7G
2016-03-25 19:09:16  26.7K 646.2K   5.1K 109.2M  21.2G  21.4G
2016-03-25 19:09:26 125.8K 609.6K   4.7K 515.2M  20.0G  19.7G
2016-03-25 19:09:36   1.7M 266.5K   2.3K   6.9G   8.7G   9.5G
2016-03-25 19:09:46 681.2K 483.5K   3.7K   2.8G  15.8G  15.4G
2016-03-25 19:09:56 720.2K 465.3K   3.7K   3.0G  15.2G  15.4G
2016-03-25 19:10:06   1.3M 312.0K   2.4K   5.5G  10.2G  10.2G
2016-03-25 19:10:16   1.3M 274.3K   1.8K   5.4G   9.0G   7.6G
2016-03-25 19:10:26 545.3K 428.2K   3.6K   2.2G  14.0G  15.2G
2016-03-25 19:10:36 115.2K 588.4K   4.6K 471.8M  19.3G  19.1G
2016-03-25 19:10:46  63.7K 604.4K   4.9K 260.8M  19.8G  20.5G
2016-03-25 19:10:56   1.2M 345.2K   2.7K   4.8G  11.3G  11.4G
2016-03-25 19:11:06   2.0M 152.6K   1.1K   8.1G   5.0G   4.4G
```

Conclusion

The Intel® Xeon® E5 v3 Series Processors and EMC DSSD D5™ Rack-Scale Flash Appliance has been proven to be

extremely beneficial for scaled SAS workloads. In summary, the faster CPU processing allows the compute layer to perform more operations per second, thus increasing the potential performance for the solution. The consistently low response times and very high throughput of the EMC DSSD D5 allow this scaled workload potential to be fully exploited.

The EMC DSSD D5™ Rack-Scale Flash Appliance is designed to be as operationally 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:

Name: Matt McDonough
Enterprise: EMC DSSD
Address: 4025 Bohannon Drive
City, State: Menlo Park CA 94025 United States
Work Phone: (408) 671-0053
Email: mattm@dssd.com

Name: Shawna Meyer-ravelli
Enterprise: Intel Corporation
Address: 2111 N.E. 25th Avenue,
City, State: Hillsboro, OR 97124
United States
Work Phone: +1 (503) 712 5520
Email: Contact_storagebuilders@intel.com

Name: Tony Brown
Enterprise: SAS Institute Inc.
Address: 15455 N. Dallas Parkway
City, State ZIP: Dallas, TX 75001
United States
Work Phone: +1(469) 801-4755
Fax: +1 (919) 677-4444
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