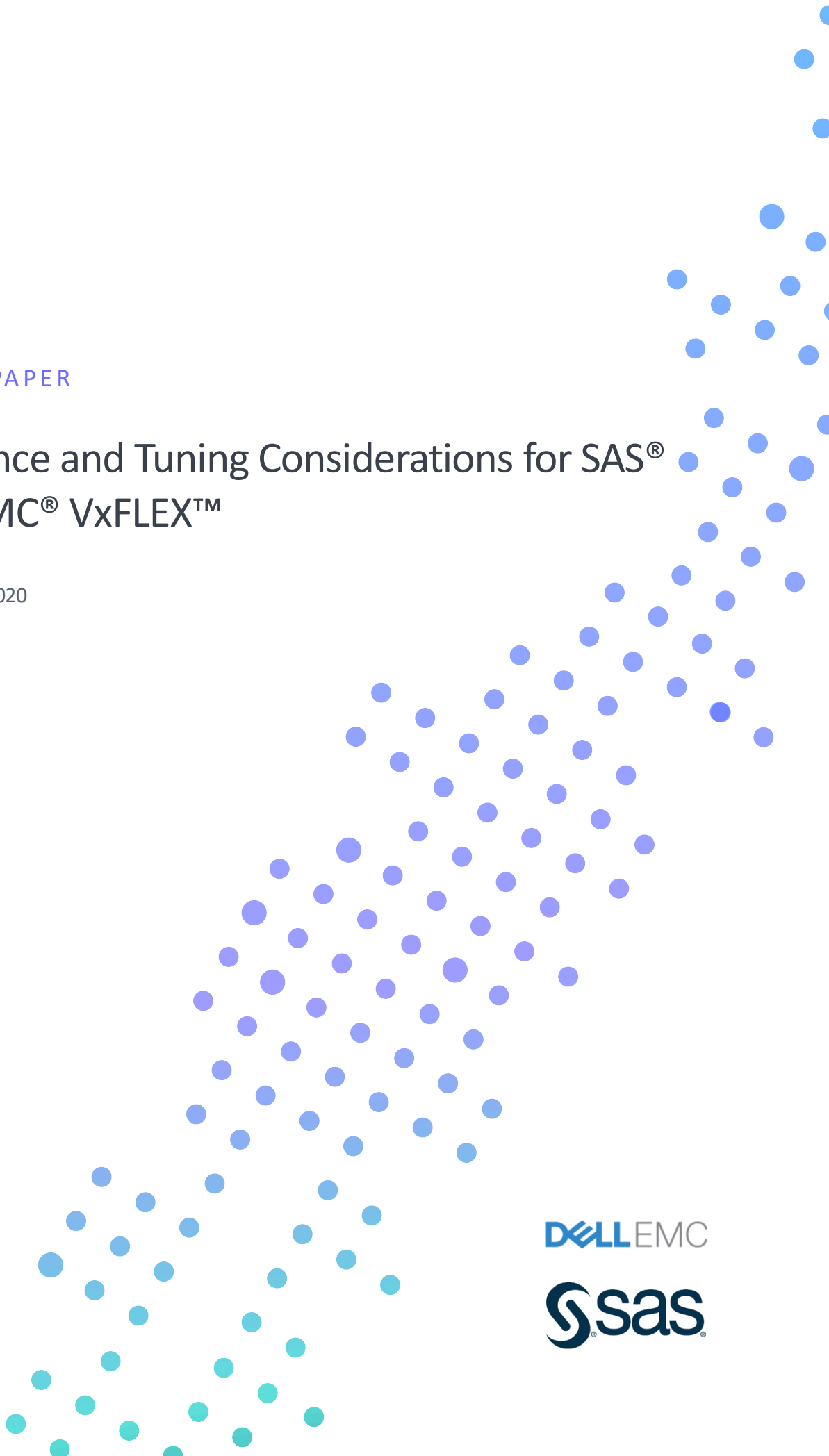
TECHNICAL PAPER

Performance and Tuning Considerations for SAS® on Dell EMC® VxFLEX™

Last update: May 2020



DELL EMC

sas

Contents

Introduction

This technical paper is about a test of the SAS® mixed analytics workload on the Dell EMC® VxFLEX™ hyper-converged infrastructure (HCI).

This effort involved a "flood test" of one, two, four, and six simultaneous Dell R640 host nodes running a SAS mixed analytics workload to determine scalability against appliance as well as uniformity of performance per node. This paper outlines performance test results performed by SAS and general considerations for configuring and tuning the Dell EMC VxFLEX HCI system for SAS application performance. The Dell EMC VxFLEX HCI is a configured appliance, unlike the VxFLEX Ready Nodes, which is building-block configurable only for various scales of operation.

An overview of the flash testing is discussed first, including the purpose of the testing. Next, detailed descriptions of the actual test bed and workload are provided along with a description of the test hardware. Test results follow, with a list of tuning recommendations. Finally, general considerations, recommendations for implementation with SAS Foundation, and conclusions are discussed.

Dell EMC VxFLEX Performance Testing

Performance testing was conducted with six Dell R640 physical host nodes running VMware ESXi 6.5. Each of these ran a virtual machine with a Red Hat Enterprise Linux (RHEL) 7.5 operating system. For the full system description, see Hardware Description. The purpose of the testing was to determine whether the VxFLEX could scale SAS large-block, sequential I/O patterns in a heavy workload with a good level of performance. In this section of the paper, we 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 SAS mixed analytics workload. This was a scaled workload of computation and I/O-oriented tests to measure concurrent, mixed job performance.

The actual workload chosen consisted of 19 individual SAS tests: 10 computational, 2 memory, and 7 I/O intensive tests. Each test consisted 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 cases, the same test (running against replicated data streams) was run concurrently, and back-to-back (or back-to-back) in a serial fashion, to achieve an average of *20 simultaneous streams of heavy I/O, computation (fed by significant I/O in many cases), and memory stress. In all, to achieve the 20-concurrent test matrix, 77 tests were launched.

An aggregate of approximately 300 Gigabytes of data and over 120 Gigabytes of data (from computation tests) were submitted to a single instance of the SAS mixed analytic workload with 20 simultaneous tests on each node. Much more data is generated from test-step activity and threaded kernel procedures such as PROC SORT (for example, PROC SORT makes three copies of the incoming file to be sorted). As stated, some of the same tests were run concurrently using different data, and some of the same tests were run back-to-back, to implement a total average of 20 tests running concurrently. This raised the total concurrent I/O throughput of the workload significantly.

SAS File Systems Used

Two primary file systems (using XFS) were used in the flash testing:

- SAS permanent data file space – SASDATA
- SAS working file space and utility data file space – SASWORK and UTILLOC

Each physical ESXi host had a single approximately 6-terabyte (TB) VMDK (Virtual Machine Disk) logical unit (LUN) that was divided into two directories: a directory for SASDATA and a directory for SASWORK and UTILLOC. NVMe storage can manage SAS block-I/O at a good performance level without traditional volume/LUN arrangements that have been used on previous storage types. The total usable storage capacity was approximately 35 TBs.

No traditional host volume arrangement such as striped logical volumes was used. The VxFLEX operating system and software-defined storage (SDS) managed the devices under their own authority for file system block management, data replication, high availability, and so on. This is not configurable by the end user. There are no compression or de-duplication operations available for the VxFLEX stored blocks. 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 to enhance SAS performance.

Hardware Description

This test bed was run against six Dell R640 host nodes using the SAS mixed analytics workload with 20 simultaneous tests.

VxFLEX Configuration

Physical Configuration

The VxFLEX infrastructure test consisted of the following:

- Machines: 6 physical ESXi hosts
- Host: Dell R640 Server
- OS: VMware ESXi 6.5
- CPU: 2 x Intel Skylake 8168 24 cores, 2.7 gigahertz (Ghz), Hyper-threading-enabled Intel Xeon Platinum 8168 Processor
- Memory: 384 GBs
- Storage: 8 x 1.5 TB NVMe Flash devices
- Fabric: 2 x 25-gigabit Ethernet (GbE)

Host OS/Virtual Configuration

Each ESXi host was divided into two VMware virtual machines:

- VM1 – 12 logical cores were needed to manage only I/O for that ESXi physical host.

- VM2 – 6 total SAS VM hosts were required for computation: VM01, VM02, VM03, VM04, VM05, and VM06.

One VM was required per physical ESXi host.

- OS: Red Hat Enterprise Linux (RHEL) 7.5
- OS Release 3.10.0-862.el7.x86_64
- CPU: 36 cores (72 logical)
- Memory: 360 GB
- SAS Version: SAS 9.4M5

Test Results

This technical paper is about a test of the SAS® mixed analytics workload on the Dell EMC® VxFLEX™ hyper-converged infrastructure (HCI).

This effort involved a "flood test" of one, two, four, and six simultaneous Dell R640 host nodes running a SAS mixed analytics workload to determine scalability against appliance as well as uniformity of performance per node.

Single Host Node Test Result

The SAS mixed analytics workload (with 20 simultaneous tests) was run in a quiet setting (no competing activity on the system) using a single node of the Dell EMC VxFLEX system. Multiple runs were executed to standardize the results.

The tuning options that were specified in VxFLEX Configuration were used. However, you must consult with your Dell EMC representative for appropriate tuning mechanisms that are applicable to your workload and system.

Table 1 shows the performance of the Dell EMC VxFLEX system.

Dell EMC VxFLEX Node	Real Time	Mean Value of CPU/Real Time Ratio	User CPU Time in Minutes — Workload Aggregate	System CPU Time in Minutes — Workload Aggregate	Average Node Real Time Minutes
Node 1	468	1.07	471	64	468

Table 1. Total Workload Elapsed Real Time, Frequency Mean Value for CPU/Real Time Ratio, and User and System CPU Time Performance Using Dell EMC VxFLEX

The third column in Table 1 shows the Mean Value of all the ratios (counting each Job Step) of total CPU time (User + System CPU) against the total Real Time (elapsed time). If the ratio is less than 1, then the CPU is spending time waiting for resources, which is usually an I/O activity. The VxFLEX system delivered an excellent 1.07 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 use more CPU cycles than wall clock time, or Real Time.

The second column shows the total elapsed run time in minutes, summed together from each of the jobs in the workload. The Dell EMC VxFLEX system, using Intel Skylake processors, executed the mixed analytics 20-simultaneous-test workload in approximately 468 minutes of aggregate job execution time.

The primary take-away from this test is that the Dell EMC VxFLEX system was able to easily provide enough throughput (with extremely consistent low latency) to fully exploit this host improvement! Its performance with this accelerated I/O demand still maintained a very healthy Mean Value of 1.07 CPU/Real Time ratio!

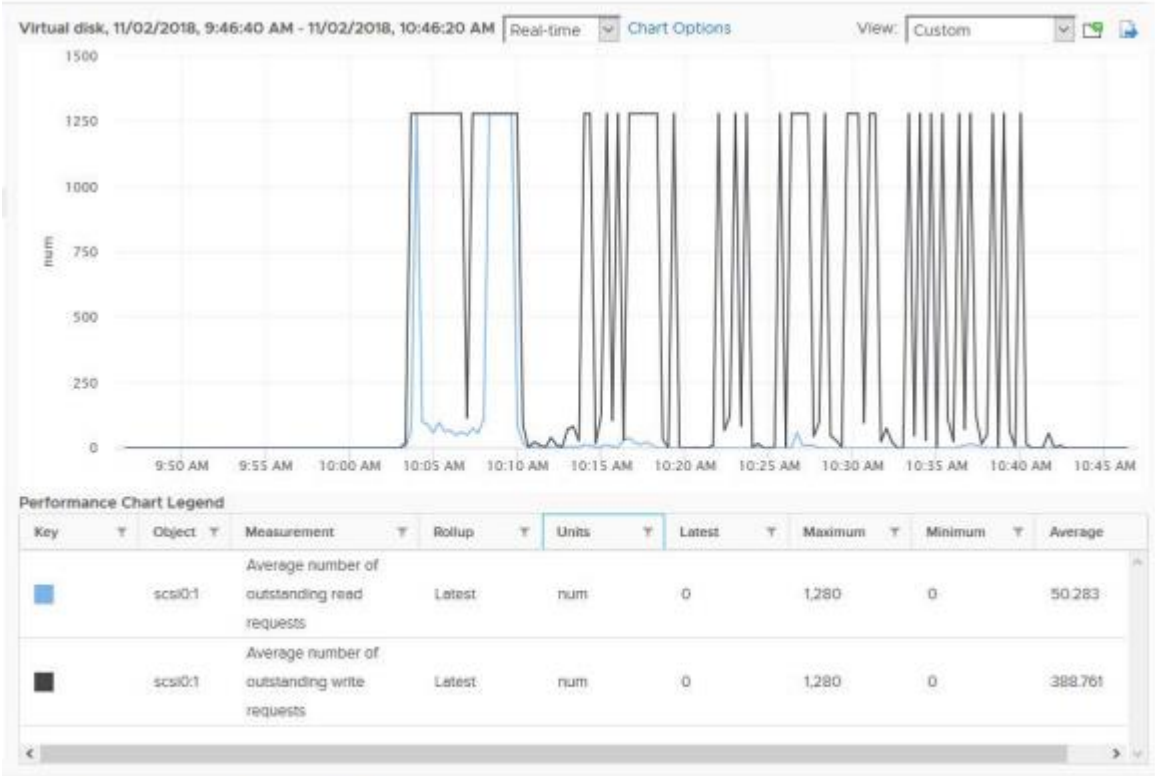
Performance Graphs

The following graphs (Graphs 1–4) show ESXi performance from a single node run and with reports on virtual disk latency, bandwidth, and transaction request rates. Write latency (from the scale-out storage devices) ranges somewhat on the high side from 12 to 18 milliseconds (ms) per transaction. This is not totally unusual for scale-out, replicated flash storage. The trade-off is acceptably commensurate with the 4.7 GBs of I/O bandwidth achieved. It takes 24 cores to run the mixed analytic 20 workload, yielding an average rate of 210 MBs per second, per core. This significantly exceeds the 150 MBs per second, per core recommendation for SAS 9.4 file systems.

Graph 1 – Read and Write Latency

Key	Object	Measure...	Rollup	Units	Latest	Maximum	Minimum	Average
	scsi0:1	Write rate	Average	KBps	0	3,260,593	0	562,274.7
	scsi0:1	Read rate	Average	KBps	0	1,702,045	0	176,583.86

Graph 2 – Read and Write Bandwidth



Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
	scsi0:1	Average number of outstanding read requests	Latest	num	0	1,280	0	50.283
	scsi0:1	Average number of outstanding write requests	Latest	num	0	1,280	0	388.761

Graph 3 – Outstanding Read and Write Requests

Graph 4 – Read and Write Requests per Second

Scaling from One Node to Two Nodes on the Dell EMC VxFLEX System

The SAS mixed analytics workload (with 20 simultaneous tests) was run in a quiet setting (no competing activity on the system) using two nodes of the Dell EMC VxFLEX system. Multiple runs were executed to standardize the results.

Table 2 shows the performance of the Dell EMC VxFLEX two-node test.

Dell EMC VxFLEX Node	Real Time	Mean Value of CPU/Real Time Ratio	User CPU Time in Minutes — Workload Aggregate	System CPU Time in Minutes — Workload Aggregate	Average Node Real Time Minutes
Node 1	477	1.07	475	65	475
Node 2	472	1.07	472	65	

Table 2. Total Workload Elapsed Time, Frequency Mean Value for CPU/Real Time Ratio, and User and System CPU Time Performance Using Dell EMC VxFLEX

During the two-node test, the Dell EMC VxFLEX system easily provided enough throughput (with extremely consistent low latency) to fully exploit this host aggregation. Its performance with this accelerated I/O demand still maintained a 1.07 ratio.

Scaling to Four Nodes on the Dell EMC VxFLEX System

The SAS mixed analytics workload (with 20 simultaneous tests) was run in a quiet setting (no competing activity on the system) using four nodes of the Dell EMC VxFLEX system. Multiple runs were executed to standardize the results.

Table 3 shows the performance of the Dell EMC VxFLEX four-node test.

Dell EMC VxFLEX Node	Real Time	Mean Value of CPU/Real Time Ratio	User CPU Time in Minutes — Workload Aggregate	System CPU Time in Minutes — Workload Aggregate	Average Node Real Time Minutes
Node 1	465	1.10	477	65	482
Node 2	489	1.06	483	66	
Node 3	486	1.06	477	64	
Node 4	487	1.06	482	65	

Table 3. Total Workload Elapsed Time, Frequency Mean Value for CPU/Real Time Ratio, and User and System CPU Time Performance Using Dell EMC VxFLEX

The Dell EMC VxFLEX system scaled from one node to four nodes with a minor 3% increase on the average total workload time. Its performance with this accelerated IO demand still maintained a 1.06 to 1.10 ratio.

Scaling to Six Nodes on the Dell EMC VxFLEX System

The SAS mixed analytics workload (with 20 simultaneous tests) was run in a quiet setting (no competing activity on the system) using six nodes of the Dell EMC VxFLEX system. Multiple runs were executed to standardize the results.

Table 4 shows the performance of the Dell EMC VxFLEX six-node test

Dell EMC VxFLEX Node	Real Time	Mean Value of CPU/Real Time Ratio	User CPU Time in Minutes — Workload Aggregate	System CPU Time in Minutes — Workload Aggregate	Average Node Real Time Minutes
Node 1	532	1.00	495	69	505
Node 2	461	1.10	475	63	
Node 3	508	1.04	480	66	
Node 4	510	1.04	488	68	
Node 5	510	1.05	490	68	
Node 6	509	1.04	487	67	

Table 4. Total Workload Elapsed Time, Frequency Mean Value for CPU/Real Time Ratio, and User and System CPU Time Performance Using Dell EMC VxFLEX

Scaling from a single node to six nodes using a heavily mixed analytic 20-simultaneous-test workload introduced a minor 7% increase to the average total workload time. The Dell EMC VxFLEX system performance with this accelerated I/O demand still maintained an average 1.05 ratio.

The mixed analytics 20 simultaneous-test workload used a mixed representation of SAS shop jobs, driving CPU core usage at or close to 100% across 144 cores. Your performance will vary according to workload differences.

General Considerations

The Dell EMC VxFLEX system can deliver significant performance for an intensive SAS I/O workload. It is very important to use the SAS tuning guides for your host operating system to optimize server-side performance with Dell EMC VxFLEX appliances, as well as any additional suggestions provided in Dell EMC VxFLEX Tuning/Provisioning Recommendations.

Previously published SAS Performance Guides on VMware and ESXi 6.0 give host-specific advice for constructing SAS file systems. The advice in those papers pertains to creating file systems on LUNs with a good level of performance, and monitoring LUN and initiator queue depths on single LUN VMDK constructions. Single LUN VMDK constructions can present queue depth limitations in heavily used systems. SSD storage can represent a wide array of physical presentations underneath the virtually abstracted file system definitions. Your attention to physical construction to support single LUN performance for VMware VMDK file systems is generally needed. (Search for VMware papers under Resources) The Dell EMC VxFLEX SSD and Flex OS, supported by Intel Skylake processors, and NVMe storagepresented no issues with queue depth or other LUN performance limitations in our testing. Our testing, as previously noted, included a single-LUN construction for SAS file systems. Your mileage can vary. Consult with your Dell EMC Representative to ensure the bandwidth sizing and configuration are correct for your workload.

Dell EMC VxFLEX Tuning and Provisioning Recommendations

The following settings were used for the Dell EMC VxFLEX OS testing in an HCI deployment.

- Jumbo frames usage
    - Refer to Dell EMC VxFLEX OS Networking Best Practices and Design Considerations.
- SDC/SDS/MDM set to a High-Performance Profile
    - Refer to the “VxFLEX OS Performance Parameters” section of the Fine-Tuning Technical Note.
- 12 vCPUs per Storage VM (SVM).
- 8 GB for DRAM provisioned per SVM.

- NMVE devices were passed through to SVM using DirectPath IO.
- All VxFLEX OS volumes were thick-provisioned and presented to worker VMs as RDM devices.
- 2 x 25 GbE connections were used for data networks using VxFLEX IP Roles.
    - Refer to Dell EMC VxFLEX OS Networking Best Practices and Design Considerations.

Conclusion

The Dell EMC VxFLEX system has been proven to be extremely beneficial for scaled SAS workloads. In summary, the faster Intel Skylake processors enable the compute layer to perform more operations per second, thus increasing the potential performance of the solution. However, it is the consistently low response times of the underlying NVMe storage layer, with the FLEX OS software defined storage system, that allows its potential to be realized.

Using the Dell EMC VxFLEX system is designed to be as straightforward as possible. It is crucial to work with your Dell EMC Storage Engineer to plan, install, and configure the hosts for the environment to attain maximum performance.

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

References

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

Release Information Content Version: 1.0 May 2020