Technical Paper

Performance and Tuning Considerations for SAS® on the Intel® Xeon® E5 v4 Series Processors and the Vexata VX-100F Storage System

Release Information Content Version: 1.0 August 2017.

Contents

Introduction 2

Intel® Xeon® E5 v4 Series Processors and Vexata VX-100F Performance Testing 2

Test Bed Description 2

Data and IO Throughput 3

Hardware Description 4

Intel® Xeon® E5 v4 Series Processors Test Host Configuration 4

Vexata VX-100F Test Storage Configuration 4

Test Results 5

Single Host Node Test Result 5

Scaling from One Node to Four Nodes on the Vexata VX-100F 6

Scaling from One Node to Eight Nodes on the Vexata VX-100F 7

General Considerations 8

Vexata Storage Systems and SAS Tuning Recommendations 9

Host Tuning 9

Vexata VX-100F Storage Tuning 9

Vexata Storage System Monitoring 9

Conclusion 10

Resources 11

Introduction

This paper represents test results for SAS® Foundation workloads on Intel servers with server board S2600WT2, Intel® Xeon® E5 v4 Series Processors and the Vexata VX-100F Storage System.

This effort includes a scaled test bed of one, four, and eight nodes simultaneously running a SAS mixed analytics workload, to determine scalability of the array as well as uniformity of performance per node. This technical paper will outline SAS performance tests results and general considerations for setup and tuning to maximize SAS Application performance with Intel® Xeon® E5 v4 Series Processors and the Vexata VX-100F Storage System.

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

Intel® Xeon® E5 v4 Series Processors and Vexata VX-100F Performance Testing

Performance testing was conducted with Intel® Xeon® E5 v4 Series Processors and the Vexata VX-100F Storage System to attain a relative measure of how well it performed with IO-heavy workloads. Of particular interest was whether the Vexata VX-100F would yield substantial benefits for SAS large-block, sequential IO patterns against the very fast Intel® Xeon® E5 v4 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 30 seconds to 54 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, 102 tests were launched per test set on each node.

*Note – Previous test papers utilized a SAS mixed analytic 20 Simultaneous Test workload in a single SMP environment. This test effort utilizes a larger SAS mixed analytic 30 simultaneous test workload against a single node, then four nodes simultaneously, and finally eight nodes simultaneously (each node running the 30 simultaneous workload). The 30-session SAS mixed analytic workload was utilized to better match the host CPU and memory resources (see Hardware Description below) of the nodes. The test sets resulted in the following numbers of overall simultaneous tests launched per test bed against the appliance:

- 1 Node – 102 Tests
- 4 Node – 408 Tests
- 8 Node – 816 Test

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 are 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 per set. This raises the total IO throughput of the workload significantly.

In the 40-minute time span of the graph below (Table 1), the 8 simultaneous node workload quickly jumps to greater than 41 GB/sec of IO throughput. The test suite is highly active for about 23 minutes and then finishes with "trail out jobs." This is a typical "SAS Shop" throughput characteristic for a single-node instance and it 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 VX-100F: SASDATA, SASWORK, and UTILLOC. Each node has its own dedicated SASDATA, SASWORK and UTILLOC file systems.

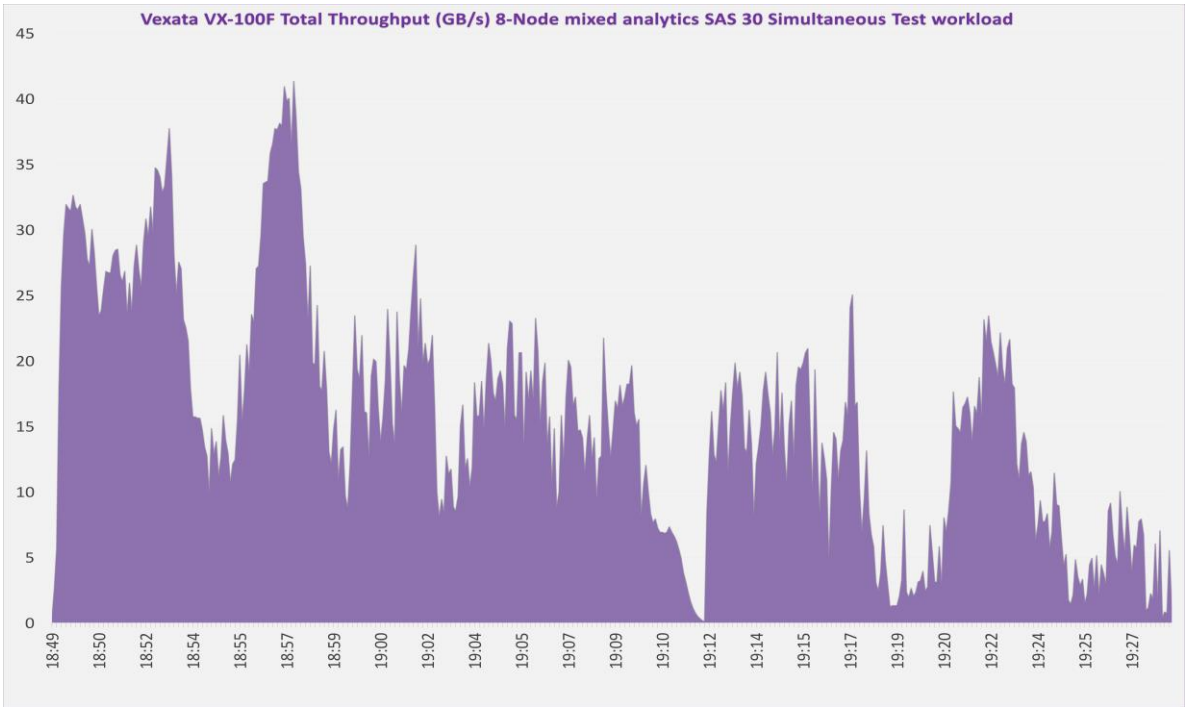


Table 1. Vexata VX-100F Storage System Test Throughput (Gigabytes/sec) for 8-Nodes, SAS mixed analytic 30 Simultaneous Test Run

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 30 Simultaneous workload's code set, data, result space, the working and utility space allocations were as follows:

- SASDATA – 1.1 Terabytes
- SASWORK – 1.1 Terabytes
- UTILLOC – 1.1 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 one, four, and eight host nodes utilizing the SAS mixed analytics 30 simultaneous test workload. All nodes were identical in configuration. The system host and storage configuration are specified below:

Intel® Xeon® E5 v4 Series Processors Test Host Configuration

The host server nodes utilized consisted of:

Host: Intel White box* servers with server board S2600WT2
Kernel: CentOS 7.2 3.10.0-327.el7.x86_64
Memory: 256 GB
CPU: Genuine Intel Xeon CPU E5-2699 v4, 2 Socket, 22 cores/socket, x86_64, 2.2GHz
HBA: Two Dual ported 32G FC Emulex HBAs LPe32002 (total of 4 ports) per host
Host Tuning:

- The following udev rules were applied:

```
# set noop scheduler and other recommended parameters for vexata disks
ACTION==”add|change”, ENV{ID_MODEL}==”VX100”, KERNEL==”sd*”, ATTR{queue/rotational}==”0”, ATTR{queue/scheduler}=”noop”,
ATTR{queue/rq_affinity}=”2”,    ATTR{queue/add_random}=”0”
ACTION==”add|change”, ENV{ID_MODEL}==”VX100”, KERNEL==”dm*”, ATTR{queue/rotational}==”0”, ATTR{queue/scheduler}=”noop”,
ATTR{queue/rq_affinity}=”2”,    ATTR{queue/add_random}=”0”
```

- The following multipath settings used:

```
devices {
    device {
        vendor              “Vexata”
        product             “VX*”
        path_grouping_policy multibus
        user_friendly_names     yes
    }
}
```

No additional settings were employed.

*White box servers are pre-production systems created in the Intel labs. Performance on productions systems available from various manufacturers may vary.

Vexata VX-100F Test Storage Configuration

Vexata VX-100F:

- 6U for the Vexata Storage System
- Additional RUs for 32G FC switch
- Vexata Storage System specifications
  - Nominal capacity of 84TB usable
  - Two Active-Active I/O Controllers
  - 16 Enterprise Storage Modules (ESMs) each with 4 x 2TB Intel® SSD DC P3700 Series drives
  - 16 ports of 32G FC
  - 4 redundant Power Supply Units (PSUs)
  - Redundant Supercap modules
  - Two 1Gb/s Ethernet Management ports
  - 60GB/s potential throughput from hosts to storage
  - Running VxOS: 2.0.2, package: v2.1.0-2
- File System Type: XFS
  - mkfs.xfs (used all defaults)
- Each host has 3 file systems each based on a 1.1TB LUN mounted on the Vexata Storage System: SASDATA, SASWORK and UTILLOC.
- No LVM used
- The SAS application used a SAS BUFSIZE of 64K

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 v4 Series Processors System utilizing the Vexata VX-100F on a single host node. Multiple runs were committed to standardize results.

Table 2 below shows the performance of the Vexata VX-100F. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 102 tests submitted. It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - Intel® Xeon® E5 v4 Series Processors / Vexata VX-100F	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	1.13	1035	58035	1181	128

Table 2. Frequency Mean Value for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using the Vexata VX-100F Storage System

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 Vexata VX-100F system delivered an excellent 1.13 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 Vexata VX-100F coupled with the faster Intel processors on the Intel compute node executes the aggregate run time of the workload in approximately 1035 minutes of total execution time.

The primary take-away from this test is that the Vexata VX-100F 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 1.13 CPU/Real Time ratio!

Scaling from One Node to Four Nodes on the Vexata VX-100F

For a fuller “flood test”, the mixed analytic 30s 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 v4 Series Processors System utilizing the Vexata VX-100F. Multiple runs were committed to standardize results.

Table 3 below shows the performance of the four host node environments attached to the Vexata VX-100F. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 102 tests submitted per node (408 in total). It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - Intel® Xeon® E5 v4 Series Processors w/ Vexata VX-100F	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	1.11	1031	58035	1163	129
Node2	1.11	963	58035	1058	112
Node3	1.12	1018	58035	1162	116
Node4	1.11	987	58035	1107	113

Table3. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using the Vexata VX-100F Storage System

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 VX-100F delivered an excellent 1.11 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 Vexata VX-100F 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 1,000 minutes per node, and 3,999 minutes of aggregate execution time for all 4 nodes.

Again, the Vexata VX-100F 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 Vexata VX-100F

The mixed analytic 30s 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 v4 Series Processors System utilizing the Vexata VX-100F. Multiple runs were committed to standardize results.

Table 4 below shows the performance of the eight-host node environments attached to the Vexata VX-100F. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 102 tests submitted per node (816 in total). It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

Storage System - Intel® Xeon® E5 v4 Series Processors w/ Vexata VX-100F	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	1011	58035	1070	116
Node2	1.09	1046	58035	1148	121
Node3	1.08	1021	58035	1075	117
Node4	1.10	1049	58035	1163	121
Node5	1.09	1008	58035	1073	114
Node6	1.09	998	58035	1059	114
Node7	1.09	1070	58035	1183	123
Node8	1.08	1035	58035	1130	116

Table4. Frequency Mean Values for CPU/Real Time Ratio, Total Workload Elapsed Time, Memory, and User & System CPU Time performance using Vexata VX-100F Storage System

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 VX-100F delivered a very good 1.089 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 Vexata VX-100F 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 1,030 minutes per node, and 8,238 minutes of aggregate execution time for all eight nodes.

The scale of eight simultaneous test workloads attained excellent CPU/Real Time ratio with only a 3% increase in average 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

Achieving read bandwidths over 40GB/s, write bandwidths over 20GB/s, and mixed read/write bandwidths of over 40GB/s, the Vexata VX-100F Flash Storage System is a transformative enterprise storage system that 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 in order to optimize server-side performance. Additional host tuning is performed as noted below.

The Vexata Storage System deploys simply and seamlessly into existing SAN storage environments and alongside any existing storage, requiring no custom host drivers, adapters or application changes. By leveraging the latest off-the-shelf NVMe SSDs from leading suppliers, the Vexata Storage System takes full advantage of advances in solid state media density, cost and performance.

Vexata Storage System (Left (rear), Right (front))

Vexata Storage Systems and SAS Tuning Recommendations

Host Tuning

The Host and Multipath tuning are listed in the Hardware section above. Single LUNs can be used for each file – SASWORK, SASDATA and UTILLOC. SAS configuration for REDHAT Systems is available at: http://support.sas.com/resources/papers/proceedings11/342794_OptimizingSASonRHEL6and7.pdf

A SAS BUFSIZE option of 64K was used for the tests. No additional application tuning was needed.

Vexata VX-100F Storage Tuning

No tuning is needed for the Vexata VX-100F to support the SAS workloads. Vexata Storage Systems are always RAID protected and can be set to either RAID6 to protect against 2 storage module failures, or RAID5 to protect against 1 storage module failure.

For SAS workloads, a fully populated Vexata VX-100F with 16 storage modules and 64 drives provides a write throughput of over 20GB/s with no tuning. Since SAS workloads are nearly 50/50 for Read/Write, this translates to over 40GB/s of total throughput.

Since Vexata Storage Systems support FC based SANs, it is generally recommended that multipathing be correctly implemented for failover. In this case, each server has 4 x 32G FC connections with the recommendation to split the ports across the IOCs and they can be mapped to two different FC switches. It is generally recommended to restrict the number of paths to each storage LUN to 16.

Vexata Storage System Monitoring

The Vexata VX-100F can be easily monitored through CLI or GUI options.

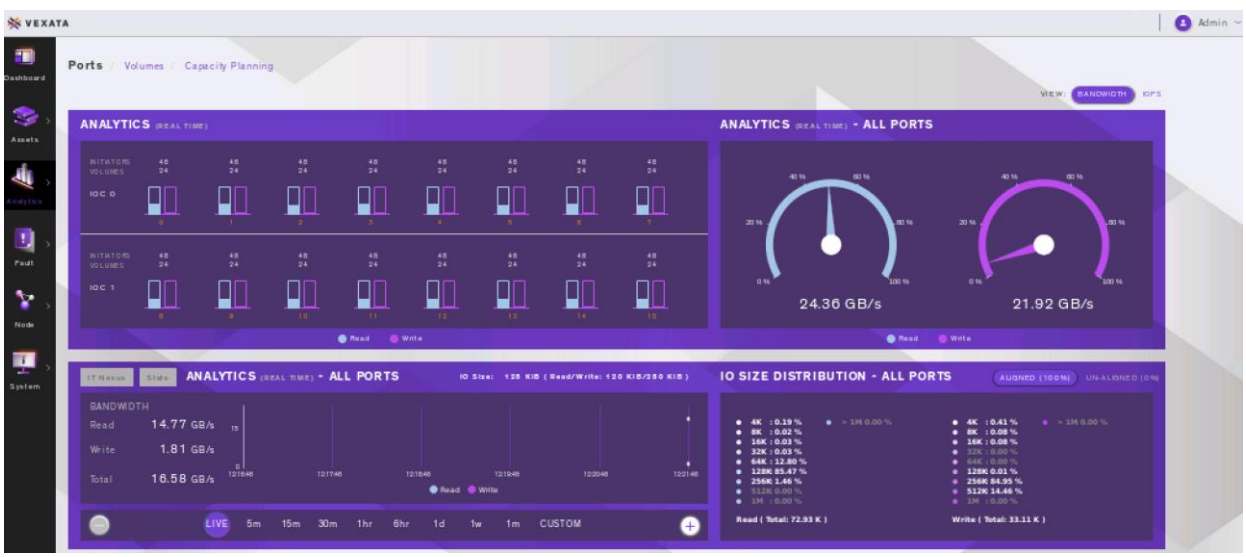
The following example shows how the CLI can be used to monitor the read/write bandwidths seen by the Vexata VX-100F.

```
Displaying throughput per second. Units - M=MB(10^6), G=GB(10^9), T=TB(10^12)
Using delay of 5 Seconds for data sampling
Will repeat  2880 times
Press <Control+C> to exit

Ioc/Port:   0/0   0/1   0/2   0/3   0/4   0/5   0/6   0/7   1/0   1/1   1/2   1/3   1/4   1/5   1/6   1/7 Total
-------------------------------------------------------------------------------------------------------------
18:57:24.288036
read     :   1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.1G  1.2G  1.2G  1.2G  1.1G  1.2G  1.2G  1.2G 18.8G
write    :   1.4G  1.4G  1.4G  1.4G  1.4G  1.4G  1.4G  1.4G  1.3G  1.4G  1.4G  1.4G  1.3G  1.3G  1.4G  1.3G 22.0G
r+w      :   2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G 40.9G
18:57:29.294553
read     :   1.1G  1.2G  1.2G  1.2G  1.1G  1.2G  1.1G  1.1G  1.2G  1.2G  1.2G  1.1G  1.2G  1.2G  1.2G  1.2G 18.5G
write    :   1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G 21.3G
r+w      :   2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G 39.8G
18:57:34.300148
read     :   1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G 20.6G
write    :   1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G 19.4G
r+w      :   2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G  2.5G 40.0G
18:57:39.305033
read     :   1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G  1.2G 19.1G
write    :   1.1G  1.0G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G 17.2G
r+w      :   2.2G  2.2G  2.3G  2.3G  2.3G  2.3G  2.2G  2.3G  2.3G  2.3G  2.3G  2.3G  2.3G  2.3G  2.3G  2.3G 36.3G
18:57:44.310131
read     :   1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G 20.7G
write    :   1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G 20.6G
r+w      :   2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G  2.6G 41.3G
18:57:49.317004
read     :   1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G  1.3G 20.6G
write    :   1.1G  1.1G  1.2G  1.2G  1.1G  1.2G  1.2G  1.2G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G  1.1G 17.9G
r+w      :   2.4G  2.4G  2.5G  2.5G  2.4G  2.5G  2.5G  2.5G  2.3G  2.3G  2.4G  2.4G  2.4G  2.4G  2.3G  2.3G 38.5G
```

Vexata bandwidth monitoring with CLI

The following example shows how the GUI can be used to monitor the read/write bandwidths seen by the Vexata VX-100F.



Vexata bandwidth monitoring with GUI

Conclusion

The Intel Xeon E5-2699 v4 Series Processors and Vexata VX-100F with 2TB Intel® SSD DC P3700 Series drives 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 Vexata VX-100F allow this scaled workload potential to be

fully exploited.

The Vexata Storage System is designed to be as operationally straightforward as possible, but to attain maximum performance, it is crucial to work with Vexata Storage 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 Vexata 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: Venkatesh Nagapudi
Enterprise: Vexata Inc
Address: 1735 Technology Dr Suite 780
City, State: San Jose, CA 95110
Work Phone: 408 218 8792
Email: venky@vexata.com

Name: Kishore Vinjam
Enterprise: Vexata Inc
Address: 1735 Technology Dr Suite 780
City, State: San Jose, CA 95110
Work Phone: 510 396 2647
Email: kishore@vexata.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