Technical Paper

Performance and Tuning Considerations for SAS® on Dell® EMC® VMAX® 250 All-Flash Array


sas
THE POWER TO KNOW.

Release Information Content Version: 1.0 April 2018

Contents

Introduction

This paper is a test of the SAS® mixed analytics workload on the Dell® EMC® VMAX® 250 All-Flash array.

This effort involves a flood test of one node and four simultaneous X-86 nodes running a SAS mixed analytics workload to determine scalability against the array and uniformity of performance per node. This technical paper outlines performance test results performed by SAS and provides general considerations for setting up and tuning the X-86 Linux host and the VMAX 250 All-Flash array for SAS application performance.

An overview of testing is discussed first, including the purpose of the testing, detailed descriptions of the actual test bed and workload, and a description of the test hardware. Test results are included, accompanied by a list of tuning recommendations. General considerations, recommendations for implementation with SAS Foundation, and conclusions are discussed.

Dell EMC VMAX Performance Testing

Performance testing was conducted with four host nodes attached via 16GB fiber channel HBA to a VMAX 250 array to deliver a relative measure of how well it performed with IO heavy workloads. Of particular interest was whether the VMAX 250 could easily manage SAS large-block sequential IO patterns. 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

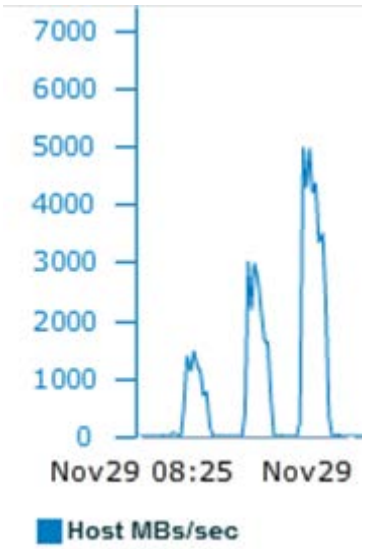
The test bed chosen for flash testing was a SAS mixed analytics workload. This was a scaled workload of computation and IO-oriented tests to measure concurrent, mixed job performance.

The actual workload was composed of 19 individual SAS tests: 10 computation, two memory, and seven IO-intensive tests. Each test was composed of multiple steps. Some tests relied on existing data stores, and other tests (primarily, computation tests) relied on generated data. The tests were chosen as a matrix of long-running and shorter-running tests (ranging in duration from approximately five minutes to one 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.

Data and IO Throughput

The IO tests input an aggregate of approximately 300GB of data, and the computation tests input more than 120GB of data for a single instance of each SAS mixed analytics workload with 20 simultaneous tests on each node. Much more data is generated as a result of test-step activity and threaded kernel procedures such as SORT (e.g., SORT makes three copies of the incoming file to be sorted). As stated, some of the same tests run concurrently using different data. Some of the same tests are run back-to-back to produce a total average of 20 tests running concurrently. This raises the total IO throughput of the workload significantly.

In Graph 1, the four-simultaneous-nodes workload quickly jumps to 5GB/sec in the initial test activity for the VMAX storage monitor in its one-hour-and-20-minute span. The test suite is highly active for about 40 minutes, and then it 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. This throughput is produced from all three primary SAS file systems: SASDATA, SASWORK, and UTILLOC. Each node has its own dedicated SASDATA, SASWORK, and UTILLOC file systems.



Graph 1. VMAX 250 IOPs and Throughput Monitor for Four-Node SAS Mixed Analytics Workload with 20 Simultaneous Tests Run

SAS File Systems

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 this workload's code set, data, result space, working space, and utility space, the following space allocations were made:

- SASDATA—4 TB
- SASWORK (and 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

This test bed was run against four host nodes using the SAS mixed analytics workload with 20 simultaneous tests. The host and storage configurations are specified in the following sections.

VMAX 250 Test Host Configuration

Here are the specifications for the eight host server nodes:

Host: Lenovo x3650 M5, Red Hat Enterprise Linux 7.2
Kernel: Linux 3.10.0-327.36.3.el7.x86_64
Memory: 256GB
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. Here is the tuned profile used for testing:

```
# 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
```

VMAX 250 Test Storage Configuration

Here are the specifications for the VMAX 250 All-Flash array:

- A dual-engine VMAX 250F All-Flash array was used.
- Each engine was equipped with dual 12-core processors and 2TB of usable cache shared. (This was mirrored between engines.)
- The array was managed by Unisphere version 8.4.0.4.
- All storage devices were placed in a single storage resource pool on thin provisioned volumes.
- Storage compression was disabled and de-duplication was unavailable.
- Dell EMC PowerPath version 6.2 was used.
- Array connection to hosts consisted of 16x16GB fiber channel over dual brocade switches (two FC connections per host).

Test Results

Single Host Node Test Results

The SAS mixed analytics workload was run in a quiet setting for the X-86 system using Dell EMC VMAX 250 on a single host node. There was no competing activity on the server or in storage. Multiple runs were performed to standardize results.

The previous tuning specifications were applied to a Linux operating system for Red Hat Enterprise Linux 7.2. Work with your Dell EMC engineer for appropriate tuning specifications for other operating systems or for the particular processors used in your operating system.

Table 1 shows the performance of the Dell EMC VMAX 250. This table shows a frequency mean value of the CPU/real-time ratio, summed from all of the 77 tests submitted. It shows summed user CPU time and summed system CPU time in minutes.

Storage System: X-86 with Dell EMC VMAX 250	Mean Value of CPU/RealTime—Ratio	Elapsed Run Time in Minutes—Workload Aggregate	User CPU Time in Minutes—Workload Aggregate	System CPU Time in Minutes—Workload Aggregate
Node1	1.04	638	620	61

Table 1. Performance Using One Node on Dell EMC VMAX 250 All-Flash Array

The second column shows the ratio of total CPU time (user + system CPU) to total real time. If the ratio is less than 1, then the CPU is spending time waiting on resources, usually IO. The VMAX system delivered a very good 1.04 ratio of real time to CPU. The natural question is, “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 from each of the jobs in the workload. The Dell EMC VMAX 250, coupled with the fast Intel processors on the Lenovo compute node, executes the aggregate run time of the workload in approximately 638 minutes of total execution time.

The primary take-away from this test is that the Dell EMC VMAX 250 easily provided 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 1.04 CPU/real time ratio!

Scaling from One Node to Four Nodes

For a fuller flood test, the SAS mixed analytics workload was run concurrently in a quiet setting for the X-86 system using Dell EMC VMAX 250 on four physically separate but identical host nodes. There was no competing activity on the server or in storage. Multiple runs were performed to standardize results.

The previous tuning specifications were applied to a Linux operating system for Red Hat Enterprise Linux 7.2.

Table 2 shows the performance of the four host node environments attached to the Dell EMC VMAX 250. This table shows a frequency mean value of the CPU/real-time ratio, summed from all of the 77 tests submitted. It shows summed user CPU time and summed system CPU time in minutes.

Storage System: X-86 with Dell EMC VMAX 250	Mean Value of CPU/Real-Time—Ratio	Elapsed Run Time in Minutes—Workload Aggregate	User CPU Time in Minutes—Workload Aggregate	System CPU Time in Minutes—Workload Aggregate
Node1	1.04	642	646	64
Node2	1.05	694	722	76
Node3	1.05	693	722	76
Node4	1.05	626	621	61

Table 2. Performance Using Four Nodes on Dell EMC VMAX 250 All-Flash Array

The second column shows the ratio of total CPU time (user + system CPU) to total real time. If the ratio is less than 1, then the CPU is spending time waiting on resources, usually IO. The VMAX system delivered a very good 1.05 ratio of real time to CPU.

The third column shows the total elapsed run time in minutes, summed from each of the jobs in the workload. The Dell EMC VMAX 250, coupled with the fast Intel processors on the Lenovo compute node, executes the aggregate run time of the workload in an average of 664 minutes per node and 2,655 minutes of total execution time for all four nodes.

The array performance peaked at approximately 5GB/sec and 45k IOPs using a SAS 64K BUFSIZE on the four-node test.

The Dell EMC VMAX 250 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 was a mixed representation of what an average SAS shop might be executing at any given time. Due to workload differences, your mileage might vary.

General Considerations

Using the Dell EMC VMAX 250 All-Flash array can deliver significant performance for an intensive SAS IO workload. It is very helpful to use the SAS tuning guides for your operating system host to optimize server-side performance with VMAX arrays. Additional host tuning is performed as noted below.

When using flash storage, the general industry recommends leaving overhead in the flash devices to accommodate garbage-collection activities and focusing on which workloads (if not all) to use flash for. Both points are discussed briefly.

- As per Dell EMC, the VMAX 250 delivers consistent performance regardless of capacity consumption. As a result, it does not require users to limit themselves to a fraction of the purchased flash capacity to enjoy the benefits of enterprise flash-based storage performance. This is the result of a number of architectural design choices used by the VMAX 250. There are no system-level garbage-collection activities to worry about (this activity is decentralized to the SSD controllers). All data-aware services (de-duplication and compression) are always on and happen in-line with no post-processing. And, the XDP (VMAX data protection) algorithm ensures minimal Write activity and locking of the SSDs.
- The VMAX All-Flash array uses front-end dynamic CPU core allocation. The VMAX administrator allocates front-end CPU cores to a shared pool. The VMAX administrator has the option of assigning CPU cores to specific front-end CPUs. Most VMAX All-Flash arrays do not directly assign CPU cores. They take advantage of a VMAX All-Flash array feature called “dynamic core allocation.” Based on actual IO activity, the VMAX All-Flash array can automatically add or remove CPU cores based on actual workload activity. The VMAX All-Flash array makes extensive use of processor cache. All VMAX IO operations are processed by VMAX cache. Read and Write IO are all cache operations. This reduces the need for specific Write considerations required by most SSD-based storage processors. When the VMAX All-Flash array receives a Write IO, the IO block is placed in VMAX cache in a cache-mirrored format (two cache copies of the IO block). A Write acknowledge is then returned to the application.

Dell EMC and SAS Tuning Recommendations

Host Tuning

It is important to study and use as an example the host and multipath tuning listed in the **Hardware Description** section. In addition, pay close attention to LUN creation and arrangements and LVM tuning. For more information about configuring SAS workloads for Red Hat Enterprise Linux systems, see http://support.sas.com/resources/papers/proceedings11/342794_OptimizingSASonRHEL6and7.pdf.

In addition, a SAS BUFSIZE option of 64K, along with a Red Hat Enterprise Linux logical volume stripe size of 64K, were used to achieve the best results from the VMAX 250 array. Testing with a lower BUFSIZE value might yield benefits on some IO tests, but might require host and application tuning changes.

VMAX Storage Tuning

The VMAX All-Flash array uses storage pools to ease VMAX administration and improve IO performance. It’s useful to think of a storage pool in terms of a file system volume manager. Physical storage is assigned to a storage pool. A thin device can then be created from the storage pool. Dell EMC calls these thin pool devices or LUNs “TDEVs.” IO to a TDEV LUN is stripped across all of the available back-end physical storage devices in the storage pool.

From a SAS perspective, use a single pool for all SAS LUNs. The TDEV LUNs can be any capacity required. There are no special VMAX All-Flash array LUN requirements for SASWORK, UTILLOC, or SASDATA. Remember, a thin pool LUN represents an IO path. Using very large thin pool LUNs reduces the number of required LUNs to meet capacity needs. It reduces the number of paths and IO queues for IO operations.

For SAS workloads, the VMAX All-Flash array automated dynamic front-end CPU core management eases SAS VMAX administration while providing high levels of IO throughput.

Conclusion

The Dell EMC VMAX 250 All-Flash array has been proven to be extremely beneficial for scaled SAS workloads when using newer, faster processing systems. In summary, the faster processor enables the compute layer to perform more operations per second, thus increasing the potential performance for the solution, but it is the consistently low response times of the underlying storage layer that allow this potential to be realized.

Operating the VMAX 250 array is designed to be as straightforward as possible, but to attain maximum performance, it is crucial to work with your Dell 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 might require additional guidance by Dell EMC and SAS engineers, depending on your host system and workload characteristics.

Resources

SAS papers on performance, best practices, and tuning are at http://support.sas.com/kb/42/197.html.

Contact Information

Name: Steven Bonuchi
Enterprise: Dell EMC Corporation
Email: steven.bonuchi@Dell.com

Name: Tony Brown
Enterprise: SAS Institute Inc.
Email: tony.brown@sas.com

Name: Jim Kuell
Enterprise: SAS Institute Inc.
Email: jim.kuell@sas.com

Name: Margaret Crevar
Enterprise: SAS Institute Inc.
Email: margaret.crevar@sas.com

To contact your local SAS office, please visit: sas.com/offices