Technical Paper

# Performance and Tuning Considerations for SAS® using Intel® Solid State DC P3700 Series Flash Storage


sas
THE POWER TO KNOW.

**Release Information** Content Version: 1.0 November 2014.

# Contents

# Introduction

The Intel® Solid State Drive Data Center Family for PCIe flash storage offers high performance, and excellent scalability for SAS® workloads. It can be configured for large SAS workloads. 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® PCIe Solid State DC P3700 Series flash storage.

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

# Intel® Performance Testing

Performance testing was conducted with the Intel® PCIe Solid State DC P3700 Series flash storage to determine a relative measure of how well it performed with IO heavy workloads compared with high-performance spinning disk. Of particular interest was whether the Intel flash storage would yield substantial benefits for SAS large-block, sequential IO patterns. 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 53 minutes, depending on hardware provisioning. Actual test times vary by hardware provisioning differences. 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, 101 tests were launched.

## Data and IO Throughput

The IO tests input an aggregate of approximately 300 Gigabytes of data, and the computation over 120 Gigabytes of data – for a single instance of each test. Much more data is generated as a result of test-step activity, and threaded kernel procedures such as SORT (e.g. SORT makes 3 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 30 tests running concurrently. This raises the total IO throughput of the workload significantly. In its run span, the workload quickly jumps to 900 MB/sec, climbs steadily to 2.0 GB/s, and achieves a peak of 4+ GB/s throughput before declining again. This is a good average “SAS Shop” throughput characteristic for a single-instance OS (e.g. non-grid). This throughput is from all three primary SAS file systems: SASDATA, SASWORK, and UTILLOC.

## SAS File Systems Utilized

There are 3 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 and utility space the following space allocations were made:

- SASDATA – 3 Terabytes
- SASWORK – 3 Terabytes
- UTILLOC – 2 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. Fallow space is generally left on the file systems to facilitate write performance and avoid write issues due to garbage collection when running short on cell space.

## Hardware Description

The host server information the testing was run performed on is as follows:

**Host:** Dell® R920
**OS:** RHEL 6.5 on 2.6.32-431.23.3.el6.x86_64 kernel (mockbuild@x86-27.build.eng.bos.redhat.com) (gcc version 4.4.7 20120313 (Red Hat 4.4.7-4) (GCC) #1 SMP Wed Jul 16 06:12:23 EDT 2014
**Memory:** 1.5 TB RAM
**CPU:** 4x Intel® CPU E7-4870 v2 2.3ghz Genuine Intel, Model 62, CPU Family 6, Stepping 7, 2300 Mhz, 30,720 KB Cache, 15 Core Processors Running Hyper-threading

## Storage

Comparative performance testing was conducted between a high-performance spinning disk array, and the Intel® P3700 PCIe Flash. A comparison of the two storage types follows:

High-performance Spinning Disk Array Definition:

- Number and types of disks / controllers: 192 x 300GB 15K rpm Fibre Channel Drives and 4 controllers, consuming 84U all together
- RAID levels: Multiple RAID 5 sets (3 data/1 parity disk per set)
- File System Type: XFS
- File System/Logical Volume Arrangement: File Systems /SASDATA, SASWORK, /UTILLOC are placed across 1 Logical Volume utilizing all 192 spindles in the array
- Fibre Channel ports: 8 x 8 Gbps Adapters with ACTIVE/ACTIVE multi-pathing
- 8 GB/sec potential throughput for the array

The Intel® PCIe Solid State DC P3700 Series flash storage:

- 5x Intel P3700 1.6TB PCIe/NVMe Flash Cards in PCIe x16 slots
- Utilizing Firmware Level 8DV10054
- File System Type: XFS
- Striped as RAID 0 via LVM

## Test Results

The Mixed Analytic Workload was run in a quiet setting (no competing activity on server or storage) for both the high-performance spinning disk storage and the Intel® PCIe Solid State DC P3700 Series flash storage. Multiple runs were committed to standardize results.

We worked with Intel storage engineers to tune the host and IO pacing (see General Considerations and Tuning Recommendations below). The Intel P3700 firmware level used in testing was 8DV10054.

The tuning options noted below apply to LINUX operating systems for Red Hat RHEL 6.5. Work with your Intel representative for appropriate tuning mechanisms for any different OS, or the particular processors used in your system.

Table 1 below shows the performance of the completely tuned environments comparing the Intel® PCIe Solid State DC P3700 Series flash storage and the high-performance spinning disk system. This table shows an aggregate SAS FULLSTIMER Real Time, summed of all the 101 tests submitted. It also shows Summed Memory utilization, Summed User CPU Time, and Summed System CPU Time in Minutes.

| Storage System | Elapsed Real Time in Minutes – Workload Aggregate | Memory Used in GB - Workload Aggregate | User CPU Time in Minutes – Workload Aggregate | System CPU Time in Minutes - Workload Aggregate |
|---|---|---|---|---|
| **High-Performance Spinning Storage** | 1208 | 56.1 | 1385 | 178 |
| **Intel® PCIe Solid State DC P3700 Series flash storage** | 1269 | 56.2 | 1467 | 214 |

| | | | | |
|---|---|---|---|---|
| **Aggregate**<br><br>**DELTA** | +61 | +0.01 | +82 | +36 |

*Table 1. Total Workload Elapsed Time, Memory, and User & CPU Time Reduction by using Intel® PCIe Solid State DC P3700 Series flash storage*

The first column in shows the total elapsed run time in minutes, summed from each of the jobs in the workload. It can be seen that the Intel® PCIe Solid State DC P3700 Series flash storage had a higher aggregate run time of the workload by approximately 61 minutes (about 5% longer), which is actually incredible, considering 5 PCIe cards are being compared to a 192 disk, 84RU external storage array!

Another way to review the results is to look at the ratio of total CPU time (User + System CPU) against the total Real time. Table 2 below 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 high-performance spinning storage array is a very large, expensive, and fast array. The Real time/CPU time ratio of the workload on the Spinning Storage array was 1.12, which is excellent.

The Intel® PCIe Solid State DC P3700 Series flash storage achieved slightly better than that performance at 1.15 with five P3700 PCIe Flash Storage cards. Both of these numbers indicate the storage was keeping up with the CPU demand and provided excellent performance. The standard deviation and spread of the ratios for spinning storage is much higher than the Intel test indicating a broader range of performance experienced by individual tests. This is indicative of greater performance consistency on the part of the Intel P3700 based system.

The 1.15 Mean Ratio associated with the Intel® PCIe Solid State DC P3700 Series flash storage indicates a good efficiency in getting IO to the CPU for service. This result was achieved using a SAS 256K BUFSIZE to present to storage. For the IO intensive SAS Application set, this is an excellent outcome. 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.

To put this into further perspective, the Spinning Disk Array and the Intel® PCIe Solid State DC P3700 Series flash storage both ran in the neighborhood of 200 minutes faster on this host with E7-4870 processors, and 500 GB more RAM than our standard lab test host. This includes comparisons against the Spinning Disk Array noted in our specifications above, and comparing to external flash arrays. There is also an inherent performance advantage in utilizing internal PCIe flash assemblies, by not having to connect to an external array via an FC link. The caveat is that there is limited space available using internal PCIe cards as compared to an external array. You must ensure your workload will fit within the scale of available card slots and capacities.

| **Storage System** | **Mean Ratio of CPU/Real-time - Ratio** | **Mean Ratio of CPU/Real-time - Standard Deviation** | **Mean Ratio of CPU/Real-time - Range of Values** |
|---|---|---|---|
| **High-Performance Spinning Storage** | 1.12 | 0.345 | 2.65 |
| ***Intel® PCIe Solid State DC P3700 Series flash storage*** | 1.15 | 0.035 | 2.15 |

*Table 2. Frequency Mean, Standard Deviation, and Range of Values for CPU/Real Time Ratios. Less than 1 in the 'Ratio' column indicates IO inefficiency.*

In short our test results were very pleasing. It showed that the Intel® PCIe Solid State DC P3700 Series flash storage, when tuned at install with recommended host-side parameters, and utilizing firmware level 8DV10054, can provide excellent performance for SAS workloads. In this case with an aggregate large workload run time only 5% longer than a large, high throughput spinning array. 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 and Tuning Recommendations

## General Notes

Utilizing the Intel® PCIe Solid State DC P3700 Series flash storage can deliver significant performance for an intensive SAS IO workload. Work with Intel and your Host Systems Administrator on install tuning, and host CSTATE tuning (if your processor model is an older version, and if IT standards allow CSTATES to be altered) to maximize performance. Processor CSTATES recommendations are typically different for different processor types, and even models. They govern power states for the processor. By not allowing processors to enter a deep "sleep" state, they are more rapidly available to respond quickly to demand loads. Generally keeping CSTATES at a 1 or 0 level works well, and for those processors supporting turbo-boost mode, it should be set to take advantage of that. On our test host, CSTATE levels did not have to be set on this newer generation of E7-4870 Intel Processors.

It is very helpful to utilize the SAS tuning guides for your operating system host to optimize server-side performance before Intel® P3700 tuning, and additional host tuning is performed as noted below. See: http://support.sas.com/kb/42/197.html

Some general considerations for using flash storage include leaving overhead in the flash devices, and considering which workloads if not all, to use flash for.

**Reads vs. Writes.** Flash devices perform much better with Reads than Writes for large-block, sequential IO (SAS). If your scale of workload dictates that you can afford flash for all your file systems that is good. If not, you may wish to bias your flash usage to file systems and data that are read-intensive to get the maximum performance for the dollar. For example, if you have a large repository that gets updated nightly, or weekly, and is queried and extracted from at a high level by users, that may be where you wish to provision your flash storage.

## Intel and SAS Tuning Recommendations

The following install tuning was performed on the host server during the installation of Intel® PCIe Solid State DC P3700

Series flash storage. These are Intel and SAS recommended tuning steps.

- Tuned storage profile was set to enterprise-storage by running "tuned-adm profile enterprise-storage".
- CSTATES were left at default on this new processor set, it is not required to change them for performance.
- Device firmware level utilized was 8DV10054. Contact Intel for the latest suggested firmware.
- The five P3700 solid-state devices were installed in PCI-E X16 slots.
- The five devices were striped using Linux LVM and then partitioned in to three file systems for /sasdata, /saswork and /utilloc.
- The partitions were formatted with XFS and mounted with the options noatime and nodiratime.
- The following commands were run and appended to /etc/rc.local for persistence across boots:

```
blockdev –setra 16384 /dev/mapper/DEVICENAME-FOR-SASWORK
blockdev –setra 16384 /dev/mapper/DEVICENAME-FOR-SASDATA
blockdev –setra 16384 /dev/mapper/DEVICENAME-FOR-UTILLOC
echo never > /sys/kernel/mm/redhat_transparent_hugepage/enabled
```

In addition a SAS BUFSIZE option of 256K was utilized to achieve the best results from Intel® PCIe Solid State DC P3700 Series flash storage.

# Conclusion

Intel® PCIe Solid State DC P3700 Series flash storage can be extremely beneficial for many SAS Workloads. Testing has shown it can significantly eliminate application IO latency, providing improved performance. It is crucial to work with your Intel Storage engineer to plan, install, and tune your host utilizing the Intel® PCIe Solid State DC P3700 Series flash storage to get maximum performance. The guidelines listed in this paper are beneficial and recommended. Your individual experience may require additional guidance by Intel 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: Allen Scheer
Enterprise: Intel
Address: 1900 Prairie City Road
City, State: Folsom, CA
United States
Work Phone: 916-356-2295
Email: William.a.scheer@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(214) 977-3916
Fax: +1 (214) 977-3921
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