Technical Paper

Performance and Tuning Considerations for SAS® on Fusion-io® ioScale™ Flash Storage

THE POWER TO KNOW®

Release Information

Content Version: 1.0 May 2014.

Trademarks and Patents

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Statement of Usage

This document is provided for informational purposes. This document may contain approaches, techniques and other information proprietary to SAS.

Contents

Introduction

The Fusion-io® ioScale™ flash storage offers high performance, ease of installation and management, and scalability for SAS® workloads. Fusion-io® ioScale™ flash storage can be installed in PCIe slots within the SAS server, and configured to serve as storage for portions of SAS workloads, or the entire workload, if scale permits. This technical paper will outline performance test results performed by SAS, and general considerations for setup and tuning to maximize SAS Application performance with Fusion 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®.

ioScale Performance Testing

Performance testing was conducted with the ioScale flash storage to garnish a relative measure of how well it performed with heavy workloads compared with traditional spinning disk. Of particular interest was whether the flash storage would yield substantial benefits for SAS large-block, sequential IO pattern. 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 SAS Mixed Analytic 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. In some instances the same test (running against replicated data streams) is run concurrently, and/or back-to-back in a serial fashion, to achieve a 30 simultaneous test-mix 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.

Data and IO Throughput

The IO tests input an aggregate of approximately 300 Gigabytes of data, 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. 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 1 hour and 45 minute span, the workload quickly jumps to 900 MB/sec, climbs steadily to 2.4 GB/s, and achieves a peak of 3.9 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 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 was left on the file systems to facilitate write performance and avoid write issues due to garbage collection when running short on available empty cells to write to.

Hardware Description

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

Host: HP DL980 G7-RB410
OS: Linux version 2.6.32-431.el6.x86_64 (mockbuild@x86-023.build.eng.bos.redhat.com) (gcc version 4.4.7 20120313 (Red Hat 4.4.7-4) (GCC)) #1 SMP Sun Nov 10 22:19:54 EST 2013
Memory: 529223432 kB RAM,
CPU: 64 x Intel(R) Xeon(R) CPU X7560 @ 2.27GHz GenuineIntel, Model 46, CPU Family 6, Stepping 6, 2266 Mhz, 24576 KB Cache

Storage

Comparative performance testing was conducted between a traditional spinning disk array, and the Fusion-io ioScale Flash storage. The traditional spinning disk array was configured with the following characteristics:

- Number and types of disks: 192x 300GB 15K rpm FC Drives
- Raid levels: Multiple raid5 sets (3data/1parity disk per set)
- File System Type: EXT4
- File System/Logical Volume Arrangement: File Systems /SASDATA, SASWORK, /UTILLOC are placed across 1 Logical Volume utilizing all 192 spindles in the array
- Host Bus Adapters: 4x 8 GB Adapters with ACTIVE/ACTIVE multi-pathing

Throughput characteristics under sustained-load testing for the spinning-disk array were as shown in Table 1:

ACTIVITY	IOPs	MB/sec Throughput	Response Time (ms)
256k Sequential Reads	32324	8474	20
256k Sequential Writes	12851	3369	20

Table 1. IOPs and Throughput Rating of Spinning Storage Used in Tests

As can been in the chart above, the average throughput of the spinning disk was appropriate for the generated workload, with 4 – 8 Gb host bus adapters.

The Fusion-io ioScale Flash cards tested are described as:

- 4 x ioScale 3.2 TB MLC Cards

Test Results

The Mixed Analytic Workload was run in a quiet setting (no competing activity on server or storage) for both the spinning disk storage and the Fusion-io ioScale Flash cards. Multiple runs were committed to standardize results. Initial runs showed erratic results, especially with very intensive IO tests running concurrently.

In the very first run, with no card or system tuning pertaining to the cards performed, some of the very intensive IO tests actually ran slower than the spinning storage, a fair amount within a 5% run-time margin, and only 9 of the tests were better by 4 – 17%. We knew that without tuning the system the performance of this very heavy IO workload would not be what we expected. We worked with Fusion-io engineers to tune the system (see General Considerations and Tuning Recommendations below), and the performance was radically better. The tuning options noted below apply to RHEL operating systems, work with your Fusion-io vendor for appropriate tuning mechanisms for any different OS.

Table 2 below shows how much faster the tuned ioScale performance was. 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	Elapsed Real Time in Minutes - (SUM)	Memory Used in GB- (SUM)	User CPU Time in Minutes- (SUM)	System CPU Time in Minutes- (SUM)
Spinning Storage	3066	56	1841	905
4 – Fusion ioScale Cards	1134	55	965	221
DELTA	1932	1	876	684

Table 2. Total Workload Elapsed Time, Memory, and User & CPU Time Reduction by using ioScale Flash

As can be seen in Table 2, the total elapsed run time, summed from each of the jobs in the workload decreased by 1932 minutes, or to a ratio of .63, which is fantastic. Memory utilization remained stable, and CPU time dropped dramatically as well, not having to wait on IO.

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 3 below shows the ratio of Total CPU Time to Real Time. If the ration is less than 1, then the CPU is spending time waiting on resources, usually IO. The 0.79 metric (e.g. less than 1.0) for the Mean of all the 102 tests for Spinning Storage indicates that there is lag between Real Time and CPU Time, indicating IO could be more efficient. The standard deviation and spread of the ratios for spinning storage is higher throughout the test load.

The 1.047 Mean Ratio associated with the Fusion-io ioScale cards indicates a much more CPU bound process (e.g. not spending time waiting on IO). For the IO intensive SAS Application set, this is the goal you wish to achieve! 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.

Storage System	Ratio of CPU/Real-time - Mean	Ratio of CPU/Real-time - Standard Deviation	Ratio of CPU/Real-time - Range of Values
Spinning Storage	0.79	0.31	2.60
4 – Fusion ioScale Cards	1.047	0.16	1.44

Table 3. Frequency Mean, Standard Deviation, and Range of Values for CPU/Real Time Ratios. Less than 1 indicates IO inefficiency.

In short our test results were very pleasing. It showed that Fusion-io ioScale Flash Cards, when properly installed, configured, and tuned, can significantly boost SAS workloads. 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. There are caveats to using internal Flash cards for workload storage and they are mentioned in the General Considerations and Tuning Recommendations that follow.

General Considerations and Tuning Recommendations

General Notes

Utilizing Fusion-io ioScale flash cards requires available PCIe slots, appropriate slot placement, and tuning as listed below. Work with your Fusion-io engineer to determine what you can configure with your available PCIe slots.

We have found flash cards very advantageous because they are inside the server, negating a TCP connection. There can be disadvantages to this approach:

- Storage cannot be shared off-board to other hosts
- Backups will impact server resources
- PCIe slot requirements may force choices for other peripherals

You must weigh the implications of having on-board, versus attached storage in your server-side resources, as part of your data safety, backup, and recovery plan.

It is very helpful to utilize the SAS Tuning Guides for your operating system host to optimize server-side performance before flash card 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 where to use flash.

Leaving overhead space on the flash devices. By not filling the capacity of the flash devices, room is left for overhead expansion. When the flash device needs to write, it will not have to perform any untended garbage collection duties to commit the writes. In very heavy write situations (e.g. SASWORK for example) this is very important. Down-formatting of cards is encouraged to maintain a high level of write performance. See your Fusion-io engineer for more details.

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.

Fusion-io and SAS Tuning Recommendations

Overview

- BIOS Tuning
- OS Tuning
- NUMA Considerations
- Enable VSL Tuning options

BIOS Tuning

Fusion-io Virtual Storage Layer (VSL) software is executed on the host system utilizing marginal CPU and Memory in order to drive best performance and management of the installed ioMemory devices. We will benefit from many BIOS updates that touch on cooling, low latency and NUMA awareness. Please contact your local Fusion-io Field Engineering to provide best guidance for your platform.

General BIOS Items that can help

- Increased Performance Profile
- Increased Cooling Profile
- Reduced C State switching level

Operating System

With each application certain tuning maybe required to gain the best performance from Fusion-io Flash devices. Many systems are optimized for disk performance which have an entirely different requirements than Flash systems. Please work with your Fusion-io Field Engineer or Customer Support to help provide guidance on what options are available to your application and OS mix.

- Some OS platforms such as Red Hat offer a tool such as ***tuned-adm*** that provide a series optimizations geared for you application and are offered as a series of application profile templates. These settings are intended to provide general profiles that can be applied to your application server OS to increase performance and reduce overhead. Different operating systems may need interrupt handler affinity and balancing tuned according to your Fusion-io engineer's recommendations.

NUMA - Simple

Forces I/O completions to happen on the same CPU that is running VSL processes for a particular device. This parameter is simple to implement (enabled or disabled), and is persistent. This parameter becomes increasingly important with the more sockets your server has populated. ***Large powerful multi-sockets boxes are those with 4 or 8 way sockets.***

Using the numa_node_forced_local Parameter

The numa_node_forced_local parameter is either enabled or disabled, and therefore does not offer as much user control as other options. It is persistent and it is enabled by modifying the /etc/modprobe.d/iomemoryvsl.

Example in iomemory-vsl.conf:

```
Options numa_node_forced_local=1
```

This parameter forces I/O completion for a particular device to happen on a CPU within the local numa_node that the other ioMemory VSL processes are running on (rather than trying to complete an I/O on the CPU that the host issued it on). Because the I/O completions are grouped with other ioMemory VSL processes, they are less likely to compete with processes from other device drivers.

This parameter may or may not improve performance depending on your configuration and workloads. You should test this parameter with your use case to determine if it improves performance.

VSL Tuning Options

VSL tuning is done by using module parameters and a table can be found in the User Guide PDF. Get more details from your local Fusion-io Field Engineer or Customer Support.

Each module parameter in the configuration file must be preceded by options iomemory-vsl.

The /etc/modprobe.d/iomemory-vsl.conf file has some example parameters that are commented out. You may use these examples as templates and/or uncomment them in order to use them.
Each module parameter in the configuration file must be preceded by options iomemory-vsl.

Description	**Module Parameter**
turn on interrupt coalescing	tintr_hw_wait=200
enable large pcie buffer	use_large_pcie_rx_buffer=1
use message signaled interrupts	disable_msi=0
bypass the kernel’s page cache	use_workqueue=0

Conclusion

Fusion-io ioScale Flash storage can be extremely beneficial for some SAS Workloads. Testing has shown it can significantly eliminate application IO latency, providing improved performance. It is crucial to work with your Fusion-io engineer to plan, install, and tune your ioScale devices to get maximum performance. The guidelines listed in this paper are beneficial, but general. Your individual experience may require additional guidance by Fusion-io depending on your system, and workload characteristics.

Resources

SAS Papers on Performance Best Practices and Tuning Guides: **http://support.sas.com/kb/42/197.html**

Contact Information:

Mauricio Borgen
Fusion-io
2855 E. Cottonwood Parkway, Suite 100
Salt Lake City, UT 84121
United States
Email: mborgen@fusionio.com

Tony Brown
SAS Institute Inc.
15455 N. Dallas Parkway
Dallas, TX 75001
United States
Work Phone: +1(214) 977-3916
Fax: +1 (214) 977-3921
E-mail: tony.brown@sas.com

Margaret Crevar
SAS Institute Inc.
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

THE POWER TO KNOW®

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2014, SAS Institute Inc. All rights reserved.