Paper 2020-2016

SAS® Grid Architecture Solution Using IBM Hardware

Whayne Rouse and Andrew Scott, Humana Inc.

ABSTRACT

This is an in-depth review of SAS® Grid performance on IBM Power8 Compute with IBM and Pure Storage™ solutions. This review spans our environment's growth over the last four years and includes the latest upgrade to our environment from the first maintenance release of SAS® 9.3 to the third maintenance release of SAS® 9.4 along with a hardware refresh.

Originally this environment was hosted on IBM compute with EMC storage and a monolithic installation of SAS® 9.2. This environment moved to a SAS® 9.3 Grid based solution in January, 2013. And now in 2016 we are launching SAS® 9.4 Grid on IBM Power8 Compute with a mix of IBM and PureStorage™ storage.

INTRODUCTION

This paper is an examination of our journey to SAS® 9.4 on new hardware, specifically IBM Power8 compute along with a mix of IBM DS8870™ and PureStorage™ Flash. The architecture places SASWORK and SASUTIL onto high-speed flash, with the main library data on conventional spinning disk. The principles and methods detailed here should apply to most upgrade projects. A good working knowledge of SAS® Administration and hardware infrastructure will be helpful in getting the most out of this paper. Two main topics are explored in this paper: First, learning about what is good or bad in the current system. It is important to understand the old before you build the new; Second, applying that understanding to building the new environment.

BEFORE YOU BUILD NEW, UNDERSTAND THE OLD

Humana's SAS® deployment grew from approximately 16 users in 2010 to over 300 in 2011. This environment was the initial installation of SAS® on AIX and introduced SAS® 9.2 to the enterprise. Scalability was already becoming an issue by the time this solution was rolled out. By the end of the 1st quarter of 2012 the user community had already grown to nearly 400 active users. SAS® 9.3 Grid was the solution picked to address these issues and allow for more scalability in the environment.

SAS®SOFTWARE REVIEW

The SAS® 9.3 Grid system was implemented to address three main items that we had not addressed in the SAS® 9.2 environment: Flexibility in the growth of the compute tier; dedicated resources for key business units; and updated SAS® configurations to better leverage the hardware..

During an upgrade that includes both hardware and software refreshes look at the latest How to Maintain Happy SAS®Users Document (Cervar 2015). This is to optimize the software configuration with the latest learnings from SAS® and to make more informed decisions in regards to hardware setup. For our solution we measured CPU, IO throughput, and active concurrent sessions. CPU and I/O throughput were monitored using Galileo Performance Explorer. Briefly, the SAS® 9.2 solution serviced about 390 interactive user sessions each day. The SAS® 9.3 solution now handles approximately 880 interactive sessions per day.

From the software layer the primary items (I/O rate and block sizes) were seen as critical areas for improvement. In the SAS® 9.2 environment the *bufsize* was set to default. The resultant behavior was SAS® picking the size of individual disk requests. In this case that size ended up being eight kilobytes (8K). We found this caused our I/O operations per second (IOPS) to be very high and contributed to, if not directly caused many issues seen in the hardware stack. After reviewing the file sizes in the system and comparing those to the SAS® recommendations we decided to use 256K as the new *bufsize*. This change reduced the number of IOPS and improved our CPU to IO throughput ratio. The net effect of these changes is described in more detail in the hardware review later in this document.

After analyzing our workload mix, 100 megabytes per second per core was used for the target I/O throughput for the SAS®9.3 Grid environment. The workload profile for the environment can be seen in Figure 1.

Percent of Work Load	Workload Description
10.19%	Advanced Analytics and Enterprise Miner Users
16.24%	Statistical Analysis
31.85%	Summary Report Builders
41.72%	Data Preparation and Distribution
100.00%	Total

Figure 1: Work Load Profile for target SAS® Environment

HARDWARE REVIEW

The original SAS® footprint at Humana in late 2011 was a single Logical Partition (LPAR) running on a Power6 based IBM model 595 computer. The LPAR started with just 8 CPUs assigned to it. By the end of 2011, demand for SAS® by the analytics community had driven the allocated CPU count to 32, with 128 gigabytes of memory assigned and sixteen four gigabit fiber channel adapters connected to an EMC Symmetrix VMAX.

Yet, with all these resources, the system struggled to attain total throughput of more than one gigabyte per second.

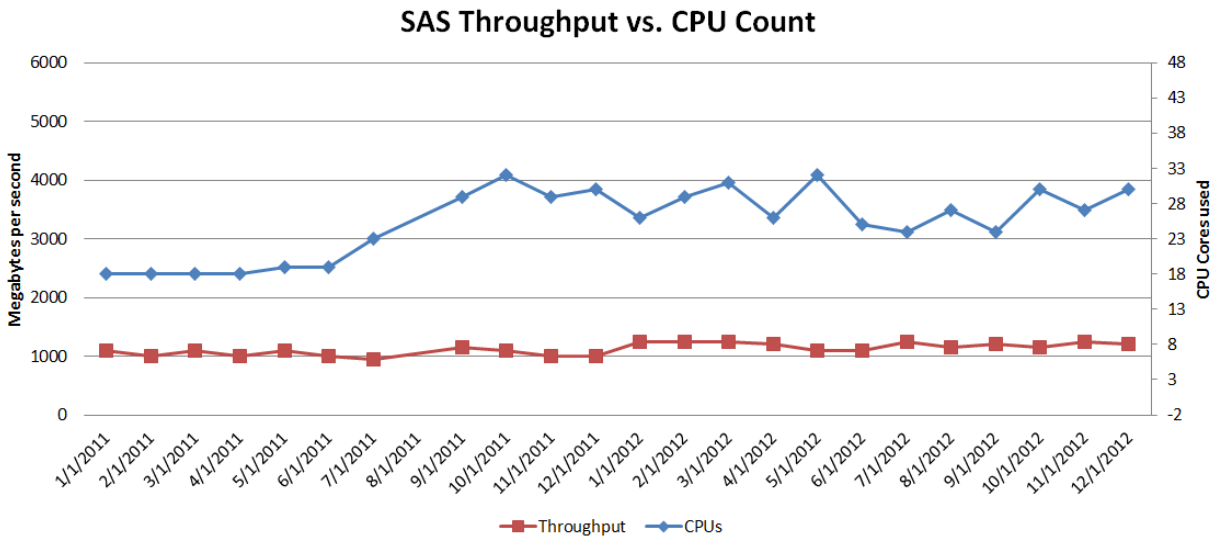

Figure 2: Throughput on single-instance SAS® partition plotted against CPU count

As you can see in Figure 2, even as CPU cores assigned to the partition increased dramatically, throughput did not increase. In fact, adding CPUs sometimes made things worse, as the system spent more time context switching than doing actual work.

Additionally, during this time, SAS® was sending work to the storage that consisted mainly of 8 kilobyte blocks with a smattering of 32K blocks being generated by SPDE. While throughput was painfully low, the high number of IOPS negatively impacted the back end storage, causing service interruptions on interactive systems that were sharing the same equipment.

The bottleneck that was capping overall throughput was AIX's own memory system combined with the sequential read behavior of SAS. AIX uses all available memory not consumed by application code for filecache. In environments with heavy sequential I/O loads filecache is quickly exhausted. The system must begin scanning the filecache for pages to discard in order to bring additional pages in from disk. The

portion of the operating system that handles this task in AIX is called the Least Recently Used Daemon (LRUD). This daemon is a multi-threaded kernel process that hunts through the system's memory for pages that can be discarded in order to make room for additional pages coming in from disk.

On this system, LRUD would consume four to eight cores by itself during prime shift. These were four to eight CPUs for which we were paying SAS® licenses that weren't actually doing any SAS® work. Additionally, all I/O read requests had to wait until LRUD could find pages to discard before they could be serviced. A better way to get data on and off the disk had to be found if SAS® was going to scale in our environment.

The solution brought to Humana was SAS® Grid for the application layer, with IBM hardware, EMC disk, and replacing the stock AIX JFS2 filesystem with IBM's General Parallel Filesystem (GPFS, now called Spectrum Scale™) version 3.5. By spreading SAS® work among many LPARs, the AIX memory subsystem bottlenecks that plagued large single-instance SAS® installations were mitigated, and GPFS's superior caching algorithms eliminated CPU waste by the LRUD.

The Grid system implemented at the beginning of 2013 was built of the following pieces:

- Two IBM Power7 based model 770 computers, each with 32 CPUs installed and 1 terabyte (TB) of memory each
- One EMC Symmetrix Disk Subsystem dedicated to SAS
- 16 active 4Gbps Fiber Channel connections (8 per 770)
- System divided into 14 Grid nodes, each with 4 CPUs and 120GB of memory
- GPFS page pool set to 64GB on each node

The effect on throughput in January of 2013 was substantial:

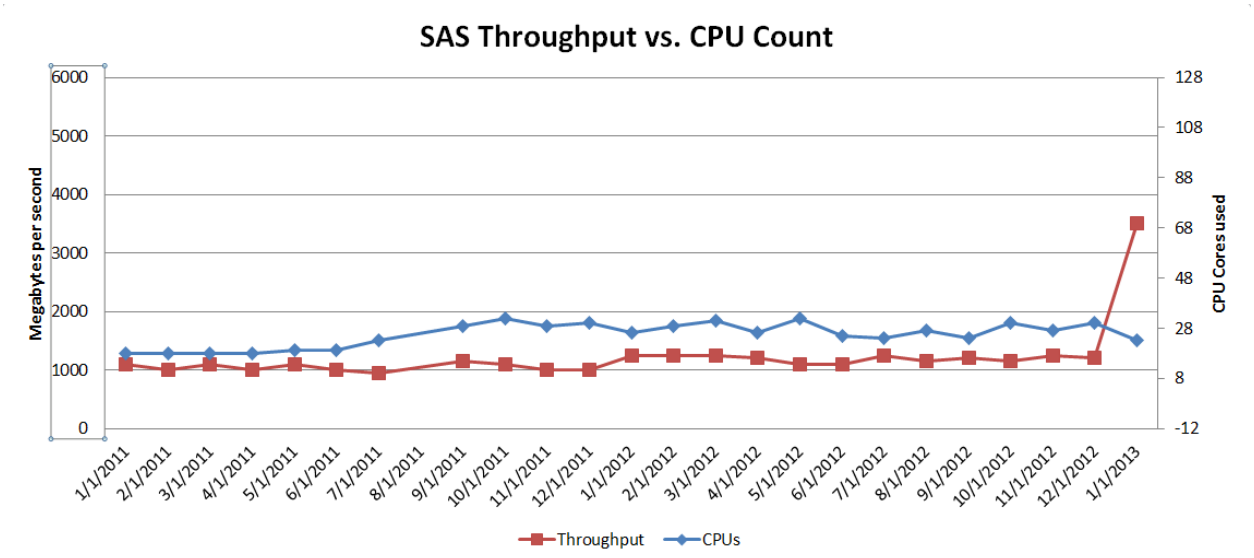

Figure 3: SAS®Throughput after implementation of GRID

As can been seen in Figure 3, after implementing grid we more than tripled I/O throughput to 3.5 GB/s while also cutting our CPU consumption from 32 to 24 cores.

SAS® usage continued to grow throughout 2013, and by mid-year, all resources on the new grid had been consumed. Additional grid nodes were added to the system, but disk throughput did not increase by a correspondent margin. The EMC storage backing the system was nearing its design limits and users were beginning to complain about jobs being suspended for long periods of time waiting for a run slot in the grid.

In late 2013, Humana acquired an IBM DS8870 disk subsystem to replace the EMC Symmetrix. This disk subsystem allowed us to run 8Gbps fiber channel connections, and provided a substantial increase in disk throughput capability. So much so that the system encountered several new bottlenecks:

1. GPFS Buffer splicing due to *scatterBuffers* being enabled – this cut all of our 256K blocks into 32 and 16K blocks, dramatically increasing IOPS and killing job throughput
2. CPU and memory locality inside the IBM 770 frames. As setting changes and dynamic operations had been performed on the LPARs that constituted the SAS® Grid nodes, the memory had become dislocated from the CPUs, which dramatically slows load/store operations from the CPU to main memory.
3. Excessive overhead inside the VIO and Hypervisor – the large amount of IOPS generated by the workload combined with the CPU load of SAS® itself created hypervisor context switching issues that hampered performance. The use of IBM's vscsi technology to virtualize the disk presented to the Grid nodes also resulted in high CPU consumption inside the virtual I/O server LPARS. We also observed high latency in the VIO disk driver layer.

To correct these issues, the following fixes were put into place:

1. Set *scatterBuffers*=no in the GPFS settings. This eliminated block splicing for sequential workloads. Our IOPs were cut by a factor of four, which reduced core consumption dramatically.
2. A full shutdown and careful startup of the two IBM 770 computers restored core-to-memory locality, ensuring each LPAR was assigned to the proper cores and the LPAR's memory was directly attached to the assigned CPUs
3. Switching of all partitions from Shared CPU allocation to Dedicated CPU allocation. This significantly reduced CPU consumption by the hypervisor and eliminated Virtual Context Switching inside the frames.

The results of these changes were encouraging. The testing in November and data migration in December produced amazing throughput numbers of nearly 10GB/s. Throughput running actual SAS®workloads settled in about 7GB/s afterwards.

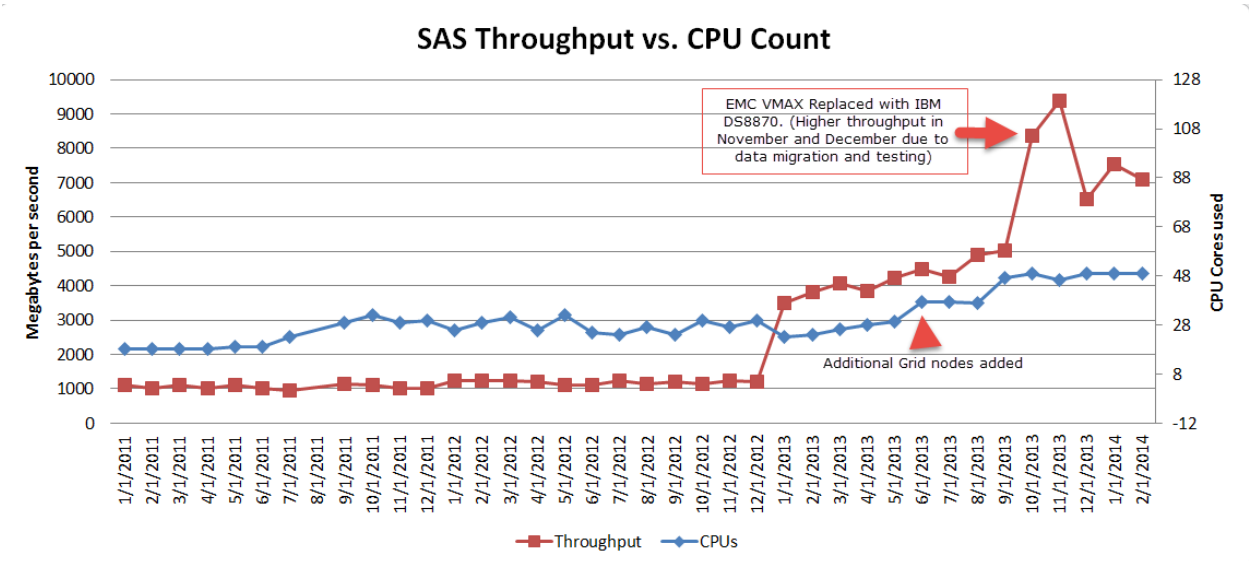

Figure 4 Throughput (in MB/s) against CPU consumption, through February, 2014

The system was now routinely exceeding 7.5GB/s of throughput using just 48 cores of CPU.

However, this capacity proved to not be enough to satisfy additional demand from the users. Design work began on a new grid that would have the capacity to meet the pent up demand from the now 600+ users

as well as move the storage architecture from a monolithic single storage device to a grid storage architecture that could grow with the compute layer.

BUILDING FOR THE FUTURE

In the prior part of this paper we discussed where we came from. Now we will share how we used SAS®9.4m3 to address the issues and continue to evolve the hardware solution.

HARDWARE MIGRATION

The new Grid design coalesced during the spring and summer of 2015. Working with SAS® and IBM, the architecture settled around a system built on IBM's new Power8 CPU architecture and a combination of DS8870 spinning disk combined with PureStorage™ FA-m50 flash arrays.

Several items were addressed:

1. Grid Node Sizing. The first iteration of our grid utilized grid nodes with four cores and 120GB of memory each. After working with IBM and SAS®, a grid node size of eight Power8 cores with 488GB of memory each was chosen. Each node would run a 256GB GPFS page pool, with the remainder available for SAS. The Nodes were placed on IBM S824 model servers, with two grid nodes per physical server. The environment was also completely de-virtualized. Each LPAR had its own physical 16Gbps fiber channel and 10Gbps network cards. The IBM Virtual I/O partitions were eliminated.
2. Filesystem Layout. In the first grid, all SAS® data, including SASWORK, SASUTIL, and base library data was lumped together in a single filesystem that was spread across the entire back-end storage array. This layout resulted in challenges for the back end storage. The workload mix prevented the DS8870 from properly recognizing sequential requests, which resulted in poor prefetch and a high cache miss rates on the storage. In the new iteration, SASWORK/SASUTIL were broken out into separate filesystems and placed on the flash arrays. The SAS®Scalable Performance Data Server® Data was also split into separate filesystems and placed alongside the base data on the DS8870 spinning disk system.
3. GPFS was used to stripe data across multiple Pure Flash Arrays, enabling us to use the filesystem to overcome bottlenecks in any individual storage device. In the future, when we add additional compute capacity to the system, we will also be able to add storage capability. The Pure Flash arrays are modestly priced for their size and capability, and allow us to grow the system along with the compute without hitting a huge step function in hardware cost in the future. GPFS Metadata replication was enabled to prevent filesystem corruption in the event one of the flash devices went out of service unexpectedly.
4. Logical Unit Number (LUN) sizing. In our first grid implementation, we used four terabyte LUNs. Since that implementation, we have learned that fewer large LUNs, while attractive from a storage management standpoint, actually hurt our performance. There are data structures and queues associated with each LUN at the OS and back-end storage level. The thinking that "the storage system will stripe me and I'll get maximum performance regardless of LUN size" is not true. To take advantage of the high degree of parallelism that modern compute and storage hardware can provide, you must break the workloads up into as many small parts as possible. In our instance, we went from 4TB LUNs in the first grid to 1TB LUNs in this system. This provided four times as many device buffers per 4TB block of storage as the first system.

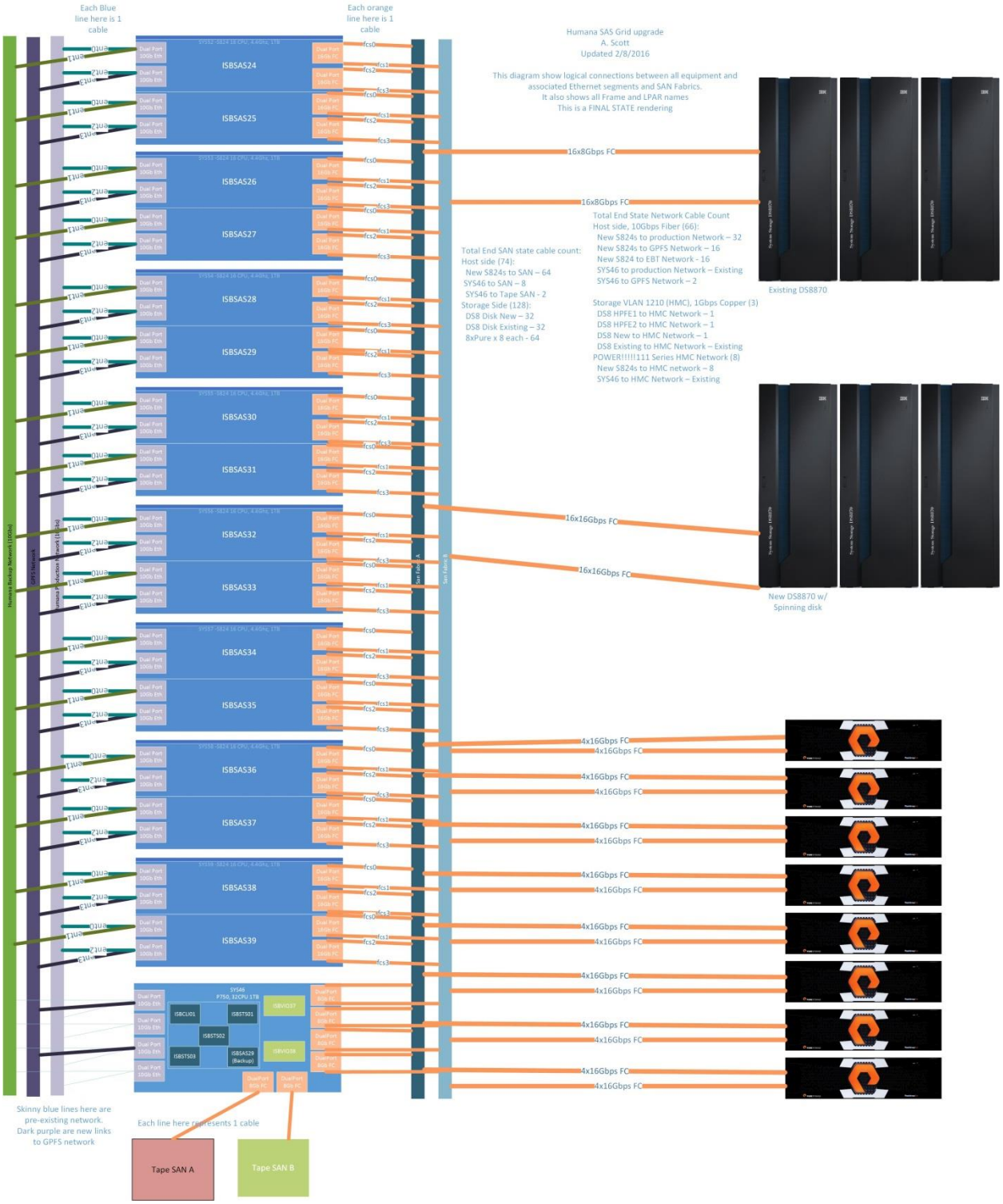

Figure 5 - New SAS®Grid architecture

Initial testing on the new platform was positive. Peak I/O rates of above 75GB/s were observed while running a benchmark detailed in the *SAS® Grid Manager – Testing and Benchmarking Best Practices for SAS®Intelligence Platform*, with steady-state throughput above 30GB/s for the duration of the test.

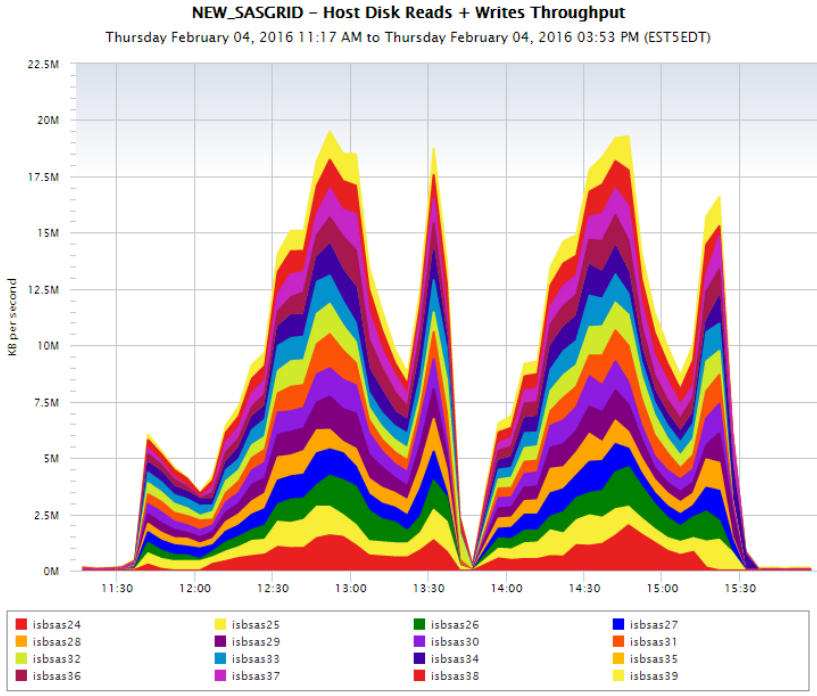

Figure 6 - Disk I/O throughput test results using SAS® benchmark

CPU usage during the test was also encouraging.

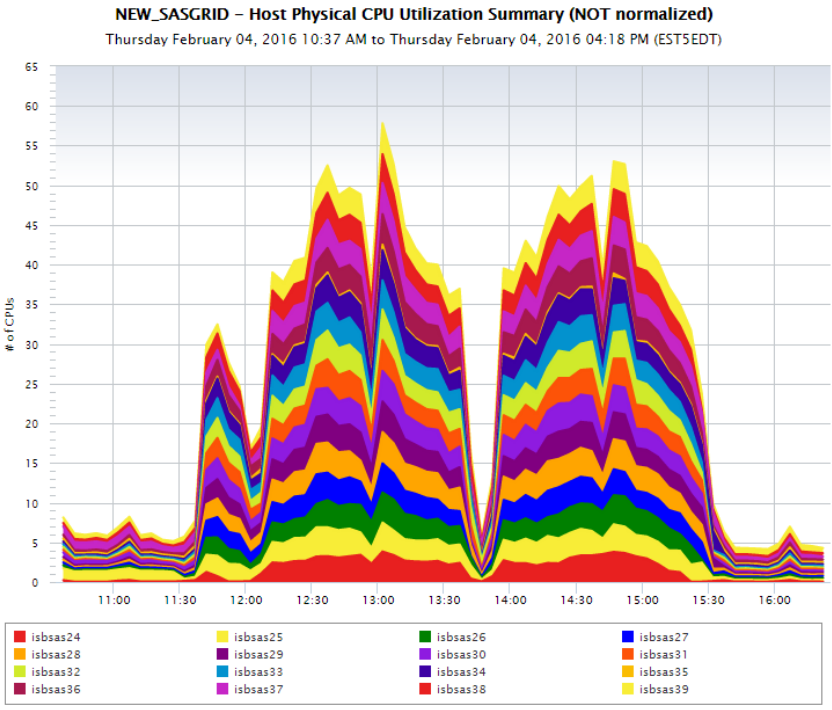

Figure 7 - SAS®CPU consumption during SAS® benchmark

As Figure 7 shows, of the 128 CPUs available, fewer than 50 were needed to generate maximum disk throughput with this particular SAS® job, indicating system overhead for 75GB/s of disk I/O running a minimally computationally intensive job was just 39% of available CPU. At maximum throughput, the grid has ample CPU headroom for actual statistical computation. Recast in a chart consistent with tracking of previous charts in this paper, the Throughput vs. CPU comparison appears in Figure 8.

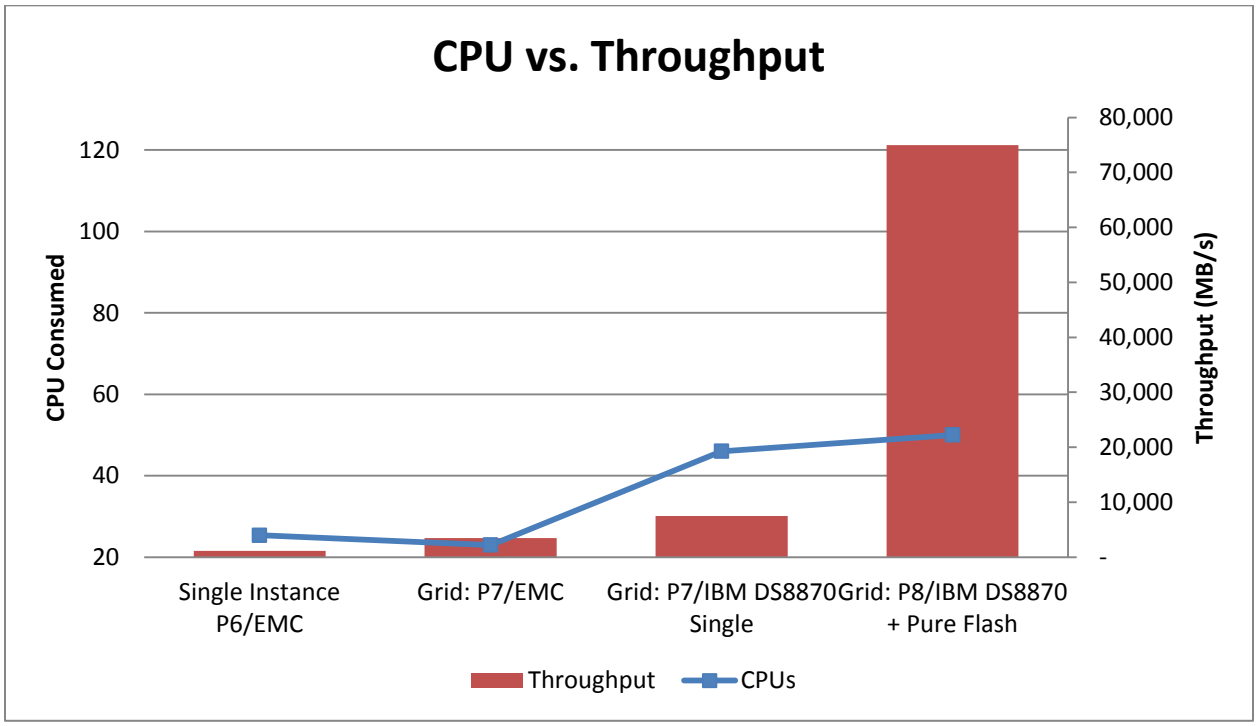

Figure 8 - Comparison of CPUs used with throughput attained in each platform

From our testing, we achieved an enormous increase in throughput with addition of the Pure Flash and Power8 hardware.

SOFTWARE AND USER MIGRATION

To move the data from the old SAS®9.3 Grid to the new SAS® 9.4 Grid we used an Open Source tool called RSYNC to move datasets. Rsync is a file synchronization tool that works at the block level. After an initial copy is made rsync allows that copy to be kept in sync with the source. Changes are replicated at the block level instead of the file level, minimizing the data that must be transferred. Rsync took approximately three weeks to move 300TB of data files into the new environment for the initial copy. After this we were able to keep the data in-sync every day using the rsync process with four of the server nodes to doing the work.

We achieved this by cross mounting the new grid environment with the old environment using GPFS to mount the legacy environment locally to the new cluster. Filesystem communication occurred over the network between the two clusters.

Rsync was then able to make direct calls through the file system. This streamlined the approach by not having to deal with SSH between servers. This also allowed us to run auditing reports against the copies in the new environment to minimize load on the old system.

Moving data using the GPFS protocol can tax a network. GPFS allows network I/O to filesystems to be spread across many individual hosts, allowing large aggregate throughput, often much higher than you can get with a simple CIFS or NFS mount. This must be taken into account before using this option.

We followed the SAS® Enterprise Miner documentation for project migration found here http://support.sas.com/kb/32/904.html. This support note is for moving Enterprise Miner projects on the same server, but the concept is the same using the rysnc process to move the project to the new location

in the new environment. Then, follow the information in step 2 in the migration document to create the Enterprise Miner project on the new SAS® Metadata server.

To move the remaining metadata we followed the Promotion Tools document in the SAS® 9.4 Intelligence Platform: System Administration Guide, Fourth Edition.

CONCLUSION

To bring things into final perspective, here is the chart we have been following through this document, with data through 2016 and the new grid's numbers from January, 2016:

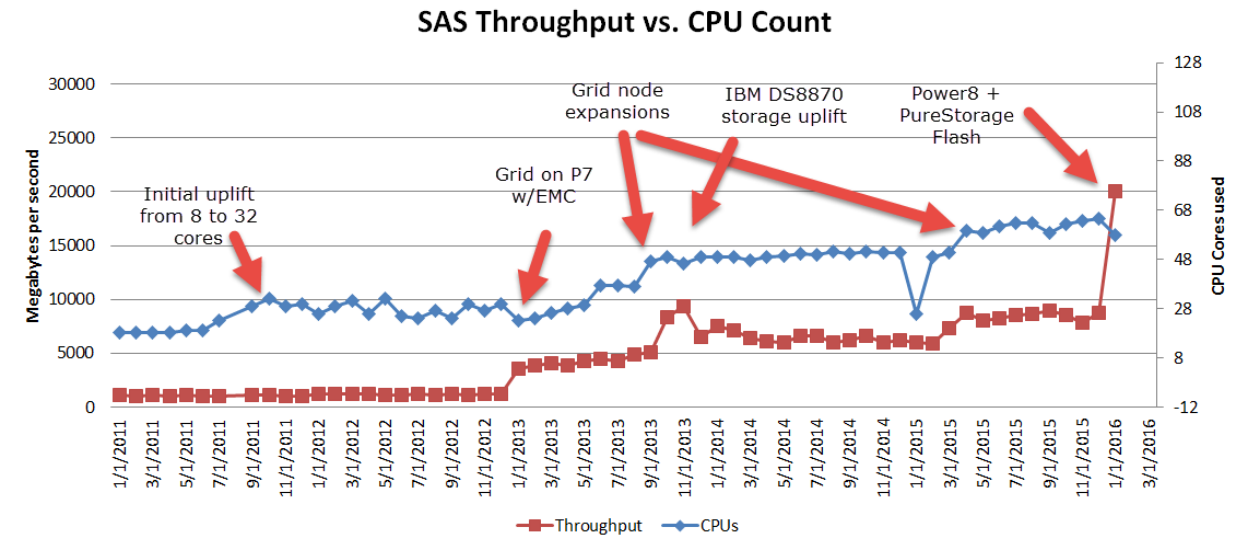

Figure 9: SAS® Throughput since 2011

Using *seven fewer* cores than we were using under Power7 with a single disk subsystem (a CPU decrease of 14% in terms of core counts) we realized a 230% increase in disk throughput for SAS® workloads. This is extremely significant when viewed in the light of SAS® software costs. The SAS® software is by far the most expensive component of this system. Wise choice of hardware and careful design can realize enormous efficiencies when running SAS® software.

By paying close attention to a few parameters, large improvements in SAS® application throughput can be attained above a "stock" configuration:

- Filesystem block size: align this with the environment's SAS® *bufnum* parameter. This will streamline I/O by not generating excessive CPU usage or I/O operations that come with having the OS handle a mismatch between the block size the application requests and the block size the filesystem is set up to handle.
- GPFS *maxMbps* parameter: This is a software throttle implemented to help GPFS not overrun back-end storage. The default value is 6000. IBM recommends setting this number to double what the node + storage combination is physically capable of. This parameter is set on a per-node basis, so be sure that the sum of this parameter for all nodes is not more than double what your back-end storage can handle.
- GPFS *scatterBuffers* parameter: For large sequential workloads like SAS, this parameter should be set to "no".

The items are relevant when running SAS® Grid on enterprise and commodity equipment alike.

On the hardware side, it is important to balance the throughput needs, software cost, and equipment cost. Commodity Intel or AMD based equipment is inexpensive on a per-unit basis, but you need a lot of units to get the kind of throughput available with higher-end hardware. Often, and especially with expensive software that is licensed per core like SAS, the cost of the software licenses to run on hundreds of

Intel/AMD cores will easily exceed the cost of better hardware that can do the same work with fewer software licenses. In our case, the IBM Power8 architecture provided an enormous boost in the amount of I/O that could move through each compute node. Each of these model S824 servers has proven the ability to run eight 16Gbps fiber channel links at wire speed simultaneously with compute capacity to spare inside the system.

TABLE OF FIGURES

REFERENCES

Margaret Crevar. 2015. *"How To Maintain Happy SAS® 9 Users." Proceedings of the SAS® Global 2015 Conference*, Cary, Nc : SAS® Institute Inc. http://support.sas.com/resources/papers/proceedings15/SAS1480-2015.pdf

SAS® Support Cary, NC. "*Usage Note 32904: Migrating SAS® Enterprise Miner™ projects to a new server on the same operating system.* October, 2015 http://support.sas.com/kb/32/904.html

SAS® Support Cary, NC. "*SAS® Grid Manager – Testing and Benchmarking Best Practices for SAS® Intelligence Platform*". https://support.sas.com/rnd/scalability/grid/grid_testingbench.pdf

SAS® Support Cary, NC. *SAS® 9.4 Intelligence Platform: System Administration Guide, Fourth Edition* http://support.sas.com/documentation/cdl/en/bisag/68240/HTML/default/viewer.htm#p04l8c3fo8tpyhn103dps7pxvru2.htm

IBM Corporation. 2012. *General Parallel Filesystem Advanced Administration Guide.* Aramonk, New York : IBM Corporation. Available at http://www-01.ibm.com/support/docview.wss?uid=pub1sc23518205

Rsync. (n.d.). Retrieved February 16, 2016, from http://rsync.samba.org/

ACKNOWLEDGMENTS

We would like to acknowledge IBM, SAS, and The ATS Group for their assistance in reviewing the design of this environment, providing feedback, and invaluable long-term performance monitoring capability

RECOMMENDED READING

- *How to Maintain Happy SAS®9 Users published by SAS® Institute Inc.* http://support.sas.com/resources/papers/proceedings15/SAS1480-2015.pdf

- IBM Corporation. 2012. *General Paralle l Filesystem Advanced Administration Guide.* Aramonk, New York : IBM Corporation. Available at http://www-01.ibm.com/support/docview.wss?uid=pub1sc23518205

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Whayne Rouse
Humana Inc.
wrouse1@humana.com

Andrew Scott
Humana Inc.
ascott12@humana.com