A new software-defined converged infrastructure for SAS Foundation mixed workloads

IBM ESS storage, IBM Spectrum Scale, IBM POWER8, and Mellanox deliver excellent performance for SAS analytics

Highlights

- IBM Elastic Storage Server delivers exceptional performance for SAS Mixed Analytics.
- Mellanox proves the viability of Ethernet as a storage fabric with same scale and performance as Fiber Channel.
- IBM Power E880 shows high performance analytics capability.
- A converged infrastructure architecture provides a cost-effective solution for SAS.
- Peak file access from storage is 16 GBps – equivalent to the disk subsystem maximum for a workload with a read:write ratio of 65:35.
- Completion time for four concurrent 30-test mixed analytics workloads can be reduced by 12.5% at 56GbE speed compared to 40 GbE speed.
- File-based storage solutions meet the demands of the large SAS Mixed Analytics workloads.
- Moving from a 40 GbE network to a 56 GbE network is a simple Mellanox switch port change and network restart. The 56 GbE capability is included at no additional charge.

Today in the analytics space, the traditional method to deliver storage is through Fiber Channel (FC), with storage area network (SAN)-attached disk, or more recently flash-based storage arrays. This paper demonstrates another option; Ethernet-attached file based storage also known as *converged infrastructure*. Teams from SAS, IBM®, and Mellanox applied a methodology to not only tune each component in the infrastructure horizontally, but also tune the solution vertically, which allowed the optimal performance to be achieved. This technical brief documents the goals, results, and supporting information.

Goals

- Demonstrate overall performance of a converged infrastructure running a demanding SAS Mixed Analytics workload.
- Prove that using Ethernet as a storage fabric is viable and helps in achieving the same or better performance when compared to traditional fabrics such as Fiber Channel or InfiniBand®.
- Show a shared file system, such as IBM Spectrum Scale™ running with IBM Elastic Storage Server (ESS) storage can optimize storage demands in a multi-host environment when applications are demanding I/O storage requests in a converged environment.
- Measure SAS Mixed Analytics workload throughput improvement when scaling from 40 Gigabit Ethernet (GbE) to 56 GbE storage connectivity.

Results (See Figure 1)

- The 20-test mixed analytics workload was not resource constrained; thus, no change was needed from 40 GbE to 56 GbE.
- 8% improvement in SAS Mixed Analytics cumulative response time with no processor utilization change at 56 GbE when 30 tests are scaled to four nodes.
- I/O throughput indicates the four-node 30-test mixed analytics spinning disk drive performance limitations.
- GbE network is capable of more than 20 GBps with no disk constraints.
- 30 tests at 56 GbE was disk constrained, while no processor, memory, or network constraints.
- Ran 30-test mixed analytics workload with 62 cores across four nodes with 96 GB memory per node, and one 40 GbE / 56 GbE network interface controller (NIC) per node.

Converged infrastructure architecture

The converged infrastructure was created using an IBM ESS GL4 storage system, a multicore IBM Power® E880 server, the IBM Spectrum Scale 4.2.1.1 file system, and the IBM 8831-NF2 Ethernet switch made by Mellanox plus the Mellanox hi-speed Ethernet adapters used between the server and storage to the switch. The main software building blocks are the IBM AIX® operating system and the IBM Spectrum Scale shared file system. The test bed employed was the SAS Mixed Analytics workload based on the SAS 9.4 M3 Foundation platform. This combination creates a powerful system with enterprise capabilities allowing for scale-out architecture, affordable storage performance, and storage fabric scalability. Refer to Figure 2 for an architecture picture depicting the software-defined converged infrastructure.

Server

The **IBM Power E880 server** is based on the IBM Power Architecture®. The architecture uses the concept of logical partitions (LPARS), which allow one or more cores in the system to be logically organized. These LPARS constitute the nodes used to run the workload. The diagram in Figure 2 has four LPARS. LPARS 1 and 2 have 16 cores each. LPARS 3 and 4 have 15 cores each. All cores have 96 GB of memory. All LPARS have one configurable 40 GbE / 56 GbE port connection. The operating system was AIX 7.2. The cores were in dedicated mode running SMT4.

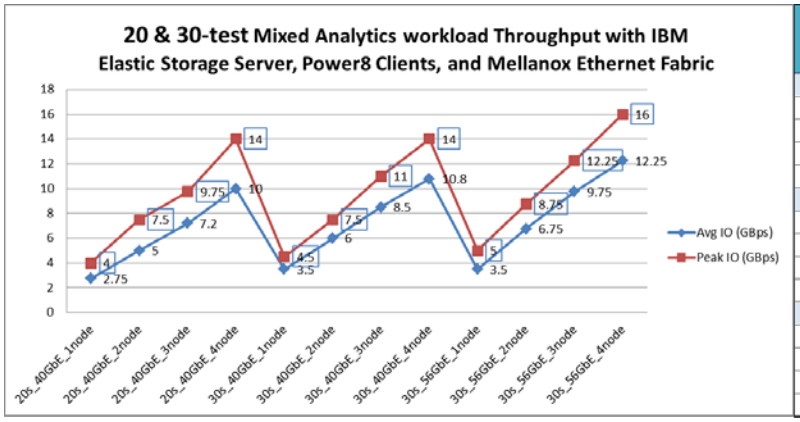
Reference architecture

Software

- SAS 9.4 M3
- IBM AIX 7.2
- IBM PowerVM
- IBM VIOS 2.2.3.50
- IBM Spectrum Scale 4.2.1.1
- IBM ESS 4.5.1
- MLNX_OS 3.3.6.1002

Hardware

- IBM Power E880 server (Model 9119-MHE)
- IBM ESS (Model 5146-GL4)



Mixed Analytics workload test streams	Test	Avg IO (GBps)	Peak IO (GBps)	CPU Utilization
	40GbE Storage Fabric			
20-test	1 node	2.75	4	60%
40-test	2 node	5	7.5	60%
60-test	3 node	7.2	9.75	60%
80-test	4 node	10 GB	14 GB	60%
	40GbE Storage Fabric			
30-test	1 node	3.5	4.5	75%
60-test	2 node	6	7.5	75%
90-test	3 node	8.5	11	75%
120-test	4 node	10.8	14	75%
	56GbE Storage Fabric			
30-test	1 node	3.5	5	75%
60-test	2 node	6.75	8.75	75%
90-test	3 node	9.75	12.25	75%
120-test	4 node	12.25	16	75%

Figure 1: 20, 30-test mixed analytics workload results at 40 GbE and 56 GbE speeds represented in graph and table formats

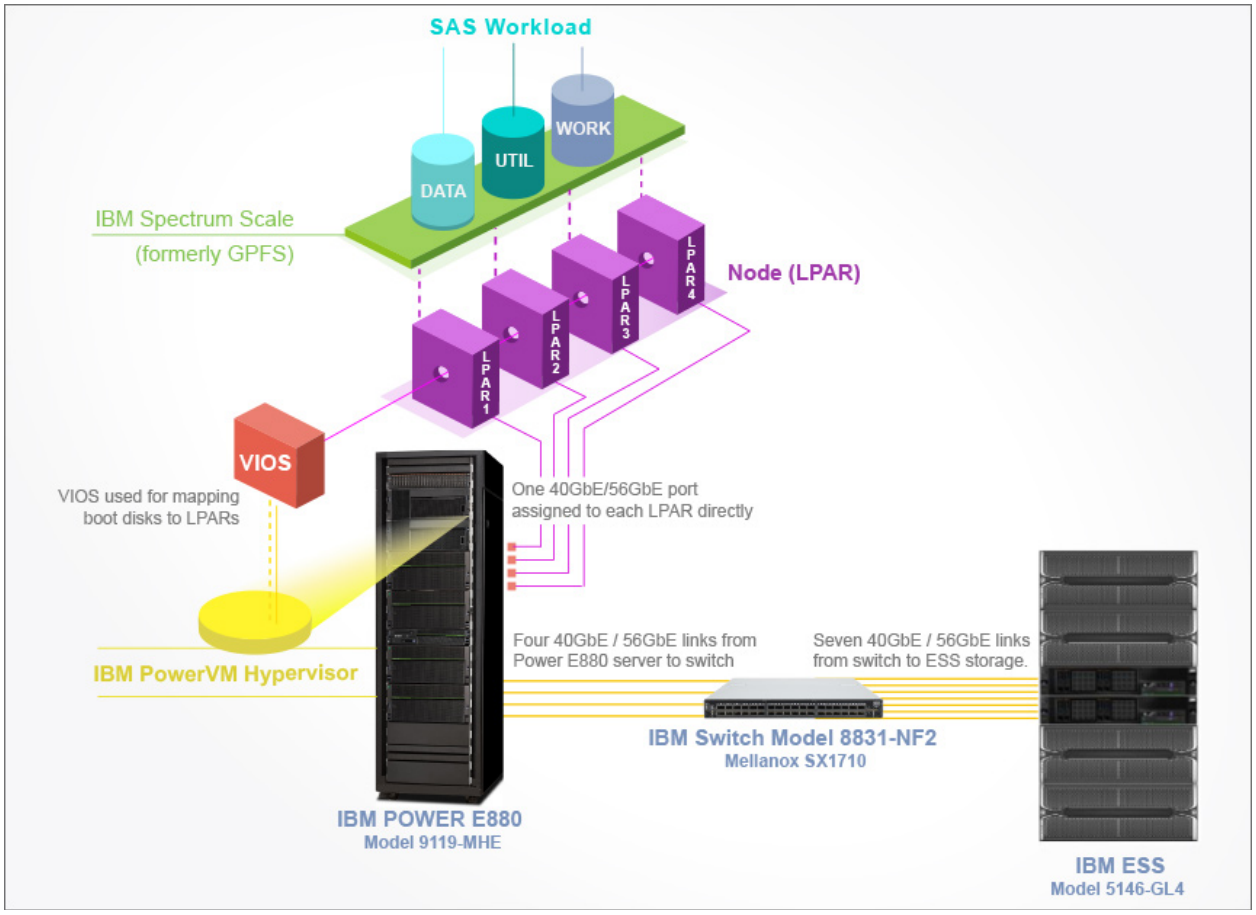


Figure 2: Converged infrastructure architecture for SAS Foundation with IBM Spectrum Scale ESS, IBM POWER8, and Mellanox Ethernet Fabric

Storage

The IBM Spectrum Scale ESS combines the processor and I/O capability of the IBM POWER8® architecture matched with IBM System Storage® assets. Together they provide a platform for a multitier storage architecture enhanced with IBM Spectrum Scale to manage block, file, and object data in a shared file system environment. The capabilities of the IBM ESS include: proprietary device pool management, software Redundant Array of Independent Disks (RAID), large cache, and scalability. IBM ESS systems are delivered as an integrated package with the hardware/software stack validated. The system comes with IBM Spectrum Scale pre-installed.

The **IBM ESS Model 5146-GL4** used in the test has four 60-drive just a bunch of disks (JBOD) enclosures. Each enclosure has 58 4.2 TB HDDs plus two 400 GB solid-state drives (SSDs) for a total of 240 drives. The predominant storage is near-line spinning disks with a raw capacity of about 246 TB. IBM ESS uses two IBM Power S822L storage servers and one IBM Power S821L management server. IBM Spectrum Scale is the storage cluster management software. Performance is more important for SAS workloads than capacity, thus 7 Gb Ethernet ports were connected to the switch.

Networks

Typically, Ethernet is not the first choice in storage fabrics. Traditionally, the choice when running an analytics style workload, such as the SAS Mixed Analytics workload, has been Fiber Channel for block I/O and possibly InfiniBand for file I/O. Some (or possibly many) have experienced difficulties with getting Ethernet working correctly as a storage fabric.

But the highly configurable IBM 8831-NF2/Mellanox SX-1710 switch has accomplished this task rather well. With its 36 ports capable of running 40 GbE and 56 GbE and latencies as low as 220 nanoseconds, it is a perfect complement to the IBM ESS GL4 with IBM Spectrum Scale.

The **IBM Switch Model 8831-NF2 / Mellanox SX1710** is capable of operating as a 40 GbE or as a 56 GbE switch. In the test, it was found that the 20-test mixed analytics workload could be run in the 40 GbE mode. The Mellanox Connect X-3 adapters in the configuration allow the fabric to be easily uplifted to a 56 GbE modality by simply changing a software-only setting. This is a benefit of a software-defined converged infrastructure. Using 56 GbE enabled the 30-test mixed analytics workload with four nodes to hit the switch limits and the IBM ESS hit its disk subsystem limits. See Figure 2 for more information on the network architecture.

Shared file system

IBM Spectrum Scale is built on the IBM General Parallel File System (IBM GPFS™) technology. It is a powerful data management system that enables the unification of block, file, and object storage into a single comprehensive solution for a project or the entire data center.

IBM Spectrum Scale 4.2.1.1 is the IBM ESS cluster management software. The Spectrum Scale file system parameters were tuned and used to create the SASWORK, SASDATA, and SASUTIL application storage space. With the SAS BUFSIZE set to 256 KB, testing the SAS workload with various file system block sizes determined that an 8 MB and 16 MB file system block size performed best. The 8 MB block size was chosen and SASWORK, SASDATA, and SASUTIL were configured with 16 TB each.

Note again that throughput that supports large block sequential IO and not capacity is the most important factor when configuring storage for SAS performance. Even though IBM ESS has hundreds of terabytes of capacity, the 16 TB sizes provided enough shared storage for four LPARS to run the workload.

SAS Foundation software

The software that was tested is SAS 9.4 M3 (the latest version at the time). The test suite that drove the work in order to measure performance was the SAS Mixed Analytics workload. The SAS Mixed Analytics workload is used to provide a means to run many tests on a system. The 20-test and 30-test mixed analytics workload scenarios were best suited for the test goals.

SAS Mixed Analytics test suite

The SAS Mixed Analytics workload consists of a mix of jobs that run in a concurrent and back-to-back fashion. These jobs stress the compute, memory, and I/O capabilities of the infrastructure. The SAS test team described the test bed they employ as a *good average SAS Shop* set of workload mix.

The SAS Mixed Analytics workload can be scaled to a number of concurrent tests to test the size of the system. For the tests discussed in this paper, a 20-test and a 30-test mixed analytics workloads were conducted. As an example, the SAS Mixed Analytics 20-test test workload consists of 20 individual SAS tests: 10 compute-intensive, 2 memory-intensive, and 8 I/O-intensive. Some of the tests rely on existing data stores and some tests rely on data generated during the test run. The tests are a mix of short-running (in minutes) and long-running (in hours) jobs. The tests are repeated to run both concurrently and in a serial fashion to achieve a 20-test workload. For the tests run in this proof of concept, when the single node 20-test workload was completed it had run a total of 71 jobs. There is of course a similar scaling of the 30-test workload where 101 jobs in total were run.

Data and I/O throughput

A single instance of the SAS Mixed Analytics 20 simultaneous test workloads on each node inputs an aggregate of about 300 GB of data for the I/O tests and about 120 GB of data for the computation tests. Much more data is generated as a result of test-step activity and threaded kernel procedures.

It is important to note that SAS I/O pattern is predominately large-block, sequential I/O. There is some random access but sequential is the dominant access pattern. When configuring for SAS I/O, there are multiple distinct patterns such as large sequential workloads in the multi-gigabyte to terabyte size, small file sequential, random access, and random data step activity. However, it is the large sequential block I/O that dominates all of these patterns. Keeping that in mind helps in configuring the file systems.

SAS filesystems utilized

There are three primary file systems involved in the testing:

- SAS permanent data file system - SASDATA
- SAS working data file system - SASWORK
- SAS utility data file system - UTILLOC

The file systems were created using a SAS BUFSIZE of 256 KB. Preliminary testing showed that IBM ESS performed best with a block size of 8 MB and also 16 MB. The 8 MB block size was chosen and each file system was created with 16 TB of capacity.

Testing the converged infrastructure

Many test cases were performed (varying the LPARs, number of tests, GbE size, and so on) and the tests were focused on the scenarios shown here.

Using a 40 GbE storage fabric:

- One-node 20 tests
- Two-node 20 tests
- Four-node 20 tests (no limits hit in storage, server, networks)
- four-node 30 tests (40 GbE network hitting limits)

Using a 56 GbE storage fabric:

- Four-node 30 tests (IBM ESS GL4 hitting limits)

This paper highlights the results from the two most important tests (the four-node tests with 40 GbE and 56 GbE). The first of these tests shows networking constraints (using 40 GbE) and the second one shows storage constraints (using 56 GbE).

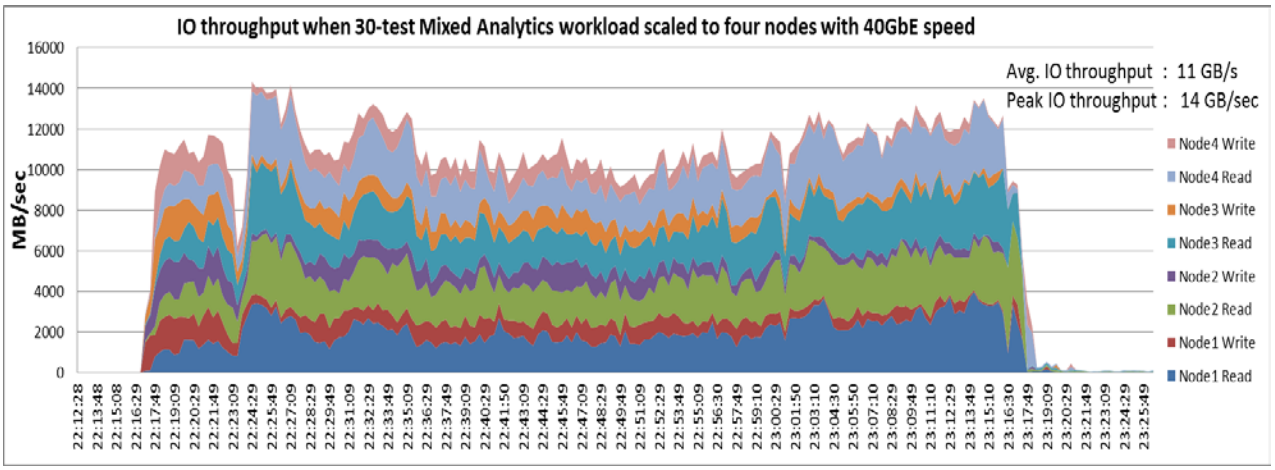


Figure 3: I/O throughput when the 30-tests mixed analytics workload scaled to four nodes (total 120 tests) with 40 GbE speed

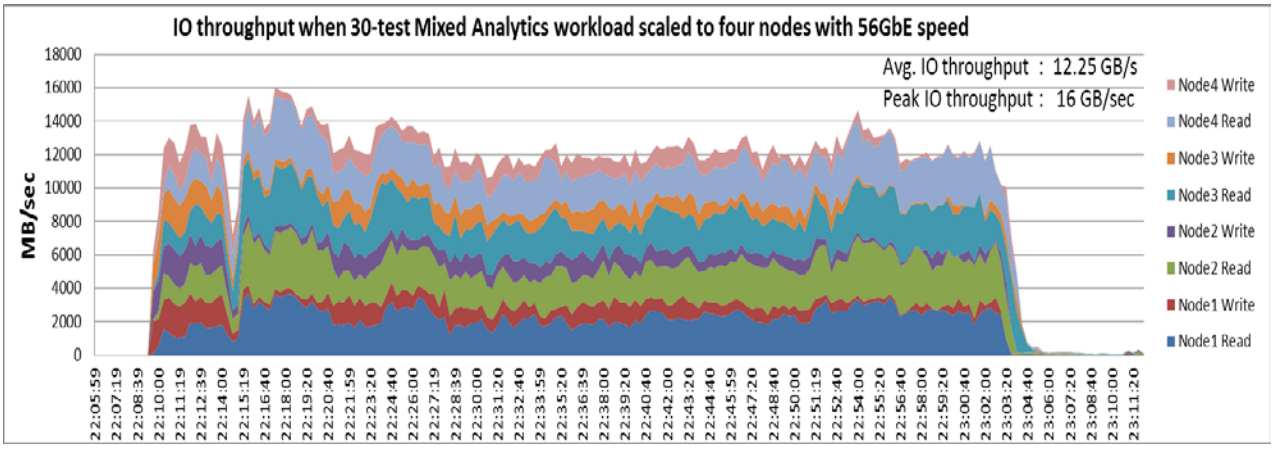


Figure 4: I/O throughput when the 30-tests mixed analytics workload scaled to four nodes (total 120 tests) with 56 GbE speed

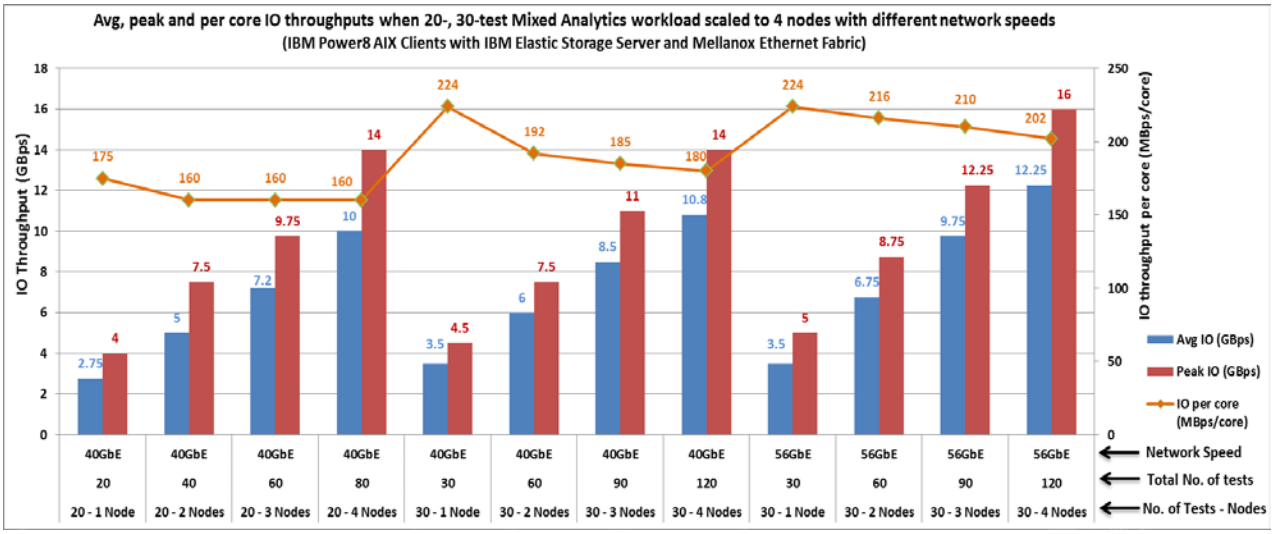


Figure 5: One to four node I/O throughput scaled (total 120 tests) from 40 GbE to 56 GbE speed with I/O throughput per core

Four-node 30-tests mixed analytics workload with 40 GbE (120 tests / 404 jobs total) 69 minutes

The I/O throughput chart from the converged infrastructure (see Figure 3) shows that a four-node workload running 30 tests completed in about 69 minutes. The workload quickly exceeds 11 GBps in total for initial input read operations and for initial SASWORK write operations. Peak I/O during the workload run was about 14 GBps. The test suite is highly active for about 60 minutes and then finishes with two low-impact, long-running *trail out* jobs. This is what the SAS team describes as a good average *SAS Shop* throughput characteristic for a single node instance that simulates the load of an individual SAS COMPUTE node. The throughput is depicted from all three primary SAS file systems SASDATA, SASWORK, SASUTIL. Estimated I/O ratio for read and write operations was 65:35.

Observation: 40 GbE network is constrained

An interesting aspect of this workload was it illustrated that the first real constraint of the system had been achieved. The 40 GbE network at a peak I/O of 14 GBps had reached its practical limits. To gain more I/O throughput, the network would need higher performance. In summary, servers and storage were not stressed and had head room for more workload. The network was stressed and was a limiting factor in the performance

Storage fabric remediation - Network constraints

The normal practice in the event of network capacity issues is to add another adapter and aggregate (bond) these links together. This requires extra slots and requires placement considerations in the server. However, because the network components were an all Mellanox solution (switch and adapters) the ability to run at 56 GbE was available. So, without any cost in money and using Mellanox products available in the IBM portfolio, the storage fabric capacity was increased by 40%. The process to enable 56 GbE was to set the speed on the interfaces to 56000 and reboot the switch and the LPARs. All ports on the ESS and the four LPAR ports were auto configured to 56 GbE in preparation for the next test. This is a great example of how SDS in practice can benefit the system.

Four-node 30-tests mixed analytics workload with 56 GbE (120 tests / 404 jobs total) 61 minutes

With a higher capacity storage network, the 30-tests workload was repeated. The I/O throughput chart (see Figure 4) shows that the four-node workload completed in about 61, minutes which is a 12.5% reduction (from 69 minutes). The workload quickly exceeds 12 GBps for initial input read operations and for initial SASWORK write operations (a 9% improvement) and has a peak I/O of 16 GBps (14% improvement). The test suite is highly active for about 55 minutes and then finishes with two low-impact, long-running *trail out* jobs. The throughput is depicted from all three primary SAS file systems SASDATA, SASWORK, SASUTIL. The estimated I/O ratio for read and write operations was 70:30 (an improvement from 65:35).

Observation: Network improved - Storage approaching limits

When the storage fabric increased in speed by 40% (40 GbE to 56 GbE) it was anticipated that the system would improve accordingly. An increase in performance was there but not as high as expected. Further investigation indicated the IBM ESS system had approached its performance limits and became the limiting factor in achieving better performance. The storage was stressed due to achieving the maximum I/O limit of the nearline SAS drives in the ESS enclosures.

Storage fabric remediation - Disk subsystem constraints

The normal practice in the event of storage performance is to add more units. This would mean upgrading IBM ESS from a GL4 model to a GL 6 model. However, the team did not have access to additional resources. It was at this point that the team determined this constraint and provided the conclusion of the testing.

Summary

Figure 5 shows a summary view of the test results for the different node configurations and GbE configurations. The test results demonstrate that this particular software-defined converged infrastructure is viable and performant when used with SAS software.

The Mellanox Ethernet high speed storage network was crucial in facilitating the IBM ESS full I/O throughput. The ability to change from 40 GbE to 56 GbE fabric at no additional cost is a key capability that can be used by SAS environments.

The IBM ESS storage server with its JBOD technology proved to be an effective storage solution for SAS workloads. Its performance was comparable to the more expensive mid-tier Fiber Channel attached flash storage but without the higher costs making it more cost effective.

The IBM Power E880 server proved to be a powerful work engine meeting the needs of the SAS Mixed Analytics workload. The processor, throughput, and memory limits were not seen during this test which shows the vitality of this system.

Get more information

To learn more about the IBM, Mellanox, and SAS products and capabilities, contact your IBM representative or IBM Business Partner, or visit the following websites:

- IBM ESS GL4 www.ibm.com/systems/storage/spectrum/ess/
- IBM Power E880 www.ibm.com/systems/power/hardware/e880/
- IBM Spectrum Scale www.ibm.com/systems/storage/spectrum/scale/
- IBM and Mellanox switches www.ibm.com/support/knowledgecenter/en/POWER8/p8hdx/p8hdx_network_switches.htm
- SAS: www.sas.com
- SAS Elastic Storage Server (ESS) battle card www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSW03542USEN&
- A deep dive into the new software-defined converged infrastructure for SAS Foundation mixed workloads www.ibm.com/developerworks/library/l-infrastructure-for-sas/

Follow us on Twitter

@IBMSystemsISVs

About the authors

Tony Brown is a software performance engineer and consultant at SAS Institute, Inc. You can contact him at Tony.Brown@sas.com or www.linkedin.com/in/tony-brown-1848753

Beth Hoffman is an IBM solution architect in IBM Cognitive Systems ISV Enablement organization. You can contact her at bethvh@us.ibm.com or www.linkedin.com/in/bethhoffmanibm

Narayana Pattipati is an IBM technical consultant in IBM Cognitive Systems ISV Enablement organization. You can contact him at npattipa@in.ibm.com or www.linkedin.com/in/narayana-pattipati-0146736

Brian Porter is an IBM storage solution architect in IBM Cognitive Systems Storage organization. You can contact him at bporter1@us.ibm.com or www.linkedin.com/in/brian-porter-a27a06b

Harry Seifert is an IBM technical sales support specialist in IBM Cognitive Systems sales organization supporting SAS deployments. You can contact him at seifert@us.ibm.com or www.linkedin.com/in/harry-seifert-329a336

Matthew Sheard is an enterprise solutions architect at Mellanox Technologies, Inc. You can contact him at matthews@mellanox.com or www.linkedin.com/in/matthew-sheard-6675602

Ben Smith is an IBM solutions architect in the IBM Systems Software Defined Infrastructure organization. You can contact him at smithbe1@us.ibm.com or www.linkedin.com/in/smithbe1/

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

Please recycle