SAS 9.3 grid deployment on IBM Power servers with IBM XIV Storage System and IBM GPFS

Narayana Pattipati

IBM Systems and Technology Group ISV Enablement

September 2012

Table of contents

Abstract

SAS Grid Computing enables organizations to create a managed, shared environment to process large volumes of data and analytic programs more efficiently. In the SAS grid environment, the SAS computing tasks are distributed among multiple nodes on a network, running under the control of SAS Grid Manager. SAS Grid Computing can leverage the processing capabilities of IBM POWER7 processor-based servers and IBM Enterprise Storage to deliver highly resilient and high-performance infrastructure required for analytics workloads. This white paper describes the architecture and best practices for the deployment of SAS Grid Manager 9.3 on an IBM Power 780 server running IBM AIX 7.1, along with IBM XIV Storage System, IBM General Parallel File System (GPFS), and IBM Platform Computing Platform Suite for SAS.

Introduction

In today’s complex business environment, enterprises deploy business intelligence (BI) and analytics solutions to improve decision making and performance. These solutions generally need a highly scalable and reliable infrastructure. The infrastructure should support more users, more compute-intensive queries, and large quantities of data.

SAS Grid Computing provides critical capabilities that are necessary for today's business analytics environments, including workload balancing, job prioritization, high availability and built-in failover, parallel processing and resource assignment, and monitoring.

The IBM® Power Systems™ family of servers includes proven workload consolidation platforms that help clients control costs while improving overall performance, availability, and energy efficiency. With these servers and IBM PowerVM® virtualization solutions, an organization can consolidate large numbers of applications and servers, fully virtualize its system resources, and provide a more flexible and dynamic IT infrastructure. In other words, IBM Power Systems with PowerVM deliver the benefits of virtualization without limits.

The IBM General Parallel File System (IBM GPFS™), a high-performance enterprise file management platform, can help customers move beyond simply adding storage to optimizing data management. GPFS is a clustered file system, which supports consistent high-performance shared file system from multiple servers. By virtualizing file storage space and allowing multiple systems and applications to share common pools of storage, GPFS provides the flexibility to transparently administer the infrastructure without disrupting applications, thus lowering storage costs.

The IBM XIV® Storage System is a high-end disk storage system designed to provide consistent, enterprise performance and exceptional ease of use. As a virtualized storage that meshes tightly with hypervisors, the XIV system offers optimal agility for cloud and virtualized environments.

The Platform Suite for SAS is delivered as part of the SAS Grid Manager product and provides enterprise scheduling across distributed machines in a grid as well as load balancing of multiple SAS applications and parallelized single applications for many SAS products and solutions. The Platform Suite for SAS consists of three platform products: Platform Process Manager, IBM Platform LSF®, and Platform Grid Management Services. All these products are required for SAS Grid Manager which includes both job scheduling and grid computing. There is also an add-on product, Platform RTM for SAS, used for grid monitoring, control, and configuration. Platform RTM for SAS is not bundled with Platform Suite for SAS but is available for download with registration.

The SAS Grid Manager deployment on IBM Power® 780 server with IBM XIV Storage, GPFS, and Platform Suite for SAS combines the flexibility of dynamic workload balancing of the SAS Grid Manager with the scalability and virtualization features of the Power hardware and high availability of the XIV Storage System. This combination provides an integrated analytical environment that can support the needs of the most-demanding organizations.

This technical white paper describes the deployment architecture for SAS Grid on IBM Power Systems with IBM GPFS and IBM XIV Storage System. The paper includes details of software, hardware, and configuration tuning options used to achieve the resulting benchmark performance. The last section includes an overview of the tools and techniques used in performance monitoring and tuning of the whole stack.

SAS 9.3 Grid Computing overview

Many SAS products and solutions have been integrated with SAS Grid Manager to allow seamless submission to the grid. SAS products and solutions that are deployed with SAS Grid Manager are able to leverage the following functionalities in a distributed grid environment:

- Enterprise scheduling of production jobs
- Workload management
- Parallel workload management – SAS applications consisting of subtasks that are independent units of work and can be distributed across a grid and executed in parallel.

Features of SAS Grid Computing:

- Grid-enabled SAS
- Managed, shared environment
- High availability
- Real-time monitoring and administration
- Flexible infrastructure

Benefits of SAS Grid Computing:

- Centralize management and reduce complexity.
- Create a highly available computing environment.
- Accelerate applications using the existing IT infrastructure.
- Future-proof and increase flexibility of your enterprise computing infrastructure by scaling out.

Deployment architecture for SAS Grid Manager on the IBM Power 780 server

This section describes the deployment architecture, configuration, and performance tuning for all the important products in the stack.

Deployment architecture product list

The deployment architecture described in this white paper has many products from SAS and IBM. The list of products includes:

SAS products

- Base SAS
- SAS/STAT
- SAS/CONNECT
- SAS Grid Manager

IBM products

- IBM Power Systems Power 780 server
- IBM XIV Storage System (Gen2)
- IBM GPFS
- IBM PowerVM
- IBM AIX® 7.1 OS (7.1.0.0)

SAS Grid on Power 780 server – Deployment architecture

This section describes the deployment architecture for deployment of SAS Grid on IBM Power 780 server with GPFS and XIV Storage System.

Software

- SAS 9.3 64-bit software
  - Base SAS 9.3
  - SAS/STAT 9.3
  - SAS/CONNECT 9.3
  - SAS® Grid Manager 9.3
- IBM AIX OS 7.1.0.0
- Virtual I/O Server (VIOS) 2.2.1.3
- IBM PowerVM Enterprise Edition
- IBM GPFS 3.4.0.7

Hardware

- IBM Power 780 Server
  - Architecture – IBM POWER7®
  - Cores - 16 (2 sockets)
  - Processor clock speed: 3864 MHz

    Power 780 server supports turbo core mode at 4.14 GHz. However, this mode is not used during the benchmark activity.
  - Simultaneous multithreading (SMT) 4 enabled
  - Memory: 256GB

    Used 80GB total memory for all the grid nodes during the benchmark activity.
  - Internal drives: Eighteen 300 GB (5.4TB)

    Used for booting logical partitions (LPARs) and VIOS, not used for SAS data. SAS data is on the XIV system.
- IBM XIV Storage System
  - XIV system version: 10.2.4.c-1
  - System ID: 1300316 (316)
  - Machine model: A14 / 2810 (Gen2)

  - Drives: 180 SATA drives each with 1 TB capacity and 7200 rpm speed
  - Usable space: 79 TB
  - Modules: 15 with 12 drives each
  - Memory / Cache: 120 GB
  - Stripe size: 1 MB (Default)
  - Six 4 Gb dual-port Fibre Channel (FC) adapters (12 ports) connected to storage area network (SAN)
- SAN connectivity
  - IBM System Storage SAN24B-4 Express Brocade Fiber Switch (NPIV enabled) (two switches)
  - Eight 8 Gb dual-port FC adapters (QLogic) connected to the Power 780 server

    Out of eight, used five 8 Gb (10 ports) FC adapters to connect the grid nodes (LPARs) to SAN for the benchmark activity.
  - Six 4 Gb FC adapters connected from SAN Switches to the XIV system (12 ports)
  - The SAN switch ports are running in auto-negotiation mode (port speed in Gbps). The ports connected to the Power780 server are running at N8 port speed and the ports connected to XIV are running at N4 port speed.

Server configurations

The Power 780 server is configured with four LPARs and each LPAR acts as a node in the SAS Grid. The LPARs are dedicated, which means that the processor resources are dedicated to an LPAR and they are not shared. Two configurations are tried in the deployment architecture:

- **VIOS with N-Port ID Virtualization (NPIV)**
  - Used for sharing physical FC adapters among LPARs. FC adapters are mapped to the VIOS and virtual FC adapters are connected to the client LPARs.
- **Non-VIOS**
  - FC adapters are directly mapped to LPARs without using VIOS.
  - FC adapters are not shared among the LPARs. Each LPAR has dedicated FC adapters mapped.

Both the deployments use dedicated LPARs for all the grid nodes. The LPARs does not use micro-partitions.

Deployment of SAS Grid on Power 780 with VIOS and NPIV

The Power 780 server is configured with four LPARs and each LPAR acts as a node in the SAS Grid. Grid Node 1 acts as Grid Control Server. The four grid nodes use dedicated LPARs. The SAS GPFS file systems are created on the LUNs mapped from XIV storage system and they are shared across the four grid nodes. Figure 1 describes the complete deployment architecture of SAS Grid on the IBM Power 780 server with XIV Storage System and GPFS, with VIOS and NPIV. Figure 1 also shows the SAN connectivity between the server and XIV Storage System.

All the five physical FC adapters (10 FC ports) are directly mapped to the VIOS. On the VIOS, twenty virtual server FC adapters are created, with each physical FC port mapping to two of the twenty virtual server FC adapters. On the VIOS, five virtual server FC adapters are mapped to each of the four grid nodes. On each grid node, five client FC adapters are created and mapped to 5 of the virtual server FC adapters created on the VIOS. This effectively virtualizes the I/O through VIOS.

The paper does not discuss about how to configure VIOS and NPIV, establish SAN connectivity between the server and the XIV system, and map storage logical unit numbers (LUNs). For those details, refer to the IBM Redbooks® *PowerVM Virtualization Managing and Monitoring* at **ibm.com/**redbooks/abstracts/sg247590.html.

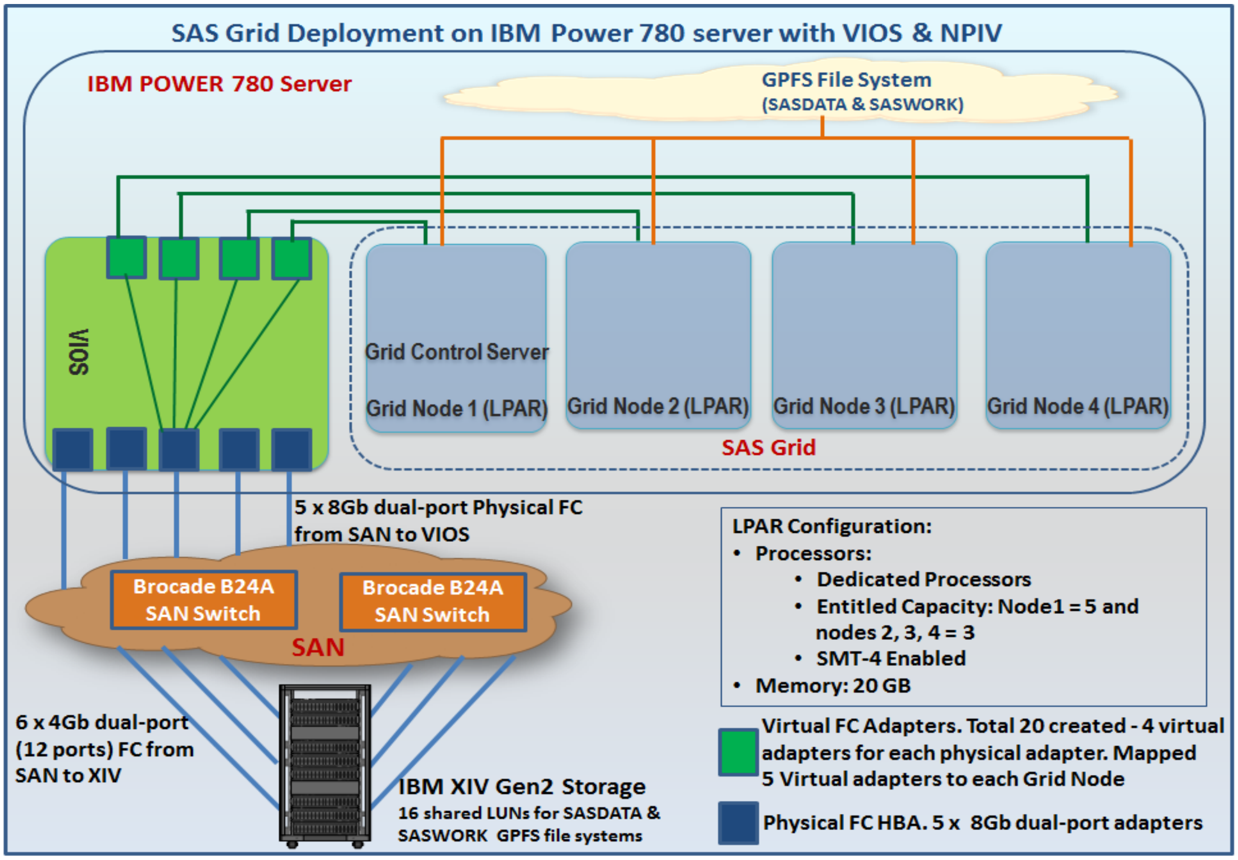

SAS Grid Deployment on IBM Power 780 server with VIOS & NPIV
IBM POWER 780 Server
GPFS File System
(SASDATA & SASWORK)
VIOS
Grid Control Server
Grid Node 1 (LPAR)
Grid Node 2 (LPAR)
Grid Node 3 (LPAR)
Grid Node 4 (LPAR)
SAS Grid
5 x 8Gb dual-port Physical FC from SAN to VIOS
Brocade B24A SAN Switch
Brocade B24A SAN Switch
SAN
6 x 4Gb dual-port (12 ports) FC from SAN to XIV
IBM XIV Gen2 Storage
16 shared LUNs for SASDATA & SASWORK GPFS file systems
LPAR Configuration:
• Processors:
• Dedicated Processors
• Entitled Capacity: Node1 = 5 and nodes 2, 3, 4 = 3
• SMT-4 Enabled
• Memory: 20 GB
🟩 Virtual FC Adapters. Total 20 created - 4 virtual adapters for each physical adapter. Mapped 5 Virtual adapters to each Grid Node
🟦 Physical FC HBA. 5 x 8Gb dual-port adapters

Figure 1: SAS grid deployment architecture on an IBM Power 780 server with VIOS and NPIV

SAS grid deployment on Power 780 without VIOS and NPIV

This configuration does not use VIOS and I/O virtualization. The five 8 Gb FC adapters are directly mapped to the grid nodes. Two 8 Gb FC adapters are mapped to Grid Node 1 and one 8 Gb FC adapter is mapped to each of the other grid nodes. In this deployment, the I/O is dedicated to each grid node.

Figure 2 describes the complete deployment architecture of SAS grid on the IBM Power 780 server with the XIV system and GPFS, which does not use VIOS and NPIV. Figure 2 also shows the SAN connectivity between the server and XIV system. The four grid nodes use dedicated LPARs.

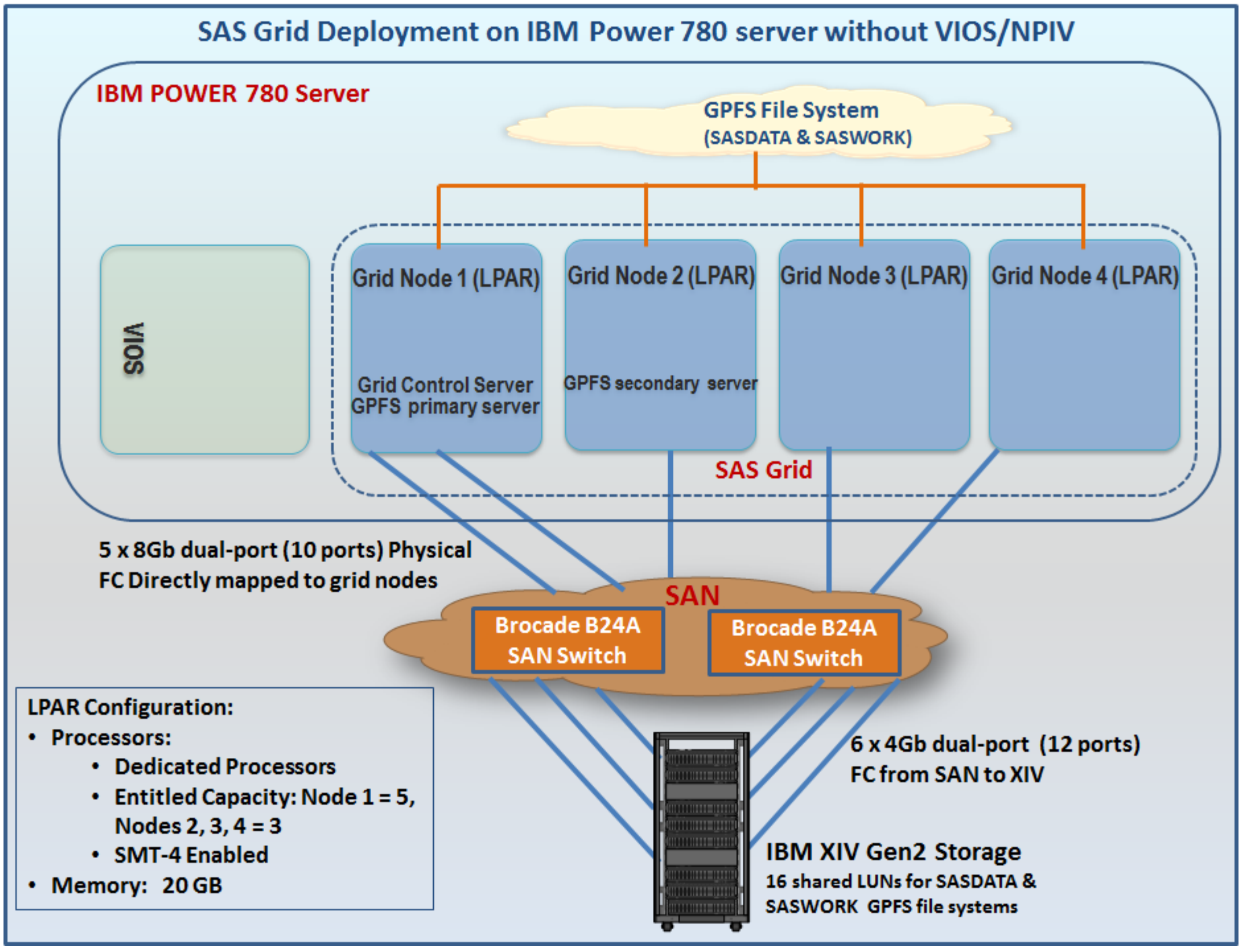


Figure 2: SAS grid deployment architecture on the IBM Power 780 server without using VIOS / NPIV

SAS Grid Manager deployment

The *SAS Grid Manager* can be deployed on any one of the deployment architectures shown in Figure 1 or Figure 2. It is deployed in the same way irrespective of whether the architecture uses VIOS / NPIV or not.

The SAS GPFS file systems, SASWORK and SASDATA, are set up and mounted across the grid. SAS Grid Manager is then installed on the SAS file systems. Figure 3 describes the SAS Grid Manager deployment on the Power 780 system, along with the components that get installed on each grid node. Grid Node 1 is given additional responsibility of Grid Control Server, in addition to being a grid node which runs SAS jobs.

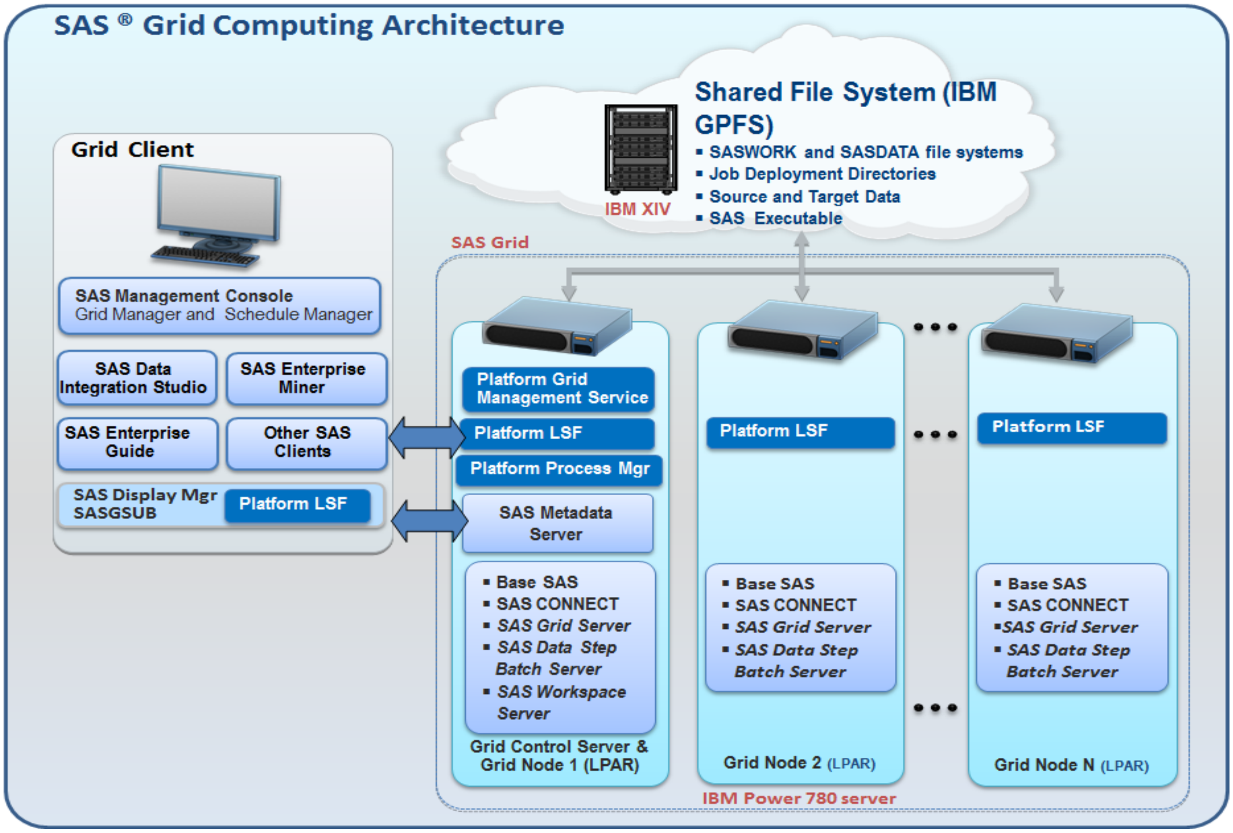


Figure 3: SAS Grid Manager and Platform Suite for SAS deployment on the IBM Power 780 server

Benchmark workload and SAS I/O characteristics

This section provides details about the workload used and the SAS I/O characteristics.

Details of the workload used

SAS created a multiuser benchmarking scenario to simulate the workload of a typical Foundation SAS customer. Each test scenario consists of a set of SAS jobs run in a multiuser fashion to simulate a typical SAS batch, SAS Enterprise Guide user environment. The individual jobs are a combination of computational and I/O-intensive SAS procedures. The jobs are started over a set interval and the scheduling and dispatch of jobs are controlled by the SAS Grid Manager. The control scripts simulate peak and off-peak workloads, which is close to real-world analytical processing. The workload has different options to run configurable number of sessions and jobs. For example, a 40 session workload runs 144 I/O, compute, and memory-intensive jobs.

The performance measurement for these workloads is response time (in seconds), which is the cumulative real time processor usage of all the jobs in the workload. The lower the response time, the better it is and workload is said to have performed better.

Details of 40 session workload are given in the following points:

- A total of 144 jobs are started over a period of time across the grid nodes
- The mixed scenario contains a combination of I/O and processor intensive jobs

- Characteristics of data used:
  - SAS data sets and text files
  - Row counts up to 90 million
  - Variable and column counts up to 297
  - 600 GB total input and output data
  - Total of 2.25 TB I/O (data read/written) by the grid

- The following table describes the typical SAS users, the jobs they submit and the data characteristics for the 40 session workload. The 40 session workload submits a total of 144 jobs.

SAS user type	Number of jobs	Data size
Base SAS statistics users	92	Less than 1 GB
	12	More than 10 GB
Advanced analytics users	36	Less than 1 GB
	4	More than 1 GB
Data mining users	4	Less than 10 GB
ETL or data integration sessions	4	More than 10 GB

Table 1: The 40 session workload details

SAS I/O characteristics

SAS has many application solutions. The foundation component, known as Base SAS, runs as a collection of processes. Here are some important I/O characteristics:

- SAS applications perform large sequential read and write operations. Some of the new SAS BI applications do some random access of data, but for the most part, the SAS workload can be characterized as predominately large sequential I/O requests with high volumes of data.
- SAS applications do not pre-allocate storage when SAS initializes or when performing write operations to a file. When SAS creates a file, it allocates a small amount of storage, but as the file grows during a SAS task, SAS extends the amount of storage needed.
- Uses OS file cache for data access. SAS does not do direct I/O by default.
- Creates large number of temporary files during long-running SAS jobs in the SAS WORK directory. These files are created, potentially renamed towards the end of the task, deleted, and manipulated potentially hundreds of times within a long-running SAS extract, transform, and load (ETL) job. The size of the files might be very small (less than 100 MB) to larger (in the 10s of GBs).

SAS file systems typically include:

- **SASDATA** is the location for the persistent data, SAS executable, input/output data files. This file system mostly gets read by the SAS applications and some occasional writes at the end of some SAS jobs.
- **SASWORK** is the temporary space for SAS sessions. The data here is available only for the duration of a SAS session and is erased when the SAS session terminates normally. This file system gets the majority of the I/O activity as this is where the temporary files are created during a SAS job.

The performance tuning at different levels should be done keeping these I/O characteristics in mind.

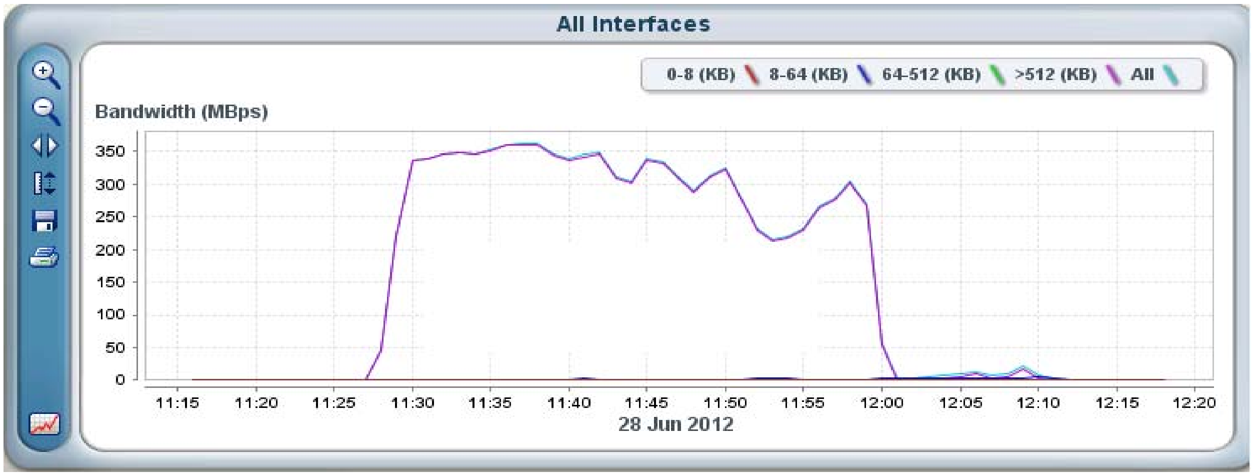


Figure 4: SAS I/O characteristics with 40 session benchmark workload (most of the I/O is 512 KB and above)

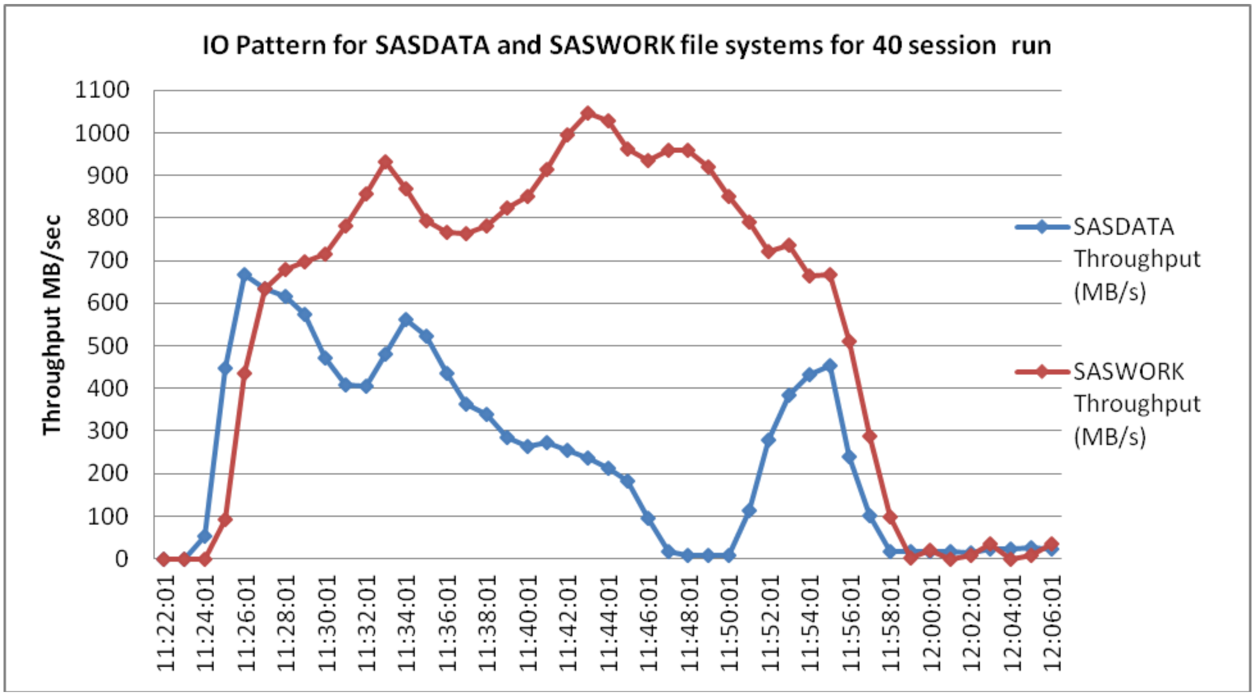


Figure 5: I/O pattern for SASDATA and SASWORK shared file systems

Detailed configuration and tuning guidelines

This section describes the configuration and tuning done to the stack to optimize SAS Grid performance on Power 780 server with XIV storage and GPFS.

AIX 7.1

Tuning guidelines for SAS 9.3 on AIX 7.1 are given in the link:
ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101529

In addition to the tuning guidelines given in the tuning guidelines document, during the benchmark testing, it was found that enabling large page support improved SAS performance. The following VMM tunable helped in achieving better performance.

vmm_mpsize_support = 2 (default value -1)

vmm_default_pspa = 100 (default value 0)

Use the following command to change the AIX OS VMM tunable.

$vmo –r –o vmm_mpsize_support = 2

XIV Storage System

From the XIV GUI, a storage pool was created and 16 volumes (LUNs), each with a size of 395 GB, were created from the pool. The volumes are then mapped to the grid cluster, consisting of the four grid nodes. From the grid nodes, the LUNs are mapped as logical disks (hdisks) and *SASWORK* and *SASDATA* GPFS file systems were created with eight LUNs each and mounted on all the grid nodes. Note that the XIV system uses the stripe size of 1 MB by default; hence, the file systems were created with 1 MB block size for optimizing the throughput. No other specific configuration or tuning is done on the XIV storage system.

GPFS

The SAS file systems ***SASWORK*** and ***SASDATA*** are based on IBM GPFS and they are shared across the four grid nodes in the SAS grid. Eight logical disks (LUNs) with a total size of 3115 GB are configured for each of file system. SAS software, input data, benchmark workload and temporary space is allocated on these file systems.

GPFS configuration information, at a high level:

- Eight LUNs per file system
- Primary server: Grid Node 1
- Secondary server: Grid Node 2
- Quorum-Managers: Grid Node1 and Grid Node2
- Grid Node 1 and Grid Node 2 contain server licenses and Grid node 3 and Grid Node 4 contain client licenses
- **File system** creation configuration and tunable:
  - **SASDATA:** File system created with eight XIV LUNs of 395 GB each
  - **SASWORK:** File system created with eight XIV LUNs of 395 GB each
  - The Network Shared Disks (NSDs) used to create the file systems have both data and metadata

```
$ mmlsdisk sasdata-gpfs
disk         driver   sector  failure  holds     holds                storage
name          type     size   group  metadata   data     status     availability   pool
------------  --------  ------ -------  --------  -----    -------------  ------------  ------------
nsd01gpfs    nsd       512      -1    yes         yes      ready         up           system
```

  - **Block size:** 1 MB

    The XIV system uses 1 MB as stripe size, by default. Hence, creating file systems with 1 MB block size gave better performance.

  - **Block allocation type:** Cluster

    The *cluster* allocation method is the default for GPFS clusters with eight or fewer nodes and for file systems with eight or fewer disks. *Cluster* block allocation type proved to be a better option for the workload run during the benchmark activity.

- **pagepool:** At least 4 GB

  The *GPFS pagepool* is used to cache user file data and file system metadata. pagepool is the pinned memory for each node. GPFS pagepool of at least 4GB of gave better performance for the workload used in the benchmark.

- **seqDiscardThreshold:** 1 MB (default)

  This parameter affects what happens when GPFS detects a sequential read access pattern. If a file is being re-read and its size is greater than 1 MB, then the file will not be cached in GPFS. However, in SAS analytics, many files are re-read and files are usually of size greater than 1 MB. Hence, increase this value based on the workload characteristics and size of the input files.

- **prefetchPct:** 20 (default)

  GPFS uses this as a guideline to limit how much pagepool space will be used for pre-fetch or write-behind buffers in the case of active sequential streams. If the workload does mostly sequential I/O, increasing it might benefit. SAS workloads predominantly do sequential I/O. Hence, increasing it to 40% might help performance.

  Note: Changing just one tunable might not help in performance. It is important to understand how the different tunable work with each other and find out the right combination by running tests. For the workload used in benchmark activity, setting *seqDiscardThreshold=1GB, pagepool=8GB and prefetchPct=40 gave* slightly better performance.

- **maxMBpS:** 5000

  The maxMBpS value should be adjusted for the nodes to match the I/O throughput that the system is expected to support. A good rule of thumb is to set the maxMBpS value to twice the I/O throughput required of a system. For example, if a system has two 4 Gb host bus adapters (HBAs) (400 MBps per HBA) maxMBpS should be set to 1600.

- **GPFS mount options**

  rw, mtime, atime, dev

- **Other GPFS cluster tunable used:**

```
autoload yes
dmapiFileHandleSize 32
maxFilesToCache 20000
prefetchThreads 72
worker1Threads 48
```

FC HBAs

For the VIOS with NPIV configuration, the five physical FC adapters (10 FC ports) are directly mapped to the VIOS. On the VIOS, twenty virtual server FC adapters are created, with each physical FC port mapping to two of the twenty virtual server FC adapters. On the VIOS, five virtual server FC adapters are mapped to each of the four grid nodes. On each grid node, five client FC adapters are created and mapped to 5 of the virtual server FC adapters created on the VIOS. This effectively virtualizes the I/O through VIOS. Following HBA tunings are done in VIOS and all the four grid nodes.

```
max_xfer_size=0x800000
lg_term_dma=0x8000000
num_cmd_elems=1024
```

The performance tunings are also applicable for deployment architecture that does not use VIOS and NPIV, as described in Figure 2. Use the following command to change the tunable.

```
$chdev –l fcs0 –a max_xfer_size = 0x800000
```

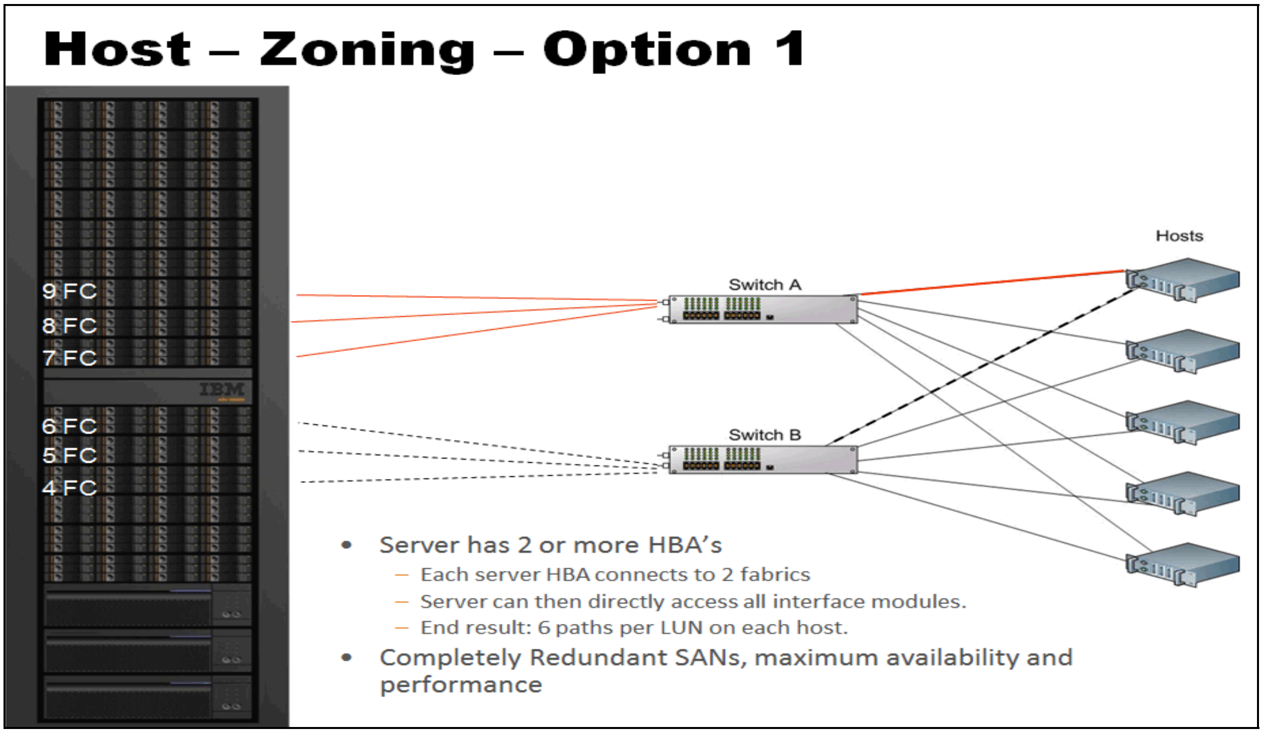
Logical disks

The following logical disk tunings are done for all the 16 logical disks in VIOS and all the four grid nodes.

```
Max_transfer = 0x800000
Queue_depth = 256

$chdev –l hdisk12 –a max_transfer = 0x800000
```

SAN Switches and host zoning with the XIV system

Two redundant Brocade B24A switches are connected in the SAN fabric. They connect the XIV system to the Power 780 server. The Power 780 server is connected to the SAN Switches by five 8 Gb dual-port FC adapters (10 ports). And the XIV system is connected to the SAN Switches by six 4 Gb dual-port FC adapters (12 ports). The ports connecting the Power 780 server to the SAN Switch are configured to have port speeds N8 (Auto) and the ports connecting the switches to the XIV system are configured to have port speeds N4 (Auto). Appropriate zoning is done on both the SAN Switches. Note that zoning should include the worldwide port names (WWPNs) from the client virtual FC adapters created on the grid nodes.



Host – Zoning – Option 1

- Server has 2 or more HBA's
  - Each server HBA connects to 2 fabrics
  - Server can then directly access all interface modules.
  - End result: 6 paths per LUN on each host.
- Completely Redundant SANs, maximum availability and performance

Figure 6: Host zoning option with SAN Fabric and XIV Storage System

SAS configuration

The following SAS options were used to optimize the workload performance.

memsize = 2048 MB

bufsize = 256 k

sortsize = 256 M

fullstimer

Platform LSF configuration

The following configuration and performance tunings are done in the Platform Computing LSF module to optimize the workload performance.

- Change the maximum number of job slots to 32 for all the nodes in the ***lsb.hosts*** file. This gave optimal performance during the benchmark testing.

```
Begin Host
HOST_NAME           MXJ   r1m    pg   ls    tmp  DISPATCH_WINDOW
p7_gridn1.unx.sas.com 32   ()      ()   ()    ()    ()
p7_gridn2.unx.sas.com 32   ()      ()   ()    ()    ()
p7_gridn3.unx.sas.com 32   ()      ()   ()    ()    ()
p7_gridn4.unx.sas.com 32   ()      ()   ()    ()    ()
End Host
```

The default value for job slots is mentioned as ( ), in which case, the job slots will be equal to the number of cpus (ncpus) shown by the LSF ***lshosts*** command. It is recommended to leave the job slots as default, unless the workload characteristics require it to be changed to a higher value.

- Change the job scheduling parameters in the lsb.params file to the following values. This particular change gave 7% to 15% improvement in the performance during the benchmark testing.

```
Begin Parameters
MBD_SLEEP_TIME = 2
SBD_SLEEP_TIME = 1
MBD_REFRESH_TIME = 15
NEWJOB_REFRESH = Y
End Parameters
```

- In the benchmark environment, Grid Node 1 acts as a Grid Control Server and GPFS cluster manager, in addition to being a node in the grid that runs jobs, which puts additional load on it. Hence, it is suggested to give more priority to the other nodes in the grid while receiving jobs from the queue. Make the following changes to the ***lsb.queues*** file to take care of the same.

```
Begin Queue
QUEUE_NAME   = normal
HOSTS = p7_gridn1+1 others+2
```

However, it is not recommended to run any workloads on the node that acts as the Grid Control Server.

Performance monitoring

Performance monitoring includes processor and memory utilization, disk, network and file system I/O, at grid node and server level.

Many tools and techniques are used to monitor performance during the benchmark activity on the SAS grid environment on the Power 780 server with GPFS and XIV Storage System. Performance monitoring was done at Platform LSF, AIX 7.1 OS, GPFS, FC adapters, SAN Switch, and XIV Storage System. There are various tools available to monitor the performance at each level, and to tune and optimize the environment. The “Appendix A: Performance monitoring tools and techniques” section describes the tools and techniques used in detail.

I/O performance benchmark results on the deployment architecture

I/O performance benchmark was done on the deployment architecture without using SAS workloads. This helped in determining the I/O throughput that the deployment architecture can produce.

Objective and I/O performance benchmark details

The objective was to find the peak sequential I/O throughput that the deployment architecture can produce with the XIV storage system.

- The I/O performance tests were done with raw disks mapped from XIV Gen2 storage.
- SASDATA GPFS file system is used to benchmark the I/O throughput
- The I/O performance tool ***ndisk64*** from the ***nstress*** suite (at: **ibm.com**/developerworks/wikis/display/WikiPtype/nstress) is used to perform the raw disk tests. The tool can generate read, write, or read-write I/O to test disks.
- ***gpfsperf*** tool (/usr/lpp/mmfs/samples/perf/gpfsperf) which is part of GPFS installation is used to test the I/O performance of GPFS with the XIV storage system.
- The I/O performance benchmark is done for 100% read, 50% read-write and 100% write, in a sequential mode.
- For raw disk I/O tests:
  - 128 ***ndisk*** processes (32 from each grid node) reading from / writing to the raw disks
  - Each node reads / writes its own raw disk (total four disks)
  - Read or Write block size is 1 MB
- For GPFS I/O tests:
  - 128 ***gpfsperf*** threads (32 from each node) reading/writing from SASDATA file system
  - Each node reads from / writes to 32 different files of 8 GB size (total 128 files)
  - GPFS block size is 1 MB and the read/write block size is 1 MB
- The deployment architecture with VIOS / NPIV, where I/O is virtualized and shared is used to perform the benchmark testing.

I/O performance benchmark summary with VIOS and NPIV

High watermark for 100% read, 100% write, 50% read-write on the grid is given in the table 2.

Operation	**Raw disk (MBps)**		**GPFS (MBps)**		**Comments**
	With VIOS / NPIV	**Without VIOS / NPIV**	**With VIOS / NPIV**	**Without VIOS / NPIV**	
100% Read	4000	4000	4100	4000	For raw disk, XIV cache hit = 100% and latency = 25 ms For GPFS, XIV cache hit = 100% and latency = 40 ms
50% Read-Write	1800	1800	1800	1820	For raw disk, XIV cache hit = 75% and latency = 60 ms For GPFS, XIV cache hit = 27% and latency = 75 ms
100% Write	1350	1350	1375	1400	For raw disk, XIV cache hit = 25% and latency = 100 ms For GPFS, XIV cache hit = 5% and latency = 120 ms

Table 2: High watermark 100% read/write, 50% read with raw disks and GPFS on deployment architecture

The I/O throughput is seen from the statistics monitor in the XIV GUI. The statistics monitor also gives details about cache hit and latency. The details of how to use the statistics monitor are given in the “Appendix A: Performance monitoring tools and techniques” section.

SAS workload benchmark results

The objective of this benchmark testing was to document the best practice tuning guidelines for achieving optimal performance of SAS grid on Power / AIX with GPFS and XIV Storage System. With optimal tuning of the entire stack, the benchmark results show how the architecture supports the computing and I/O requirements of the SAS workloads running on the grid.

The benchmark was run on the grid to simulate the workload of a typical Foundation SAS customer that uses grid computing. The goal was to evaluate the multiuser performance of the SAS software on the IBM Power platform. The workload has an emphasis on I/O, which SAS software uses heavily while running a typical Foundation SAS program execution.

The following workload scenarios were run as part of the benchmark testing:

- 40 session workload
- Scalability of workloads on grid
- 80 session workload

40 session benchmark with VIOS and NPIV

System configuration and tuning used for the benchmark run:

- The benchmark uses SAS grid deployment on a Power 780 server with VIOS and NPIV, described in Figure 1.
- The performance tuning guidelines described in the earlier sections holds good for the benchmark run.
- The benchmark testing uses dedicated processors for LPARs (grid nodes) as described in Figure 1. This means, the grid has 14 cores assigned (five cores to node 1 and three cores each to nodes 2, 3, and 4).

40 session benchmark results summary

Here is the summary of the 40 session workload benchmark run on the SAS grid deployment on the Power 780 server with GPFS and XIV Storage System.

- The deployment architecture delivered a peak IO throughput of 1375 MBps and sustained throughput of around 1300 MBps, seen from XIV GUI.
- Around 90 MB per core per sec sustained SAS I/O throughput, considering the grid has only 14 cores assigned to the nodes in dedicated mode.
- The clock time of the workload is 43 mins and cumulative processor real time of all the 144 jobs is 68412 seconds.
- Processor usage of the grid during the workload is 45%, processor wait is 10%, and processor idle is 45%.
- SASDATA and SASWORK GPFS file systems together reported sustained throughput of 1500 MBps and peak throughput of 1800 MBps. This is higher than the throughput reported by the XIV Storage System. This indicates that some of the I/O requests hit the GPFS cache.

40 session benchmark detailed results

Table 3 gives detailed statistics of the 40 session benchmark run on the deployment architecture with VIOS and NPIV. Figure 6 shows the I/O throughput achieved for the run at GPFS and XIV Storage System. Figure 7 shows the I/O throughput for individual nodes and Figure 8 shows the processor usage pattern for the grid.

Performance statistics		Value	Measurement
XIV Gen2 (statistics collected from XIV GUI)	Sustained throughput	1300	MBps
	Peak throughput	1375	MBps
	Sustained IOPS	1450	-
	Peak IOPS	1563	-
	Average latency	33	ms
	Peak latency	53	ms
	Cache hit percentage	15	-
	Cache miss	85%	
	Read-Write ratio	1.1	
	I/O block size	Greater than 512	KB
GPFS (collected using a tool based on mmpmon)	Sustained throughput	1450	MBps
	Peak throughput	1879	MBps
Processor (collected using nmon)	User + Sys usage	45%	6.3 cores out of total 14 assigned to grid nodes
	Wait	10%	1.4 cores out of total 14 assigned to grid nodes
	Idle	45%	6.3 cores out of total 14 assigned to grid nodes

Table 3: Performance statistics for the 40 session benchmark run with VIOS / NPIV

In Table 3, the XIV system data is collected from XIV GUI. The GPFS data is collected by running the ***gpfs_perf*** Perl script on the SASWORK and SASDATA file systems on all the grid nodes and consolidating the I/O performance numbers. The processor usage information is collected from the nmon charts developed using ***nmon_analyzer***. Usage of these tools is described in the "Appendix A: Performance monitoring tools and techniques" section.

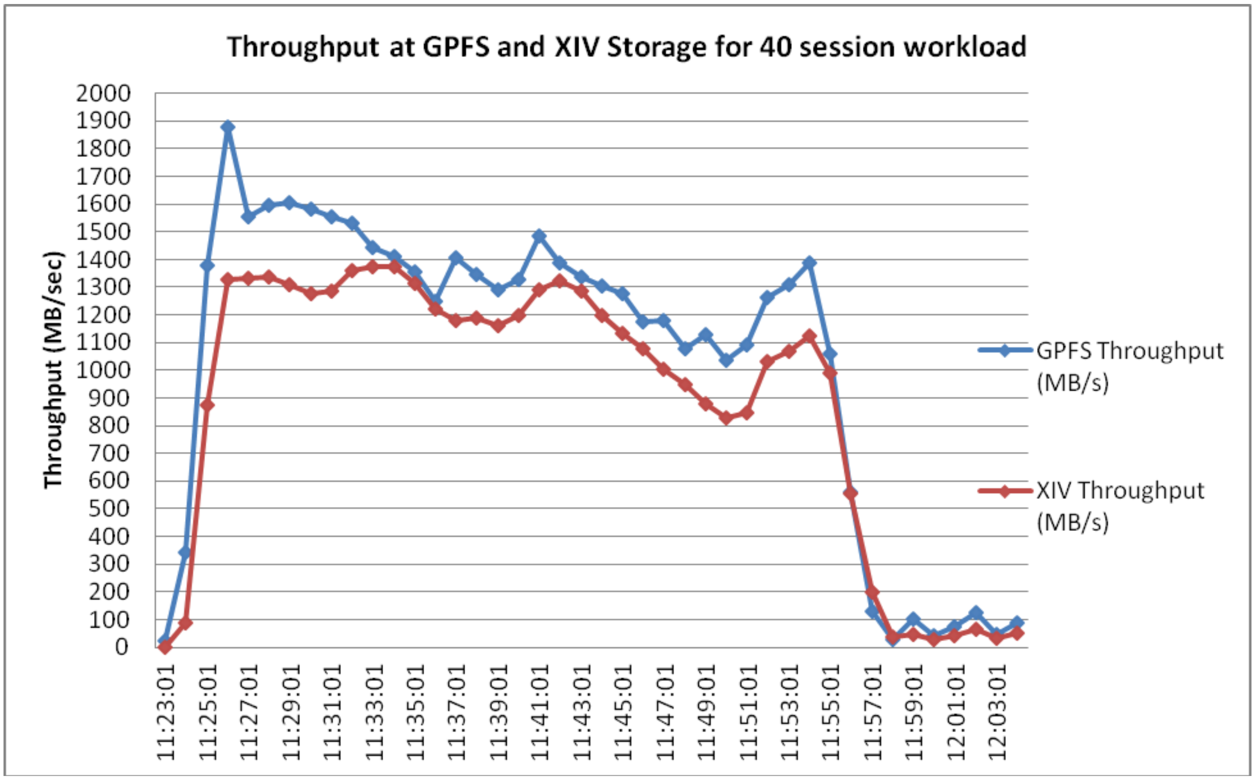


Figure 7: SAS I/O throughput for the 40 session with VIOS and NPIV at GPFS file system and XIV system levels

The graph in Figure 7 is plotted with I/O bandwidth data from IBM XIV GUI and GPFS I/O bandwidth from gpfs_perf Perl script.

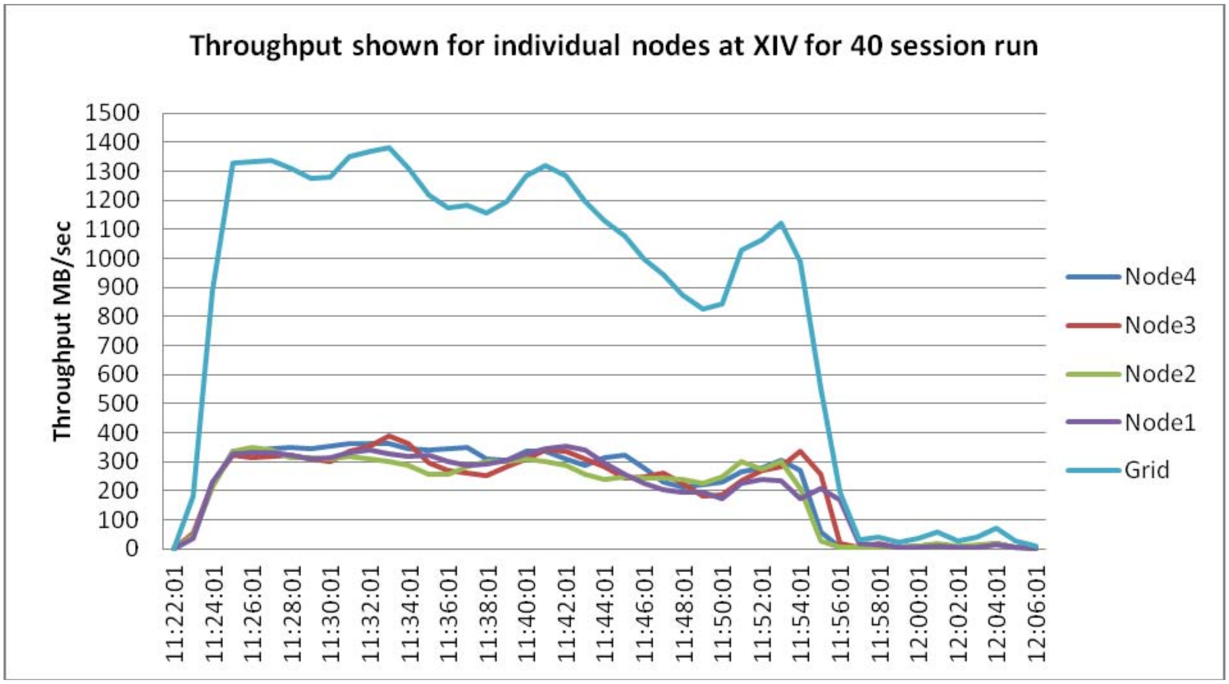


Figure 8: SAS I/O throughput shown for individual grid nodes and for the entire grid at the XIV system

The graph in Figure 8 is plotted using node wise IO BW data collected from XIV GUI.

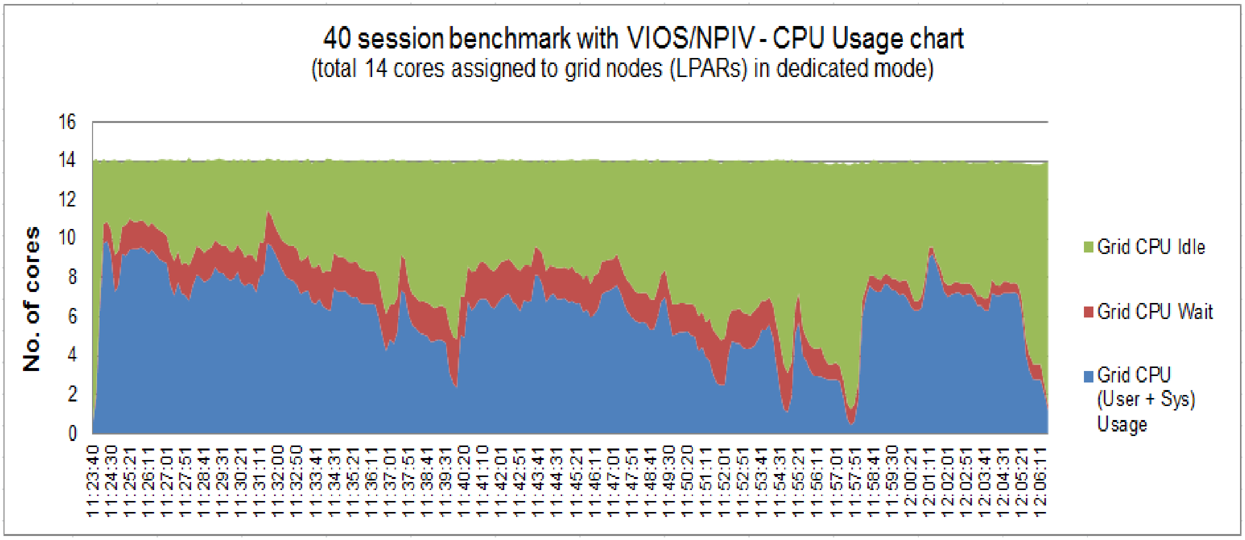


Figure 9: Grid processor usage, idle, and wait information for the 40 session workload with NPIV

The graph in Figure 8 is plotted by collating the processor usage (user + Sys), processor idle, and processor wait information collected from the nmon charts created by ***nmon analyzer.***

SAS grid scalability on Power 780 benchmark with VIOS and NPIV

The objective of this benchmark exercise is to demonstrate the scalability of the SAS grid on an IBM Power 780 server.

Grid scalability testing details

Starting with 1 node and 10 workload sessions, scalability is benchmarked by adding additional nodes and running additional work sessions: 20 sessions on 2 nodes, 30 sessions on 3 nodes, and 40 sessions on 4 nodes.

Number of nodes	Number of Sessions	Number of jobs
1	10	36
2	20	72
3	30	108
4	40	144

Table 4: The grid scalability testing details with a number of nodes, sessions, and jobs

System configuration and tuning used for the benchmark run:

- The benchmark uses SAS grid deployment on a Power 780 server with VIOS and NPIV, as described in Figure 1.
- The performance tuning guidelines described in the earlier sections holds good for the benchmark run.
- The benchmark uses dedicated processors for LPARs (grid nodes) as described in Figure 1.
- The back-end XIV storage capacity remains the same as you add more nodes. No additional capacity is added.

Grid scalability test results

The grid scaled reasonably linear; as nodes are added and additional jobs are scheduled, as shown in Figure 10.

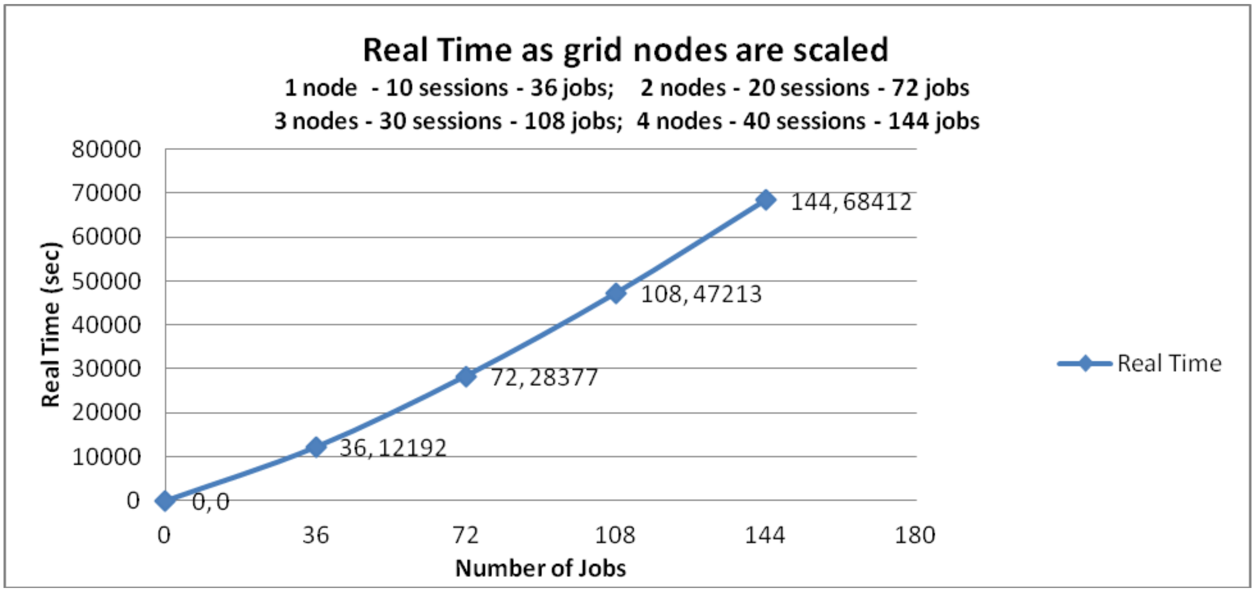


Figure 10: Graph shows scalability of grid nodes on the Power 780 server as nodes, sessions are added to the grid

80 session benchmark with VIOS and NPIV

Details of the 80 session workload:

- A total of 288 jobs are started over a period of time across the four grid nodes.
- The mixed scenario contains a combination of I/O-intensive jobs and processor-intensive code.

System configuration and tuning used for the benchmark run:

- The benchmark uses SAS grid deployment on the Power 780 server with VIOS and NPIV, as described in Figure 1.
- The performance tuning guidelines described in the earlier sections holds good for the benchmark run.
- The benchmark testing uses dedicated processors for LPARs (grid nodes), as described in Figure 1.
- The back-end storage remains same as in the case of a 40 session workload. No additional capacity is added.

Summary of the benchmark results:

- The deployment architecture delivered a peak I/O throughput of 1.5 GBps and sustained throughput of 1.45 GBps for the 80 session workload
- More than 100 MB per core per sec sustained SAS I/O throughput, considering that the grid has only 14 cores assigned to the nodes in dedicated mode.
- The processor usage of the grid during the workload is 61%, processor wait is 11%, and processor idle is 28%.

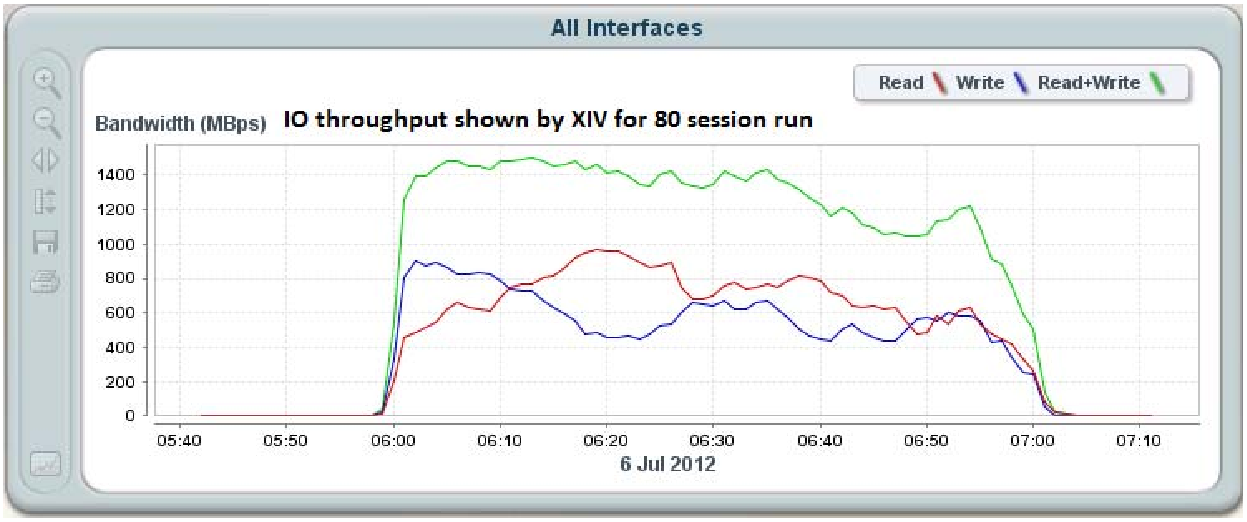


Figure 11: I/O throughput shown by the XIV system for the 80 session workload with VIOS and NPIV

40 session benchmark without VIOS and NPIV

The 40 session benchmark is run to understand how the workload behaves on the deployment architecture without VIOS and NPIV. As described in Figure 2, the FC adapters are directly mapped to the grid nodes, without virtualizing through VIOS.

System configuration and tuning used for the benchmark run:

- The benchmark uses SAS grid deployment on the Power 780 server without VIOS and NPIV, as described in Figure 2.
- The performance tuning guidelines described in the earlier sections holds good for the benchmark run.
- The benchmark testing uses dedicated processors for LPARs (grid nodes), as described in Figure 2.

Summary of the benchmark results:

- The deployment architecture delivered a peak I/O throughput of 1362 MBps and a sustained throughput of 1300 MBps for the 40 session workload.
- This translates to 93 MBps per core SAS IO throughput, considering the fact that the grid has only 14 cores assigned in the dedicated mode
- The clock time of the workload is 43 min and cumulative processor real time of all the 144 jobs is 68030 seconds.

Comparison of the 40 session results between configurations with and without VIOS/NPIV

The 40 session performance benchmark tests show that both the configurations delivered similar I/O throughput and response times. This indicates that the I/O virtualization through VIOS/NPIV does not add any additional performance impact on the Power server. The I/O virtualization using VIOS/NPIV benefits the customer without any performance and throughput impact. Table 5 and Figure 11 give details of the benchmark run.

40 Session workload performance statistics	With NPIV and VIOS	Without NPIV and VIOS
Processor real time (response time)	68412	68030
Sustained throughput (MBps)	1300	1300
Peak throughput (MBps)	1375	1362
Sustained IOPS	1450	1450
Peak IOPS	1563	1547
Average latency (ms)	33	33
Processor (user + system) usage	44%	43.5%
Processor wait	10%	10%
Processor idle	46%	46.5%

Table 5: Comparison of the 40 session benchmark on deployment architectures with and without VIOS/NPIV

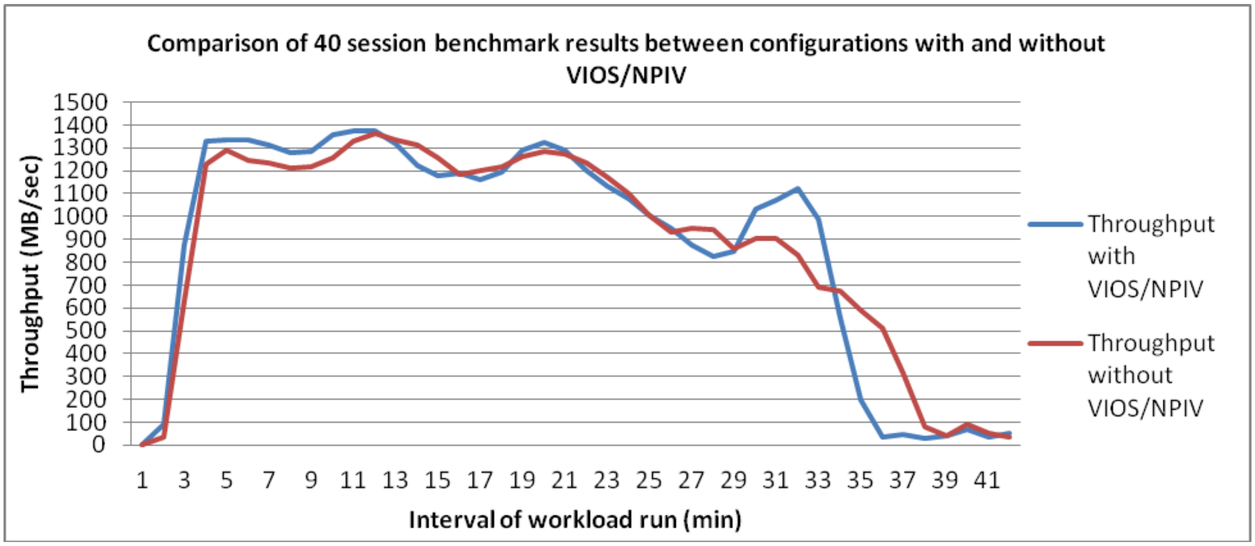


Figure 12: Comparison of I/O throughput for the 40 session with and without VIOS/NPIV

Summary

The paper demonstrates SAS 9.3 grid deployment on the IBM Power 780 server with the XIV Storage System and GPFS. It describes how to tune and optimize SAS grid on Power servers. IBM Power Systems and Storage, combined with GPFS and the Platform Computing technology have a proven history of providing the intense compute and I/O that is needed to run the most-demanding SAS customer analytic workloads. Additionally, this demonstrates the powerful workload management and job scheduling capabilities of SAS grid and the solutions that power it, including IBM Platform LSF and IBM Platform Process Manager.

Appendix A: Performance monitoring tools and techniques

This section describes the performance monitoring tools, usage, and techniques.

AIX tools to monitor performance

The following tools were used during the performance benchmark activity on SAS grid on Power server.

Tool	Description
Vmstat	Monitor overall system performance in the areas, such as processor, virtual memory manager activity, and I/O.
Svmon	Monitor the detailed memory consumption on real and virtual memory.
Ps	Monitor process / thread status and memory consumption as well.
Iostat	Monitor overall I/O status including disks loads or adapters, and system throughput.
sar	Report the per-processor, disk, run queue statistics.
netstat	Report network and adapter statistics.
lparstat	Report logical partition-related information. For example, partition configuration, hypervisor call, and processor utilization statistics.
mpstat	Report logical processor information in logical partition. For example, simultaneous multithreading (SMT) utilization, detailed interrupts, detailed memory affinity, and migration statistics for AIX threads, and dispatching statistics for logical processors.
topas	Report the local system's statistics, including: processor, network, I/O, processes, and workload management classes utilization.
nmon	A commonly used freeware tool for capturing AIX performance data: **ibm.com**/developerworks/aix/library/au-analyze_aix/ nmon can run on the console showing performance statistics real-time or it can log the performance statistics into an output file. Run nmon on all the nodes in the grid. For example, the following command runs nmon for about an hour and generates output under the /tmp/ directory. *$ nmon -f -s 10 -c 360 -t –o /tmp/* The nmon analyzer can be used to analyze the nmon output file and automatically creates dozens of graphs reflecting key system performance characteristics. Refer to the *SAS Performance Monitoring – A Deeper Discussion* paper at: http://www2.sas.com/proceedings/forum2008/387-2008.pdf. This is a SAS Global Forum 2008 paper for procedure of collecting nmon trace.
nmon analyzer	A commonly used tool to produce performance reports from the raw files generated by

	the nmon tool. It generates performance reports in a very nice graphical format. Refer to the following URL for more information: **ibm.com**/developerworks/aix/library/au-nmon_analyser/
nmon consolidator	Reads in nmon output files from several AIX nodes (including LPARs) to produce consolidated performance reports. **ibm.com**/developerworks/wikis/display/WikiPtype/nmonconsolidator
ndisk64	ndisk64 from nstress suite can be used to drive I/O by reading, writing large files from raw disks or file systems. The tool can be used to measure the I/O throughput that an environment can generate. The tool provides many options to read from one or more disks/files and the I/O can be 100% read, 100% write, or any mix of read-write (for example 60% read and 40%write). More details about the tool available at the following URL: **ibm.com**/developerworks/wikis/display/WikiPtype/nstress For example, if you want to measure 100% read throughput with a raw disk using the following command. It does 100% read from raw disk hdisk24 for 120 seconds with 32 threads in parallel. *$ /nstress/ndisk64 -f /dev/rhdisk24 -S -s 32768M -r 100 -t 120 -b 1M -M 32*

GPFS specific performance monitoring tools and techniques

mmpmon command

GPFS provides the ***mmpmon*** command to monitor the file system IO performance. The tool provides two types of I/O statistics:

- fs_io_s

  Displays the I/O statistics for each mounted file system, which helps in proper tuning of the GPFS cache and thread configuration.

- io_s

  Displays the I/O statistics for the entire system.

The type of I/O statistics collection can be passed in a command file to ***mmpmon***. For example,

```
[root@gridn1] / ==> # cat /tmp/mmpmon.txt
fs_io_s
[root@gridn1] / ==> # mmpmon -d 500 -r 2 -i /tmp/mmpmon.txt
mmpmon node 10.xx.xx.xx name gridn1 fs_io_s OK
cluster:        grid-cluster.unx.com
filesystem:     sasdata-gpfs
disks:                8
timestamp:      1342591726/838936
bytes read:         23147
bytes written:          0
opens:              493
closes:             493
reads:               21
writes:               0
readdir:            972
inode updates:    257
```

gpfs_perf Perl script

This is a Perl script written using the **mmpmon** command to collect the I/O statistics for a given file system in a node. Download it from the following URL: **ibm.com**/developerworks/wikis/display/hpccentral/mmpmon+scripts. For example, to know the I/O statistics for the ***saswork-gpfs*** file system on Grid Node1, run the following command on the node:

```
[root@gridn1] / ==> # perl gpfs_perf.pl  saswork-gpfs
Starting GPFS performance data gathering on host gridn1.
Duration is 300 seconds.

Wed Jul 18 02:24:40 EDT 2012
mmpmon node 10.16.0.8 name gridn1 rhist reset OK
mmpmon node 10.16.0.8 name gridn1 reset OK
mmpmon node 10.16.0.8 name gridn1 rhist on OK
Gathering Performance data for [saswork-gpfs] GPFS device.
 read      write
MB/sec  MB/sec    fopens  fclose     reads writes inodes
10.9     199.0        3       3       11   3454      1
 33.5     194.9        7       9       35   3451      1
 73.3     202.7       24      22      663   3733      0
 85.5     213.9        6       8     1190   3909      2
 90.6     126.4       27      27     1255   2394      2
157.8     116.4       15      13     1113   2309      1
```

```
42.0    212.3       14      17      1025   3716      0
```

The script runs for 1 minute, by default. The script can be changed to run for a longer time. If you want to monitor the I/O statistics of a file system on entire cluster, then the script needs to be run on all the cluster nodes for the file system. And, the data needs to be collated to understand the I/O statistics at cluster level for the file system.

GPFS diagnosis tools

GPFS provides the **mmdiag** and **mmfsadm** commands to identify any performance bottlenecks. These tools can be used along with performance monitoring tools to understand the bottlenecks and tune the GPFS cluster.

For example, the most common useful indicator of performance bottlenecks to the GPFS is the waiter information. This can be monitored by using the **mmdiag** command.

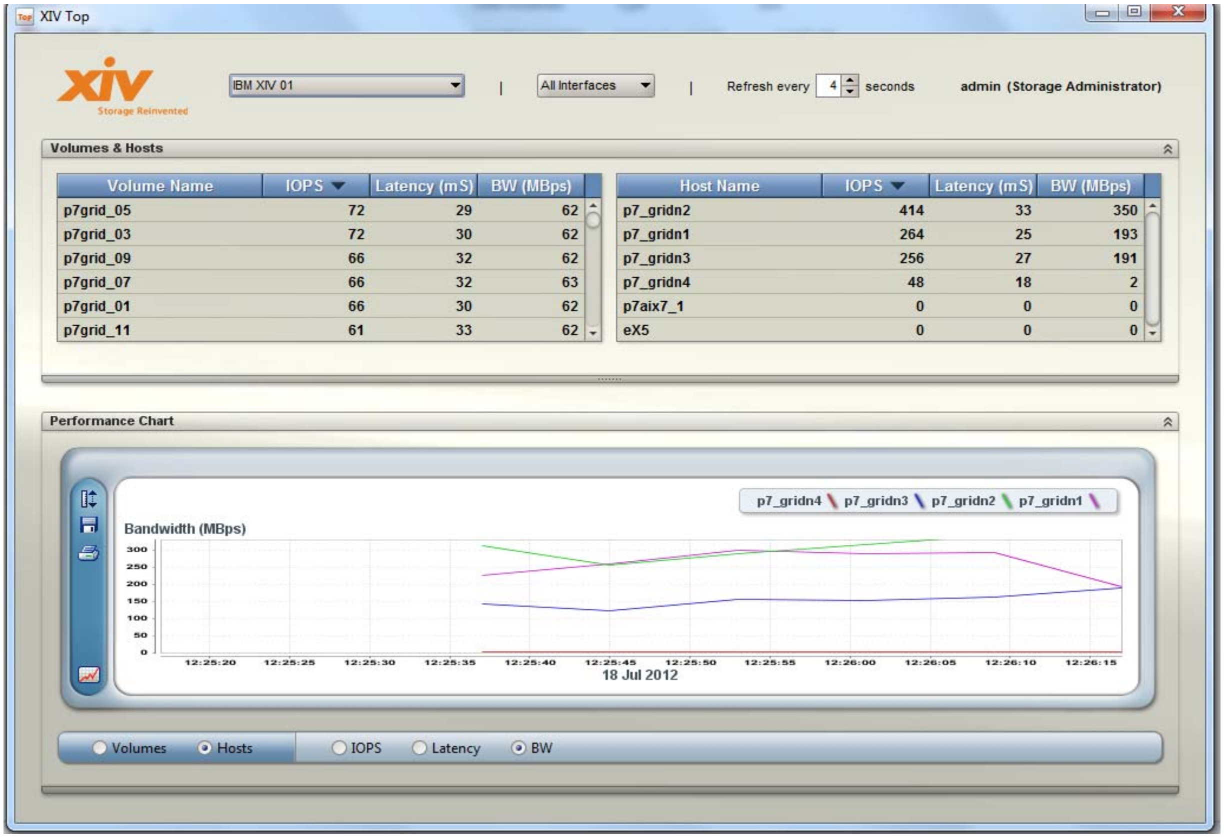
```
[root@ gridn1] / ==> # mmdiag --waiters

=== mmdiag: waiters ===

0x111BD22F0 waiting 0.072270834 seconds, WritebehindWorkerThread: for I/O completion

0x111BCC7D0 waiting 0.013595178 seconds, WritebehindWorkerThread: for I/O completion

0x111BDC410 waiting 0.018315647 seconds, PrefetchWorkerThread: for I/O completion

0x111A53650 waiting 0.029956725 seconds, PrefetchWorkerThread: for I/O completion

0x111B9BCB0 waiting 0.018127772 seconds, PrefetchWorkerThread: for I/O completion

0x110E0D170 waiting 0.040269190 seconds, WritebehindWorkerThread: for I/O completion

0x1110FA270 waiting 0.000892459 seconds, FileBlockRandomReadFetchHandlerThread: for I/O completion
```

XIV storage monitoring

The following tools were used to monitor the XIV Storage System during the benchmark activity: XIV TOP, XIV GUI, and XCLI. All the three tools are available as part of IBM XIV Management Tools package.

XIV TOP

XIV Top allows real-time monitoring for volumes and hosts simultaneously. The following screen capture describes the capabilities of XIV Top.



IBM XIV 01 | All Interfaces | Refresh every 4 seconds admin (Storage Administrator)

Volumes & Hosts

Volume Name	IOPS ▼	Latency (mS)	BW (MBps)
p7grid_05	72	29	62
p7grid_03	72	30	62
p7grid_09	66	32	62
p7grid_07	66	32	63
p7grid_01	66	30	62
p7grid_11	61	33	62

Host Name	IOPS ▼	Latency (mS)	BW (MBps)
p7_gridn2	414	33	350
p7_gridn1	264	25	193
p7_gridn3	256	27	191
p7_gridn4	48	18	2
p7aix7_1	0	0	0
eX5	0	0	0

Performance Chart

○ Volumes ● Hosts ○ IOPS ○ Latency ● BW

Figure 13: XIV Top view

XIV GUI

The XIV GUI acts as the management console for the XIV Storage System. A simple and intuitive GUI enables storage administrators to manage and monitor all the system aspects easily. The monitoring functions available from the XIV GUI allow the user to easily display and review the overall system status, events, and several performance statistics. The tool is installed as part of IBM XIV Management Tools package available for Microsoft® Windows® PC.

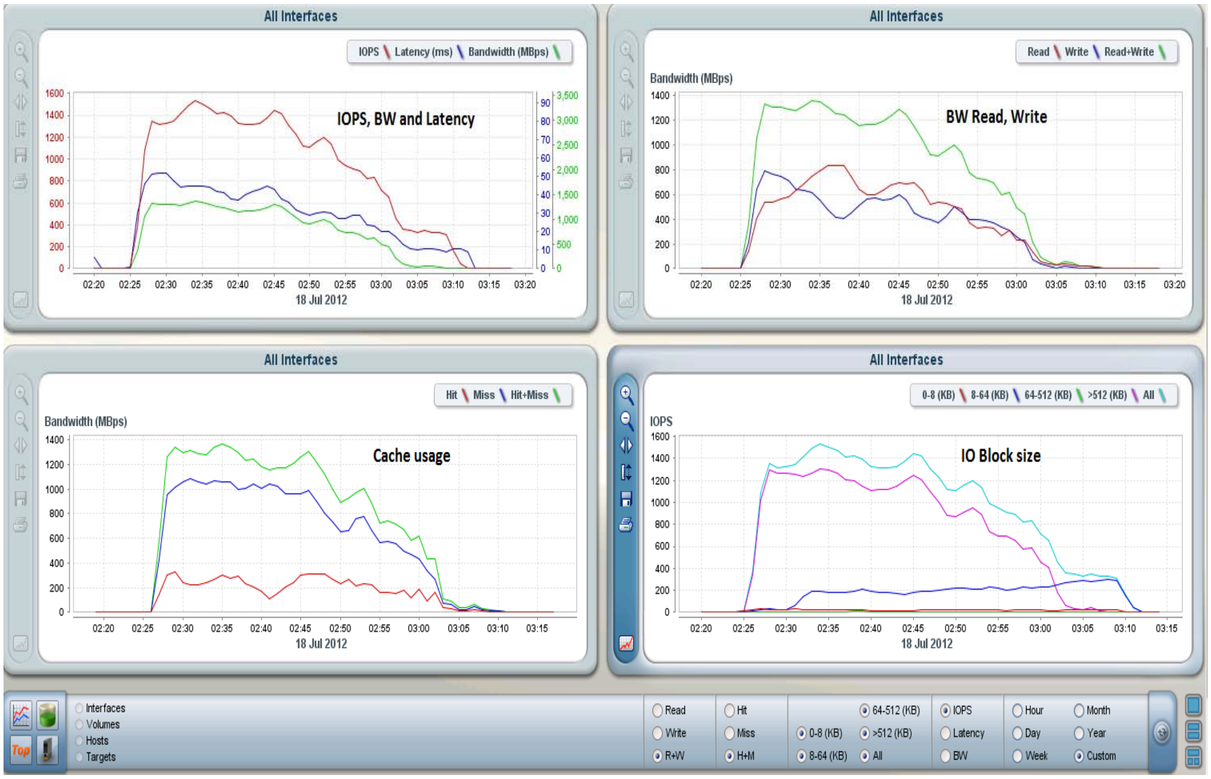
The Statistics Monitor provides intuitive interface to collect the following statistics:

- IOPS, BW, latency, block size details
- Read-Write I/O mix
- Cache hit/miss information

The statistics can be collected based on:

- Entire XIV system, or individual nodes
- Volumes
- FC interfaces

The statistics monitor allows saving the statistics in the CSV format to allow users to analyze the data at a later point. And it also allows users to view and save statistics for an earlier date and time.



○ Read ○ Hit ● 64-512 (KB) ● IOPS ○ Hour ○ Month
○ Write ○ Miss ● 0-8 (KB) ● >512 (KB) ○ Latency ○ Day ○ Year
● R+W ● H+M ● 8-64 (KB) ● All ○ BW ○ Week ● Custom

Figure 14: The XIV statistics monitor view

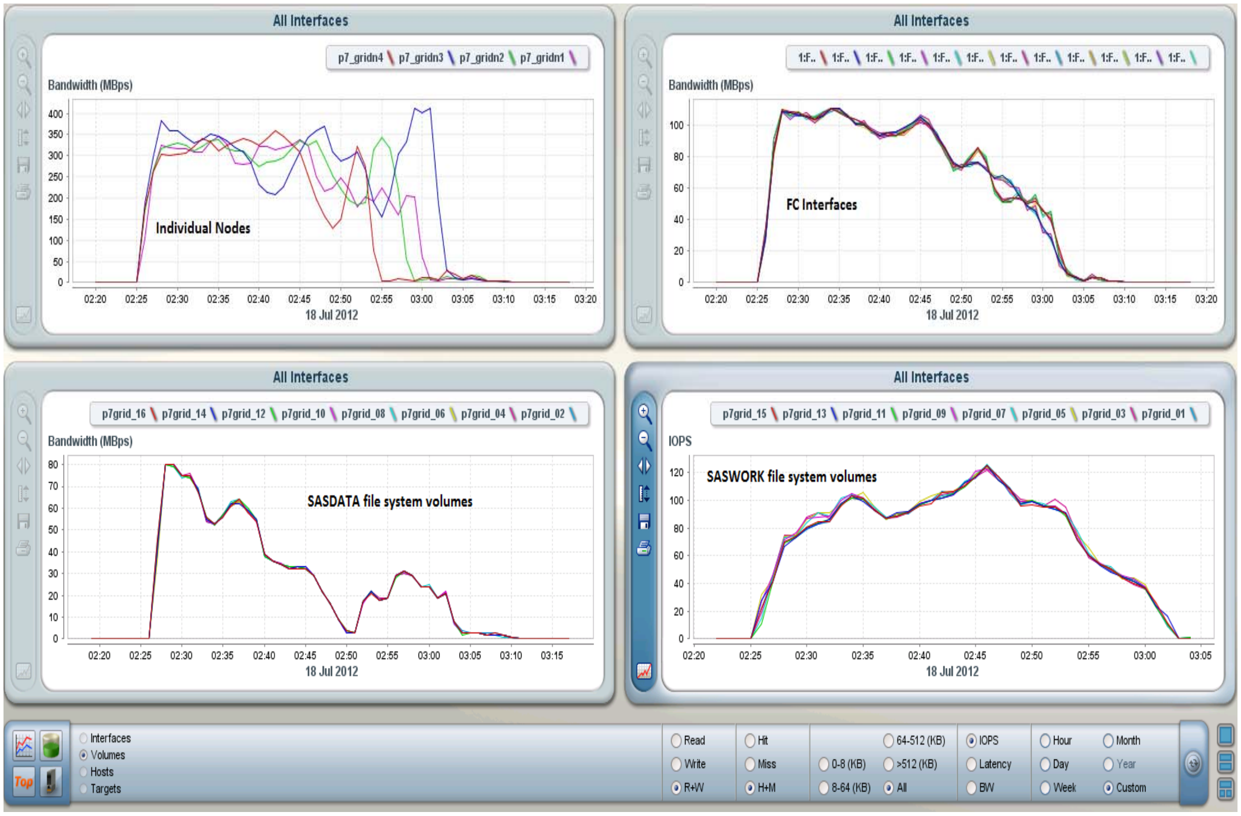


Figure 15:Another view of the XIV statistics monitor

XCLI

The IBM XIV Storage System command-line Interface (XCLI) provides several commands to monitor the XIV Storage System and gather real-time system status, monitor events, and retrieve performance statistics. XCLI **statistics_get** command retrieves performance statistics from the XIV system. XCLI provides detailed performance statistics. XCLI is installed as part of the IBM XIV Management Tools package, which can be installed on a Microsoft® Windows® PC.

For example, to retrieve XIV I/O performance statistics for the 40 session benchmark, run the following command from Windows or the Cygwin command prompt. The command retrieves 60 performance statistics starting from 2012-07-18.02:25:00, with an interval of 1 min.

```
# /cygdrive/c/Program\ Files\ \(x86\)/XIV/GUI10/xcli.exe -m 127.0.1.27 -u admin -p xxxxxx -s statistics_get start=2012-07-18.02:25:00 interval=1 resolution_unit=minute count=60 > tmp/xiv-perf-stats.csv
```

The performance output can be stored in a CSV file, which can later be imported into Microsoft® Excel sheet for further analysis. XCLI generates detailed data and users are advised to use Microsoft® Excel macros to derive cumulative I/O statistics, including: throughput, IOPS, latency, and cache usage. With the XCLI 2.4.4 version, the benchmark team imported the CSV file generated with above command into a Microsoft® Excel sheet and used the following macros to derive the cumulative I/O throughput.

Cumulative read throughput (Kbps): = SUM(D2,G2,J2,M2,P2,S2,V2,Y2)

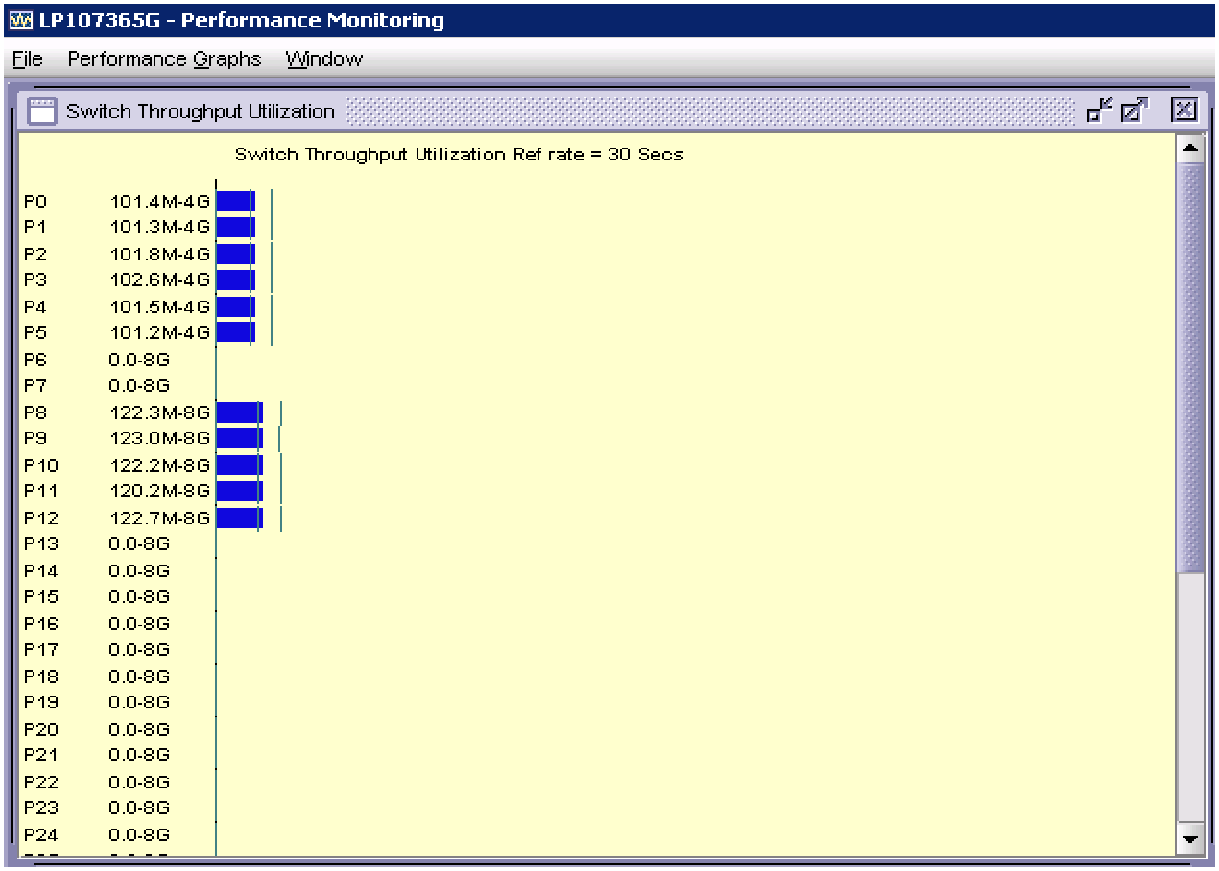
Cumulative write throughput (Kbps): = SUM(AB2,AE2,AH2,AK2,AN2,AQ2,AT2,AW2)

Aggregate throughput at the XIV system for the workload (Kbps): = SUM(AY3,AZ3)

Performance monitoring at SAN Switch

It is advised to monitor the I/O throughput through the SAN Switch performance monitor to ensure that I/O is distributed through all the connected ports in the switch. And, if redundant SAN switches are used, monitor both the switches.

The deployment architecture used for benchmark has two SAN switches. The following figure shows the performance monitoring view from the Switch Explorer interface provided by the Brocade switches.


LP107365G - Performance Monitoring
File
Performance Graphs
Window
Switch Throughput Utilization

Switch Throughput Utilization Ref rate = 30 Secs

P0	101.4M-4G
P1	101.3M-4G
P2	101.8M-4G
P3	102.6M-4G
P4	101.5M-4G
P5	101.2M-4G
P6	0.0-8G
P7	0.0-8G
P8	122.3M-8G
P9	123.0M-8G
P10	122.2M-8G
P11	120.2M-8G
P12	122.7M-8G
P13	0.0-8G
P14	0.0-8G
P15	0.0-8G
P16	0.0-8G
P17	0.0-8G
P18	0.0-8G
P19	0.0-8G
P20	0.0-8G
P21	0.0-8G
P22	0.0-8G
P23	0.0-8G
P24	0.0-8G

Figure 16: Switch Throughput view for each port

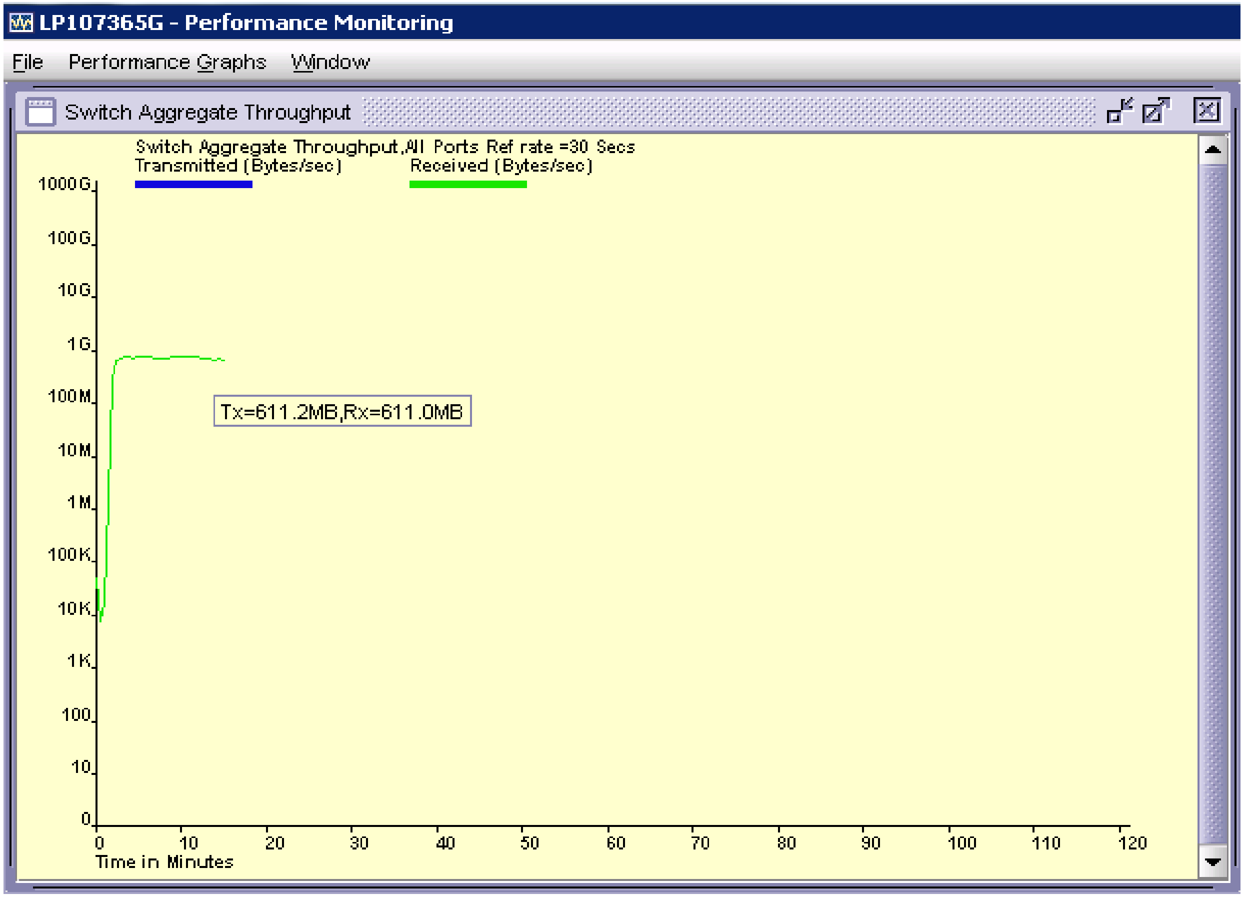


Figure 17: Switch Aggregate throughput view

Resources

These web resources provide useful references to supplement the information contained in the paper.

- IBM Power Systems Information Center
  http://publib.boulder.ibm.com/infocenter/pseries/index.jsp
- Power Systems on IBM PartnerWorld®
  ibm.com/partnerworld/systems/p
- AIX on IBM PartnerWorld
  ibm.com/partnerworld/aix
- IBM Systems on IBM PartnerWorld
  ibm.com/partnerworld/systems/
- IBM Publications Center
  www.elink.ibmlink.ibm.com/public/applications/publications/cgibin/pbi.cgi?CTY=US
- IBM Redbooks
  ibm.com/redbooks
- IBM developerWorks®
  ibm.com/developerworks
- AIX 7.1 tuning guide
  ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101529
- Running SAS applications on Virtual I/O based storage configuration
  ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101664
- IBM PowerVM Virtualization Managing and Monitoring
  ibm.com/redbooks/abstracts/sg247590.html
- Storage best practices: SAS 9 with IBM System Storage and IBM System P
  http://www.sas.com/partners/directory/ibm/SAS_IBM_Storage_Best_Practices0311.pdf
- IBM XIV Storage System: Architecture, Implementation, and Usage
  ibm.com/redbooks/abstracts/sg247659.html
- IBM XIV Storage System: Host Attachment and Interoperability
  ibm.com/redbooks/abstracts/sg247904.html
- General Parallel File System - Document Library:
  http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html
- General Parallel File System FAQs (GPFS FAQs):
  http://publib.boulder.ibm.com/infocenter/clresctr/topic/com.ibm.cluster.gpfs.doc/gpfs_faqs/gpfs_faqs.html

- GPFS Concepts, Planning, and Installation Guide
  http://publib.boulder.ibm.com/epubs/pdf/a7604135.pdf
- GPFS Administration and Programming Reference
  http://publib.boulder.ibm.com/epubs/pdf/a2322215.pdf
- GPFS Tuning Parameters
  ibm.com/developerworks/wikis/display/hpccentral/GPFS+Tuning+Parameters
- A survey of shared file systems – SAS
  http://support.sas.com/rnd/scalability/papers/SurveyofSharedFilepaperApr25final.pdf
- Grid Computing in SAS 9.3
  http://support.sas.com/documentation/cdl/en/gridref/64808/PDF/default/gridref.pdf
- SAS Grid Computing
  - http://www.sas.com/resources/factsheet/sas-grid-computing-factsheet.pdf
  - http://www.sas.com/technologies/architecture/grid/index.html
- Administering Platform LSF - SAS
  http://support.sas.com/rnd/scalability/platform/PSS5.1/lsf7.05_admin.pdf
- Running Jobs with Platform LSF
  http://support.sas.com/rnd/scalability/platform/PSS5.1/lsf7.05_users_guide.pdf

Acknowledgements

I would like to thank Frank Battaglia, Harry Seifert, Jerrold Heyman, David Quenzler, Adam Diaz, Scott Fadden, Glenn Fujimoto of IBM for guidance and review of the paper. I would also thank Margaret Crevar of SAS for reviewing the paper.

About the author

Narayana Pattipati works as a technical consultant in the IBM Systems and Technology Group ISV Technical Enablement team. He has more than 12 years of experience in the IT industry. He currently works with software vendors to enable their software solutions on IBM platforms. You can reach Narayana at npattipa@in.ibm.com.

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

ACCELERATE INTELLIGENCE®, ACCELERATING INTELLIGENCE®, LSF®, and LSF ACTIVECLUSTER® are trademarks or registered trademarks of Platform Computing, an IBM Company.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.