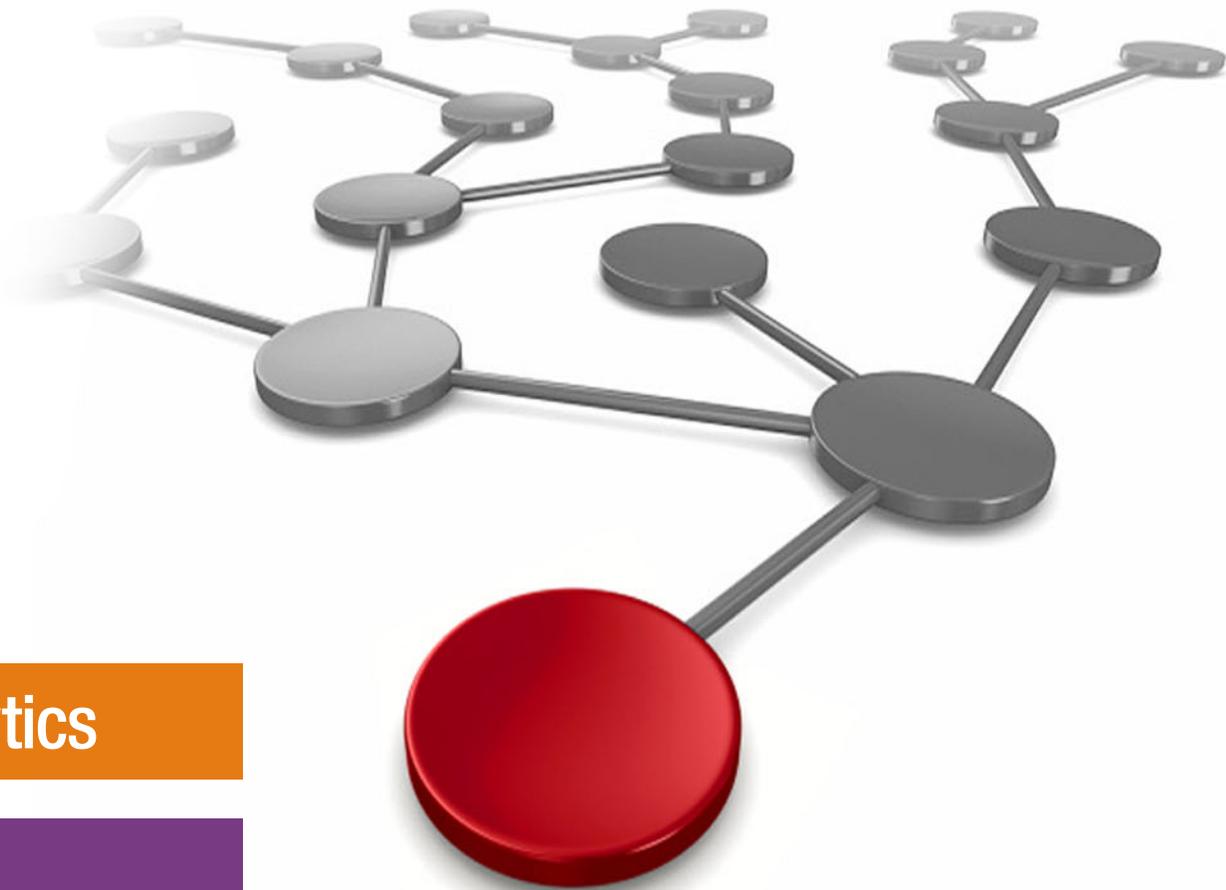


IBM Storage Solutions for SAS Analytics using IBM Spectrum Scale and IBM Elastic Storage System 3000 Version 1 Release 1

Sanjay Sudam



 Analytics

Storage



Introduction

This IBM® Redpaper® publication is a blueprint for configuration, testing results, and tuning guidelines for running SAS workloads on Red Hat Enterprise Linux that use IBM Spectrum® Scale and IBM Elastic Storage® System (ESS) 3000. IBM lab validation was conducted with the Red Hat Linux nodes running with the SAS simulator scripts that are connected to the IBM Spectrum Scale and IBM ESS 3000.

Simultaneous workloads are simulated across multiple x-86 nodes running with Red Hat Linux to determine scalability against the IBM Spectrum Scale clustered file system and ESS 3000 array. This paper outlines the architecture, configuration details, and performance tuning to maximize SAS application performance with the IBM Spectrum Scale 5.0.4.3 and IBM ESS 3000.

This document is intended to facilitate the deployment and configuration of the SAS applications that use IBM Spectrum Scale and IBM Elastic Storage System (ESS) 3000.

The information in this document is distributed on an “as is” basis without any warranty that is either expressed or implied. Support assistance for the use of this material is limited to situations where IBM Spectrum Scale or IBM ESS 3000 are supported and entitled and where the issues are specific to a blueprint implementation.

Scope

This blueprint guide provides a solutions architecture and related solution configuration workflows, with the following essential components:

- ▶ IBM Spectrum Scale
- ▶ IBM Elastic Storage System 3000
- ▶ SAS Analytics
- ▶ Mellanox Ethernet Switches

This technical report does not replace any official manuals and documents that were produced by:

- ▶ IBM
- ▶ SAS Institute Inc.
- ▶ Mellanox Technologies

Prerequisites

This technical paper assumes that the user has basic knowledge of the following technology:

- ▶ IBM Elastic Storage System 3000
- ▶ IBM Spectrum Scale
- ▶ SAS
- ▶ Mellanox Switches
- ▶ IP networking

Solution architecture and components

This section describes the solution building blocks that are used for validating the solution in the lab.

The IBM Elastic Storage System 3000 that is shown in Figure 1 combines the performance of NVMe storage technologies with the reliability and the rich features of IBM Spectrum Scale, along with several high-speed attachment options, such as 100 Gbps Ethernet and InfiniBand, all in a powerful 2U storage system.

With each of these drive options, IBM Spectrum Scale on NVMe is the market leader in all-flash performance and scalability with a bandwidth of approximately 40 GBps per NVMe all-flash appliance and 100 microseconds latency.

Providing data-driven multicloud storage capacity, the NVMe all-flash appliance is deeply integrated with the software defined capabilities of IBM Spectrum Storage™ to seamlessly plug it into Analytics workload.



Figure 1 IBM Elastic Storage System 3000

Table 1 lists the key specifications of the IBM Elastic Storage System 3000.

Table 1 IBM Elastic Storage System 3000 Key Specifications

Building blocks	Specifications
System Features	<ul style="list-style-type: none"> ▶ Dual 2-socket Storage Controllers, Active/Active ▶ 384 GB or 768 GB memory per controller ▶ De-Clustered RAID supporting erasure coding schemas: 3-way replication, 4-way replication, 4+2P, 4+3P, 8+2P, and 8+3P
Performance	<ul style="list-style-type: none"> ▶ Sequential read performance up to 42GBps ▶ Sequential write performance up to 32GBps
Networking	<ul style="list-style-type: none"> ▶ EDR InfiniBand, up to 12 ports ▶ 100G Ethernet, up to 12 ports
Drive Support	12 or 24 NVMe SSDs (1.92 TB, 3.84 TB, 7.68 TB, or 15.36 TB)

For more information about ESS 3000 specifications, see [this web page](#).

IBM Spectrum Scale and IBM Spectrum Scale RAID

IBM Spectrum Scale is a high-performance, highly available, clustered file system that is available on various platforms, including the public cloud service providers. It provides concurrent access to a single file system or set of file systems from multiple nodes. A key ability that IBM Spectrum Scale provides is a single namespace (or data plane) so that each data source can add data to the repository by using NFS, SMB, Object, or a POSIX interface.

Another key feature of IBM Spectrum Scale is that it enables data to be tiered automatically and transparently to and from more cost-effective storage, including hard disk drives (HDD), tape, and cloud. Over the course of a decade, this feature saves customers an order of magnitude of cost and provides simplified access at the same time.

IBM Spectrum Scale RAID is a software implementation of storage erasure code technologies within IBM Spectrum Scale that provides sophisticated data placement and error-correction algorithms to deliver high levels of storage performance, availability, and reliability. Shared file systems are created from the Network Shared Disks (NSD) that are defined with IBM Spectrum Scale RAID. This file system can be accessed concurrently by all compute nodes in the configuration to efficiently meet the capacity and performance requirements of modern scale out applications, such as AI.

Network - Mellanox switches and adapters

Typically, Ethernet is not the first choice in storage fabrics. Traditionally, the choice when running an analytics workload, such as the SAS Mixed Analytics workload, was Fibre Channel for block I/O and possibly InfiniBand for file I/O. However, the improved 100 Gbps Ethernet networks that are built with the Mellanox switches accomplished the task well.

The SN2100 switch is an ideal spine and top of rack (ToR) solution. It offers maximum flexibility with port speeds of 10 Gbps - 100 Gbps per port and port density that enables full rack connectivity to any server at any speed. The uplink ports realize various blocking ratios that suit any application requirement.

The Mellanox ConnectX-5 Ethernet adapters provide high performance and flexible solutions with up to two ports of 100 GbE connectivity. These adapters enabled the high-performance network for running the SAS mixed analytics and deliver the sustained throughput requirements of larger systems with the higher CPU cores.

Solution lab validation

This section describes the lab installation, architecture, configuration, and validation of the SAS mixed workloads with the IBM Spectrum Scale and IBM ESS 3000.

The converged Ethernet infrastructure was created by using an IBM ESS 3000, Mellanox 100 Gbps Ethernet switches, and x86 servers with connect-x5 adapters running with the Red Hat Enterprise Linux 8.1 version for the SAS workloads. The main building block of the storage systems is IBM Spectrum Scale high-speed clustered shared file system. Figure 2 shows the infrastructure that is used in the lab validation.

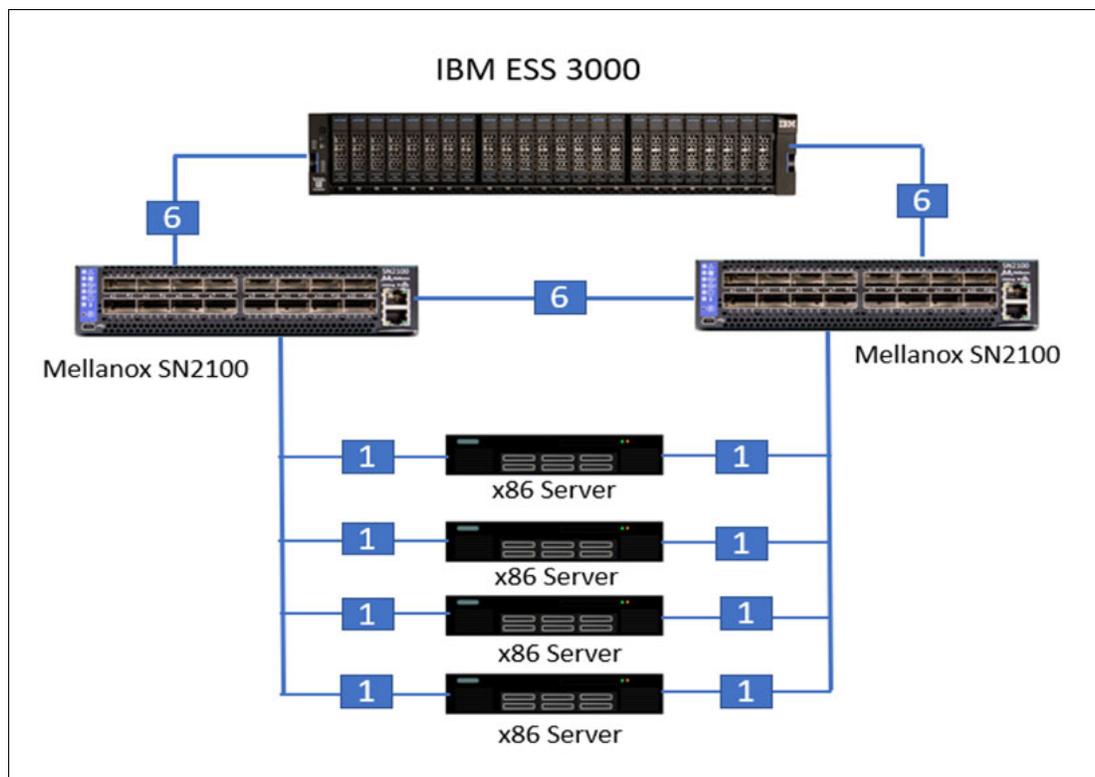


Figure 2 Ethernet converged architecture for SAS mixed workloads

Storage configuration

The lab storage featured the following configuration:

- ▶ IBM Elastic Storage System 3000 with 24 x 7.6 TB NVMe drive, 12 x 100 Gbps Ethernet ports
- ▶ IBM Spectrum Scale RAID Code
- ▶ IBM Spectrum Scale 5.0.4.3

Host configuration

The lab host features the following configuration:

- ▶ x86 Server with Intel Xeon Platinum 8160 CPU @ 2.10 GHz
- ▶ Memory: 256 GB

- ▶ Operating system level: Red Hat Enterprise Linux 8
- ▶ Network: Mellanox Connectx-5 dual port adapter with MLNX_OFED_LINUX-4.7-3.2.9.0

Ethernet switches

The Mellanox SN2100 with 16 ports Ethernet each with 100 Gbps Ethernet switches were used.

Networking configuration

The lab networking featured the following configuration:

- ▶ Ethernet MTU was changed to the 9000 on the Red Hat Linux client nodes and IBM ESS 3000.
- ▶ Ethernet MTU=9216 was configured for each switch port.
- ▶ On the ESS 3000 nodes, network tunable parameters are preset and configured part of the ESS 3000 installation process.
- ▶ The following Ethernet adapter parameters are configured on the Linux nodes:

```
ethtool -G ens6f0 rx 8192 tx 8192
ethtool -G ens6f1 rx 8192 tx 8192
mlnx_tune -r -c
ethtool -K ens6f0 tx-nocache-copy off
ethtool -K ens6f1 tx-nocache-copy off
```

- ▶ Ethernet network bonding is recommended on the requiring high availability of the networks.
- ▶ The following bonding configuration of the Linux nodes was used:

```
nmcli con add type bond con-name bond0 ifname bond0
nmcli con add type bond-slave ifname ens5f0 master bond0
nmcli con add type bond-slave ifname ens5f1 master bond0
nmcli connection modify bond0 bond.options
"miimon=100,mode=4,xmit_hash_policy=layer3+4"
nmcli con mod bond0 ipv4.method manual ipv4.address "192.0.2.61/24"
```

IBM Spectrum Scale configuration

The following settings were applied to the IBM Spectrum Scale 5.4.0.3 installation, in varying parameter combinations to determine optimal performance for this validation in the lab:

```
Ethernet Fabric = 100GbE
maxFilestoCache=50000
maxMBpS=10000 on the Linux nodes
maxMBpS=24000 on the ESS 3000 nodes
workerThreads=1024
Pagepool Size=128GB Linux nodes
prefetchPct=40 on Linux nodes
```

Pagepool parameter for the ESS 3000 nodes are configured to optimal value during the installation process. Only Linux client nodes Pagepool values are changed, not the ESS nodes during the lab validation.

IBM recommends a large page pool space for IBM Spectrum Scale implementations with SAS workloads. The Pagepool size depends upon the available system memory per client node in the IBM Spectrum Scale cluster and the specific SAS workload requirements. For this workload, the test team configured it to 128 GB.

IBM Spectrum Scale tuning was based on the previous SAS solution validation that was done by the IBM team for IBM Spectrum Scale running on the older generation of the Elastic Storage System with SAS MA20 workloads, which is available at [web page](#).

Shared file system

IBM Spectrum Scale 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 version 5.4.0.3 is the building platform of this solution. The IBM Spectrum Scale file system parameters were tuned for throughput that supports large block sequential I/O, which is the most important factor when configuring storage for SAS performance.

Data and I/O throughput

The SAS I/O pattern is predominately large-block, sequential I/O. Some random access does occur, but sequential is the dominant access pattern. When configuring for SAS I/O, 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, are used. However, it is the large sequential block I/O that dominates all of these patterns. This fact helps in configuring the file systems (see Figure 3).

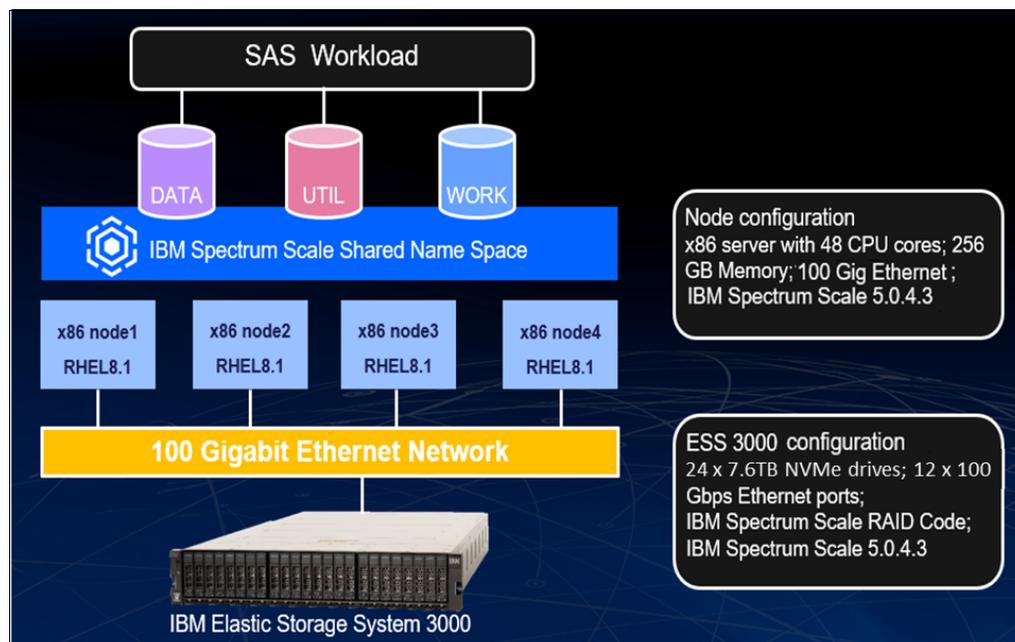


Figure 3 IBM ESS 3000 lab configuration for SAS mixed workloads

The following primary file systems are involved in the SAS configurations¹:

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

The I/O throughput requirements for the SAS compute tier are as follows²:

- ▶ For the WORK/UTILLOC file system, SAS requires a *minimum* I/O throughput rate of 125 MBps per physical CPU core.

¹ <http://support.sas.com/resources/papers/performance-tuning-sas-red-hat-using-ibm-spectrum-scale.pdf>

² SAS Usage Note 59680: Testing throughput for your SAS 9 File Systems: The rhel_iotest.sh script

- ▶ For permanent SAS data files, SAS requires a *minimum* I/O throughput rate of 75 - 100 MBps per physical CPU core.
- ▶ Overall, I/O throughput should be at a *minimum* of 100 - 125 MBps per physical CPU core.

Testing the converged infrastructure

Remote direct memory access (RDMA) over Converged Ethernet (RoCE) is a network protocol that allows RDMA over an Ethernet network. RoCE enables RDMA's efficient data transfer over Ethernet networks to enable transport offload with hardware RDMA engine implementation and superior performance.

Note: Consider the following RDMA configuration guidelines:

- ▶ An RPQ is required for RoCE support with the IBM ESS 3000. For more information about the RPQ or SCORE process, contact your sales representative and ask them to contact IBM Spectrum Scale development.
- ▶ Two different network subnets are recommended at the IBM Spectrum Scale cluster:
 - One subnet for IBM Spectrum Scale daemon network, which is a regular TCP/IP network for IBM Spectrum Scale cluster daemon communication usage.
 - Second subnet for RoCE configuration, which is data access network to the IBM ESS 3000 for the superior I/O performance.

RoCE configuration on the nodes

Two ports on each IO node are configured for the RoCE configuration, one from each adapter. A total of 4 x 100 GbE ports from the IBM ESS 3000 are configured for the RoCE RDAM usage at the IBM Spectrum Scale cluster level:

- ▶ Configure I/O node 1

Adapters enp94s0f1 and enp216s0f1 are used for the RoCE usage in the lab testing validation:

```
[root@io1-ess3k ~]# ibdev2netdev
mlx5_0 port 1 ==> enp94s0f0 (Up)
mlx5_1 port 1 ==> enp94s0f1 (Up)
mlx5_2 port 1 ==> enp216s0f0 (Up)
mlx5_3 port 1 ==> enp216s0f1 (Up)
```

- ▶ Configure buffers and transmit queue length parameters of the Ethernet interfaces by way of udev rules in /etc/udev/rules.d:

```
[root@io1-ess3k rules.d]# cat 99-ibm-network-custom.rules
KERNEL=="enp94*", RUN+="/sbin/ethtool -G %k rx 8192" , RUN+="/sbin/ethtool -G %k tx 8192" , RUN+="/sbin/ip link set %k txqueuelen 10000"
KERNEL=="enp21*", RUN+="/sbin/ethtool -G %k rx 8192" , RUN+="/sbin/ethtool -G %k tx 8192" , RUN+="/sbin/ip link set %k txqueuelen 10000"
[root@io1-ess3k rules.d]# udevadm control --reload-rules
[root@io1-ess3k rules.d]# udevadm trigger
```

- ▶ Configure the MTU and IPV6 parameters

Add the following parameters to the interface file in the /etc/sysconfig/network-scripts folder:

```
MTU=9000
IPV6_ADDR_GEN_MODE=eui64
```

```
[root@io1-ess3k network-scripts]# cat ifcfg-enp94s0f1
```

```

TYPE=Ethernet
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=none
DEFROUTE=yes
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=yes
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
NAME=enp94s0f1
UUID=eb863062-b136-420f-982c-82c0affe1702
DEVICE=enp94s0f1
ONBOOT=yes
MTU=9000
IPADDR=192.0.3.101
PREFIX=24
IPV6_ADDR_GEN_MODE=eui64

```

```

[root@io1-ess3k network-scripts]# cat ifcfg-enp216s0f1
TYPE=Ethernet
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=none
DEFROUTE=yes
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=yes
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
NAME=enp216s0f1
UUID=eac9626b-d572-4829-80a0-f9d8d1995243
DEVICE=enp216s0f1
ONBOOT=yes
IPADDR=192.0.3.102
PREFIX=24
MTU=9000
IPV6_ADDR_GEN_MODE=eui64

```

- Configure RoCE related parameters on the system

Add the following entries to the `/etc/sysconfig` configuration file and run `sysctl -p`:

```

net.ipv4.tcp_ecn=1
net.ipv4.conf.default.arp_filter=1
net.ipv4.conf.all.arp_announce=2
net.ipv4.conf.default.arp_announce=2
net.ipv6.conf.enp216s0f1.disable_ipv6=0
net.ipv6.conf.enp94s0f1.disable_ipv6=0

```

- Because we use many IP addresses in the same subnet for RoCE, configure the routes manually so that both of the interfaces are used for network traffic.

Add routing table `t1` and `t2` entries to the `/etc/iproute2/rtables`:

```

[root@io1-ess3k ~]# cat /etc/iproute2/rt_tables
#
# reserved values
#

```

```

255    local
254    main
253    default
0      unspec
#
# local
#
#1     inr.ruhep
200    t1
201    t2

```

- Define the routes and rules by using the tables that are defined in the previous step:

```

[root@io1-ess3k ~]# ip route add 192.0.3.0/24 dev enp94s0f1 src 192.0.3.101
table t1
[root@io1-ess3k ~]# ip route add table t1 default via 192.0.3.101 dev enp94s0f1
[root@io1-ess3k ~]# ip rule add table t1 from 192.0.3.101

```

```

[root@io1-ess3k ~]# ip route add 192.0.3.0/24 dev enp216s0f1 src 192.0.3.102
table t2
[root@io1-ess3k ~]# ip route add table t2 default via 192.0.3.102 dev enp216s0f1
[root@io1-ess3k ~]# ip rule add table t2 from 192.0.3.102

```

```

[root@io1-ess3k ~]# ip route show table t1
default via 192.0.3.101 dev enp94s0f1
192.0.3.0/24 dev enp94s0f1 scope link src 192.0.3.101
[root@io1-ess3k ~]#
      [root@io1-ess3k ~]# ip route show table t2
default via 192.0.3.102 dev enp216s0f1
192.0.3.0/24 dev enp216s0f1 scope link src 192.0.3.102
[root@io1-ess3k ~]#

```

```

[root@io1-ess3k ~]# ip route show
192.0.2.0/24 dev enp94s0f0 proto kernel scope link src 192.0.2.11 metric 107
192.0.3.0/24 dev enp94s0f1 proto kernel scope link src 192.0.3.101 metric 108
192.0.3.0/24 dev enp216s0f1 proto kernel scope link src 192.0.3.102 metric 109
192.168.20.0/24 dev enp29s0f1 proto kernel scope link src 192.168.20.21 metric
104

```

To make the routes and rules persistent across the restart, create the corresponding interface files in `/etc/sysconfig/network-scripts` or to add a start script so that the commands are run during the start process.

- Configure the QoS parameters for RoCE:

```

sysctl -w net.ipv4.tcp_ecn=1

mlnx_qos -i enp216s0f1 --trust dscp
mlnx_qos -i enp94s0f1 --trust dscp
mlnx_qos -i enp216s0f1 --pfc 0,0,0,1,0,0,0,0
mlnx_qos -i enp94s0f1 --pfc 0,0,0,1,0,0,0,0

ifdown enp216s0f1
ifdown enp94s0f1
ifup enp94s0f1
ifup enp216s0f1

```

```
echo 106 > /sys/class/infiniband/mlx5_1/tc/1/traffic_class
echo 106 > /sys/class/infiniband/mlx5_3/tc/1/traffic_class
cma_roce_tos -d mlx5_1 -t 106
cma_roce_tos -d mlx5_3 -t 106
```

To make QoS parameters persistent across the restart, add a start script with these commands.

Repeat the previous steps on the second ESS 3000 IO node and Red Hat Linux nodes to enable RoCE configuration.

- Configure the RDMA parameters at the IBM Spectrum Scale cluster:

```
[root@io1-ess3k ~]# mmchconfig verbsRdma=enable -N
ess3k1a-hs,ess3k1b-hs,isv517-hs,isv650-hs,isv650_02-hs
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
[root@io1-ess3k ~]# mmchconfig verbsRdmaCm=enable -N
ess3k1a-hs,ess3k1b-hs,isv517-hs,isv650-hs,isv650_02-hs
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
[root@io1-ess3k ~]#

[root@io1-ess3k ~]# mmchconfig verbsPorts="mlx5_1/1 mlx5_3/1" -N
ess3k1a-hs,ess3k1b-hs
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
[root@io1-ess3k ~]# mmchconfig verbsPorts="mlx5_2/1" -N isv650-hs
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
[root@io1-ess3k ~]#

[root@io1-ess3k ~]# mmchconfig verbsRdmasPerConnectionOverride=4 -N
ess3k1a-hs,ess3k1b-hs,isv517-hs,isv650-hs,isv650_02-hs
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
[root@io1-ess3k ~]#
```

Solution testing and results

Many test cases were performed (varying the nodes, number of tests, GbE size, and IBM Spectrum Scale cluster configuration parameters) and the tests were focused on the following scenarios:

- Using 100 GbE TCP/IP network
- Using RDMA over Converged Ethernet (RoCE)

This publication highlights the results from the SAS workload simulator by using the I/O shell script utility for Red Hat Enterprise Linux environments³.

³ [SAS Usage Note 59680](#): Testing throughput for your SAS 9 File Systems: The rhel_iotest.sh script

SAS I/O shell script

SAS provides an automated utility that uses Linux `dd` commands to measure the I/O throughput of a file system in a Red Hat Enterprise Linux (RHEL) environment. This utility mimics the behavior of the SAS workloads and helps in estimating the I/O throughput from the storage systems.

For more information about configuring the utility on the Red Hat Linux nodes, see [this web page](#).

During the lab validations, we configured the utility across four RHEL nodes and ran the following tests:

- ▶ Single node
- ▶ 2 nodes
- ▶ 4 nodes

Single-node simulation

Testing was conducted from a single x86 node with 48 CPU cores, 256 GB Memory, 100 GbE, and Red Hat Enterprise Linux 8.1 operating system. The node was configured with the RoCE for accessing the ESS 3000. The script results are shown in Figure 4.

```
[root@isvmlnx650-2 singlenode]# cat rhel_iotest.results
-----
RESULTS
-----
INVOCATION:  rhel_iotest -t /gpfs/Test_8M/isv650-02/singlenode

TARGET DETAILS
directory:   /gpfs/Test_8M/isv650-02/singlenode
df -k:      Test_8M          64375783424 7733248 64368050176  1% /gpfs/Test_8M
mount point: Test_8M on /gpfs/Test_8M type gpfs (rw,nodev,relatime,seclabel)
filesize:   251.33 gigabytes

STATISTICS
read throughput rate:  197.33 megabytes/second per physical core
write throughput rate: 244.81 megabytes/second per physical core
-----
```

Figure 4 Single node SAS IO script results

Figure 5 shows IBM ESS 3000 performance throughput (GBps [Y] over time [X]) during the single node validation for both read and write operations from the IBM ESS 3000 Web GUI console.

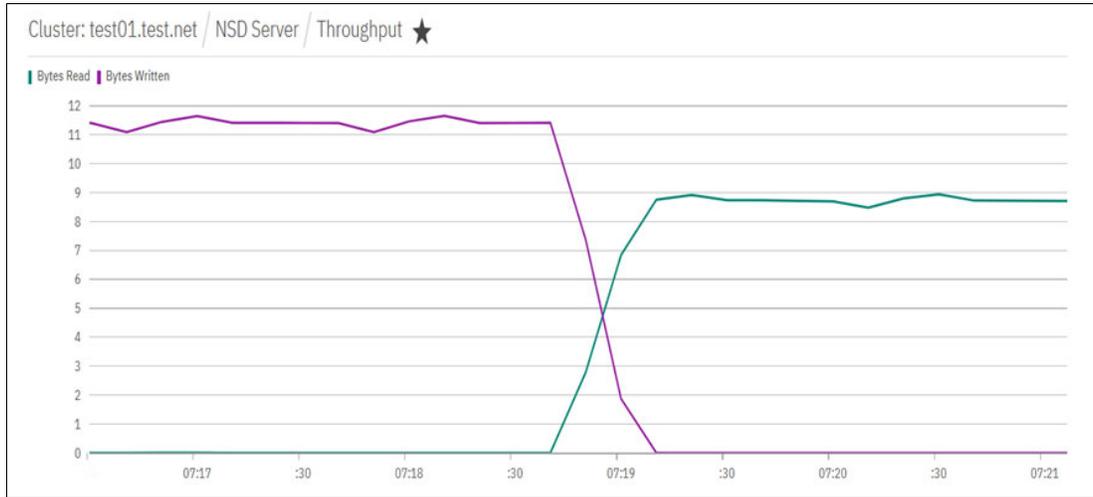


Figure 5 IBM ESS 3000 throughput during single node validation

4-Node simulation

During this run (see Figure 6), four nodes are configured with the SAS I/O Shell script simulator to mimic the SAS workloads on the IBM Spectrum Scale file system.

```
[root@isvmlnx650-2 isv650-02]# cat rhel_iotest.results
-----
RESULTS
-----
INVOCATION:  rhel_iotest -t /gpfs/Test_8M/isv650-02

TARGET DETAILS
  directory:  /gpfs/Test_8M/isv650-02
  df -k:      Test_8M      64375783424 7725056 64368058368 1% /gpfs/Test_8M
  mount point: Test_8M on /gpfs/Test_8M type gpfs (rw,nodev,relatime,seclabel)
  filesize:   251.33 gigabytes

STATISTICS
  read throughput rate:  197.93 megabytes/second per physical core
  write throughput rate: 162.63 megabytes/second per physical core
-----

[root@isvmlnx650-2 isv650-02]#
```

Figure 6 4-node concurrent I/O throughput

IBM ESS 3000 approached its performance limits during the 4-node validation. Figure 7 shows GBps (Y) over time (X).

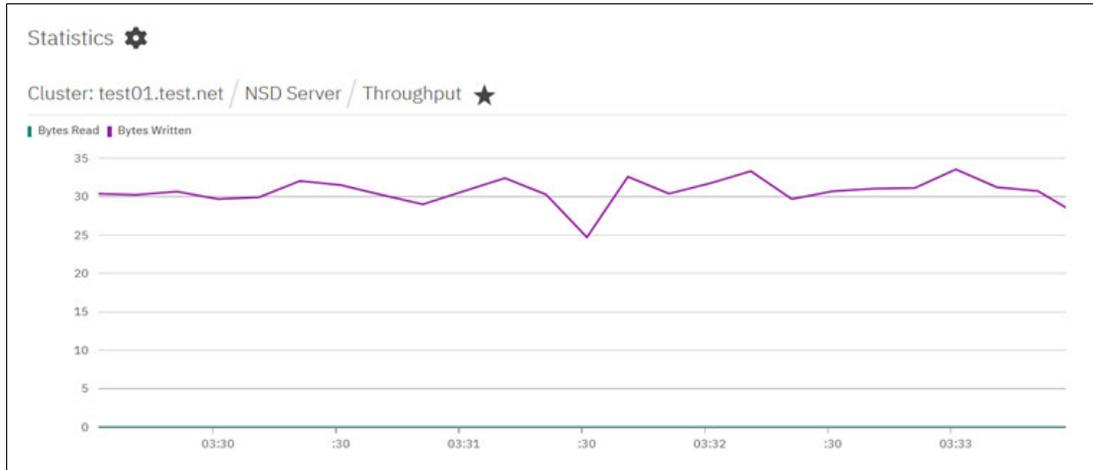


Figure 7 IBM ESS write throughput during 4 node validation

Table 2 lists the test results for the different node configurations with the IBM ESS 3000 system. The test results demonstrate that this converged infrastructure delivers the highest performance that is required with the larger CPU core machines that are used with SAS software.

Table 2 IBM ESS 3000 validation summary: Nodes with 48 CPU cores and 256 GB memory

Throughput	Single validation		4-node validation	
	Write	Read	Write	Read
Total I/O	11.4 GBps	9.3 GBps	30.4 GBps	36.9 GBps
I/O per Node	11.4 GBps	9.3 GBps	7.6 GBps	9.23 GBps
I/O per CPU core	244 MBps	198.4 MBps	163 MBps	196 MBps

IBM also offers other models of ESS systems, which can be used for the environments that require fewer performance throughputs compared to the NVMe based flash systems. Depending on the workload requirements, IBM offers various ESS models to suit your needs.

ESS includes the following models (see Figure 8):

- ▶ Types of hardware: SAS, NL-SAS, SSD, or NVMe
- ▶ Various sizes of disk, SSD, and NVMe



Figure 8 IBM ESS storage system models

Each building block contains the following components:

- ▶ A pair of IBM Spectrum Scale I/O NSD data servers
- ▶ ESS 5000, which is a fully integrated storage building block that features a pair of IBM POWER9™ data servers
- ▶ ESS 3000, which is a fully integrated 2U storage building block that includes a pair of x86 data servers
- ▶ A POWER-based ESS management server (one per IBM Spectrum Scale cluster)

IBM Spectrum Scale is the platform that is used across all the ESS models and provides a single global namespace for all the data, which offers a single point of management. Data can be tiered in differentiated classes of storage with global access, which ensures that data is always available in the right place, at the right time.

IBM Elastic Storage Servers are the physical building blocks and you can start as a small initial system then grow elastically to enterprise scale that is based on workload requirements. Each ESS is a storage building block for IBM Spectrum Scale. You can flexibly combine different sizes and models to suit your SAS workload characteristics.

Summary

The IBM Spectrum Scale with IBM ESS 3000 architecture over a 100 Gigabit Ethernet fabric provides a leading-edge performance with high bandwidth and low latencies that are required for the SAS Analytics workload.

The test results demonstrate that this converged infrastructure is viable and provides the required high performance when used with SAS software.

The Mellanox Ethernet high-speed storage network was crucial in facilitating the IBM ESS 3000 full I/O throughput. The ability to change from TCP/IP to RoCE at no extra cost is a key capability that can be used by SAS environments for improved performance.

For more Information

For more information about the IBM, Mellanox and SAS products and capabilities, contact your IBM representative or IBM Business Partner, or see the following websites:

- ▶ [IBM Spectrum Scale](#)
- ▶ [IBM Elastic Storage System](#)
- ▶ [SAS](#)

Appendix

The following scripts and parameters were used for the RoCE validation with the ESS 3000 system in the lab:

- ▶ QoS parameter configuration start script on the ESS 3000 node1:

```
[root@io1-ess3k ~]# cat roce_config.sh
#!/bin/bash
mlnx_qos -i enp216s0f1 --trust dscp
mlnx_qos -i enp94s0f1 --trust dscp
sysctl -w net.ipv4.tcp_ecn=1
mlnx_qos -i enp216s0f1 --pfc 0,0,0,1,0,0,0,0
mlnx_qos -i enp94s0f1 --pfc 0,0,0,1,0,0,0,0

ifdown enp216s0f1
ifdown enp94s0f1

ifup enp94s0f1
ifup enp216s0f1

sleep 15
echo 106 > /sys/class/infiniband/mlx5_1/tc/1/traffic_class
echo 106 > /sys/class/infiniband/mlx5_3/tc/1/traffic_class
cma_roce_tos -d mlx5_1 -t 106
cma_roce_tos -d mlx5_3 -t 106
[root@io1-ess3k ~]#
```

- ▶ Custom script for configuring the multiple routes in the same subnet on the ESS 3000 nodes:

```
[root@io1-ess3k ~]# cat route_config.sh
#!/bin/bash

ip route add 192.0.3.0/24 dev enp94s0f1 src 192.0.3.101 table t1
ip route add table t1 default via 192.0.3.101 dev enp94s0f1
ip rule add table t1 from 192.0.3.101

ip route add 192.0.3.0/24 dev enp216s0f1 src 192.0.3.102 table t2
ip route add table t2 default via 192.0.3.102 dev enp216s0f1
ip rule add table t2 from 192.0.3.102
[root@io1-ess3k ~]#
```

- ▶ IBM Spectrum Scale cluster configuration parameters:

```
[root@io1-ess3k ~]# mmlsconfig
Configuration data for cluster test01.test.net:
-----
clusterName test01.test.net
```

```

clusterId 11235395819543164259
dmapiFileHandleSize 32
minReleaseLevel 5.0.4.0
ccrEnabled yes
cipherList AUTHONLY
maxblocksize 16m
[ess_x86_64_mmvdisk_78E0656]
nsdRAIDTracks 131072
nsdRAIDEventLogToConsole all
nsdRAIDBlockDeviceMaxSectorsKB 0
nsdRAIDBlockDeviceNrRequests 0
nsdRAIDBlockDeviceQueueDepth 0
nsdRAIDBlockDeviceScheduler off
nsdRAIDSmallThreadRatio 2
nsdRAIDDefaultGeneratedFD no
nsdRAIDThreadsPerQueue 16
nsdRAIDSSDPerformanceShortTimeConstant 2500000
panicOnIOHang yes
maxStatCache 128k
pitWorkerThreadsPerNode 32
[common]
autoload no
[ems,ess_x86_64_mmvdisk_78E0656]
maxFilesToCache 128k
[isv517-hs,isv650_02-hs,isv650-hs]
maxFilesToCache 200000
[common]
maxReceiverThreads 32
ignorePrefetchLUNCount yes
[ess_x86_64,ess_x86_64_mmvdisk_78E0656]
nspdBufferMemPerQueue 24m
nspdThreadsPerQueue 2
nsdRAIDMaxPdiskQueueDepth 248
nsdRAIDMasterBufferPoolSize 2G
nspdQueues 120
[ems]
maxMBpS 20000
[ess3k1a-hs,ess3k1b-hs,ess_x86_64_mmvdisk_78E0656]
maxMBpS 24000
[isv517-hs,isv650_02-hs,isv650-hs]
maxMBpS 10000
[ess_x86_64_mmvdisk_78E0656]
nsdMinWorkerThreads 3842
nsdMaxWorkerThreads 3284
[isv517-hs,isv650_02-hs,isv650-hs,ems,ess_x86_64_mmvdisk_78E0656]
numaMemoryInterleave yes
workerThreads 1024
[ess_x86_64_mmvdisk_78E0656]
pagepool 485982036787
[isv650_02-hs,isv650-hs,ems]
pagepool 128G
[isv517-hs]
pagepool 96G
[isv517-hs,isv650_02-hs,isv650-hs,ems]
prefetchPct 40

```

```

[isv517-hs, isv650_02-hs, isv650-hs, ems, ess_x86_64_mmvdisk_78E0656]
nsdSmallThreadRatio 1
[ess3k1a-hs, ess3k1b-hs, isv517-hs, isv650_02-hs, isv650-hs, ems]
verbsRdma enable
verbsRdmaCm enable
[isv517-hs, isv650_02-hs, ems]
verbsPorts mlx5_0/1
[isv650-hs]
verbsPorts mlx5_2/1
[ess3k1a-hs, ess3k1b-hs]
verbsPorts mlx5_1/1 mlx5_3/1
[common]
verbsRdmaperConnectionOverride 4
adminMode central

```

► TCP/IP networking parameters on the Red Hat Enterprise Linux nodes:

```

[root@isv650 ~]# sysctl -p
net.ipv4.tcp_rfc1337 = 1
net.ipv4.tcp_max_tw_buckets = 1440000
net.ipv4.tcp_mtu_probing = 1
net.ipv4.tcp_window_scaling = 1
net.ipv4.tcp_adv_win_scale = 2
net.ipv4.tcp_low_latency = 0
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 1
net.core.netdev_budget = 600
net.ipv4.tcp_max_syn_backlog = 4096
net.ipv4.tcp_fin_timeout = 30
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.core.rmem_default = 16777216
net.core.wmem_default = 16777216
net.core.optmem_max = 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 87380 16777216
net.core.somaxconn = 8192
vm.min_free_kbytes = 512000
kernel.sysrq = 1
kernel.shmmax = 137438953472
net.core.netdev_max_backlog = 250000
net.ipv4.tcp_ecn = 1
net.ipv4.conf.default.arp_filter = 1
net.ipv4.conf.all.arp_announce = 2
net.ipv4.conf.default.arp_announce = 2
net.ipv6.conf.ens6f0.disable_ipv6 = 0 # disable interface in bond from ipv6
[root@isv650 ~]#

```

Author

This paper was produced by a team of specialists from around the world working with the IBM Redbooks, Tucson, Arizona, US.

Sanjay Sudam is the Senior Solution Architect for the Data and AI solutions using IBM Storage systems. He is responsible for creating the reference architectures and solution blueprints with the IBM Storage portfolio and external ISV partners solutions. Sanjay has created end-to-end reference architectures for AI, analytics, cloud data, data protection, digital video surveillance, and media entertainment solutions with the IBM portfolio.

Thanks to the following people for their contributions to this project:

Larry Coyne
IBM Redbooks®, Tucson, AZ, US

Jennifer Chen
Udayasuryan Kodoly
Kedar Karmarkar
Olaf Weiser
IBM Systems

Margaret Crevar
SAS Institute Inc.

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

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.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) ®
IBM®
IBM Elastic Storage®

IBM Spectrum®
IBM Spectrum Storage™
POWER9™

Redbooks®

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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

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



REDP-5609-00

ISBN 0738459100

Printed in U.S.A.

Get connected

