Redbooks
ibm.com/redbooks

IBM Spectrum Scale Deployment on IBM SoftLayer

Nikhil Khandelwal
John Lewars
Gautam Shah
Larry Coyne

Cloud

Infrastructure Solutions

In partnership with
IBM **Academy of Technology**

IBM

Redpaper

Overview

This IBM® Redpaper™ publication describes various activities that are necessary to deploy IBM Spectrum Scale™ on IBM SoftLayer®. It includes the following topics:

- Sizing and configuration guidance
- Cluster deployment
- Monitoring
- Storage options on IBM SoftLayer
- Network considerations on SoftLayer

Sizing and configuration guidance

There are many factors that must be taken into account when configuring a IBM Spectrum Scale cluster on any hardware. Desired performance, cost, cluster size, and workload can influence any hardware and configuration decisions. The following sections provide guidance to help you navigate the many options available when choosing hardware and deploying a cluster. Contact IBM for assistance in sizing and configuring a cluster that meets your requirements.

Getting started with IBM SoftLayer

IBM SoftLayer offers several virtual and bare metal server options. This paper is centered around bare metal storage deployments. However, you can use virtual systems and even mix virtual and physical systems into large clusters. To get started, visit the SoftLayer website at:

http://www.SoftLayer.com

Also, you can contact SoftLayer sales at sales@SoftLayer.com.

Bare metal systems can be provisioned using a simple web interface and are typically billed on a monthly basis. For large clusters containing hundreds of nodes, meeting network requirements and ensuring enough inventory is available might require additional notification and planning. Contact your SoftLayer representative for assistance in planning for large clusters.

If a managed high-performance computing (HPC) cloud solution is required, IBM Platform Computing offers managed cloud deployments that include IBM Spectrum Scale. For details about these managed options, go to:

http://www.ibm.com/systems/platformcomputing/solutions/hpccloud.html

Or you can contact your IBM sales representative.

Configuration options for IBM Spectrum Scale

SoftLayer offers bare metal servers with two, four, 12, and 36 drive bays. Order servers used as nodes offering storage on IBM Spectrum Scale with 12 or 36 bays. You can run compute or application only nodes on any hardware or on virtual systems. Because there are no shared storage options, deployments on SoftLayer use the shared nothing cluster (SNC) model.

Applications can be run using two methods:

- Applications can be run on the same hardware as the nodes that provide storage.Based on application usage, file placement optimization can be configured to reduce latency between an application and its data.
- Alternatively, applications can be run on dedicated client nodes which access data stored within the cluster. This model allows CPU and disk resources to be scaled independently.

Storage

Replicate data and metadata between two or more servers using file system data and metadata mirroring. Based on the size of the cluster and the fault tolerance required, 2-way or 3-way file system mirroring can be used. File system mirroring duplicates data and metadata across two or three servers in the cluster, providing redundancy in case of node failure. For

clusters with more than 10 servers, use 3-way mirroring to reduce the chances of data becoming unavailable in the case of nodes being taken down.

Use two drives in a RAID 1 configuration for the operating system. You can configure the remainder of the drives in the server as a RAID array or, in some cases, leave them as individual disks with no RAID array configured. There are many factors that can influence the decision on drive configuration.

If no RAID array is configured, enable 3-way mirroring on the file system to protect against drive loss. Do not use Serial Advanced Technology Attachment (SATA) drives with no RAID protection due to the drive's higher failure rate and slower rebuild time. Also, note that the potential performance impact of an individual drive failure on a cluster with no RAID protection will be greater than on a cluster with RAID protection because of the need to copy data between multiple servers.

Important: If using a RAID array, the decision on what type of RAID protection to deploy is important. Typical RAID levels used are RAID 6 (8+2p) arrays consisting of 10 drives, RAID 5 (8+1p) arrays consisting of 9 drives, and RAID 5 (4+1p) arrays consisting of 5 drives each. RAID 6 offers greater protection against multiple drive failures; however, write performance for RAID 6 arrays is slower than RAID 5. If using SATA disks, use RAID 6 to protect against multiple drive failures. If using SAS or solid-state drives (SSDs), RAID 5 is an option; however, configure at least one hotspare drive in the chassis. For additional guidance about RAID and disk options on SoftLayer bare metal systems, read "Storage options on IBM SoftLayer" on page 17.

Table 1 lists sample capacities for RAID 6 deployments. Note that the capacity is prior to file system mirroring (2- or 3-way).

Table 1 High capacity

Chassis	Drive type	RAID configuration	Usable capacity per node
12-bay	2 TB SATA	1 RAID 6 (8+2p)	16 TB
	4 TB SATA	1 RAID 6 (8+2p)	32 TB
	6 TB SATA	1 RAID 6 (8+2p)	48 TB
	8 TB SATA	1 RAID 6 (8+2p)	64 TB
36 bay	2 TB SATA	3 RAID 6 (8+2p) 4 hotspares	48 TB
	4 TB SATA	3 RAID 6 (8+2p) 4 hotspares	96 TB
	6 TB SATA	3 RAID 6 (8+2p) 4 hotspares	144 TB
	8 TB SATA	3 RAID 6 (8+2p) 4 hotspares	192 TB

Table 2 shows sample high performance considerations.

Table 2 High performance

Chassis	Drive type	RAID configuration	Usable capacity per node
12-bay	800 GB SSD	1 RAID 6 (8+2p)	6.4 TB
		1 RAID 10 (5+5)	4 TB
	960 GB SSD	1 RAID 6 (8+2p)	7.6 TB
		1 RAID 10 (5+5)	4.7 TB
	1.2 TB SSD	1 RAID 6 (8+2p)	9.6 TB
		1 RAID 10 (5+5)	6 TB
36-bay	800 GB SSD	3 RAPD 6 (8+2p) 4 Hotspares	19.2 TB
		4 RAID 10 (4+4) 2 Hotspares	12.8 TB
	960 GB SSD	3 RAID 6 (8+2p) 4 Hotspares	23.8 TB
		4 RAID 10 (4+4) 2 Hotspares	15.3 TB
	1.2 TB SSD	3 RAID 6 (8+2p) 4 Hotspares	28.9 TB
		4 RAID 10 (4+4) 2 Hotspares	19.2 TB

Performance note: The 960 GB SSDs have fewer write cycles, which might not be ideal for high-write workloads.

For large clusters, use metadata only disks. These disks should ideally be SSD-based for the best performance. A single server can contain a metadata-only disk and data disks. If metadata only disks are used, a minimum of three disks are required, in three separate servers for redundancy. Typically, nodes with metadata-only disks should also be used as IBM General Parallel File System (GPFS™) file system managers. In general, metadata capacity should be approximately 3% - 5% of the total file system capacity. For clusters of up to 10 nodes, three metadata disks are sufficient. It is recommended to add one additional node with a metadata disk for every 10 additional servers in the cluster. For further guidance about metadata versus data disks, see the File Placement Optimizer (FPO) white paper available at:

http://ibm.co/2fZ4q7k

Mixing SSD and spinning hard disk within the same node might be useful to help satisfy metadata requirements or, in some cases, to provide separate tiers of storage with different performance characteristics. It is recommended that the drive type and array configuration remain identical for all drives assigned to a single storage pool.

Network

All data is replicated to other nodes in this cluster via a private network connection. A 10Gb network is required meet basic bandwidth requirements for most applications. This type of network is available only in certain data centers and can be specified on the system order page. Public network connectivity can be sized as required for the application. Several data centers have optional dual 10Gb network connectivity for additional bandwidth as well. These adapters are typically bonded via Link Aggregation Control Protocol (LACP) for increased availability and higher aggregate bandwidth.

InfiniBand can be used for inter-node communication, depending on availability from SoftLayer. The InfiniBand high bandwidth and low latency can significantly improve IBM Spectrum Scale performance. Enable Remote Direct Memory Access (RDMA) in GPFS if using an InfiniBand connection to maximize the effectiveness of the InfiniBand network.

Memory

IBM Spectrum Scale requires memory for page pool and caching operations. Increasing the memory given to the page pool can improve performance. A page pool of 2 - 4 GB is ideal for many typical workloads, but the size can be increased or decreased based on application use, performance requirements, and workload.

Software

The installation cookbook is tailored to a Red Hat 7 installation. GPFS software supports other Linux distributions, such as SUSE Linux Enterprise Server SLES and Ubuntu.

Installation note: There are restrictions on some operating systems and specific functions. File placement optimization is currently not supported on Debian. So do not use this operating system (OS).

See the GPFS frequently asked questions (FAQ) in IBM Knowledge Center for more information:

https://www.ibm.com/support/knowledgecenter/en/SSFKCN/com.ibm.cluster.gpfs.doc/gpfs_faqs/gpfsclustersfaq.html

Other software and monitoring tools can be selected as required for a particular application.

Cluster deployment

The following sections discuss the activities required for cluster deployment.

Configure host environments

It can be useful to add the GPFS `bin` directory to the path on one or more of the cluster nodes:

```
# vim /root/.bash_profile
```

Look for the `PATH` line and add `GPFS` to it:

```
PATH=$PATH:$HOME/bin:/usr/lpp/mmfs/bin
```

Tune Linux and network performance settings

Updating some settings in `/etc/sysctl.conf` might improve network utilization and cluster throughput. You can find recommendations at IBM developerWorks®:

http://ibm.co/2fByQMy

Because data is replicated across multiple nodes, network performance is a key component of overall system performance. This is particularly true for SSD-based systems, as the high speeds of the system disk amplify any network latency. In cases such as this, it might be useful to look also at settings information available at IBM developerWorks:

http://ibm.co/2gydP9K

In addition, increasing buffers on the network adapter to a higher number has also been found to improve performance in certain cases. The available settings vary, depending on the adapter type, but can be queried and set via the `ethtool` command, with the `-g` or `-k` flags.

For tuning recommendations for file placement optimization systems and systems, for which all storage is single-tailed (each storage component is connected only to one node or machine), refer to the following information available at IBM developerWorks:

http://ibm.co/2fZ2z2v

Set up /etc/hosts files

Gather the system name and private IP address from each system that will be used in this cluster. Add the names and addresses to `/etc/hosts` to allow hostname lookups. Use only the private IP addresses for this step, because use of the public IP can result in large data charges.

```
# vim /etc/hosts
```

Add an entry for each system and name. For example, replacing `fpo-nodeX` with the IP and `hostname` in the configuration, as shown in Example 1.

Example 1 Replace fpo-nodeX with the IP and hostname

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
192.168.122.10 fpo-node1
192.168.122.11 fpo-node2
192.168.122.12 fpo-node3
192.168.122.13 fpo-node4
192.168.122.14 fpo-node5
192.168.122.15 fpo-node6
192.168.122.16 fpo-node7
192.168.122.17 fpo-node8
```

Repeat on each host in the cluster.

Configure password-less SSH

On each host in the cluster, generate a new SSH key, and add it to the authorized_keys file as follows:

1. Execute the following command:

   ```
   # ssh-keygen -q -t rsa -f /root/.ssh/id_rsa -N ""
   ```

2. Gather the contents of each host's `/root/.ssh/id_rsa.pub` file and concatenate it into a single file, named `/root/.ssh/authorized_keys`.

3. Copy the `/root/.ssh/authorized_keys` file to each host in the cluster. From each host, run the following command:

   ```
   # ssh localhost
   ```

4. Enter `yes` at the following prompt if it displays:

   ```
   The authenticity of host “localhost (::1): can't be established. ECDSA key
   fingerprint is 7f:0f:75:a2:d0:e4:0f:7e:e9:91:e8:37:5f:ce:c6:2c. Are you sure
   you want to continue connecting (yes/no)?
   ```

   Connect via SSH from each host to every other host in the cluster, by entering `yes` at this prompt as required.

5. Ensure that each host can connect via SSH to every other host without any prompt or password being required.

Prepare disks for use

The SoftLayer provisioning process can create and mount file systems on disks that will be used for GPFS (Example 2). Unmount or remove these file systems.

Example 2 Display file systems

```
# df
Filesystem        1K-blocks    Used    Available Use% Mounted on /dev/sda3
286661072 1779964    270319572   1% /
tmpfs              16421272       0     16421272   0% /dev/shm
/dev/sda1            253871   58783       181981  25% /boot
/dev/sdb1      109364387840   36064 109364351776   1% /disk1
```

In Example 2, `sdb1` is mounted as `/disk1`. To remove these file systems, unmount them and remove any entries from `/etc/fstab` that corresponds to these file systems (Example 3).

Example 3 Remove file system

```
# unmount /disk1
# rmdir /disk1
# vi /etc/fstab
```

Look for corresponding entries, such as:

```
LABEL=/disk1            /disk1                  xfs     defaults        1 2
```

Comment or remove any such lines from this file.

Determine which SCSI device is used by the operating system (typically `/dev/sda`) by running the **`df`** or **`mount`** commands. All other disks will be used as Network Shared Disk (NSD) devices by IBM Spectrum Scale. If a system was ordered with SSD and spinning disk, determine which device is on SSDs and note this for later. You can use the **`lsscsi`** command to assist with this detection. (The **`lsscsi -s`** command displays the size of disks.)

Configure Linux firewall

All systems must be able to communicate via TCP port 1191. The Linux firewall must either be disabled or port 1191 must be allowed to pass through. In addition, the performance monitoring tool requires port 4739 on TCP and User Datagram Protocol (UDP).

There are multiple valid ways to open the required ports. Example 4 adds the private network interface to the internal network zone and allows ports 1191, 8889, and 10080 to communicate on this zone. To do so, it determines the adapter name that is used for the private network on the provisioned system, which will vary depending on the hardware configuration on the provisioned system. Example 4 also uses the `ens3` interface, which you replace with the interface name on the hardware.

Example 4 Establish interface name on the hardware

```
# firewall-cmd --zone=public --remove-interface=ens3 -permanent
# firewall-cmd --zone=internal --add-interface=ens3 -permanent
# firewall-cmd --zone=internal --add-port=1191/tcp -permanent
# firewall-cmd --zone=internal --add-port=4739/tcp -permanent
# firewall-cmd --zone=internal --add-port=4739/udp -permanent
# firewall-cmd --reload
```

By default, IBM Spectrum Scale encrypts traffic between nodes, however if public network interfaces are installed it is recommended to use a firewall to block IBM Spectrum Scale and any other unnecessary ports from access over the public network.

Deploy GPFS

You can complete the remainder of the deployment using one of the following methods:

- The IBM Spectrum Scale protocol installer
- A manual installation

If the system is to be completely RHEL-7 based, on similar hardware, you can use the IBM Spectrum Scale protocol installer. This protocol installer guides you through the remainder of the steps, including RPM installation as well as Cluster and NSD creation. The installer is included in the IBM Spectrum Scale protocols package. To use it, extract the protocols package and run. To use the toolkit, review IBM Knowledge Center instructions that are available at:

`http://ibm.co/2foxbcc`

The sections that follow provide instructions on the manual steps of installing and configuring IBM Spectrum Scale. If you use the protocol installer, it is helpful to review the following section for guidance regarding NSD creation as well as quorum and manager configuration. The protocol installer is not capable of setting up FPO volumes. So you need to apply these settings to storage pools after the protocol installer is complete. Also, ensure that the proper licenses are configured after installation, as FPO licenses cannot be applied.

Deploy GPFS RPMs

To deploy a GPFS Red Hat RPM Package Manager (RPM), complete the following steps:

1. Extract the install package:

   ```
   # ./Spectrum_Scale_Advanced-4.2.0.0-x86_64-Linux-install
   ```

2. Install the required RPMs:

   The `net-tools` and `ksh` packages are required by IBM Spectrum Scale and are not part of most default installation packages. Ensure that they are installed:

   ```
   # yum -y install ksh net-tools
   ```

3. Install the GPFS packages.

   Change to the directory that contains the extracted RPMs and install:

   ```
   # cd /usr/lpp/mmfs/4.2.0.0/
   # rpm -ihv gpfs.base-4.2.0-0.x86_64.rpm gpfs.docs-4.2.0-0.noarch.rpm
   gpfs.ext-4.2.0-0.x86_64.rpm gpfs.gpl-4.2.0-0.noarch.rpm
   gpfs.gskit-8.0.50-47.x86_64.rpm gpfs.msg.en_US-4.2.0-0.noarch.rpm
   ```

4. Repeat the previous steps for all systems in the cluster.

Build and install GPFS portability layer

The GPFS portability layer can be built on a single system and distributed to the remainder of the cluster, or it can be built individually on each system. Complete the following steps:

1. Pre-requisites must be installed:

   ```
   # yum -y install gcc gcc-c++ kernel-devel make rpm-build
   ```

2. Build the portability layer:

   ```
   # mmbuildgpl
   ```

These steps should be done on all nodes in the cluster. Optionally, the portability layer can be created as an RPM on a single node with the `-build-package` option, and the RPM can be distributed to the remaining nodes in the cluster.

Create and configure GPFS cluster

Create a cluster definition file that contains the names of all nodes and their respective roles.

Quorum nodes

In most clusters, three quorum nodes are sufficient. In clusters containing more than 136 nodes, the recommended number of quorum nodes can be determined by the following formula:

```
Quorum nodes = 3 + 2 * ((#cluster nodes - 135) / 90)
```

The number of quorum nodes should always be an odd number. There should not be fewer than three or more than seven quorum nodes defined in the cluster.

Manager nodes

In most clusters, three manager nodes are sufficient. These might or might not be assigned to the same nodes as the quorum nodes in order to reduce the number of server licenses that are required by a cluster. If the cluster has nodes with metadata-only disks configured, set the manager nodes to use these nodes. If there are more than 136 nodes in a cluster, the recommended number of manager nodes can be determined by the following formula:

```
Manager nodes = 3 + (#cluster nodes -135) / 45
```

Define the cluster

Create a new file `cluster.cfg` file, and specify all nodes in the cluster and their roles, separated by a colon (Example 5).

Example 5 Create new file

```
fpo-node1:quorum
fpo-node2:quorum-manager
fpo-node3:
fpo-node4:manager
fpo-node5:quorum
fpo-node6:manager
fpo-node7:
fpo-node8:
```

Run the `mmcrcluster` command to create the cluster (Example 6).

Example 6 Create the cluster

```
# mmcrcluster -N cluster.cfg -r /usr/bin/ssh -R /usr/bin/scp -C fpo-cluster
mmcrcluster: Performing preliminary node verification ...
mmcrcluster: Processing quorum and other critical nodes ...
mmcrcluster: Processing the rest of the nodes ...
mmcrcluster: Finalizing the cluster data structures ...
mmcrcluster: Command successfully completed
mmcrcluster: Warning: Not all nodes have proper GPFS license designations.
    Use the mmchlicense command to designate licenses as needed.
mmcrcluster: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
```

Accept the licenses on all nodes. All quorum and manager nodes must be configured with a server license. Remaining nodes should be configured with an FPO license.

```
# mmchlicense server --accept -N fpo-node1,fpo-node2,fpo-node4,fpo-node5,fpo-node6
```

Example 7 shows nodes designated as possessing server licenses.

Example 7 Nodes possessing server licenses

```
fpo-node1
    fpo-node5
    fpo-node2
    fpo-node6
    fpo-node4
mmchlicense: Command successfully completed
mmchlicense: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
# mmchlicense fpo --accept -N fpo-node3,fpo-node7,fpo-node8
```

Example 8 shows the nodes designated as possessing FPO licenses.

Example 8 FPO licenses

```
fpo-node3
    fpo-node7
    fpo-node8
mmchlicense: Command successfully completed
mmchlicense: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
```

Apply GPFS configuration

Example 9 shows the GPFS parameters recommended for FPO configurations on SoftLayer.

Example 9 GPFS parameters for FPO configurations on SoftLayer

```
autoload=yes
readReplicaPolicy=local
restripeOnDiskFailure=yes
unmountOnDiskFail=meta
disableInodeUpdateOnFDatasync=yes
forceLogWriteOnFdatasync=no
minMissedPingTimeout=60
leaseRecoveryWait=65
worker1Threads=72
nsdMinWorkerThreads=48
nsdInlineWriteMax=1M
nsdSmallThreadRatio=2
nsdThreadsPerQueue=10
maxFilesToCache=100000
maxStatCache=512
prefetchAggressivenessRead=2
prefetchAggressivenessWrite=0
maxMBpS=4000
```

Settings note: The `maxMBps=4000` setting is intended for nodes with one 10 Gbps link. You can adjust this setting according to the number of links. For example, you can set it to `8000` if you are using two-mode, four-bonded 10 Gbps links.

For nodes with disk caches enabled but with no battery protection present, enable the following configuration option:

```
dataDiskCacheProtectionMethod=2
```

In non-RAID adapter configurations, check whether a cache is enabled using **`sdparm --long /dev/<diskname> | grep WCE`**.

For nodes with at least 32 GB of memory, use a page pool of 4 GB. For certain workloads, increasing the page pool to larger values might improve performance, which can be tested and changed as needed.

For IBM Spectrum Scale clusters running 4.2.0.3 or newer code, increasing the `workerThreads` setting to a value higher than the default of 48 might improve performance. The setting can be changed to a common value across all nodes, such as 512. For further

tuning and to increase efficiency, this value can be set individually to eight times the number of processor cores for each node, assuming at least six processor cores in each node.

You can use the **mmchconfig** command to change these settings (Example 10).

Example 10 The mmchconfig command

```
# mmchconfig
readReplicaPolicy=local,\
restripeOnDiskFailure=yes,\
unmountOnDiskFail=meta,\
disableInodeUpdateOnFDatasync=yes,\
forceLogWriteOnFdatasync=no,\
minMissedPingTimeout=60,\
leaseRecoveryWait=65,\
maxMBpS=4000,\
worker1Threads=72,\
nsdMinWorkerThreads=48,\
nsdInlineWriteMax=1M,\
nsdSmallThreadRatio=2,\
nsdThreadsPerQueue=10,\
maxFilesToCache=100000,\
maxStatCache=512,\
prefetchAggressivenessRead=2,\
prefetchAggressivenessWrite=0
```

These recommendations are a summary of key recommendations at the time this paper was written. Refer to the *IBM Spectrum Scale Tuning Recommendations for Shared Nothing Environments* topic for tuning recommendations for FPO systems and systems for which all storage is single-tailed (each storage component is connected only to one node or machine):

http://ibm.co/2fZ2z2v

Start GPFS

After applying all the settings and setting the page pool, start GPFS using the following command:

```
# mmstartup -a
```

Wait approximately 1 minute, and then verify that GPFS is up on all nodes (Example 11).

Example 11 Verify GPFS is up on all nodes

```
# mmgetstate -a
Node number  Node name        GPFS state
-----------------------------------------
1      fpo-node1        active
2      fpo-node2        active
3      fpo-node5        active
4      fpo-node3        active
5      fpo-node4        active
6      fpo-node6        active
7      fpo-node7        active
8      fpo-node8        active
```

Define NSD

This section discusses defining NSD.

Storage pools

If you ordered separate metadata and data disks for this system, there should be at least two storage pools: system and data. The system pool contains all of the metadata disks and does not have FPO enabled. The data pool contains all of the data disks and enables FPO. FPO is enabled by the following settings:

- `allowWriteAffinity=yes`
- `writeAffinityDepth=1`
- `blockGroupFactor=128`

These FPO settings, along with a 1 MB block size are optimal for many workloads. Set the usage size for metadata disks in the system pool to `metadataOnly` and set the usage for disks in the data pool to `dataOnly`.

If you ordered a single type of disk for the system, there will be a single storage pool: system. The system pool contains all disks, and the usage of each disk is set to `metadataAndData`. FPO behavior is enabled on this pool.

Failure groups for FPO-enabled pools are defined with three numbers:

- The first number indicates the frame number in which the system resides.
- The second number indicates whether the system resides in the top of the rack (1) or the bottom of the rack (0).
- The third number indicates the slot number in which the system resides.

For example, `2,1,5` indicates a system in frame 2, top of the rack, slot 5. In SoftLayer, the nodes might be spread at locations throughout the data center. So these numbers have less meaning. However, each node must still have all of its disks in a unique failure group in order for GPFS to replicate the data. With a `writeAffinityDepth` of `1`, the system writes the first copy of the data to the local node. The second copy goes to a node in a different rack. The third copy goes to a node in the same rack but in a different half. Therefore, if the file system will be configured with two replicas for data, there must be systems in at least two separate “frames.” If there will be three copies of data, there must be systems in two frames and in both halves of each frame.

FPO has several additional options that might be useful for certain types of applications or for analytics. For further details about FPO, and as guidance for metadata and data disks, see the white paper *File Placement Optimizer (FPO)* that is available at:

`http://ibm.co/2fZ4q7k`

Example 12 is a sample NSD configuration file for a system with a separate metadata and data tier.

Example 12 Sample NSD configuration file

```
%pool: pool=system
       blockSize=256K
       layoutMap=cluster
       allowWriteAffinity=no
%pool: pool=data
       blockSize=1M
       layoutMap=cluster
```

```
       allowWriteAffinity=yes
       writeAffinityDepth=1
       blockGroupFactor=128
%nsd: nsd=meta1nsd
      device=/dev/sda
      servers=fpo-node1
      usage=metadataOnly
      failureGroup=101
      pool=system
%nsd: nsd=meta2nsd
      device=/dev/sdb
      servers=fpo-node2
      usage=metadataOnly
      failureGroup=102
      pool=system
%nsd: nsd=meta3nsd
      device=/dev/sda
      servers=fpo-node3
      usage=metadataOnly
      failureGroup=103
      pool=data
%nsd: nsd=data4nsd
      device=/dev/sda
      servers=fpo-node4
      usage=dataOnly
      failureGroup=1,0,1
      pool=data
%nsd: nsd=data5nsd
      device=/dev/sda
      servers=fpo-node5
      usage=dataOnly
      failureGroup=2,0,1
      pool=data
```

If there will be no separate metadata and data nodes, the NSD configuration might look as shown in Example 13.

Example 13 NSD configuration

```
%pool: pool=system
       blockSize=1M
       layoutMap=cluster
       allowWriteAffinity=yes
       writeAffinityDepth=1
       blockGroupFactor=128
%nsd: nsd=node1nsd
      device=/dev/sda
      servers=fpo-node1
      usage=dataAndMetadata
      failureGroup=1,0,1
      pool=system
%nsd: nsd=node2nsd
      device=/dev/sdb
      servers=fpo-node2
      usage=dataAndMetadata
      failureGroup=1,1,1
```

```
      pool=system
%nsd: nsd=node3nsd
      device=/dev/sda
      servers=fpo-node3
      usage=dataAndMetadata
      failureGroup=2,0,1
      pool=system
%nsd: nsd=node4nsd
      device=/dev/sda
      servers=fpo-node4
      usage=dataAndMetadata
      failureGroup=2,1,1
      pool=system
```

After the NSD configuration file is written, create the NSDs using the following command:

```
# mmcrnsd -F nsds.txt
```

Create and mount GPFS file system

Create and mount the file system. Select your desired name for the file system and the metadata and data replication parameters that are needed:

- `-m` specifies the number of metadata replicas
- `-d` specifies the number of data replicas

In this example, the file system block size is defined in the `%pool` section of the NSD file. A 1 MB block size is useful for many typical workloads; however, application workload should be carefully considered. The block size can have a major effect on performance and storage efficiency, and cannot be changed after file system creation.

Example 14 creates three metadata and data copies in a file system named `gpfs-fpo`.

Example 14 Create three metadata and data copies

```
# mmcrfs gpfs-fpo -F nsds.txt -m 3 -M 3 -r 3 -R 3 -A yes -Q no -S relatime -E no
# mmmount all -a
```

Validate GPFS configuration

Verify all disks are up on the file system and the file system is mounted on all nodes (Example 15).

Example 15 Verify disks

```
# mmcrfs gpfs-fpo -F nsds.txt -m 3 -M 3 -r 3 -R 3 -A yes -Q no -S relatime -E no
# mmmount all -a
```

Monitoring

This section discusses monitoring of monitoring of cluster and storage status.

Check system status

You can run hardware checks using standard SoftLayer tools, such as the Nimsoft monitoring service. These checks can ensure that RAID arrays are fully redundant and that other hardware is operating normally.

To ensure GPFS is in an optimal state during normal operation, periodically run checks to ensure that all cluster nodes and disks are active and up.

Example 16 shows how to check the node status.

Example 16 Check node status

```
# mmgetstate -a
  Node number  Node name       GPFS state
-----------------------------------------
       1      fpo-node1        active
       2      fpo-node2        active
       3      fpo-node5        active
       4      fpo-node3        active
       5      fpo-node4        active
       6      fpo-node6        active
       7      fpo-node7        active
       8      fpo-node8        active
```

Example 17 shows how to check the disk status.

Example 17 Check disk status

```
# mmlsdisk gpfs-fpo
disk        driver   sector     failure holds    holds                            storage name
type       size       group metadata data  status        availability pool
------------ -------- ------ ----------- -------- ----- ------------- ------------ ------------
gpfs1nsd     nsd         512          101 Yes      No    ready         up           system
gpfs2nsd     nsd         512          102 Yes      No    ready         up           system
gpfs3nsd     nsd         512          103 Yes      No    ready         up           system
gpfs4nsd     nsd         512       1,0,1 No       Yes   ready         up           data
gpfs5nsd     nsd         512       1,0,2 No       Yes   ready         up           data
gpfs6nsd     nsd         512       1,0,3 No       Yes   ready         up           data
gpfs7nsd     nsd         512       2,0,1 No       Yes   ready         up           data
gpfs8nsd     nsd         512       2,0,2 No       Yes   ready         up           data
```

If the message `Due to an earlier configuration change the file system may contain data that is at risk of being lost.` displays, some data on the file system might not be replicated and might be at risk of a node failure. The **`mmrestripe -R`** command corrects any such issues, although it can cause performance of the system to temporarily decrease while it runs, depending on the amount of data that needs to be replicated.

System recovery

This section discusses system recovery considerations.

Down disks

In the event of a hardware failure or in certain abnormal operations, a disk might become unavailable (shown as *down* in `mmlsdisk` output). After the hardware is recovered, a disk can be restarted using the following command:

```
# mmchdisk <filesystem> start -d <disk name>
```

Node failures

If a node fails or is reboot for any reason, it automatically rejoins the GPFS cluster during the next boot. If GPFS does not autostart, issue the **`mmstartup`** command to start GPFS on a single node. After GPFS has started on a node, determine the status of rejoining the cluster by running the **`mmgetstate`** command.

After a node reboots, it is advisable to check the disk status on that node using the **`mmlsdisk`** command.

Storage options on IBM SoftLayer

When choosing a storage option, it is important to consider availability, cost, and performance of the solution. The solution cost is driven by the drive type and utilization of the underlying storage. Table 3 lists utilization of common disk configurations.

Table 3 Utilization of common disk configuration

RAID type	2-way replication	3-way replication
JBOD only	NOT RECOMMENDED	33%
RAID1	25%	16.66%
RAID 5 (4+1)	40%	26.66%
RAID 5 (8+1)	44.44%	29.66%
RAID 6 (8+2)	40%	26.66%

As clusters increase in size, the chances for multiple nodes failures increase. If there are 10 nodes providing storage in the cluster or more, strongly consider 3-way replication. Consider RAID 6 protection for high capacity SATA and NL-SAS disks due to the long rebuild times if and when disks fail.

Performance of the solution depends greatly on the workload being used on the cluster. Typically, RAID 5 and 6 perform best on sequential I/O workloads and large-block workloads due to the higher potential of full stripe writes. RAID 1 and JBOD solutions are ideal for small block and random I/O workloads. On cloud deployments in IBM SoftLayer, the additional replication factor across the network does add some complexity.

For large block and sequential I/O, all RAID solutions typically run into network bottlenecks for replication. This issue is especially true for an SSD-based solution. As a result, there is little difference in performance on any RAID level.

Random and small block I/O have different performance characteristics for SSD and spinning disk configurations. For SSD configurations, the network latency between nodes becomes a limiting factor on extremely small I/Os (0-128 KB). As a result, there is typically little difference in overall performance between the various RAID levels. For I/Os from 128 KB to the file system block size, RAID1 and JBOD configurations might have a small speed advantage over RAID 5 or RAID6. On spinning disk, RAID1 can have a more noticeable speed increase over RAID5 or RAID6 for small block and random I/O.

If RAID 5 or RAID 6 protection is chosen, it is best to match the RAID stripe size with the file system block size for the best performance. Because the stripe size of the RAID array is configured during system provisioning, this might need to be coordinated, because the systems are being ordered. The stripe size might depend on hardware availability and the adapter in the system. Most RAID adapters can be queried to determine the current stripe size. For example, the LSI adapters used in some systems can be queried via the `storcli` CLI command (installed in `/opt/MegaRAID/storcli/`). Other adapters will have their own management CLIs. For several adapters, a 256 KB strip size is set as a default, so 8+1 or 8+2 RAID schemes will have a 2 MB stripe size (8 * 256 KB).

Network considerations on SoftLayer

IBM SoftLayer maintains high-speed networking within its data center. However, several factors can make networking on cloud systems more challenging, including system location within the data centers and shared connections between networking equipment. It can be typical to see some variability in network performance during different times of day or even from connection to connection due to the shared nature of the network infrastructure. Thus, consider network performance when completing operations, such as system benchmarks, which can vary by 5% - 10% as they are run during different times of day.

Prior to using a system, run a brief network test after a system is provisioned to verify the network is performing as expected. Tools such as `iperf` or `nsdperf` (in `/usr/lpp/mmfs/samples/net`) can be useful to run these tests.

Tools such as `iperf` or `nsdperf` are used to send traffic from one system to another. It is best to run these tests at least five times in order to ensure that all routes are being used. Due to bonding and load balancing, you might see different results from test to test. It is also useful to run these tests in each direction to ensure that each system can send and receive data properly.

Sample nsdperf usage

Start `nsdperf` as a server on systems to test (Example 18).

Example 18 Start nsdperf

```
ssh spec1 /usr/lpp/mmfs/samples/net/nsdperf -s &
ssh spec2 /usr/lpp/mmfs/samples/net/nsdperf -s &
ssh spec3 /usr/lpp/mmfs/samples/net/nsdperf -s &
```

Run `nsdperf` as a client on one system and as a server on other systems (Example 19).

Example 19 Run nsdperf

```
nsdperf> client spec1
Connected to spec1
nsdperf> server spec2 spec3
Connected to spec2
Connected to spec3
nsdperf> test read write
Connected to spec2
Connected to spec3
1-2 read 1090 MB/sec (259 msg/sec), cli 2% srv 1%, time 10, buff 4194304
1-2 write 1170 MB/sec (280 msg/sec), cli 2% srv 1%, time 10, buff 4194304
```

Sample iperf3 usage

Start `iperf` as a server on one system using the following command:

```
ssh spec2 iperf3 -s &
```

Start `iperf` as a client on a second system as shown in Example 20.

Example 20 Start iperf as a client

```
iperf3 -c spec2
Connecting to host spec2, port 5201
[  4] local 10.152.11.57 port 45160 connected to 10.152.11.8 port 5201
[ ID] Interval           Transfer     Bandwidth       Retr  Cwnd
[  4]   0.00-1.00   sec  1.10 GBytes  9.43 Gbits/sec   19    650 KBytes
[  4]   1.00-2.00   sec  1.10 GBytes  9.42 Gbits/sec    0    658 KBytes
[  4]   2.00-3.00   sec  1.10 GBytes  9.42 Gbits/sec    0    658 KBytes
[  4]   3.00-4.00   sec  1.10 GBytes  9.42 Gbits/sec    0    660 KBytes
[  4]   4.00-5.00   sec  1.10 GBytes  9.42 Gbits/sec    0    660 KBytes
[  4]   5.00-6.00   sec  1.10 GBytes  9.41 Gbits/sec    0    665 KBytes
[  4]   6.00-7.00   sec  1.09 GBytes  9.41 Gbits/sec    0    666 KBytes
[  4]   7.00-8.00   sec  1.10 GBytes  9.42 Gbits/sec    0    666 KBytes
[  4]   8.00-9.00   sec  1.10 GBytes  9.42 Gbits/sec    0    666 KBytes
[  4]   9.00-10.00  sec  1.10 GBytes  9.42 Gbits/sec    0    666 KBytes
```

Authors

This paper was produced by a team of specialists from around the world working with the International Technical Support Organization, Tucson Center.

Nikhil Khandelwal is a Senior Engineer with the IBM Spectrum Scale development team. He has over 15 years of storage experience on network-attached storage (NAS), disk, and tape storage systems. He has led development and worked in various architecture roles. Nikhil is currently part of the IBM Spectrum Scale client adoption and cloud teams.

John Lewars is a Senior Technical Staff Member with the IBM Spectrum Scale development team. He has been with IBM for about 18 years, starting first on the high performance computing service team, working primarily on performance problems and code fixes, and then moving on to work in communications protocols and network management development. John has an extensive background in large parallel systems delivery and bring-up and led the technical computing performance development team for years before moving to his current assignment working with large customers and deployments of IBM Spectrum Scale in cloud environments.

Gautam Shah is a Senior Technical Staff Member with the IBM Spectrum Scale development team. He recently led the delivery of the protocol functionality that is integrated with IBM Spectrum Scale and now leads the team helping with delivery of IBM Spectrum Scale in cloud environments. He has been with IBM over 20 years with experience in various roles in the high-performance computing, including communication protocols and clustered file system development and deployment of large-scale clusters. He also worked on assignment at the IBM Systems Lab in Pune, India, on the NAS offering based on General Parallel File System. He is a member of the IBM Academy of Technology. He received a B.Sc. in Physics from Loyola College, Chennai and a B.E in Computer Science and Automation from Indian Institute of Science, Bangalore, and a Ph.D. in Information and Computer Science from Georgia Institute of Technology.

Larry Coyne is a Project Leader at the International Technical Support Organization, Tucson Arizona center. He has 35 years of IBM experience with 23 in IBM storage software management. He holds degrees in Software Engineering from the University of Texas at El Paso and Project Management from George Washington University. His areas of expertise include client relationship management, quality assurance, development management, and support management for IBM storage software.

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

LindaMay Patterson
International Technical Support Organization, Poughkeepsie Center

Dean Hildebrand
IBM Research

Pallavi Galgali
Wei Gong
Theodore Hoover
Connie Woodward
IBM Systems

Xiang Zhan
IBM Sales and Distribution (S&D), Systems Hardware Sales

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 ITSO 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 IBM 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)  ®
developerWorks®
GPFS™
IBM®
IBM Spectrum Scale™
Redbooks®
Redpaper™

The following terms are trademarks of other companies:

SoftLayer, and SoftLayer device are trademarks or registered trademarks of SoftLayer, Inc., an IBM Company.

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

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

REDP-5410-00

ISBN 0738455806

Printed in U.S.A.