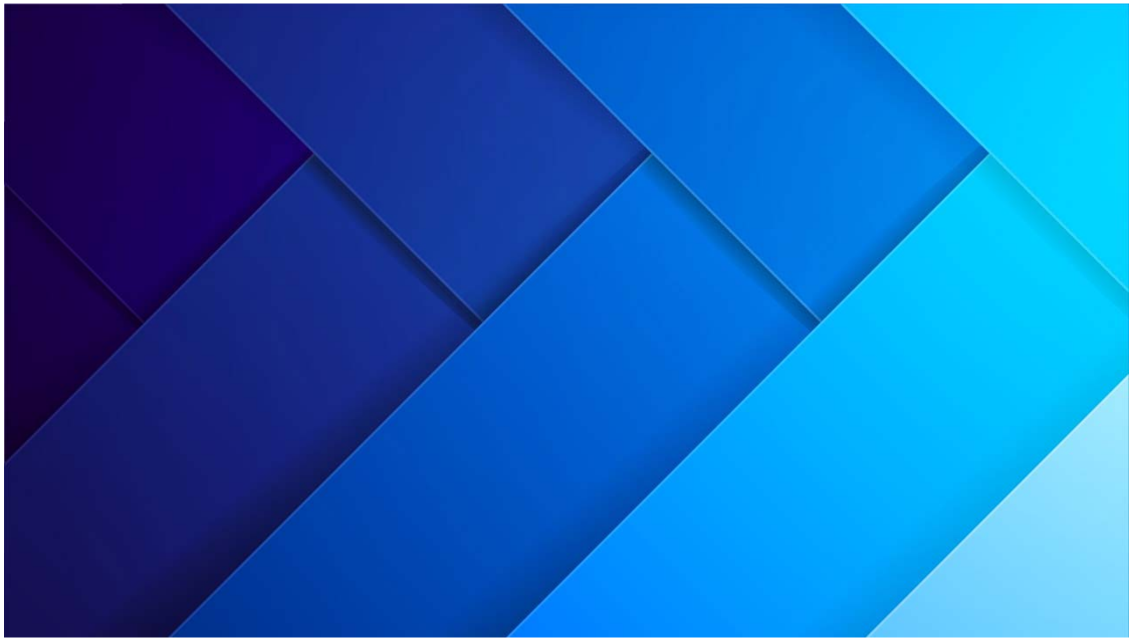
IBM Systems MEDIA

What’s New and Updated With Spectrum Scale 5.0.4.4

AIX expert Jaqui Lynch provides a primer on Spectrum Scale 5.0.4.4.



By Jaqui Lynch

06/15/2020

Spectrum Scale (formerly known as GPFS) is a non-blocking filesystem that is used in high performance computing clusters, Oracle environments and many other environments where you want a high performance, shared filesystem that goes beyond what NFS or Samba can provide. This article discusses recent updates for my Spectrum Scale systems from 4.2.3 and 5.0.3.2 to 5.0.4.4.

What’s New in Spectrum Scale 5.0.4

IBM has recently announced a Scale developer edition. This is limited to 12TBs which is enough to build a robust test environment. It is free for non-production use such as test and upgrade preparation. It can be downloaded from the Spectrum Scale Try and Buy page.

5.0.4 includes a lot of new functionality, including support for Redhat Enterprise Linux 8, the ability to do reclaims on NVMe devices and improved recovery of CCR enabled clusters. There is also an RPQ available to provide thin provisioning support for file system data and metadata. One of the improvements in the network environment is provided on github—there is a C program called nsdperf that can be used for stress testing. Additionally, mmhealth has been updated to monitor more events. New monitoring and health events have been added to monitor SSD wear, firmware, nameserver issues and critical threads in mmfsd that may be overloaded or hung. There are also new thresholds for DiskIoLatency_read, DiskIoLatency_write and MemoryAvailable_percent. There are also significant updates to SMB and NFS support as well as to AFM (active file management).

Prior to 5.0.3 there was a mutex contention issue with the SGInodeMapMutex that impacted Spectrum Scale file create performance as the number of threads creating files on a given node increased. This issue has been mitigated in 5.0.3 with fixes plus the use of maxFilesToCache and by also setting maxInodeDeallocHistory to 0 (or just reducing it from its default of 4096). 5.0.3 also resolves some of the issues around token contention when moving directories out of inodes.

In 5.0.4.2 enhancements were made to add an option to execute small sequential AIO/DIO writes as buffered I/O. This allows multiple small writes to be combined into a single larger I/O. The parameters involved are dioSmallSeqWriteBatching and dioSmallSeqWriteThreshold and they are set on the client nodes.

In 5.0.4.3 (APAR IJ22412) significant improvement was made to resolve locking issues with mmap read performance. Performance is dramatically improved when multiple threads are reading the same file.

Cluster Upgrades

The initial cluster was just a single node that we were about to expand to add four x86 Linux nodes and one Power Linux node. It was running Spectrum Scale 4.2.3.12. The initial step was to upgrade this node to 5.0.3.2 and then bring in the new client nodes, which would be installed at 5.0.3.2. Several months later we then went through the update process to go to 5.0.4.4.

The cluster has a single AIX node that is connected via fibre to all the disk LUNs. The five client nodes are all RHEL 7 (four are x86 and one is Linux on POWER). They are connected using the network (NSDs) to the storage. This, unfortunately, rules out rolling updates as the AIX node owns all the disks. In the future this will change, but for this upgrade series the upgrade was disruptive.

Upgrade Steps

The first step is to download the software from Fix Central. It is a separate download for each operating system: I had to download the AIX, Linux 64 bit x86 and Linux Power PC 64 Little Endian versions. In all cases a backup was take prior of the operating system.

4.2.3.12 to 5.0.3.2 on AIX

Since this was still a single node this was a very straight forward upgrade. In AIX it is a 2 stage upgrade – first you have to upgrade to the 5.0.3.0 level and then to the 5.0.3.2 level. The cluster was shut down and smitty update_all (install) was used to upgrade from 4.2.3.12 to 5.0.3.0. I then checked the level using lslpp:

lslpp -l | grep gpfs showed the level as 5.0.3.0.

Then smitty update_all was used again to update to 5.0.3.2 and the levels were checked again

```
lslpp -l | grep gpfs
  gpfs.adv                   5.0.3.2  COMMITTED  GPFS Advanced Features
  gpfs.base                  5.0.3.2  COMMITTED  GPFS File Manager
  gpfs.crypto                5.0.3.2  COMMITTED  GPFS Cryptographic Subsystem
  gpfs.ext                   5.0.3.2  COMMITTED  GPFS Extended Features
  gpfs.gskit               8.0.50.86  COMMITTED  GPFS GSKit Cryptography
  gpfs.license.dm            5.0.2.0  COMMITTED  GPFS Data Management Edition
  gpfs.msg.en_US             5.0.3.2  COMMITTED  GPFS Server Messages - U.S.
  gpfs.base                  5.0.3.2  COMMITTED  GPFS File Manager
  gpfs.docs.data             5.0.3.2  COMMITTED  GPFS Server Manpages and
```

The cluster was then started and tested.

I then updated the cluster so would support the new functionality in 5.0.3.2 using:

```
mmchconfig release=LATEST
```

Prior to running that command mmlsconfig showed:

```
minReleaseLevel 4.2.3.0
```

After the command was run the system showed:

```
minReleaseLevel 5.0.3.0
```

The additional nodes were then brought into the cluster, all at 5.0.3.2.

5.0.3.2 to 5.0.4.4 on the whole cluster

Several months later 5.0.4.4 came out with some features that made upgrading worthwhile. After downloading the correct software to each node and taking backups, the cluster was shutdown completely.

```
mmshutdown -a
```

mmgetstate -as was used to check everything was down

The AIX node was updated first, then the Power Linux node and finally the x86 nodes.

AIX Update

As with the earlier update smitty update_all was used to first update the node to 5.0.4.0 and then to update it to 5.0.4.4. Each time there were six filesets to go on. At the end lslpp showed:

```
lslpp -l | grep gpfs
  gpfs.adv                   5.0.4.4  APPLIED    GPFS Advanced Features
  gpfs.base                  5.0.4.4  APPLIED    GPFS File Manager
  gpfs.crypto                5.0.4.4  APPLIED    GPFS Cryptographic Subsystem
  gpfs.ext                   5.0.4.4  APPLIED    GPFS Extended Features
  gpfs.gskit               8.0.50.86  COMMITTED  GPFS GSKit Cryptography
  gpfs.license.dm            5.0.2.0  COMMITTED  GPFS Data Management Edition
  gpfs.msg.en_US             5.0.4.4  APPLIED    GPFS Server Messages - U.S.
  gpfs.base                  5.0.4.4  APPLIED    GPFS File Manager
  gpfs.docs.data             5.0.4.2  APPLIED    GPFS Server Manpages and
```

Spectrum Scale was then started just on the AIX node to make sure there were no issues and that the filesystems mounted, etc.

Linux Updates

The next node to be updated was the Linux on Power node followed by the x86 Linux nodes. The process was the same for all of these but we had a different file to use for the update.

For Linux on power the file was:

Spectrum_Scale_Advanced-5.0.4.4-ppc64LE-Linux-install

For Linux on x86 the file was:

Spectrum_Scale_Advanced-5.0.4.4-x86_64-Linux-install

The Linux install differs from the AIX one in that you do not have to go to the 5.0.4.0 release prior to going to 5.0.4.4. You go straight to 5.0.4.4.

The first step after ensuring Scale is down is to unload the kernel

On Linux on Power you will see:

```
mmfsenv -u

Unloading modules from /lib/modules/3.10.0-1062.4.1.el7.ppc64le/extra

Unloading module tracedev
```

On Linux on x86 you will see:

```
 mmfsenv -u

Unloading modules from /lib/modules/3.10.0-957.el7.x86_64/extra

Unloading module tracedev
```

The Linux-install file should then be made executable and run – this extracts the Scale code for the actual install. You will need to accept the license agreement during that step.

The files and rpms are extracted into: /usr/lpp/mmfs/5.0.4.4

```
cd /usr/lpp/mmfs/5.0.4.4/gpfs_rpms
```

This is an upgrade so the command issued will be something like:

```
rpm -Uvh gpfs.base*.rpm gpfs.gpl*rpm gpfs.license*rpm gpfs.msg*rpm gpfs.compression*rpm gpfs.adv*rpm gpfs.crypto*rpm
gpfs.java*rpm
```

You can then run rpm -qa | grep gpfs to check the levels and on Power Linux you should see:

```
rpm -qa | grep gpfs

gpfs.adv-5.0.4-4.ppc64le

gpfs.gpl-5.0.4-4.noarch

gpfs.crypto-5.0.4-4.ppc64le

gpfs.java-5.0.4-4.ppc64le

gpfs.gskit-8.0.50-86.ppc64le

gpfs.base-5.0.4-4.ppc64le

gpfs.license.adv-5.0.4-4.ppc64le

gpfs.compression-5.0.4-4.ppc64le

gpfs.msg.en_US-5.0.4-4.noarch

#
```

On x86 it will have x86_64 where it says ppc64le above.

You will then need to build the portability layer:

```
/usr/lpp/mmfs/bin/mmbuildgpl
```

Once that is done you can bring the node up and run your tests.

Once all the nodes have been upgraded and tested you can then update the release to LATEST. This is done using:

```
mmchconfig release=LATEST
```

This will update minReleaseLevel

```
minReleaseLevel 5.0.4.0
```

Finishing Touches

The final step is updating the filesystems to the latest level. In Scale 5.0.4.0 new filesystems are created at format level 22.0. 5.0.2.0 is level 20.01 and 5.0.3.0 is level 21.00. Current filesystems will remain at the level they were created at. Once all the nodes in the cluster have been upgraded then all the filesystems should be upgraded to level 22.0. This can be done dynamically without an outage by running the following command against each filesystem.

```
mmchfs  filesystem -V full
```

Level 22.0 adds support for thin provisioned storage devices, NVMe SSDs and some additional AFM functionality. Once you update to level 22.0 only nodes at 5.04 or higher can access the filesystems.

I usually wait a week before running the mmchfs against each filesystem. While it is highly unlikely that I would revert to the old cluster level I like to keep that open as an option. Once you update the filesystem you can’t go back.

You can check the current filesystem level by running mmlsfs against the filesystem. About half way down you will see lines like:

```
-V              21.00 (5.0.3.0)        Current file system version
                20.01 (5.0.2.0)        Original file system version
```

Summary

In this article we have looked at what is new in 5.0.4.4 and on what is involved in updating a Spectrum Scale cluster to 5.0.4.4. While we were unable to perform a rolling upgrade that is certainly an option where you have redundancy in the servers that own the disks. Because this is a mixed cluster it was not possible to have the AIX node and the Linux nodes all be fibre connected. The upgrade itself was very straight forward and very easy to do.

References

For more information:

1. Spectrum Scale Try and Buy

https://www.ibm.com/products/scale-out-file-and-object-storage (https://www.ibm.com/products/scale-out-file-and-object-storage)

2. NSDPerf C program

https://github.com/IBM/SpectrumScale_NETWORK_READINESS/blob/master/nsdperf.C (https://github.com/IBM/SpectrumScale_NETWORK_READINESS/blob/master/nsdperf.C)

3. Scale 5.0 administration guide

https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.0/com.ibm.spectrum.scale.v5r00.doc/pdf/scale_adm.pdf (https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.0/com.ibm.spectrum.scale.v5r00.doc/pdf/scale_adm.pdf)

4. Fix Central Spectrum Scale Software

https://www.ibm.com/support/fixcentral/options?selectionBean.selectedTab=find&selection=System+Storage%3bStorage+software%3bSoftware+defined+storage%3bibm%2fStorageSoftware%2fIBM+Spectrum+Scale (https://www.ibm.com/support/fixcentral/options?selectionBean.selectedTab=find&selection=System+Storage%3bStorage+software%3bSoftware+defined+storage%3bibm%2fStorageSoftware%2fIBM+Spectrum+Scale)

5. Spectrum Scale 5.0.4.4 AIXRelease Notes

https://www.ibm.com/support/pages/readme-and-release-notes-release-5044-ibm-spectrum-scale-5044-spectrumscalestandard-5044-ppc64-aix-update-readme (https://www.ibm.com/support/pages/readme-and-release-notes-release-5044-ibm-spectrum-scale-5044-spectrumscalestandard-5044-ppc64-aix-update-readme)

6. Spectrum Scale 5.0.4 Knowledge center

https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.4/ibmspectrumscale504_welcome.html (https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.4/ibmspectrumscale504_welcome.html)

7. Spectrum Scale User Group Presentations

https://www.spectrumscaleug.org/presentations/ (https://www.spectrumscaleug.org/presentations/)

About the author
Jaqui Lynch has over 38 years of experience working with a projects and OSes across vendor platforms, including IBM Z, UNIX systems and more.

Related Content

Application development (/application-development) Service Programs and Signatures → (https://ibmsystemsmag.com/Power-Systems/10/2003/service-programs-signatures)
Systems management (/systems-management) A Look at File Systems → (https://ibmsystemsmag.com/Power-Systems/09/2004/file-systems-commands)
Systems management (/systems-management) Accessing the Data in Core Dumps → (https://ibmsystemsmag.com/Power-Systems/01/2006/core-dumps-data-access)

IBM Systems MEDIA