

Spectrum Scale 5.0.2 Updates

Christopher D. Maestas



Please Note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

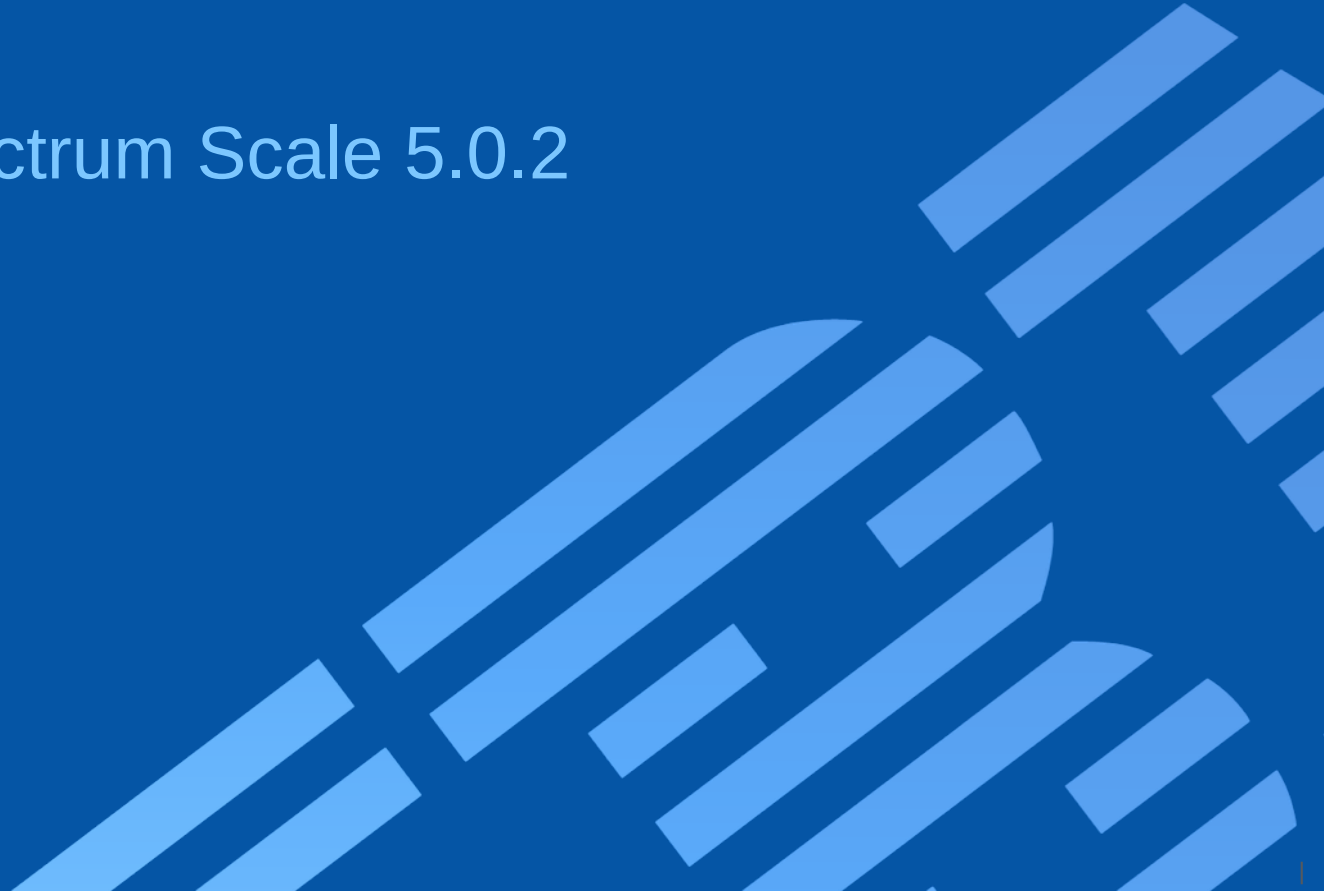
Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

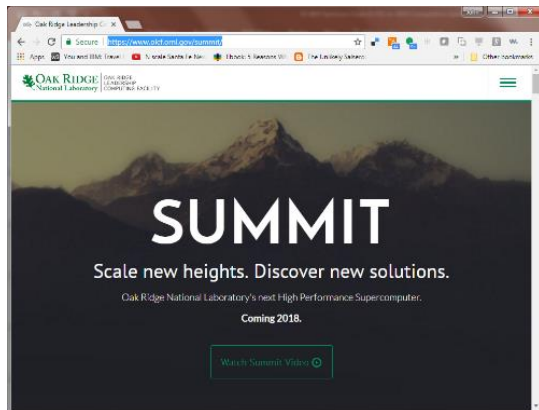


New in IBM Spectrum Scale 5.0.2

Performance!



Performance engineering matters



<https://www.olcf.ornl.gov/summit/>

Imagine you need to deliver:

- 2.5 TB/sec single stream IOR
as requested from ORNL
- 1 TB/sec 1MB sequential read/write
as stated in CORAL RFP
- Single Node 16 GB/sec sequential read/write
as requested from ORNL
- 50K creates/sec per shared directory
as stated in CORAL RFP
- 2.6 Million 32K file creates/sec
as requested from ORNL

What innovations in storage would this require?

Quick side not about performance

IBM has a benchmark center in Poughkeepsie

On the truck code of Scale and ESS

Take what research and performance teams do and replicate

Publish externally like

- IOR runs on ESS GL6S and GS4S with **Scale 5.0.1.1**

May have some more data before SC18!

- POSIX transactions per second

maxActiveIallocSegs enhancement

A single node has created a large number of files in multiple directories

Processes and threads on multiple nodes are now concurrently attempting to delete or unlink files in those directories.

Configuration parameter– *maxActiveIallocSegs*

Specifies the number of active inode allocation segments maintained

The default value is 8 on file systems that are created at file system format version 5.0.2 or later, otherwise it is 1

- change of this attribute is not effective until after the file system is remounted.
- Not dependent on fs version format

If equal nodes creating and deleting, no BIG difference between 5.0.1 and 5.0.2

maxStatCache enhancement

Spectrum Scale < 5.0.2, the stat cache is not effective on the Linux platform

`maxStatCache=0 || LROC (man mmchconfig)`

Spectrum Scale >= 5.0.2 stat cache is effective on the Linux platform for all configurations

Configuration parameter – `maxStatCache`

maintains only enough inode information to perform a query on the file system.

file and dir stat operation performance may be improved when the inode is in the stat cache.

If not set, $\text{maxStatCache} = 4 * \text{maxFilesToCache}$

“`mmcachectl show`” can be used to verify if file inode is in the stat cache

Commands: `ls -l` and `mdtest`
have shown improvement.

<i>FileType</i>	<i>NumOpen Instances</i>	<i>NumDirect IO</i>	<i>Size (Total)</i>	<i>Cached (InPagePool)</i>	<i>Cached (InFileCache)</i>
<i>file</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>C</i>
<i>file</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>C</i>
<i>file</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>C</i>

IBM Spectrum Scale 5.0.2

Operational Efficiencies



Rebuild GPL module if new kernel detected

autoBuildGPL configuration option.

Before starting GPFS, if the kernel module is missing, automatically call *mmbuildgpl* to build the GPL if *autoBuildGPL* parameter is configured.

```
mmchconfig autoBuildGPL={no|yes|quiet|verbose|quiet-verbose|verbose-quiet}
```

Where:

no This is the default. No action will be taken if no kernel module is found

yes *mmbuildgpl* will be called to build the GPL if the kernel module is missing

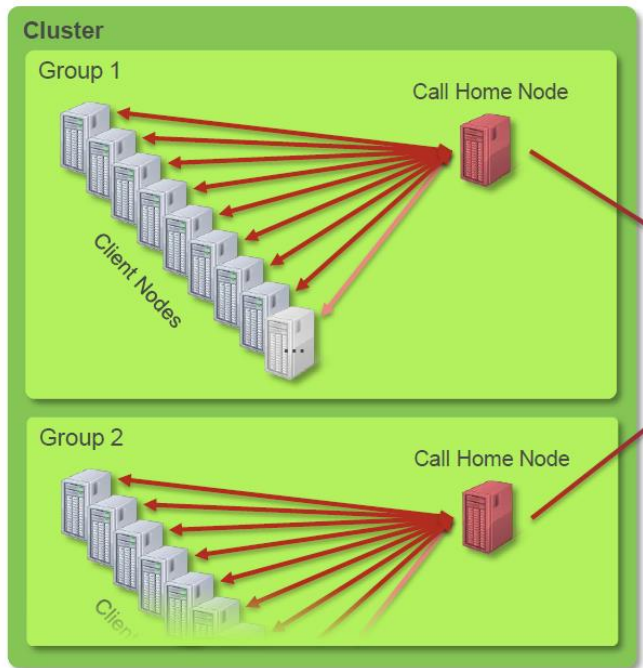
quiet Same as yes. The *mmbuildgpl* command will be called with *--quite* option.

verbose Same as yes. The *mmbuildgpl* command will be called with *-v* option.

quiet-verbose or verbose-quiet

Both *--quite* and *-v* will be passed to *mmbuildgpl*

Proactive Services - callhome



Can group nodes by class

Can find configuration challenges and recommend changes

Uses the `-Y` flag for mm commands

Looking at pagepool - mmcachectl

mmcachectl show --show-filename

```
[root@ScaleGUILabCentOS7 ~]# mmcachectl show --show-filename | head
```

FSname	Fileset ID	Inode	SnapID	FileType	NumOpen Instances	NumDirect IO	Size (Total)	Cached (InPagePool)	Cached (InFileCache)	FileName
guilabfs1	1	165893	0	directory	0	0	3872	0	FD	/ibm/guilabfs1/obj_default/o/zldevice6
guilabfs1	1	165909	0	directory	0	0	3872	0	FD	/ibm/guilabfs1/obj_default/o/zldevice22
guilabfs1	1	165985	0	directory	0	0	3872	0	FD	/ibm/guilabfs1/obj_default/o/zldevice98
cesSharedRoot	0	50705	0	file	0	0	1636	0	FD	/ibm/cesSharedRoot/object/keystone/pg_ident.conf
guilabfs1	3	429168	0	directory	0	0	3872	0	FD	/ibm/guilabfs1/obj_N0ufoguidemo/s13651809171zldevice113
guilabfs1	1	161796	0	directory	0	0	3872	0	FD	/ibm/guilabfs1/obj_default/ac/zldevice5
cesSharedRoot	0	50440	0	directory	0	0	16384	16384	F	/ibm/cesSharedRoot/object/keystone/base/1

```
[root@ScaleGUILabCentOS7 ~]#
```

```
[root@ScaleGUILabCentOS7 ~]# mmcachectl show | head
```

FSname	Fileset ID	Inode	SnapID	FileType	NumOpen Instances	NumDirect IO	Size (Total)	Cached (InPagePool)	Cached (InFileCache)
guilabfs1	1	165893	0	directory	0	0	3872	0	FD
guilabfs1	1	165909	0	directory	0	0	3872	0	FD
guilabfs1	1	165985	0	directory	0	0	3872	0	FD
cesSharedRoot	0	50705	0	file	0	0	1636	0	FD
guilabfs1	3	429168	0	directory	0	0	3872	0	FD
guilabfs1	1	161796	0	directory	0	0	3872	0	FD
cesSharedRoot	0	50440	0	directory	0	0	16384	16384	F

Create maintenance period on NSD disks, server or entire cluster

Users can still by pass this, but disks may be marked as down and *mmchdisk* to start these down disks could take a long time

```
[root@ScaleGUILabCentOS7 ~]# mount | grep gpfs
cesSharedRoot on /ibm/cesSharedRoot type gpfs (rw,relatime,seclabel)
guilabfs1 on /ibm/guilabfs1 type gpfs (rw,relatime,seclabel)
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmchfs guilabfs1 --maintenance-mode yes
Failed to enable maintenance mode for this file system.
Maintenance mode can only be enabled once the file system has been unmounted
everywhere. You can run the mmlsmount <File System> -L command to see which
nodes have this file system mounted. You can also run this command with the
"--wait" option, which will prevent new mounts and automatically enable
maintenance mode once the unmounts are finished.
mmchfs: tschfs failed.
mmchfs: Command failed. Examine previous error messages to determine cause.
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmumount guilabfs1
Wed Sep 19 12:18:40 UTC 2018: mmumount: Unmounting file systems ...
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmchfs guilabfs1 --maintenance-mode yes
[root@ScaleGUILabCentOS7 ~]# mmlsfs guilabfs1 --maintenance-mode
flag                value                description
-----
--maintenance-mode Yes                Maintenance Mode enabled?
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmmount guilabfs1
Wed Sep 19 12:19:21 UTC 2018: mmmount: Mounting file systems ...
mount: permission denied
mmm mount: Command failed. Examine previous error messages to determine cause.
[root@ScaleGUILabCentOS7 ~]#
```

Check remote clusters

mmnetverify now can check remote clusters for host-name resolution, network connectivity via ping, and GPFS daemon connectivity.

***mmnetverify* -cluster NAME**

nsdperf - bandwidth verification is still recommended

Callbacks for IO hangs (man mmaddcallback)

diskIOHang callback add notification and datacollection scripts to analyze a local I/O request pending for more than 5 minutes.

panicOnIOHang panics the node kernel when a local I/O request pends for more than five minutes.

Remember deadlocks? Don't do it like that right out of the gate!



Estimate an offline mmfsck

- New mmfsck option: --estimate-only
- Displays estimation of offline fsck run time for given mmfsck options
- Does not scan the file system
- Can be run when file system is online or offline
- Works for offline fsck only
- Participating nodes must be at 5.0.2 or later
- The estimate is based on mmfsck command line options, configuration of the target file system and average disk and network I/O throughput of the participating nodes

```
mmfsck fs1 -nv --estimate-only
```

```
Checking "fs1"
```

```
  FckFlags                0x2000009
```

```
...
```

```
Estimating fsck run time
```

```
Measuring disk stat...
```

```
Measuring RPC stat...
```

```
Estimating bytes to scan...
```

```
Fck will complete in 0 hours 0 minutes 58 seconds  
(+/- 4 seconds)
```

```
Note that this estimate does not factor in any CPU  
processing overhead and assumes balanced scan  
workload across all threads and nodes
```

```
...
```

```
File system is clean.
```

```
Fck completed in 0 hours 0 minutes 0 seconds
```

Network PD improvement –

dump the TCP_INFO when disk lease overdue occurs (Linux only)

- Is it a GPFS problem or network problem by looking at fields of TCP_INFO

Network resiliency enhancement - prioritize commMsgCheckMessages RPC to avoid RPC time-out node requested expels

- When sending commMsgCheckMessages RPC could be blocked because of heavy TCP connection (lots of NSD read and write RPC), and if the wait time of exclusive use exceeds 300s, this could cause expel even if the network is good though it's just slow.

Network resiliency enhancement - when CM pings a node near to being expelled, due to a lease timeout, ensure take into account the subnets configuration if set.

- When doing ping check, such as disk lease overdue, current design is to do ping check on the primary address, then cannot detect network problems on the subnets IP address, so check subnets IP address

5.0.2 Spectrum Scale GUI –What's new

- Remote Cluster Capacity data for Filesets and File Systems
- Remote Cluster Quota info
- Node Class Management
- CES IP Health status, Preferred CES nodes and non-hostable nodes exposed
- File Audit Log enable/disable
- Extended Legend in Dashboard views
- More lines in charts (up to 20)
- ClusterName in banner
- Filtered views by health state
- Enhanced event filtering

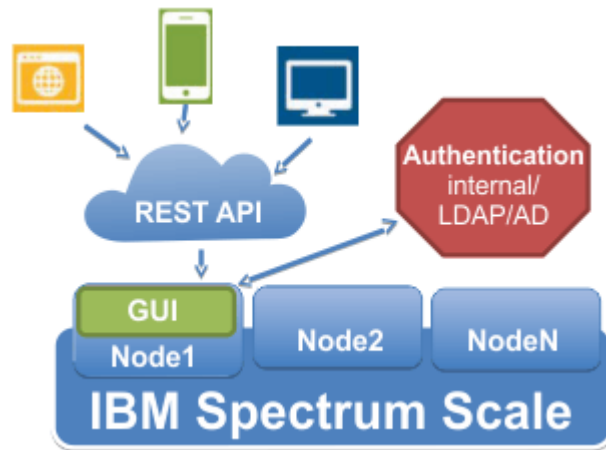
GUI and the REST API

Driven by same WebSphere server

Authentication shared between GUI and REST API

THE strategic interface for integrating with 3rd party customer applications, automation or monitoring

[https://\[GUI_NODE\]:443/ibm/api/explorer/#!/Spectrum_Scale_REST_API_v2/](https://[GUI_NODE]:443/ibm/api/explorer/#!/Spectrum_Scale_REST_API_v2/)



Liberty REST APIs

API Discovery : APIs available from the API Discovery feature

ShowHide : List Operations : Expand Operations

Spectrum Scale REST API v1 : DEPRECATED

ShowHide : List Operations : Expand Operations

Spectrum Scale REST API v2 : APIs for managing a Spectrum Scale cluster

ShowHide : List Operations : Expand Operations

GET	/nscalemgt/v2/cluster	Get details about your access token. If you are a security administrator a list of all access tokens is returned.
POST	/nscalemgt/v2/cluster	Request access to this API.
GET	/nscalemgt/v2/cluster/status	Get status about your access token.
GET	/nscalemgt/v2/cluster/addresses	Get listing of CDS Addresses.
GET	/nscalemgt/v2/cluster/addresses/{cddAddress}	Get detailed information about a CDS Address.
GET	/nscalemgt/v2/cluster/services	Get listing of CDS Services.
GET	/nscalemgt/v2/cluster/services/{serviceName}	Get detailed information about a CDS Service.
GET	/nscalemgt/v2/cluster	Get current configuration information.
GET	/nscalemgt/v2/cluster/config	Get cluster config.
GET	/nscalemgt/v2/filesystems	List of filesystems in the cluster.
GET	/nscalemgt/v2/filesystems/{filesystemName}	Get detailed information about a filesystem.
GET	/nscalemgt/v2/filesystems/{filesystemName}/acl/path	Get access control list of filesystem.
POST	/nscalemgt/v2/filesystems/{filesystemName}/acl/path	Write access control list of filesystem.
GET	/nscalemgt/v2/filesystems/{filesystemName}/info/state	List all state in a filesystem.
GET	/nscalemgt/v2/filesystems/{filesystemName}/info/data	Get listing of disks.
GET	/nscalemgt/v2/filesystems/{filesystemName}/disks/{diskName}	Get detailed information about a disk.
GET	/nscalemgt/v2/filesystems/{filesystemName}/filesystems/filesystems	Get listing of filesystems.
POST	/nscalemgt/v2/filesystems/{filesystemName}/filesystems	Create a new filesystem.
DELETE	/nscalemgt/v2/filesystems/{filesystemName}/filesystems/{filesystemName}	Delete a filesystem.

REST API - Extra endpoints in 5.0.[1,2]

URL	Operation
/cliauditlog	GET
/config	PUT
/filesystems/{filesystemName}/filesets/{filesetName}/afmctl	POST
/filesystems/{filesystemName}/policies	GET, PUT
/nodes/{name}/services	GET, PUT
/perfmon/sensors	GET, PUT

URL	Operation	Description
/filesystems/{filesystemName}/audit	PUT	Enable/Disable File Audit Logging (mmaudit)
/smb/shares/{shareName}/acl	DELETE, GET, PUT	SMB Share ACL management

GUI optimizations

- Reduce call to mmhealth
- Reduce to 2 CPU cores for JAVA and postgres
- Reduce local I/O on GUI node
- Reduce memory on GUI node
- Should help with ESS EMS



System health

mmces address list can see who is preferred (**--extended-list**) and who cannot host (**-- full-list**)

mhealth --show-state-changes can display state change

```
2018-09-17 15:37:24.191434 UTC      node_state_change      INFO      The state of this node changed to TIPS.
2018-09-17 15:50:53.965807 UTC      component_state_change  INFO      The state of component NFS changed to STOPPED.
2018-09-17 15:50:58.807964 UTC      component_state_change  INFO      The state of component NFS changed to CHECKING.
2018-09-17 15:51:09.222651 UTC      component_state_change  INFO      The state of component NFS changed to DEGRADED.
2018-09-17 15:51:09.245315 UTC      node_state_change      INFO      The state of this node changed to DEGRADED.
2018-09-17 15:52:08.360062 UTC      component_state_change  INFO      The state of component NFS changed to HEALTHY.
2018-09-17 15:52:08.387570 UTC      node_state_change      INFO      The state of this node changed to TIPS.
2018-09-17 15:55:08.597398 UTC      component_state_change  INFO      The state of component NFS changed to STOPPED.
2018-09-17 15:55:16.914209 UTC      component_state_change  INFO      The state of component NFS changed to CHECKING.
2018-09-17 15:55:23.831495 UTC      component_state_change  INFO      The state of component NFS changed to DEGRADED.
2018-09-17 15:55:23.852630 UTC      node_state_change      INFO      The state of this node changed to DEGRADED.
2018-09-17 15:56:24.013737 UTC      component_state_change  INFO      The state of component NFS changed to HEALTHY.
2018-09-17 15:56:24.047747 UTC      node_state_change      INFO      The state of this node changed to TIPS.
```

When unmounting CES FS, error if CES services are running

```
[root@ScaleGUIlabCent057 ~]# mmumount cesSharedRoot
Mon Sep 17 17:48:32 UTC 2018: mmumount: Unmounting file systems ...
umount: /ibm/cesSharedRoot: target is busy.
(In some cases useful info about processes that use
the device is found by lsOF(8) or fuser(1))

cesSharedRoot device umount failed.
Please suspend the CES node(s) ScaleGUIlabCent057 using --stop flag first to release the shared root /ibm/cesSharedRoot before unmounting.
If needed check for other processes locking the file system.
mmumount: Command failed. Examine previous error messages to determine cause.
```

Install Toolkit 5.0.2 New Features

Recall install toolkit introduced in 4.1.1.0

Mark nodes offline during upgrade

Do an offline upgrade for entire cluster

Exclude nodes from upgrade

(upgrade subset of nodes)

Resume a previously failed upgrade

Enhanced node listing

(NSD, client, protocol, audit, callhome ...)

Enhanced CES shared root creation and detection (populate)

Ability to specify broker nodes for
File Audit Logging

Removal of gpfs.ext on upgrade
(consolidated into gpfs.base)
(works with rpm/yum update too)

Upgraded chef for orchestration

Support Ubuntu 18.04 and 18.04.1
s390x installation support

Watch Folder installation
(via key enablement)

Windows 10 support! Pro and Enterprise

Both heterogeneous and homogeneous clusters

Currently, Secure Boot must be disabled on Windows 10 nodes

FAQ update: **A14.7: Windows 10 related advisories and recommendations:**

1. *User Access Control (UAC) must not be disabled on latest Windows versions such as Windows 10.* GPFS now runs with UAC enabled (default OS setting).
2. Latest versions of Windows such as Windows 10 now come with a built-in antivirus component known as **Windows Defender**. While performing real-time scanning of files, Windows Defender may memory-map these files even when they are not in use by any user application. This memory-mapping of files on GPFS filesystems by Windows Defender in the “background”, can result in performance degradation. *Therefore, it is recommended that GPFS drives/volumes be “Excluded” from Windows Defender scans all together.*
3. Windows 10 version 1803, now incorporates a native secure shell ‘**OpenSSH** for Windows’. GPFS requires 'OpenSSH for Cygwin', especially if the Windows node(s) join a GPFS cluster having Linux/AIX nodes. Therefore, before operating a Windows 10 node in a mixed GPFS cluster, please ensure that the Windows native 'OpenSSH SSH Server' is not enabled/running and that the 'Cygwin sshd' service is working reliably. Additionally, it is recommended that the Windows Subsystem for Linux (WSL) feature not be installed to avoid potential conflicts with Cygwin.

IBM Spectrum Scale 5.0.2

Other Protocols



“mmuserauth” enhancement for password

Example for FILE authentication

```
mmuserauth service create --type ad --data-access-method file  
--netbios-name test --user-name administrator --idmap-role  
master --servers myADServer --pwd-file fileauth
```

Contents of fileauth saved at /var/mmfs/ssl/keyServ/tmp/ are:

```
%fileauth:  
password=Passw0rd
```

Example for OBJECT authentication

```
mmuserauth service create --type ad --data-access-method  
object --base-dn "dc=example,DC=com" --servers myADserver  
--user-id-attr cn --user-name-attr SAMAccountName --user-  
objectclass organizationalPerson --user-dn  
"cn=Users,dc=example,dc=com" --pwd-file objectauth
```

Contents of fileauth saved at /var/mmfs/ssl/keyServ/tmp/
are:%objectauth:

```
password=Passw0rd  
ksAdminPwd=Passw0rd1  
ksSwiftPwd=Passw0rd2
```

For FILE authentication now validates DNS records to AD servers as well

Samba update

- Allow user to change min and max SMB protocols
- Reduce load on cache generation if a lot of idmap lookups occur
- Graceful behavior of ctdb during OOM
- Log memory, change to unhealthy if swap > 95% used

Spectrum Scale Release	General Availability	Samba Version	Platform Support (accum.)
4.1.1	2Q15	4.2	x86_64/RHEL7
4.2.0	4Q15	4.3	ppc64/RHEL7
4.2.1	2Q16	4.3	x86_64/SLES12
4.2.2	4Q16	4.4	ppc64le, ppc64, x86_64 / RHEL7.2
4.2.3.0 - 4.2.3.8	2Q17	4.5	x86_64, ppc64, ppc64le / RHEL 7.3, 7.4
5.0.0	4Q17	4.6	x86_64/Ubuntu 16.04.2
5.0.1	1Q18	4.6	RHEL 7.5 (5.0.1.1)
5.0.2 >= 4.2.3.9	3Q18	4.6	+ Ubuntu 18.04

Ganesha NFS update

- Restructure code to “maybe” support more exports per filesystem
- Pseudo path for export at creation time*
- Performance counters (ganesha_stats)*

Integration with mm and GUI coming soon

```
[root@ScaleGUILabCent057 ~]# mmnfs export add /ibm/guilabfs1/cdm --pseudo /cdm [2/1132]
ccess_type=RW)"
mmnfs: The NFS export was created successfully
mmnfs: Restarting NFS services.
[root@ScaleGUILabCent057 ~]# mmnfs export list
```

Path	Delegations	Clients
/ibm/guilabfs1/cdm	NONE	*

```
[cdmaestas@oc0873784061 scale_GUI_lab]$ mount | grep nfs4
192.168.123.10:/cdm on /mnt type nfs4 (rw,relatime,vers=4.0,rsiz=1048576,wsiz=104857
```

Object Release Overview

Spectrum Scale		Openstack
4.1.1		Kilo
4.2.1		Liberty
4.2.2		Mitaka
5.0.2		Pike

Spectrum Scale		swift3
4.1.1		1.7
4.2.0		1.8
4.2.1		1.10
5.0.2		1.12

Objectizer changes

- Used to convert files into objects in Swift on File environments
- Process containers in parallel
- Remove limitation on number of accounts that could be processed
- Handle filenames with UTF or Unicode characters



AWS Quick Starts

IBM Spectrum Scale on AWS

High-performance storage solution for managing data at scale

Deploy on AWS into a new VPC

or deploy into an existing VPC

[View deployment guide](#)

Two models of deployment with **a good deployment guide!**

virtual private cloud (VPC) that spans two Availability Zones in your AWS account.

- Can build a new VPC for IBM Spectrum Scale, or
- Deploy the software into your existing VPC

Deployment and configuration tasks are automated by AWS CloudFormation templates

- Customizable prior to launch

What Are We Doing on Amazon?

AWS Marketplace offering that allows customers with a supported way of using their Spectrum Scale Data Management edition licenses on AWS.

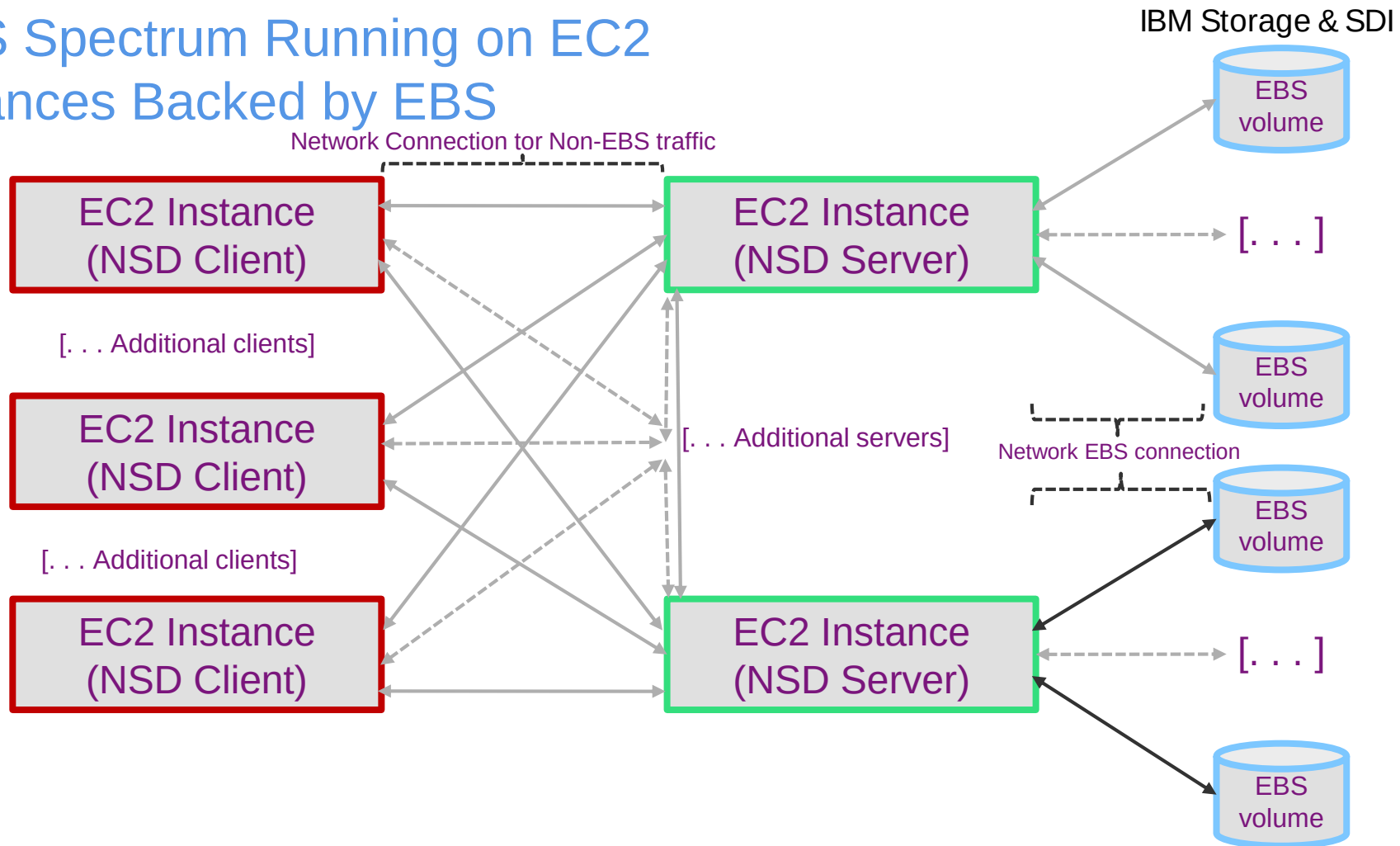
Provide customers with an easy way to deploy Spectrum Scale, a deployment method we understand and have tested. Benefits include:

- starting from a well tested configuration with some tuning done*
- limiting the different types of configurations that we have to support on AWS

Customers can get support for production Spectrum Scale clusters running on AWS, provided they:

1. Use only the features that we currently support on AWS
2. Deploy/create Spectrum Scale clusters on AWS using the IBM Spectrum Scale Marketplace deployment method (IBM Spectrum Scale team's cloud formation templates)
3. Use the Amazon Machine Image (AMI) that we provide in our IBM Spectrum Scale Marketplace offering

AWS Spectrum Running on EC2 Instances Backed by EBS



mmaws - Managing IBM Spectrum Scale workflows on AWS

Usage:

<code>mmaws add_nodes</code>	Add compute/server nodes
<code>mmaws remove_nodes</code>	Removing compute/server nodes
<code>mmaws list_instances</code>	Listing instances in the vpc
<code>mmaws start_nodes</code>	Starting nodes
<code>mmaws stop_nodes</code>	Stopping compute/all nodes
<code>mmaws create_lambda_functions</code>	Create Lambda functions
<code>mmaws collect_debug_data</code>	Collect AWS debug data

optional arguments:

<code>-?, -h, --help, help</code>	show this help message and exit
-----------------------------------	---------------------------------

New in IBM Spectrum Scale 5.0/5.0.1

Security and Compliance!



DEFAULTNISTSP800131AFast Encryption enhancement

DEFAULTNISTSP800131AFast uses **128-bit key length** and 128-bit keys are secure according to NIST publication SP 800.131A.

DEFAULTNISTSP800131AFast can provide 5-20% speed up for certain I/O workloads (e.g. large block random reads, direct I/O) compared to DEFAULTNISTSP800131A

Encryption ALGO value – **DEFAULTNISTSP800131AFast**
Maps to 'AES:128:XTS:FEK:HMACSHA512'

Sample Encryption policy

```
RULE 'EncPolicyGeneratorRule2' ENCRYPTION 'EncPolicyGenerator2' IS  
ALGO 'DEFAULTNISTSP800131AFast'  
KEYS('KEY-ABC..XYZ:sklmnRKM')  
RULE 'EncPolicyGeneratorFileRule2' SET ENCRYPTION 'EncPolicyGenerator2'  
FOR FILESET('encryptedFSet_FAST_NIST')
```

For I/O > 2 MiB Write (> 15%) and Read (> 3%) performance is faster versus **DEFAULTNISTSP800131A**

Alert for Certificate Expiration in keystore

IBM Storage & SDI

Problem: Spectrum Scale does not alert when client or key-server certificate in keystore is going to expire

Solution: Periodically check validation of all certificates in keystore. (Including client and key server certificates); generate alert and dump it into GPFS log when detect coming expiration, for example, in next 6 month.

Watch Folders 101 - /usr/lpp/mmfs/samples/util/tswf.C

Take actions based on filesystem events

- Run against folders, filesets (independent too)
- Modeled after Linux inotify, but works with clustered filesystems, and supports recursive watches for filesets (independent too)

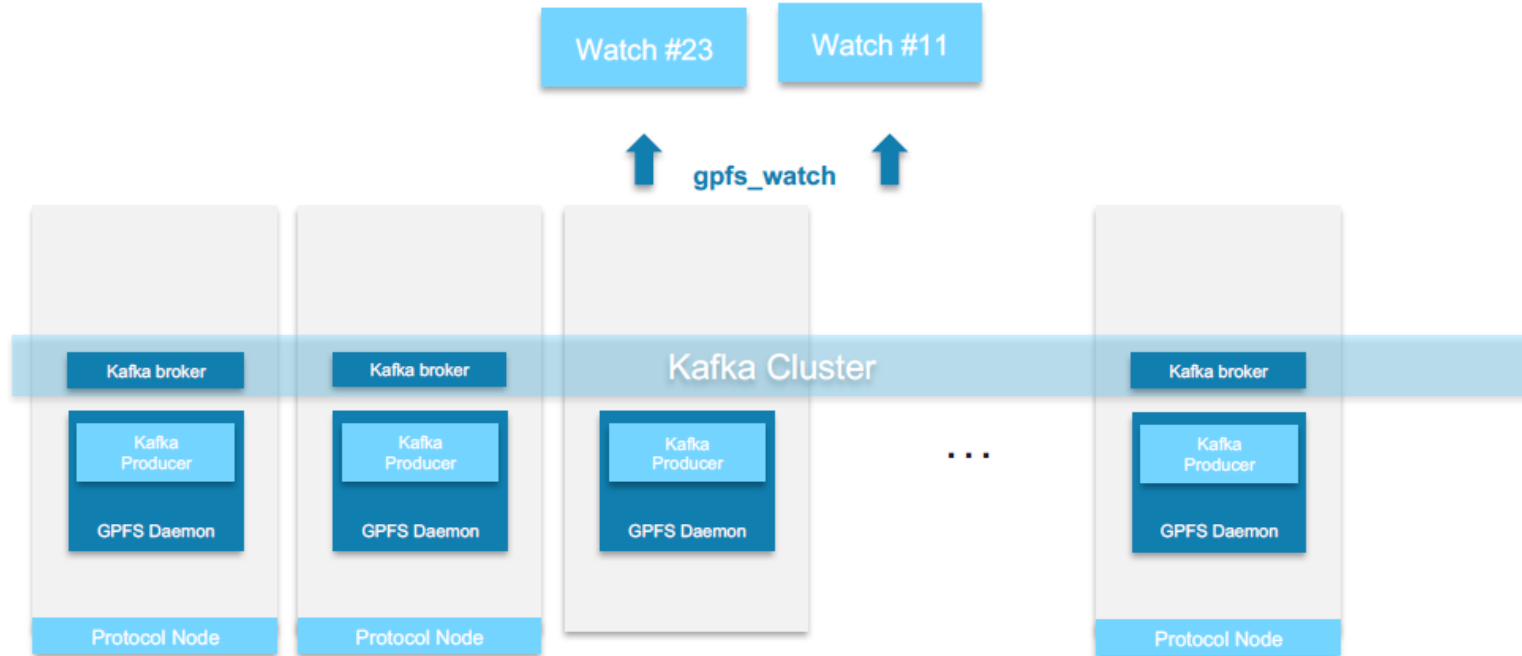
2 primary components

- GPFS API (included within <gpfs_watch.h>)
- **mmwatch**— provides information of all watches running within cluster

A watch folder application uses API as a C program on cluster

- Utilizes message queue to receive events from multiple nodes and consume from the node running the program
- Events come in from all eligible nodes within cluster and from accessing clusters

Watch Folders using LWE and Message queue (Kafka)



Limitations and Requirements `#include <gpfs_watch.h>`

➤ **Requires key enablement in 5.0.2**

- Development/Sales will provide approved use cases with a hidden configuration variable
- All Clusters and file system format
 - code level $\geq 5.0.2$
- Message queue must be enabled on owning cluster of filesystem
 - Minimum 3 Linux quorum nodes and 3 nodes for brokers
 - Data Management Edition (DME)
(yes advanced too)
- 25 watches per file system
 - 3 GB per watch of local disk space per watch
- 100 watches per cluster

Watch Folder Troubleshooting

mmwatch –

verify information about all currently running watches

/var/adm/ras/mmwf.log –

primary log file for watch API and mmwatch command

/var/adm/ras/mmfs.log

(major problems with policy, watches, etc.)

/var/adm/ras/mmsgqueue.log

(problems with the message queue)

Watch Folder Performance

Streaming I/O is fine

Lots of reads (70/30) is fine

Lots of metadata performance, it depends

mmmsgqueue updates

mmmsgqueue status -v

Troubleshoot message queue issues

Show: status and uptime for brokers and zookeepers

Shows which brokers haven't registered with zookeepers

mmmsgqueue config --add-nodes -N

Add cluster nodes to message queue without disabling FAL or the message queue

Integrated checks for proper packages, local disk space and removal of old Kafka logs

For FAL - adds new partitions, redistributes partitions and starts consumer processes

File Audit Logging (FAL) updates

mmaudit all list –events

Default is “all” but choose from:

1. ACLCHANGE,
2. CLOSE,
3. CREATE,
4. DESTROY,
5. GPFSATTRCHANGE,
6. OPEN,
7. RENAME,
8. RMDIR,
9. UNLINK,
10. XATTRCHANGE

Update monitored events without disabling/re-enabling

Support remote clusters

- Producers on remote clusters can send events to file system cluster
- [mmaudit/mmsgqueue](#) run from file system cluster

Space Requirements

- Broker node – localdisk 20 GB

Support callhome

- [mmwatch all list -Y](#)
- [mmsgqueue list –topics](#)

New in IBM Spectrum Scale 5.0.2

Data Movement (Compression, AFM and TCT)

AFM Performance improvements:

- User defined gateway mapping with `afmHashVersion=5`
Assign at fileset create or modify after `afmGateway=NODENAME`

AFM prefetch enhancements:

- Get statistics of transfer during pre-fetch
- `--enabled-failed-file-list`
- `--retry-failed-file-list`
- `--directory # build a list!`
- `--policy # policy syntax`



Transparent Cloud Tiering enhancements

IBM Storage & SDI

Remote mounted filesystem support

Clients can access tiered files on a remotely mounted filesystem

Ability to tier different **filesets** to different cloud containers

Yes, can now be **fileset** focused!

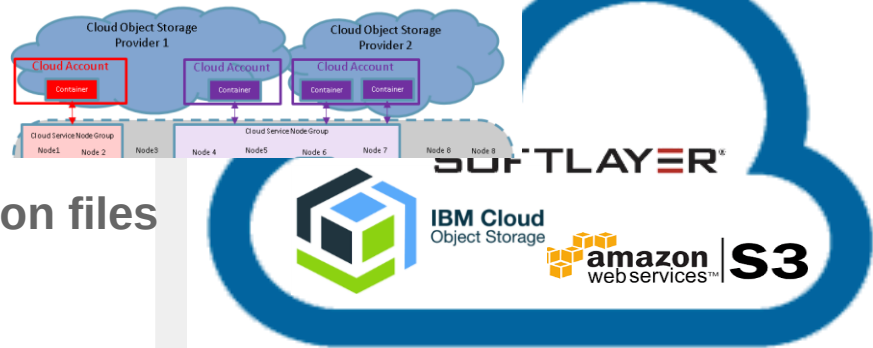
Enhanced support for multiple cloud accounts and containers

Pull and push to different cloud providers

Container spillover in same fileset > 100 Million files

Usability improvements

CLI has been enhanced for ease of deployment with multiple cloud accounts



Big Data and Analytics Enhancements

IBM Storage & SDI

Spectrum Scale Certification with HortonWorks Data Platform (HDP) 3.0

Certified on both Power8 and x86 platforms

Certified with Ambari 2.7 for rapid deployment

Integration with Spectrum Conductor for Spark

For further details, see the Redpaper:

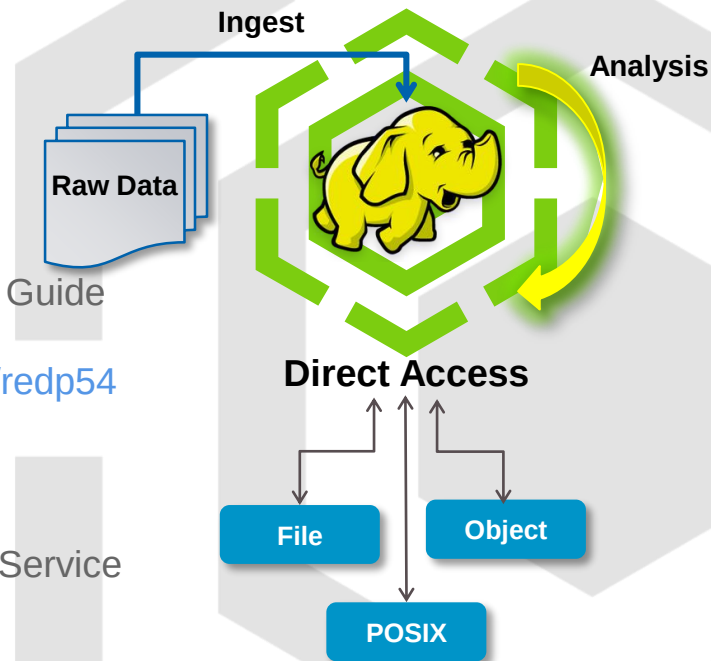
Hortonworks Data Platform with IBM Spectrum Scale: Reference Guide
for Building an Integrated Solution

<https://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5448.html?Open>

Hortonworks Data Platform (HDP) Solution Brief:

Hortonworks Data Platform on IBM Power Systems for Financial Service

<https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=POS03163USEN>



Big Data and Analytics Enhancements

IBM Storage & SDI

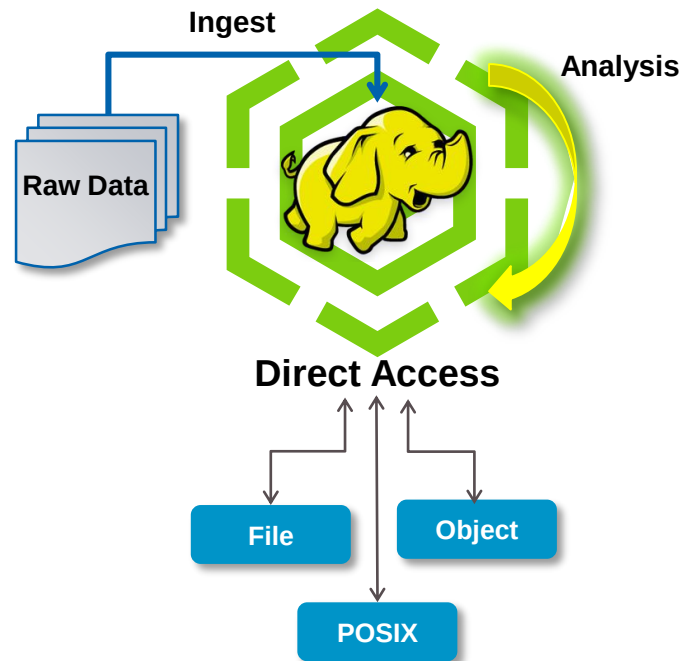
FPO v5.0.2

Try and resume suspended disks if requested

Check for replica mismatch `mmrestripefile -c --read-only`

HDFS Transparency v3.0.0-0 GA

- Supports HDP 3.0 and Mpack 2.7.0
- Supports Apache Hadoop 3.0.x
- Support native HDFS encryption
- Spectrum Scale Configuration now in:
 - `/var/mmfs/hadoop/etc/`
 - `/var/log/transparency`



Thank You.
IBM Storage & SDI

