Spectrum Scale 5.0.3 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.

IBM Spectrum Scale

Summary!

Use Cases for Spectrum Scale and the Elastic Storage Server (ESS)

1. Back-up / Restore
2. Archive
3. Information Life Cycle Management
4. Unified Storage view in your “Data Ocean”
5. Big Data and Analytics
6. Data-intensive Technical Computing
7. Spectrum Storage for AI
8. Selected Solutions
   - Industry Solutions
   - ISV Solutions and Offerings

Spectrum Scale Parallel Architecture

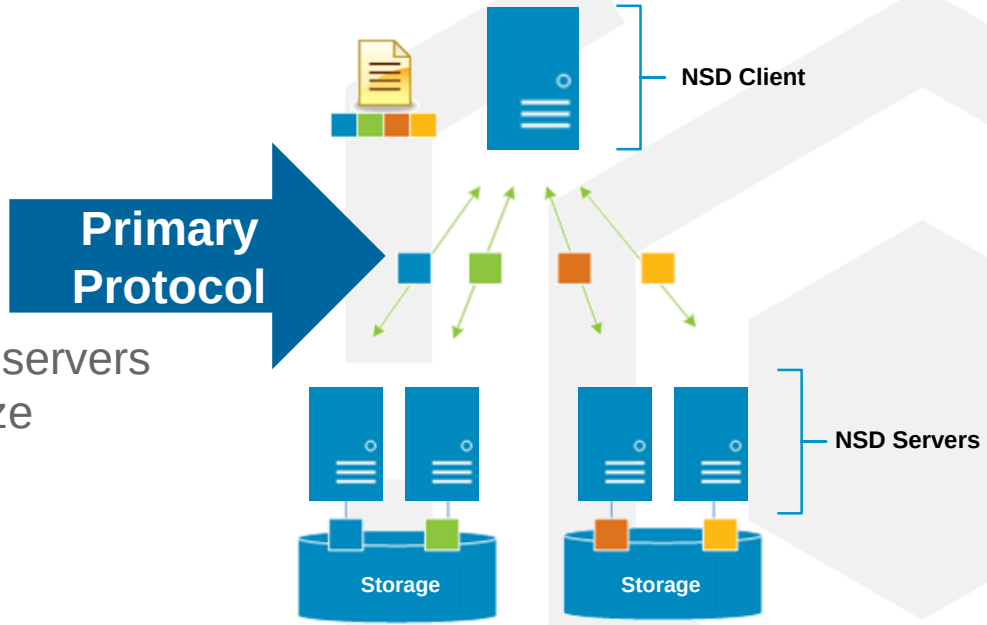
No Hot Spots

All NSD servers export to all clients in active-active mode

Spectrum Scale stripes files across NSD servers and NSDs in units of file-system block-size

File-system load spread evenly

Easy to scale file-system capacity and performance while keeping the architecture balanced



NSD Client does real-time parallel I/O
to all the NSD servers and storage volumes/NSDs

POLL: What Versions, What Benchmarks

3.X or 4.1.X please no?	nsdperf/gpfsperf
4.2.1/2	IOR/mdtest
4.2.3	iozone
5.0.1	fio
5.0.2	vdbench
OTHER?	OTHER?

New in IBM Spectrum Scale 5.0.3

Performance!

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, maxStatCache = 4 * maxFilesToCache, if < 10k

"*mmcachectl show*" can be used to verify if file inode is in the stat cache

Commands: ls -l and mdtest have shown improvement.

FileType	NumOpen Instances	NumDirect IO	Size (Total)	Cached (InPagePool)	Cached (InFileCache)
file	0	0	0	0	C
file	0	0	0	0	C
file	0	0	0	0	C

QOS improvements in large clusters

QoS node collects stats X time and sends to QoS manager node Y time

Report via **mmlsqos**

Large clusters => more communication, performance degradation

Set new defaults based number of mounts

Allow changes

stat-slot-time : QoS collects

stat-poll-interval : QoS -> QoS manager

Table 1. Default intervals for collecting and sending statistics

Number of nodes that have mounted the file system	Interval between collecting statistics, in milliseconds	Interval between sending statistics to the QoS manager, in seconds
< 32	1000	5
< 64	2000	10
< 128	3000	15
< 256	4000	20
< 512	6000	30
< 1024	8000	40
< 2048	10000	50
< 4096	12000	60
< 8192	12000	60
< 16384	12000	60
16384 or more	24000	120

***mmchqos** Device --enable [--stat-poll-interval Seconds] [--stat-slot-time Milliseconds]*

IBM Spectrum Scale 5.0.3

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
```

COMING in a PTF near you – Upgrade gpfs.gplbin without downtime

mmgetstate

mmshutdown

mmgetstate

  make sure it is down or will fail

rpm –ivh gpfs.gplbin.XYZ.rpm

mmstartup

mmgetstate

Set ENV variable

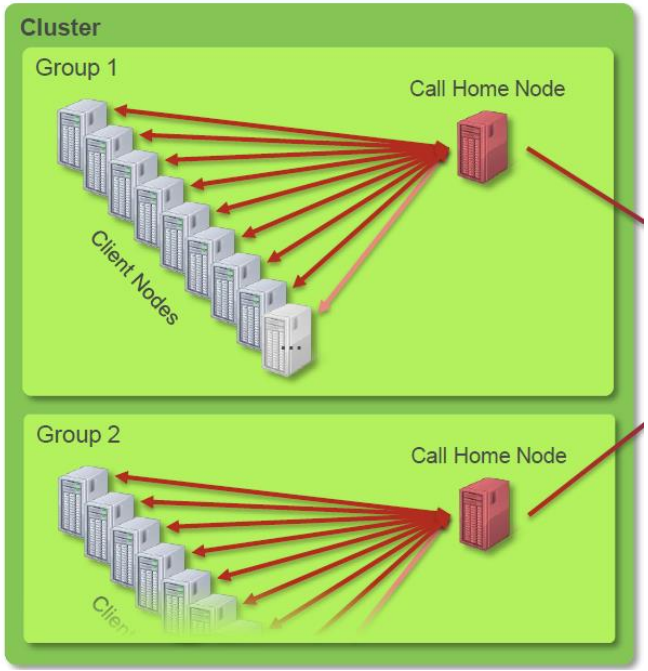
  Export MM_INSTALL_ONLY=1

rpm –ivh gpfs.gplbin.XYZ.rpm

mmgetstate

Supported in a future PTF

Proactive Services - callhome



less perl dependencies!

faster data collection

basic snaps, License, OS, CPU arch, memory, config

mmhealth reports if ECuRep connection is down

event based uploads automatically processed

Network enhancements

mmnetverify

mmhealth checks
availability,
port state,
link state
but not connectivity

check RDMA connectivity (between nodes)

nsdperf **for “stress” testing**

Attempt to reconnect socket before expel

- Only on Linux

mmchconfig *proactiveReconnect=yes*

Raise network reconnects to **mmhealth**

```
# mmhealth node eventlog
Timestamp       Event Name       Severity        Details
2019-TIME TZ    reconnect_start  WARNING         Attempting to …
2019-TIME TZ    reconnect_done   INFO            Reconnected to …
```

```
# mmhealth node eventlog
Timestamp       Event Name         Severity      Details
2019-TIME TZ    reconnect_start    WARNING       Attempting to …
2019-TIME TZ    reconnect_failed   ERROR         Reconnect … failed
2019-TIME TZ    reconnect_aborted  INFO          Reconnect … aborted
```

Spectrum Scale misc.

Deprecate primary and backup server
mm[cr,ch]cluster

Grab security/encryption and/or network related data for **gpfs.snap**

Display certificate expiration
mmkeyserv *server show*

designate license with **mmaddnode**
mmaddnode *–N name:manager:name-a:server --accept*

Log better CCR messages
6027-4200 [E] Maximum number of retries reached
6027-4201 [B] Version mismatch on conditional put
6027-4202 [B] Version match on conditional get
6027-4203 [B] Invalid version on put
6027-4204 [E] Not enough CCR quorum nodes available
6027-4205 [E] ccr.nodes file missing or empty
6027-4206 [E] CCR is already initialized
6027-4207 [E] Unable to reach any quorum node (Check your firewall or network settings)

CCR recovery options

*1) # **mmsdrrestore** -p <QNODE_WITH_GOOD_CCR_COPY>*
*2) # create **mmsdrbackup** && **mmsdrrestore** -F /x/f –a*
3) # mmsdrrestore --ccr-repair

There is a dry-run mode
GPFS must be down

Spectrum Scale – mmfsck and mmbackup

Issues with block allocation map corruption, have to do offfline **mmfsck**

Added capability to do online

Cannot detect and repair non-structural corruptions (bad allocation map bits – marked free but in use)

Ability to *--use-stale-replica* if no chance of recovery disks with higher *failure-config-version*

Today **mmbackup** expires, selects new and gets incremental changes

Allow granular tuning when exclusive lock issues due to massive incremental changes

mmbackup {Device | Directory} [-t {full | incremental}] [-N {Node[,Node...] | NodeFile | NodeClass}] [-g GlobalWorkDirectory] [-s LocalWorkDirectory] [-S SnapshotName] [-f] [-q] [-v] [-d] [-a IscanThreads] [-n DirThreadLevel] [-m ExecThreads | [[--expire-threads ExpireThreads] [--backup-threads BackupThreads | **[[--selective-backup-threads selBackupThreads] [--incremental-backup-threads incBackupThreads]]]]]** [-B MaxFiles | [[--max-backup-count MaxBackupCount | **[[--max-incremental-backup-count MaxIncBackupCount] [--max-selective-backup-count MaxSelBackupCount]]]** [--max-expire-count MaxExpireCount]]] [--max-backup-size MaxBackupSize] [--qos QosClass] [--quote | --noquote] [--rebuild] [--scope {filesystem | inodespace}] [--backup-migrated | --skip-migrated] [--tsm-servers TSMServer[,TSMServer...]] [--tsm-errorlog TSMErrorLogFile] [-L n] [-P PolicyFile]

5.0.3 Spectrum Scale GUI –What's new

No default admin user

*# **/usr/lpp/mmfs/gui/cli/mkuser** ADMINUSER -g SecurityAdmin*

Configure LDAP for GUI USER from GUI

- Can test connectivity

Manage Quotas

- user, group, fileset
- capacity and inode quotas
- any other setting
  - (scope, grace time)
- Email daily quota reports

NFS client management

Support pseudo paths NFSv4

Better monitoring for NFS exports and SMB shares

Manage NFS/SMB authentication!

Migrate policy to external pool with best practice excludes

- .ltfsee, ,snapshots, .mmbackup
- small files
- recently access files
- migrated files

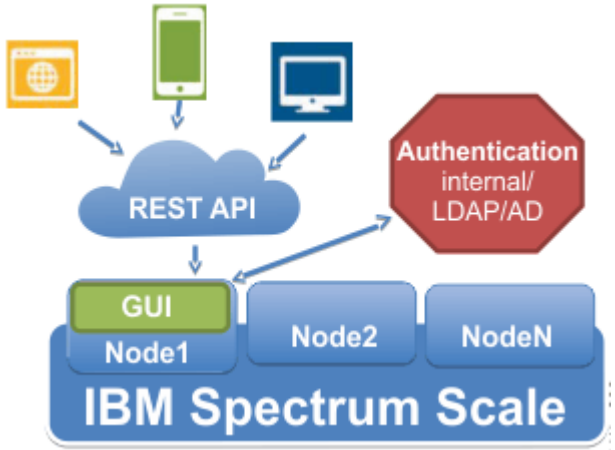
GUI and the REST API

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://<yourguihost>/ibm/api/explorer




Liberty REST APIs

API Discovery : APIs available from the API Discovery feature

Spectrum Scale REST API v1 : DEPRECATED

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

GET /scalemgmt/v2/access
POST /scalemgmt/v2/access
GET /scalemgmt/v2/access/status
GET /scalemgmt/v2/ces/addresses
GET /scalemgmt/v2/ces/addresses/{cesAddress}
GET /scalemgmt/v2/ces/services
GET /scalemgmt/v2/ces/services/{service}
GET /scalemgmt/v2/cluster
GET /scalemgmt/v2/config
GET /scalemgmt/v2/filesystems
GET /scalemgmt/v2/filesystems/{filesystemName}
GET /scalemgmt/v2/filesystems/{filesystemName}/acl/{path}
PUT /scalemgmt/v2/filesystems/{filesystemName}/acl/{path}
GET /scalemgmt/v2/filesystems/{filesystemName}/afm/state
GET /scalemgmt/v2/filesystems/{filesystemName}/disks
GET /scalemgmt/v2/filesystems/{filesystemName}/disks/{diskName}
GET /scalemgmt/v2/filesystems/{filesystemName}/filesets
POST /scalemgmt/v2/filesystems/{filesystemName}/filesets
DELETE /scalemgmt/v2/filesystems/{filesystemName}/filesets/{filesetName}

REST API - Extra endpoints in 5.0.3

PUT – enable/disable, POST – set, GET - view

Quota management

PUT FSNAME/quotamangement

POST FSNAME/quotagracedefaults
GET FSNAME/quotagracedefaults

PUT FSNAME/quotadefaults
POST FSNAME/quotadefaults
GET FSNAME/quotadefaults

PUT FSNAME/filesets/FSNAME/quotadefaults
PUT FSNAME/filesets/FSNAME/quotadefaults
GET FSNAME/filesets/:all:quotadefaults

Filesystem

PUT FSNAME/mount
PUT FSNAME/unmount
GET FSNAME
Status, RO, RW…

Updates mmhealth

CES with
SUDO wrapper and SE-Linux

Colorized output

Thresholds monitor, which collector
*# **mmhealth** cluster show threshold -v*

Determine CES IP failover
*# **mmhealth** node eventlog | grep move_cesips*

ESS monitoring of
pdisk, fan speed, new enclosures

NVME monitoring

Watchfolder monitoring

If in doubt with node state
*# **mmhealth** node show --resync*

Install Toolkit 5.0.3 New Features

Recall install toolkit introduced in 4.1.1.0

GUI installtoolkit is being deprecated

Upgrade flow changes to minimize I/O disruptions

Use to change product editions

Mixed O/S support

Pre-checks what packages must be upgrade (and if missing dependency)
Post-checks ensure all upgrades successful

Product edition change path	Installation toolkit (cluster online or offline)	Manual node by node (cluster online)	Manual all nodes (cluster offline)
Standard Edition to Data Access Edition	Yes	Yes	No
Standard Edition to Data Management Edition	Yes	No	Yes
Standard Edition to Advanced Edition	Yes	No	Yes
Advanced Edition to Data Management Edition	Yes	Yes	No

 # cd /usr/lpp/mmfs/VERSION/installer
*# ./**spectrumscale** config populate*

IBM Spectrum Scale 5.0.3

Other Protocols

“mmuserauth” enhancements

***mmuserauth** service list* – updates for USER_NAME

report the user name that connects to the DC, not always Administrator

***mmuserauth** service check* – report which DC CES is connected to

userauth file check on node: NODENAME

…

NETLOGON connection: OK, connection to DC: *SERVERNAME*

***mmuserauth** service create* – detect if big clock skew with cluster and DC

*$ **mmuserauth** service create --data-access-method file --type ad --netbios-name NAME --servers SERVER --user-name Administrator --pwd-file fileauth.pwdfile --idmap-role master*

WARNING: Time difference between current node and domain controller is 538 seconds. It is greater than max allowed clock skew 300 seconds.

File authentication configuration completed successfully.

Samba update

SMB 3.1.1

Some work to improve DNS responses by DNS caching for winbind

If you have long VFS calls, check for ILM, backups, snapshots running

Log message if export runs and there is no fs mount

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
5.0.3	2Q19	4.9	RHEL 7.6 (bringing mutex fixes)

NFS Ganesha

Stats, stats, stats!

1. Every operation of NFSv3 and NFSv4
2. RPC queue statistics (receive and send queue)
3. Recall **5.0.2** could get data from FSAL (GPFS Layer)

ganesha_stats enhanced to support these features

Leverage when observe slowness for data access over NFS to inspect each layer

Object Release Overview

Spectrum Scale	Openstack
4.1.1	Kilo
4.2.1	Liberty
4.2.2	Mitaka
5.0.3	Pike

Spectrum Scale	swift3
4.1.1	1.7
4.2.0	1.8
4.2.1	1.10
5.0.3	**2.15.1**

Spectrum Scale Offerings on AWS

Marketplace Offering With BYOL (Sep 2018)
https://aws.amazon.com/marketplace/pp/B07DRLMG2W

Provides an AMI (boot image) with Spectrum Scale Data management edition installed on RHEL

Automated deployment

Targeted for HPC use on AWS

BYOL License Support (Bring Your Own License)
Customer still has to pay Amazon for AWS resources used and RHEL and other software they will consume.

Spectrum Scale version 5.0.2.1

AWS Quickstart (90 days, avail Sep 2017)

A cluster of 16 EC2 instances can be launched and configured with a shared filesystem mounted

on all nodes in less than an hour (& does not require any Spectrum Scale Admin Skills).

New in IBM Spectrum Scale 5.0.3

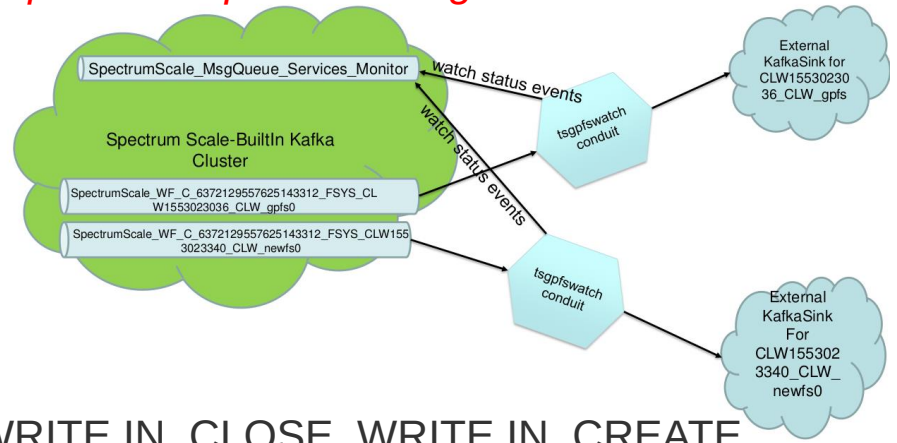
Security

Clustered Watch

Captures file system activity

- Generates an event notification for that activity
- Streams the notifications to topics within the message queue
- Events are consumed by a conduit
- Conduit sends these events to an external 'sink'

External sink should be a Kafka message queue setup and managed independently by the customer



Watch file operations across nodes in cluster

- **mmwatch**

Watch

- entire file system, fileset or inode space

Events: {IN_ACCESS,IN_ATTRIB,IN_CLOSE_NOWRITE,IN_CLOSE_WRITE,IN_CREATE, IN_DELETE,IN_MODIFY,IN_MOVED_FROM,IN_MOVED_TO, IN_MOVE_SELF,IN_OPEN}

New in IBM Spectrum Scale 5.0.3

Data Movement (Compression, AFM and TCT)

New File Compression Algorithms

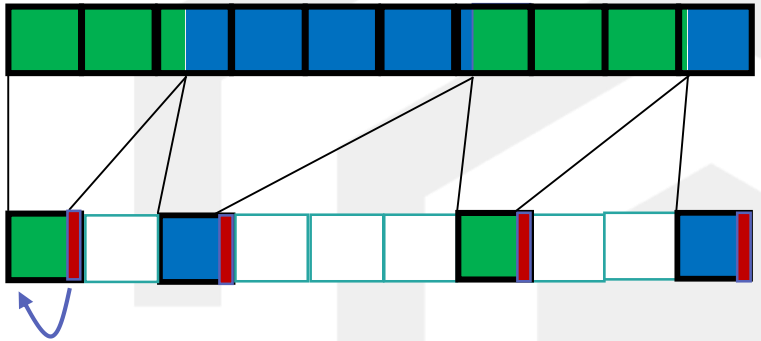
Genomics compression methods added in release 5.0.3 are:

- ***alphae***
- ***alphah***
- ***zfast***

Run via ILM policy or

mmchattr --compress …

Still have ***zlib*** and ***lz4***



Advanced File Management (AFM) enhancements

Kerberos V5 support in AFM remote mounts to secure NFS traffic.

afmEnableNFSSec

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**

Advanced File Management (AFM) enhancements

Resync Performance Enhancement

reduce delay in reading data under various directory hierarchies

Async Re-Validation to improve application performance during readdir/lookups

IW mode target

queue async lookups to gw node

mmchconfig *afmRefreshAsync=yes*

AFM DR tried with 100 filesets and 1 Billion files aggregate

Tech paper under review with sales/support team before release

Transparent Cloud Tiering enhancements

Support for Azure Cloud Storage

Shift to Amazon SDK for AWS-S3 and IBM COS

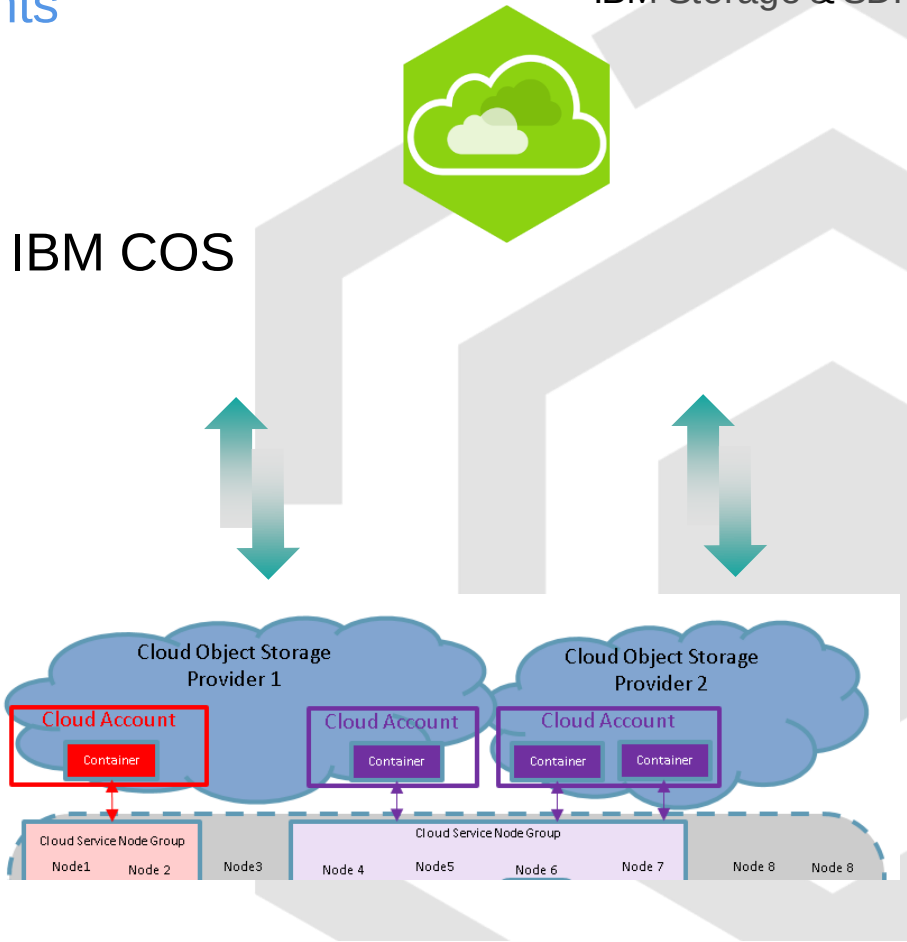
Automatic Container Spillover

Client Assist Recall

Simplified SOBAR backup-restore

zLinux support

Quota Support and more



Big Data and Analytics Enhancements

HDFS Transparency v3.1.0-1 GA (2019-Mar-29)

Spectrum Scale Certification with HortonWorks Data Platform (HDP) 3.X

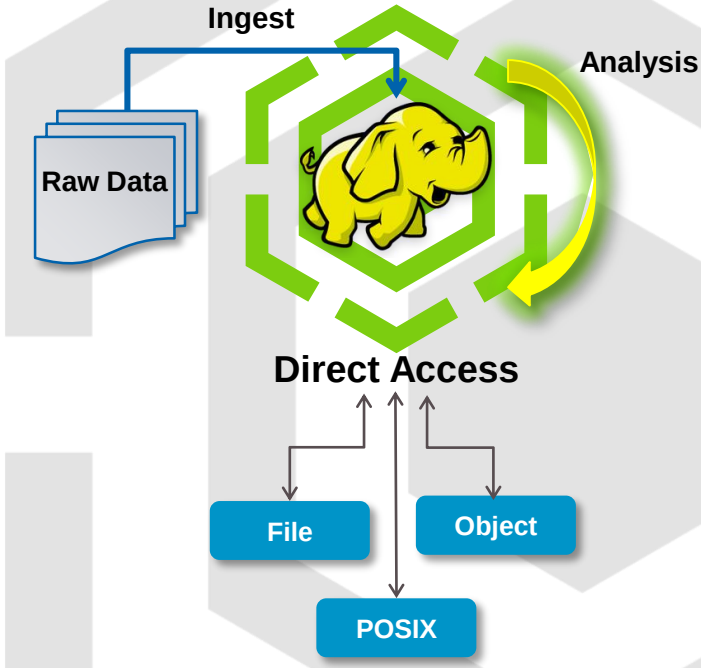
- Certified on both Power8 and x86 platforms
- Certified with Ambari 2.7 for rapid deployment

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



Thank You.

IBM Storage & SDI