What’s new in Spectrum Scale and the Elastic Storage System (ESS)?

September 20, 2022

Chris Maestas, Chief Architect,
Storage for Data and AI Solutions
cdmaestas@us.ibm.com

Disclaimer

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. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

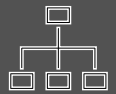
IBM Global Data Platform for Unstructured File & Object Data

Unstructured Data Services Framework

Applications and Workloads

Data Access Services

Data Caching and Core Services



Data Management Services

Data Security Services

IBM Storage / © 2022 IBM Corporation

Featured Updates

Data Access Services - GPU Direct Storage (GDS) on **write Tech Preview**, High Performance Object (HPO)

Data Caching and Core Services – ability to route over multiple network interfaces without bonding using **Multi-Rail Over TCP (MROT)**

Data Caching and Core Services - Enhanced scalability for independent filesets

Data Security Services –**Safe Guarded Copy (SGC)** support protect data in IBM Spectrum Scale file systems!

Data Security Services – **Remote Fileset Access Control (RFAC)** that allows restricted views of projects on remote clusters.

Shift to Quarterly Release Cadence

Spectrum Scale

Survey – tell me about upgrades?

Why?

- To address requests for quarterly updates to bring new features out more rapidly

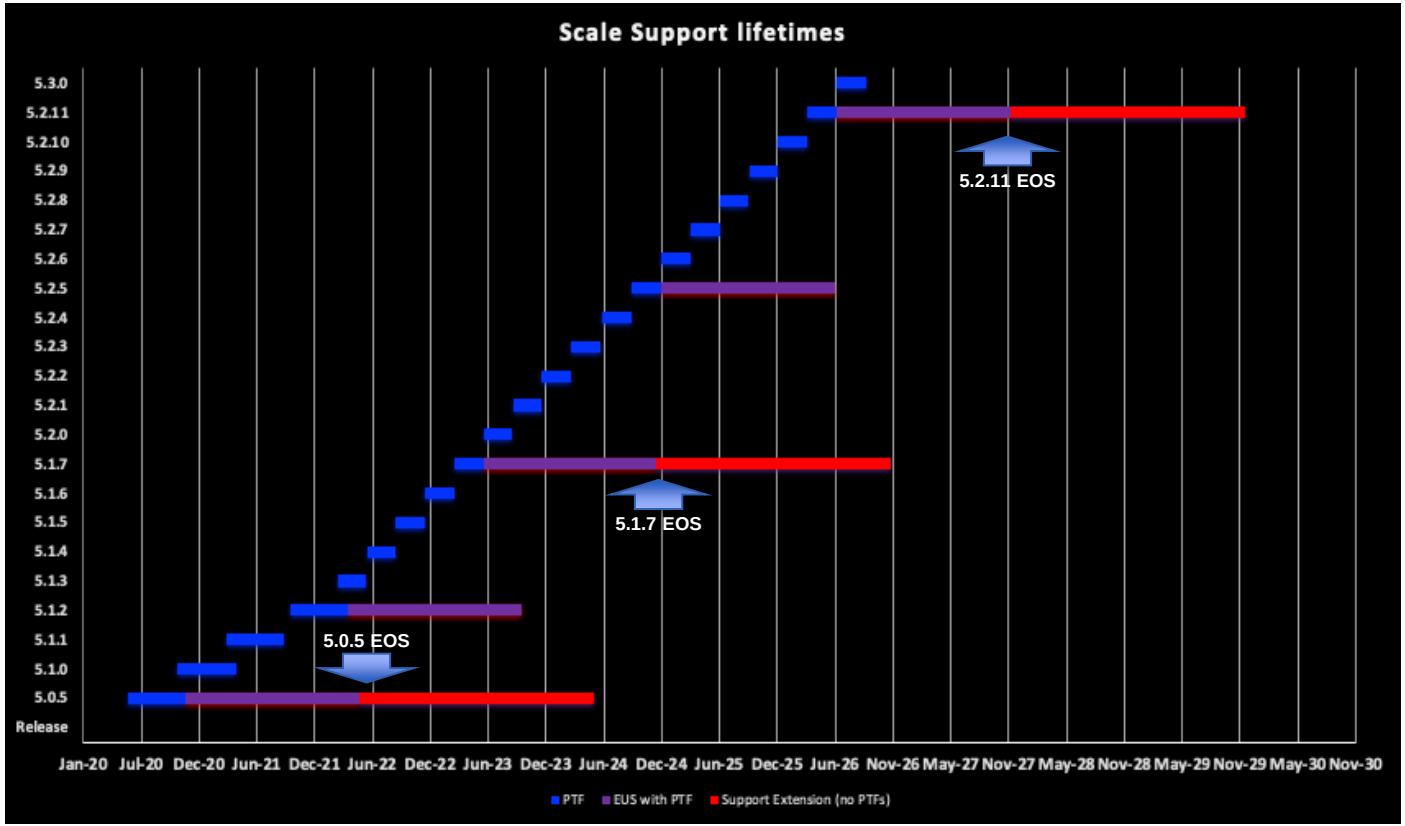
Maintain Extended Update Support concept

EUS with PTFs every 18 months

Extended support on last EUS within a release (example: V.R.x, 4.2.3, 5.1.2, 5.1.last)

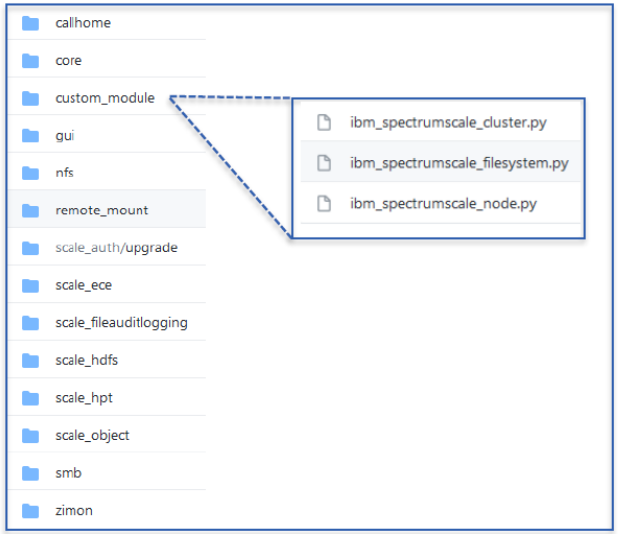
Increase the number of Modification levels with new function

Note: Version numbers and release timing are for example purposes to demonstrate the goal of EUS every 18 months and do **not** represent a commitment to deliver a specific version or on a specific timeline



Data Management Services – Ansible Toolkit

- Modified the command to enable upgrade workload prompt at a node level to allow administrators to stop and migrate workloads before a node is shut down for upgrade.
- Several optimizations in the install and upgrade path that is resulting in faster install and upgrades.
- Scalability improvements and OS currency support (RHEL 8.6, Ubuntu 22.04.x, SLES 15 SP4)
- Ansible collection support



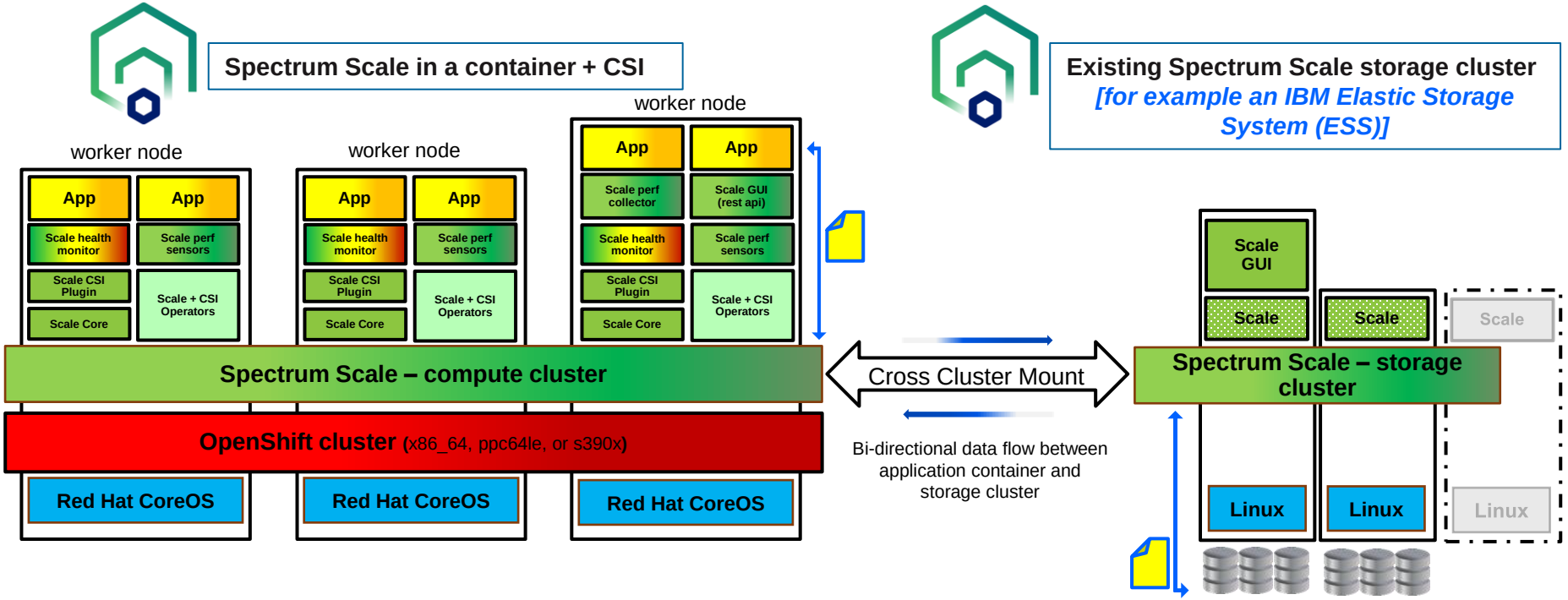
Spectrum Scale deployment is open sourced on Github

Ansible Playbooks:
https://github.com/IBM/ibm-spectrum-scale-install-infra

Bundle the CLI toolkit into packages but a user can deploy their own orchestration utilizing the eternal github playbooks.

Data Access Services – IBM Spectrum Scale Container Native Storage Access (CNSA)

Cluster Overview



Data Access Services – Container Native Storage Access

Improvements introduced in CNSA 5.1.4
https://www.ibm.com/docs/en/scalecontainernative?topic=overview-supported-features

Wider support to use the latest CNSA functionality.

- Support for upgrading IBM Spectrum Scale Container Native Storage Access (CNSA) from v5.1.4 to 5.1.4.1
- Planned support for RedHat OpenShift Container Platform 4.11
- CNSA images now hosted on the entitled IBM Cloud Container Registry.
- Automated deployment of the CSI driver
- Support for storage cluster encryption
- Rolling upgrade of IBM Spectrum Scale is supported
- Support for a limited set of IBM Spectrum Scale configuration settings to be set directly
- Grafana support
- Support for X86, Power and Z.
- Direct storage attachment on x88, power and Z
- Automatic quorum selection is Kubernetes topology aware.

Data Access Services – Container Storage Interface

Improvements introduced in CSI 2.5

Upgrades for OpenShift, Kubernetes and Ansible as well as improved functionality that support simpler administration and configuration.

- Planned support for Red Hat OpenShift 4.11 and Kubernetes 1.23.
- Upgraded CSI specification from 1.3.0 to 1.5.0
- Added support for Consistency Group (**version**=*2*)
- Support to enable the compression for persistent volumes
- Support to enable the tiering for persistent volumes
- Increased attacher statefulset's replica count to two for high availability of attached volumes
- Upgraded Kubernetes CSI sidecar containers
- Migrated from CSI Ansible® operator to CSI Go operator

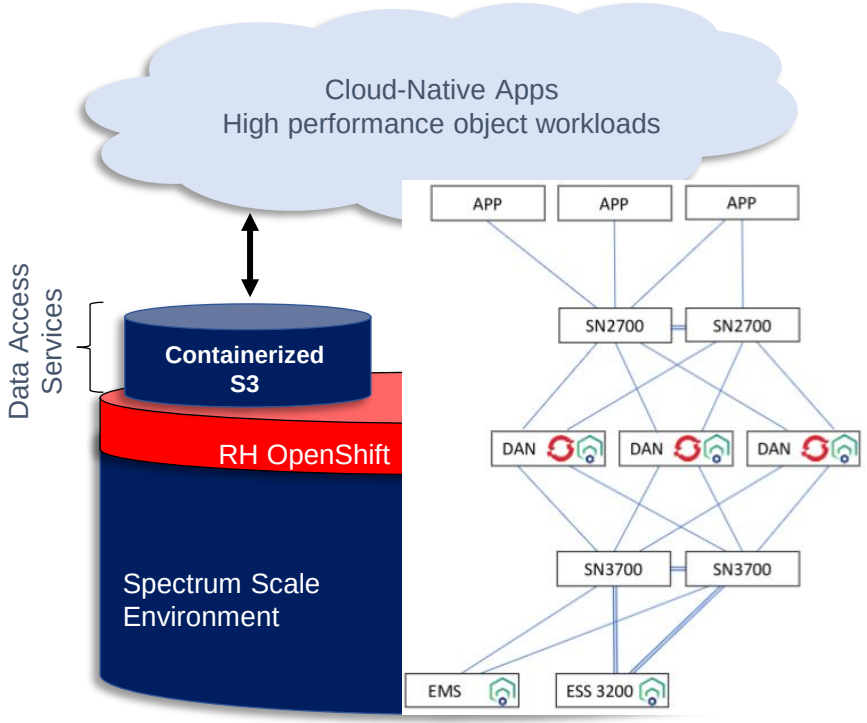
Data Access Services – S3 object access

Containerized S3 object access integrated within Spectrum Scale delivering high performance object for AI and analytics workloads

Customer Requirements & DAS S3 Dependencies:

- Spectrum Scale 5.1.3.1: DAE, DME, ESS for DAE, ESS for DME, ECE (future)
- OpenShift 4.9.31 → dedicated OpenShift Cluster
- CNSA 5.1.3.1 / CSI 2.5.1
- ESS models at GA, followed by any storage supported by CNSA

Performance: 60 GB/s w/ 3 DAN (Data Access) nodes on vanilla ethernet and scales linearly



Data Access Services – GPU Direct Storage (GDS)

Spectrum Scale

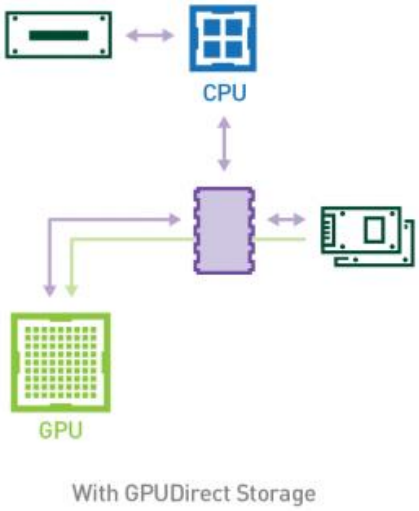
GPU Direct Storage Write – Tech Preview in Spectrum Scale 5.1.5!

Understand how to get GDS and the requirements.

Spectrum Scale Knowledge Center:
https://www.ibm.com/docs/en/spectrum-scale/5.1.5?topic=summary-changes
https://www.ibm.com/docs/en/spectrum-scale/5.1.5?topic=architecture-gpudirect-stora
support-spectrum-scale

Nvidia GDS Documentation:
https://docs.nvidia.com/gpudirect-storage/index.html
https://developer.nvidia.com/gpudirect-storage

For help getting started: scale@us.ibm.com
* For details on supported versions, refer to the Spectrum Scale FAQ



Hardware

- x86 client with GPU that supports GDS (refer to NVIDIA doc)

Storage server: traditional NSD and ESS

- RDMA capable fabric:
  - NIC: Mellanox CX5 and CX6
  - Switch: IB and Ethernet (RoCE)

Spectrum Scale:

- 5.1.2: Read, IB
- 5.1.3: RoCE, GDS write in compat mode)
- 5.1.5: accelerated GDS write

Client O/S:

- RHEL 8.6
- Ubuntu 20.04

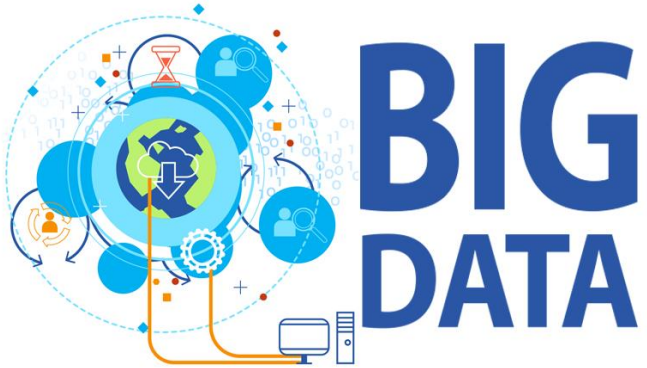
MOFED

- Mellanox OFED stack
- Current recommendation: MLNX_OFED_LINUX-5.4-1.0.3.0, 5.6-2.0.9.0

CUDA (client only)

- CUDA 11.4.2, 11.5.1, 11.6.2, 11.7
- CUDA 11.8 available early 4Q (required for accelerated GDS write)
- CUDA C/C++ program
- NVIDIA DALI (data loading library)

Data Access Services – Big Data & Analytics and Traditional File Services



Support and Currency:

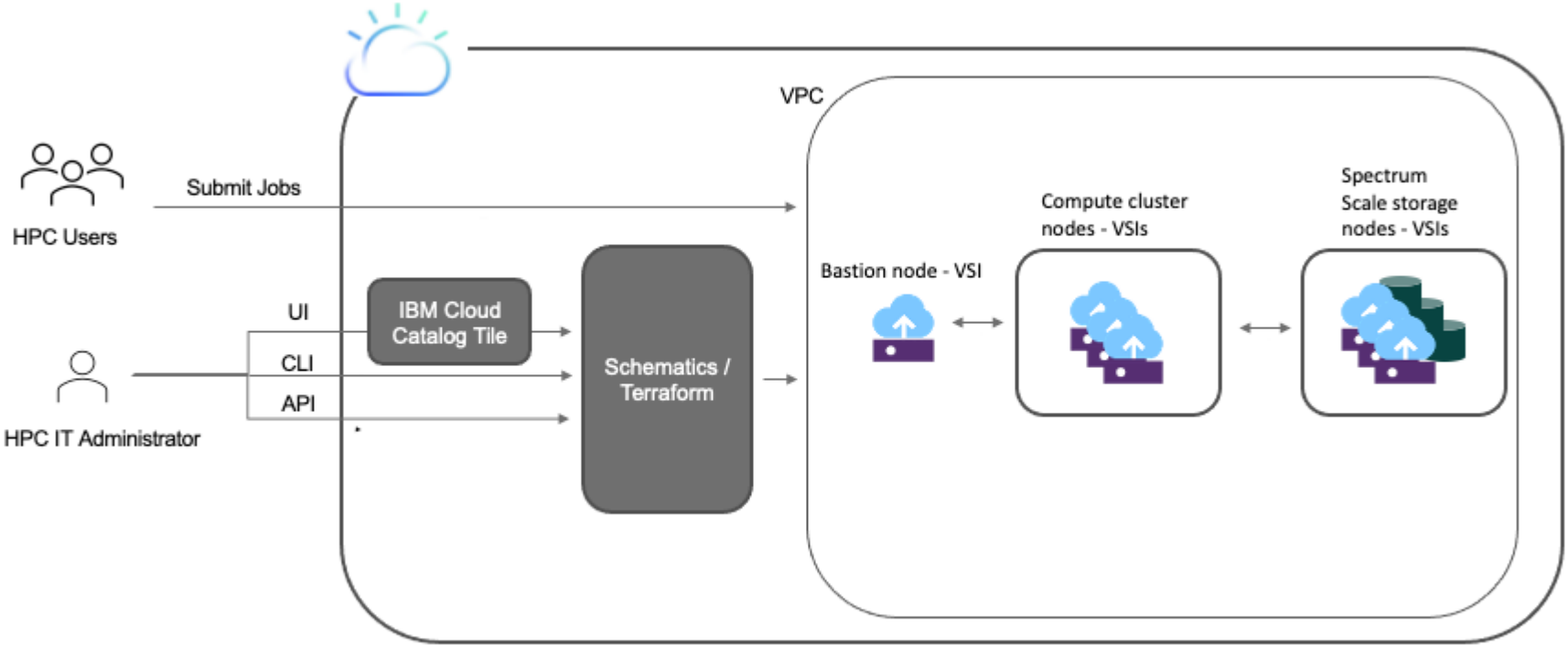
- Cloudera Data Platform (CDP) Private Cloud Base is certified with IBM Spectrum Scale on x86_64 and ppc64le since December 2020.
- Cloudera Hortonworks Data Platform (HDP) 3 and HDFS Transparency 3.1.0 end of service on December 31st, 2021.
- Opensource Hadoop 3.2.2
- Includes HDFS Transparency 3.1.1-10, HDFS Transparency 3.2.2-1 and HDFS Transparency 3.3.0-2.
- NFS-Ganesha support for 3.5 code base

Improved performance:

- Improved memory efficiency for HDFS Transparency NameNode.
- Optimized parallelism for DataNode request processing via delete, du and list configuration options.
- NFS - Added new config parameter (**readdir_res_size**) to improve readdir performance and other critical ganesha defects
- SMB – introduced **wide links** parameter to control following links

Data Access Services – Spectrum Scale on IBM Cloud!

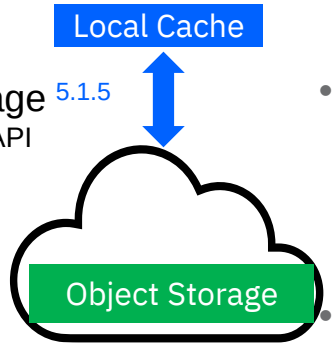
Similar to AWS experience - https://www.ibm.com/cloud/hpc



Data Caching Core Services – Active File Management (AFM)

Continued testing of other Cloud Object Storage environments

- IBM Cloud Object Storage 5.1.0
- Amazon S3 5.1.0
- Microsoft Azure Blob storage using S3 Gateway 5.1.3
- Minio 5.1.3
- Google Cloud Platform 5.1.4
- Seagate Lyve Cloud Object Storage 5.1.5
  Lyve cloud APIs are almost similar with S3 API



- IPV6 support! 5.1.5
- Support of creating and upload objects for empty directories in AFM to cloud object storage – 5.1.4
- Support of marking files and directories as local in AFM to cloud object storage fileset 5.1.3

```
#mmafmctl fs setlocal -j AFMtoCOS --path /ibm0/fs/AFMtoCOS/file1
```

- Support of adding user defined prefix in AFM to cloud object storage fileset. 5.1.4

```
#mmafmcosconfig fs1 afmbktprefix1 --endpoint https://region@endpoint --object-fs \
--xattr--prefix dir1  --bucket bkt1 --acls--mode sw
```

- *Manual Update (MU)* mode to support manual replication of files using a file list or ILM 5.1.3

Data Caching and Core Services – Spectrum Scale Core Improvements

Spectrum Scale

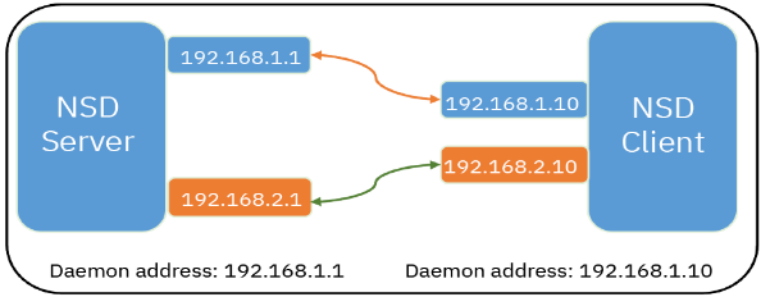
Network flexibility with Multi-Rail over TCP (MROT) and Multiple Connections over TCP (MCOT)

MROT - Concurrent use of multiple physical network interfaces without requiring bonding configuration

- Use **mmchconfig** command and the *subnets* attribute to add more IP addresses for daemon communication
- **maxTcpConnsPerNodeConn** controls the total number of TCP connections between a pair of node (valid values = 1-16, default = 2)
  - Smaller values may be needed with large clusters!
- Both nodes in each connecting pair need to be running at Spectrum Scale **5.1.5** and works with remote cluster mounts
- MCOT/MROT still is TCP/IP!
  - communications still go through kernel stack
    - Results: reduced bandwidth and higher latencies vs RDMA
    - But: full storage bandwidth can be achieved with fewer clients
    - MROT provides High Availability (HA) by failing over from one network interface to another

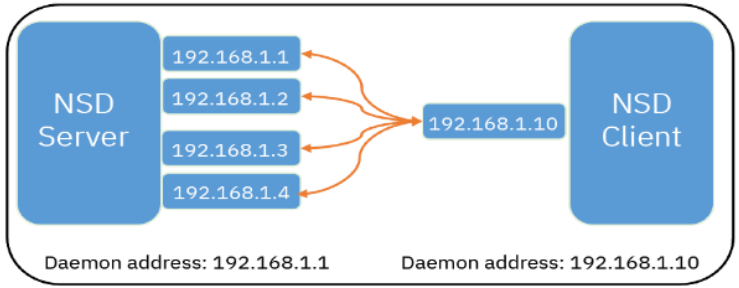
Configuring N:N connection model

Figure 1. Configuring the N:N connection model



Configuring M*N connection model

*Figure 2. Configuring the M*N connection model*



Data Caching and Core Services - Spectrum Scale Core Improvements

Spectrum Scale

Features that allow you to improve your resource performance.

- Allow **mmfsd** to dedicate specific TCP connections exclusively for ‘small message’ and ‘large message’ use.

  For example, a commonly used command to watch for changes generates lots of small messages for metadata:

  - # watch –n 5 “ls –ltr /fs1/lots_o_files_dir/”

- preferDesignatedMnode parameter – control metanode placement on a manager node, which is usually the same node as token server for that file/

- New workload solutions
  - gpfsFineGrainReadSharing (FGRS)

    optimizes performance of applications which run on multiple nodes where tasks issue small strided reads that are less than a full block
  - gpfsFineGrainWriteSharing (FGWS) hint

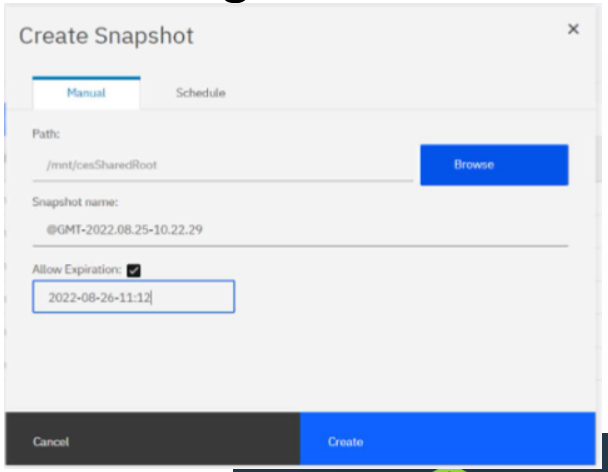
    performance of non-overlapping small strided writes to a shared file from a parallel application can now be optimized

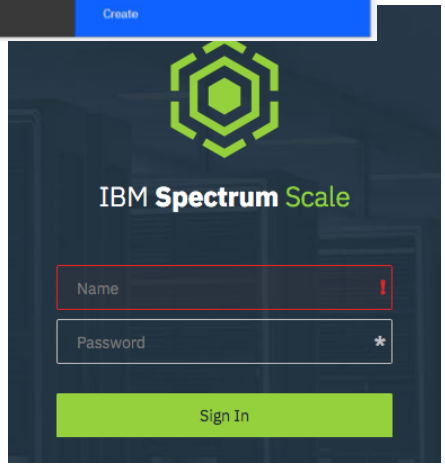
Data Management Services - GUI/API Changes

Spectrum Scale

Administration and reliability

Simpler management.





- Create Safe Guarded Snapshots (SGC)!
- Support Data Access Services (DAS) operations for High Performance Object (HPO)
- Updates to cache tables on AFM management pages
- Ensure High Availability for GUI/REST API
  - Replay logged jobs if failure occurs

Data Management Services – Monitoring, Availability & Proactive Services (MAPS) Updates

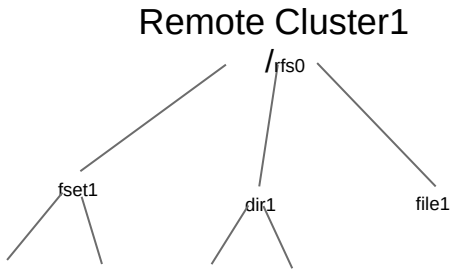
System Health & Monitoring

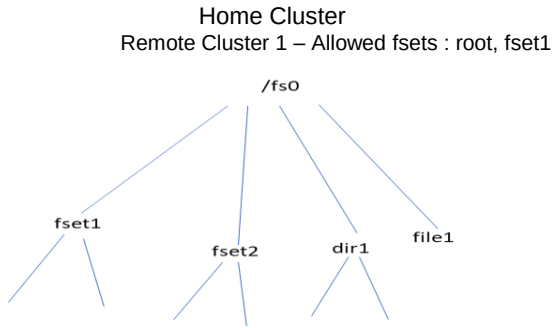
Enhanced awareness on the status of your system components

- Spot node troubles faster with: **mmdiag --network** by looking for abnormal pending RPC messages
- Enhanced stretch cluster monitoring via a new **STRETCHCLUSTER** component
- Improve **mmsysmoncontrol** starting conditions to check for invalid conditions and report them to the console
- Monitor AFM memory queue alerts in **mmhealth**.

Data Security Services
Remote Fileset Access Control (RFAC)

- No changes to CLI used for configuring remote mounts on remote cluster (Remote cluster is unaware of RFAC being enforced by home cluster)
- New syntax can be used to allow access to only a subset of filesets
- "root" fileset must be specified as one of the allowed filesets, and can't be removed from the list later.
- "grant" and "deny" commands can be used multiple times to edit the list of allowed filesets.
- if a child fileset is allowed, parent filesets should be allowed too for child fileset to be accessible.





Data Security – Security and Resiliency

Scale on Z Systems and encryption support for other key management software

- **Updates to Erasure Code Edtion (ECE) guidance and usage for Spectrum Scale in Linux on Z**
- IBM z16 support
- z/OS NFS client support

- Support CipherTrust Manager 2.8 for encryption

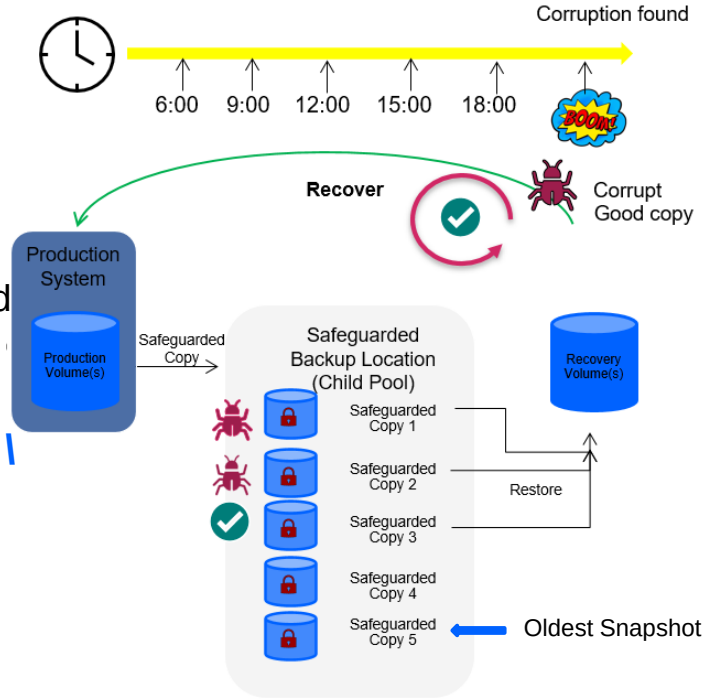
Data Security – Spectrum Scale Core Improvements

Spectrum Scale

Immutable snapshots

quickly create cyber-resilient point-in-time copies of the file system data and prevent this copy from being deleted through user errors, malicious actions, or ransomware attacks.

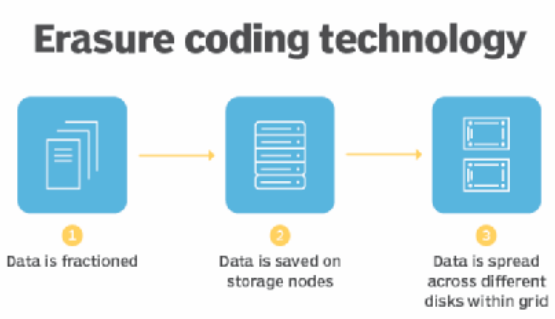
- Safe Guarded Copy (SGC) is really just an immutable snapshot of the data
  - since it remains online it also requires some degree of retention to prevent deletion via:
    1) user errors, 2) malicious actions, or 3) ransomware attacks
  - Expiration time has been introduced to snapshots and snapshots cannot be deleted until expiration time has elapsed
- Option for
  **mmcrsnapshot <device> <snapshotName> [-j fileset] **
  [--expiration-time YYYY-MM-DD-HH:MM[:SS]]
- Spectrum Scale GUI able to schedule SGC periodically



Data Security – Resiliency – Spectrum Scale Erasure Code Edition Changes

Spectrum Scale

- Support RoCE on a lossless network.
- 3-node ECE deployment
  - Minimal 3 to maximal 32 servers per RG
  - Support GNR 3- or 4-way mirroring but not 4+2p, 4+3p, 8+2p or 8+3p
- Support on background reclaim
  - User friendly automatic free space reclaim with trimming, instead of manually reclaim **customer test before using it in production**
- KVM virtio disk support
  - Start support from Alibaba cloud
  - Technically support other virtual environments but check with IBM first via RPQ request
- Dell PERC SAS Adapter Support
  - Dell RAID controller part: 12Gb/s PowerEdge RAID Controller: PERC H730P Mini, PERC H745 Front SAS, and PERC H755 Front SAS, managed by PercCLI utility.
  - Need RPQ to work with IBM to certify other types of adapters before production



Stabilized Features

https://www.ibm.com/docs/en/STXKQY/pdf/scale_deprecated_features.pdf

Category	Stabilized functionality	Recommended Action	Since Version
cNFS	All	IBM®'s strategic path is to invest in User Space solutions for NFS support of Scale workloads. Once User Space performance and function are considered to be sufficient to replace cNFS, anticipate that the support for cNFS is deprecated.	5.0.5

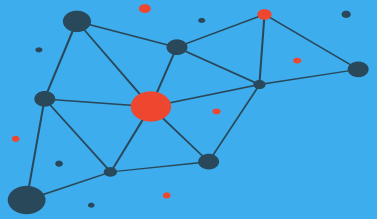
Deprecated Features

Spectrum Scale

Category	Deprecated functionality	Recommended Action	Since Version
Block size	The --metadata-block-size option of mmcrfs command is deprecated. This option is used for defining metadata blocks to a different size than the data blocks.	Only a single definition for the number of subblocks per block exists per file system. Selecting a smaller metadata block size has the unintended side effect of increasing the subblock size for data blocks. Although it is supported to set metadata blocks to a different size than data blocks by using the --metadatablock-size parameter, it is not recommended to use that option. This option is currently being deprecated and it will be removed in a future release. For more information, see the topic mmcrfs command in the IBM Spectrum Scale: Command and Programming Reference.	5.1.2
TCT	All	TCT can continue to be used for existing purposes. There are no plans to extend its purpose to more use cases.	5.0.5
FPO	All	FPO and SNC remain available. However, it is recommended to limit the size of deployments to 32 nodes. There are no plans for significant new functionality in FPO nor increases in scalability. The strategic direction for storage using internal drives and storage rich servers is IBM Spectrum Scale Erasure Code Edition (ECE)	5.0.5

Log your IDEA!

https://ibm-sys-storage.ideas.ibm.com/ideas



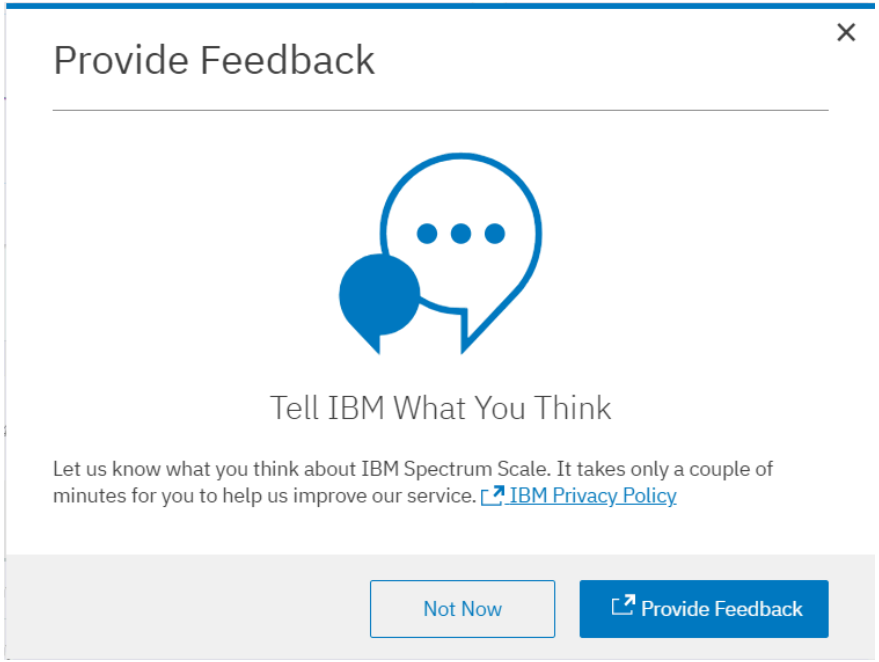
Check https://www.spectrumscaleug.org/experttalks for charts, notes and upcoming talks

- Past talks:
  - 001: What is new in Spectrum Scale 5.0.5?
  - 002: Best practices for building a stretched cluster
  - 003: Strategy update
  - 004: Update on performance enhancements in Spectrum Scale (file create, MMAP, direct IO, ESS 5000)
  - 005: Update on functional enhancements in Spectrum Scale (inode management, vCPU scaling, NUMA considerations)
  - 006: Persistent Storage for Kubernetes and OpenShift environments
  - 007: Manage the lifecycle of your files using the policy engine
  - 008: Multi-node scaling of AI workloads using Nvidia DGX, OpenShift and Spectrum Scale
  - 009: Continental: Deep Thought – An AI Project for Autonomous Driving Development
  - 010: Data Accelerator for Analytics and AI (DAAA)
  - 011: What is new in Spectrum Scale 5.1.0?
  - 012: Lenovo - Spectrum Scale and NVMe Storage
  - 013:Event driven data management and security using Spectrum Scale Clustered Watch Folder and File Audit Logging
  - 014: What is new in Spectrum Scale 5.1.1?
  - 015: IBM Spectrum Scale Container Native Storage Access

Thank you!

Please help us to improve Spectrum Scale with your feedback

- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply



Spectrum Scale User Group

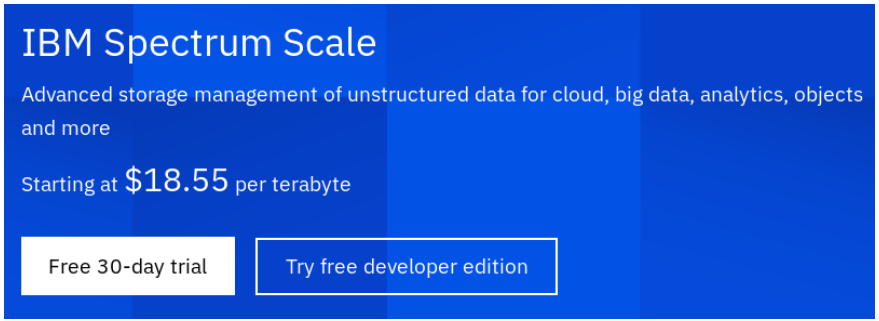
The Spectrum Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Spectrum Scale.

The format of the group is as a web community with events held during the year, hosted by our members or by IBM.

See our web page for upcoming events and presentations of past events. Join our conversation via mail and Slack.

www.spectrumscaleug.org

Spectrum Scale Developer Edition!



Fully functional!

- Based on first PTF of a release
- Derived from **Data Management Edition (DME)**
- Limited to 12 TBs:
  enough for a small test cluster
- Available from the Scale ”try and buy” page on ibm.com

Free for non-production use, e.g. test, learning, upgrade prep…

- If you have to ask, it’s probably not permitted

Not formally supported

Spectrum Scale on GitHub!
https://github.com/IBM/SpectrumScaleTools

- IBM Spectrum Scale Bridge for Grafana
- IBM Spectrum Scale cloud install
- IBM Spectrum Scale Container Storage Interface driver
- IBM Spectrum Scale install infra
- IBM Spectrum Scale Security Posture
- Oracle Cloud Infrastructure IBM Spectrum Scale terraform template
- SpectrumScale_ECE_CAPACITY_ESTIMATOR
- SpectrumScale_ECE_OS_OVERVIEW
- SpectrumScale_ECE_OS_READINESS
- SpectrumScale_ECE_STORAGE_READINESS
- SpectrumScale_ECE_tuned_profile
- SpectrumScale_NETWORK_READINESS

Find open source tools that are related with IBM Spectrum Scale.

Unless stated otherwise, the tools compiled in this list come with no warranty of any kind from IBM.

Check out the FAQ!

https://www.ibm.com/support/knowledgecenter/en/STXKQY/gpfsclustersfaq.html
https://www.ibm.com/support/knowledgecenter/STXKQY/gpfsclustersfaq.pdf?view=kc
https://www.ibm.com/support/knowledgecenter/SSYSP8/gnrfaq.html

HTML or PDF

Spectrum Scale version compatibility with OS or kernels

Updated regularly!