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

London - June 30th, 2022

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

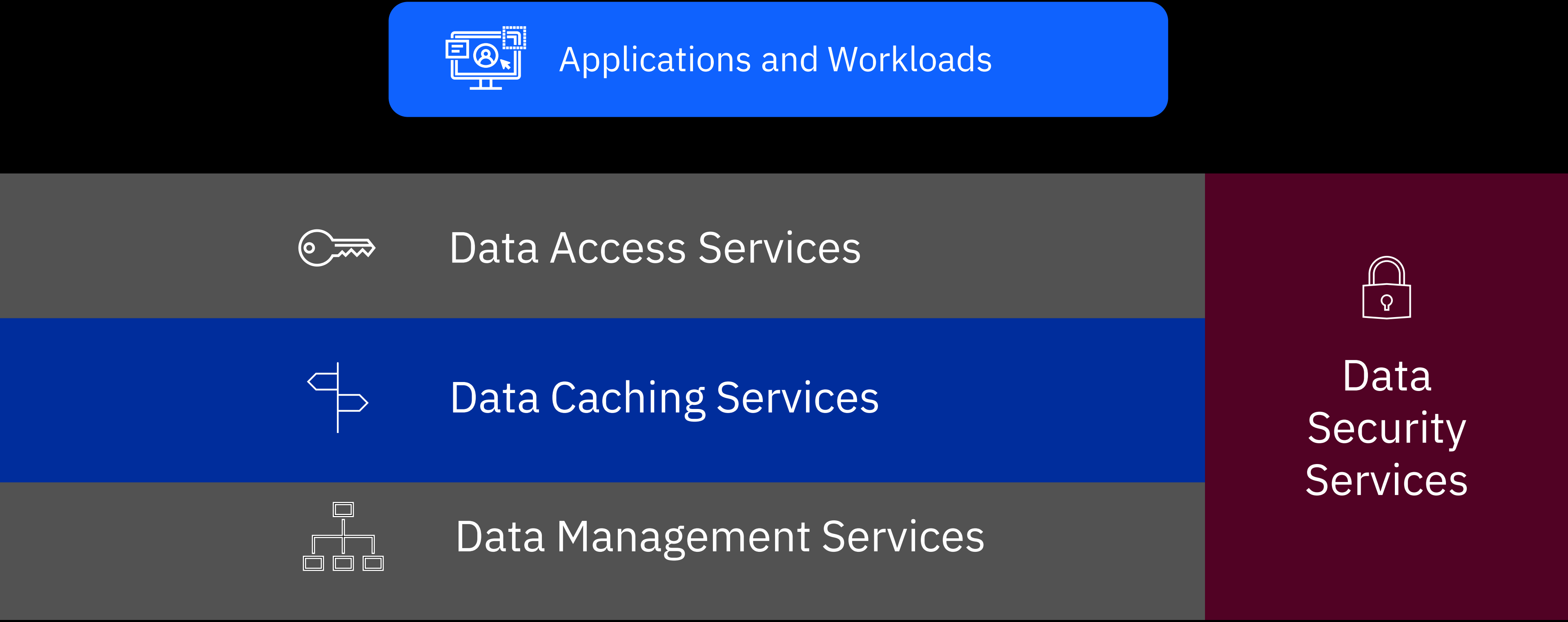
Disclaimer

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 Services

Data Management Services

Data Security Services

Featured Updates

Data Access Services - GPU Direct Storage (GDS) on **RoCE** environments, High Performance Object (HPO)

Data Caching Services – expanded caching support for **Azure and Google** clouds with more control

Data Management Services - Enhanced scalability for independent filesets (1000 -> 3000)

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?

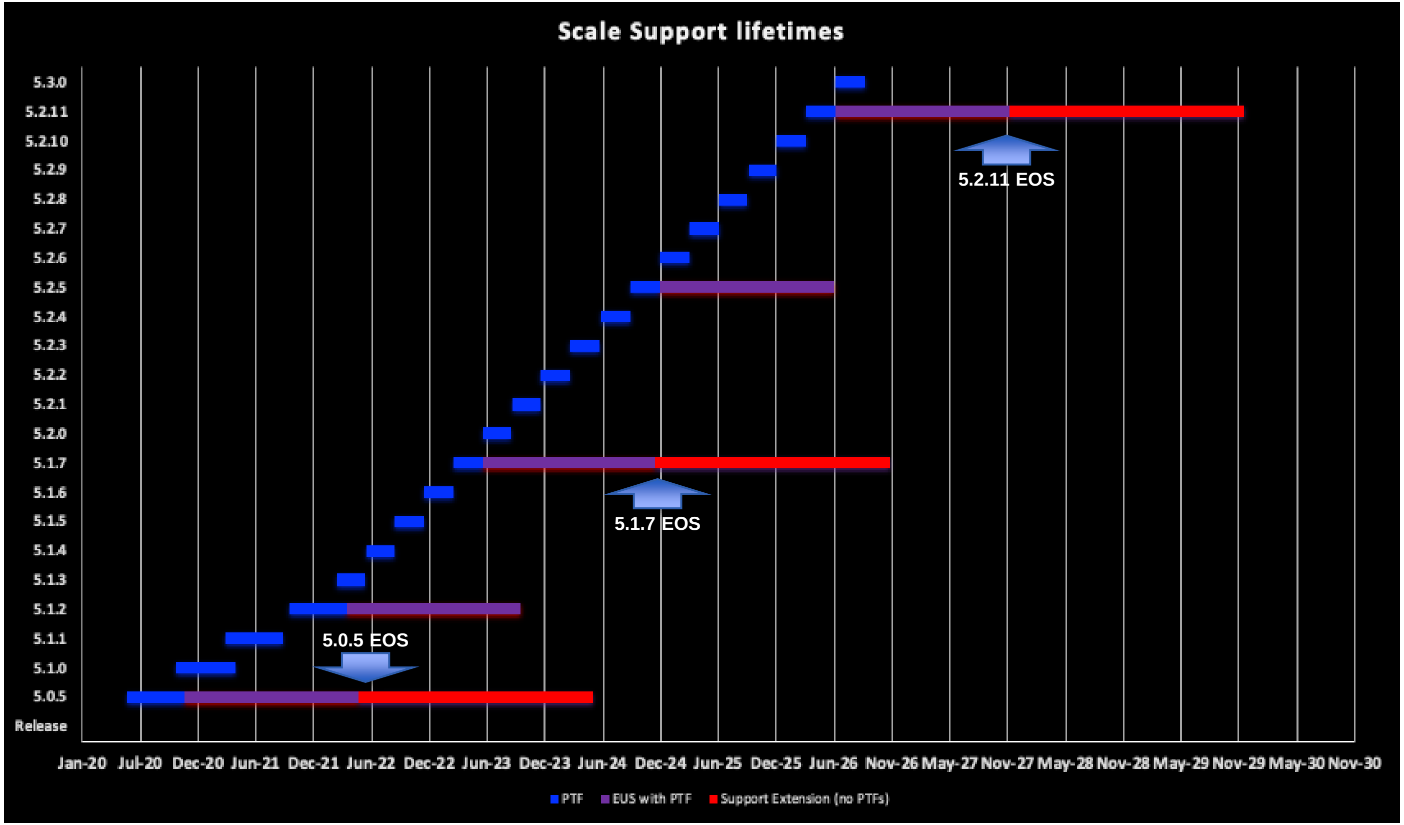
- To address requests for quarterly updates to bring new features out more rapidly

Maintain Extended Update Support concept

EUS with PTFs every 18 months

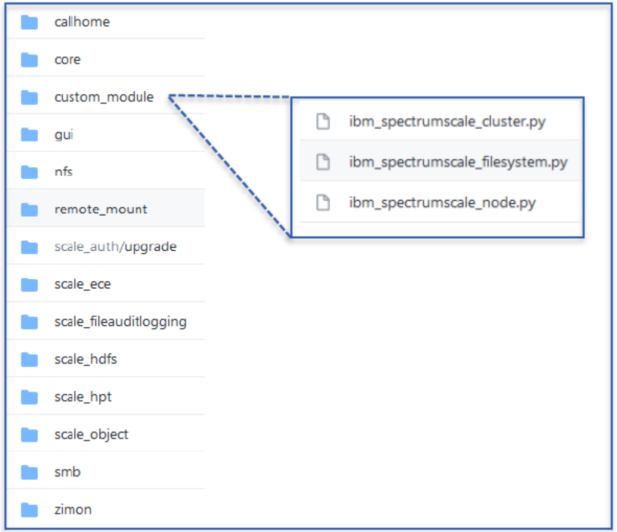
Extended support on last EUS within a release (example: V.R.x, 4.2.3, 5.1.4, 5.1.last)

Increase the number of Modification levels with new function



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



callhome
core
custom_module
gui
nfs
remote_mount
scale_auth/upgrade
scale_ece
scale_fileauditlogging
scale_hdfs
scale_hpt
scale_object
smb
zimon

ibm_spectrumscale_cluster.py
ibm_spectrumscale_filesystem.py
ibm_spectrumscale_node.py

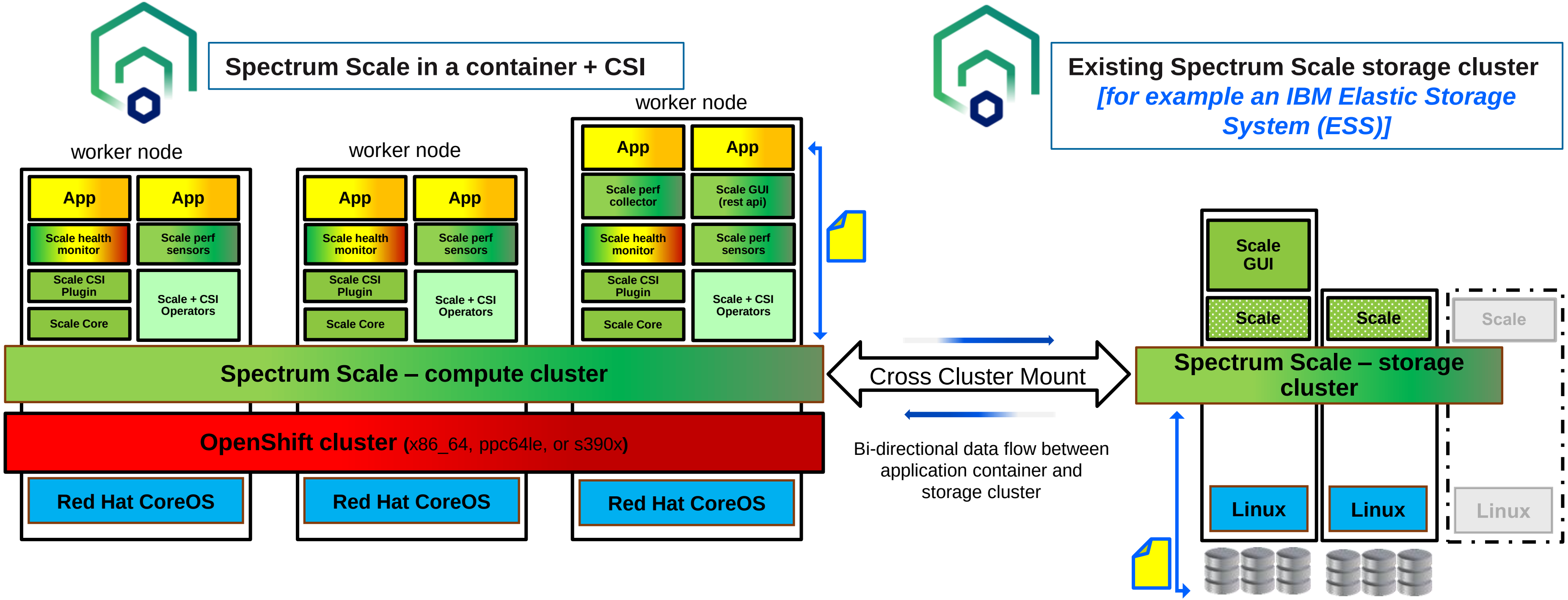
Spectrum Scale deployment is open sourced on Github

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

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



Spectrum Scale in a container + CSI

worker node

App | App

Scale health monitor | Scale perf sensors

Scale CSI Plugin | Scale + CSI Operators

Scale Core

worker node

App | App

Scale health monitor | Scale perf sensors

Scale CSI Plugin | Scale + CSI Operators

Scale Core

worker node

App | App

Scale perf collector | Scale GUI (rest api)

Scale health monitor | Scale perf sensors

Scale CSI Plugin | Scale + CSI Operators

Scale Core

Spectrum Scale – compute cluster

OpenShift cluster (x86_64, ppc64le, or s390x)

Red Hat CoreOS | Red Hat CoreOS | Red Hat CoreOS

Cross Cluster Mount

Bi-directional data flow between application container and storage cluster

Existing Spectrum Scale storage cluster

[for example an IBM Elastic Storage System (ESS)]

Scale GUI

Scale | Scale | Scale

Spectrum Scale – storage cluster

Linux | Linux | Linux

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 v5.1.4.1 to 5.1.4
- Support for RedHat OpenShift Container Platform 4.10

- 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 x86 and power servers.
- 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.

- Support for Red Hat OpenShift 4.10 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

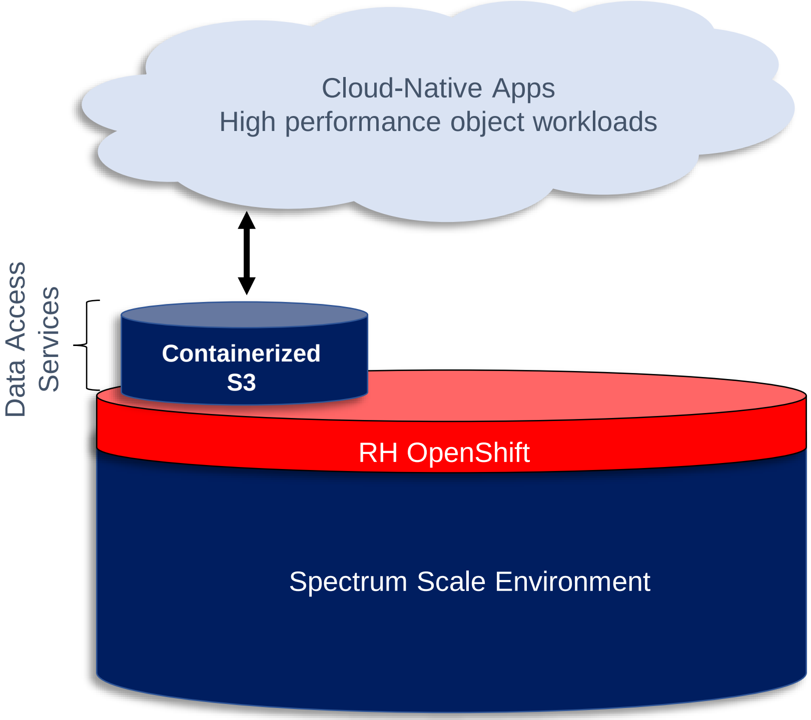
Data Access Services – S3 object access

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: MVP baseline 60 GB/s w/ 3 DAN (Data Access) nodes on vanilla ethernet, scales linearly, increased performance with each release as well as S3 functionality.



Data Access Services – GPU Direct Storage (GDS)

Spectrum Scale

Scale with NVIDIA

Understand how to get GDS and the requirements.

Spectrum Scale Knowledge Center:
https://www.ibm.com/docs/en/spectrum-scale/5.1.4?topic=summary-changes
https://www.ibm.com/docs/en/spectrum-scale/5.1.4?topic=architecture-gpudirect-storage-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

Which GDS Release*?

- CUDA 11.4 or later
- CUDA 11.5 for RoCE

Supported Storage*

- Spectrum Scale 5.1.4 and newer
- ESS or any NSD client-server storage model

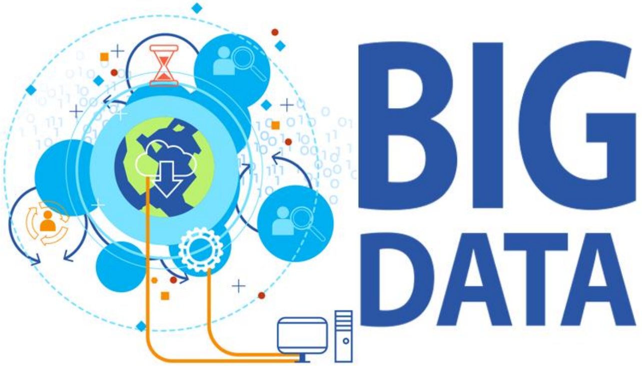
Supported Network*

- Infiniband (RDMA)
- Ethernet (RoCE)

GPUs*

- NVIDIA Ampere (e.g. NVIDIA A100)

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



Support and Currency:

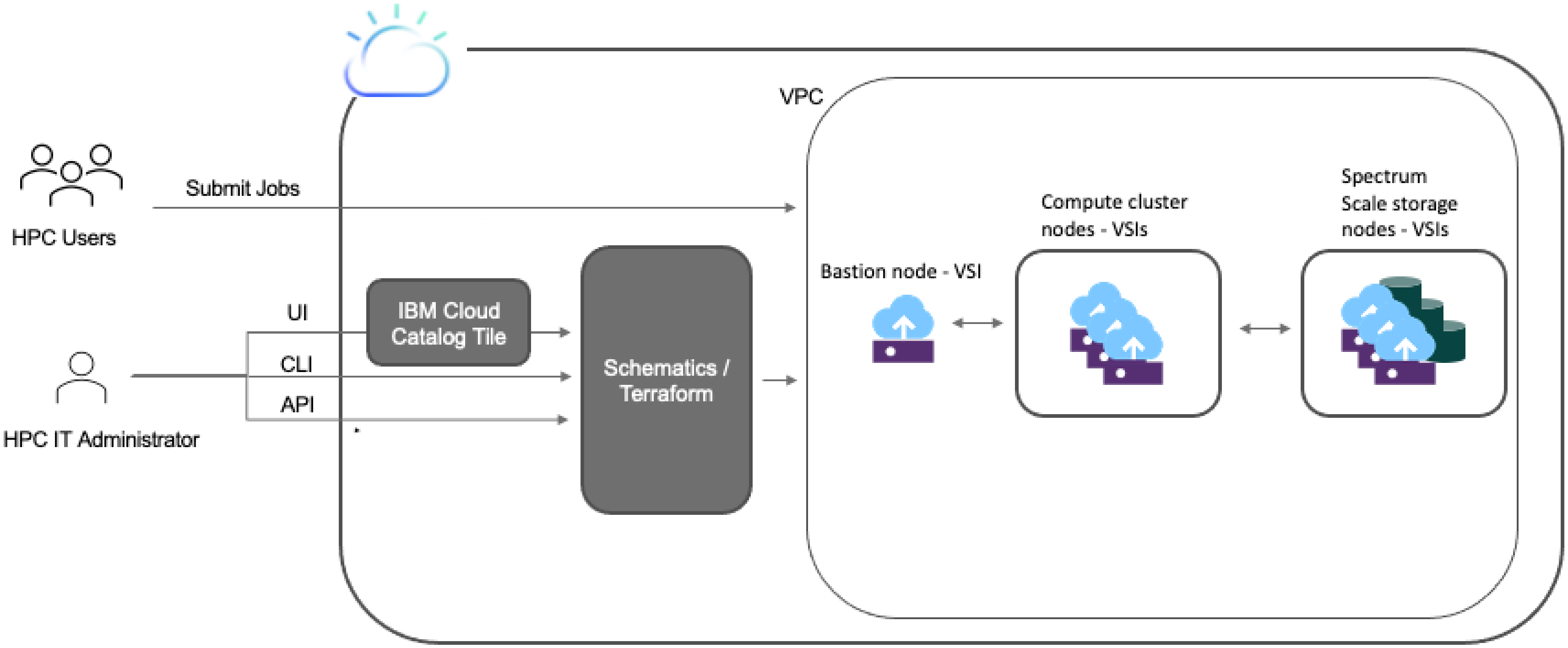
- Cloudera Data Platform (CDP) Private Cloud Base is certified with IBM Spectrum Scale on x86_64 and ppc64le since December 2020.
- Cloudera Hortonworks Data Platform (HDP) 3 and HDFS Transparency 3.1.0 end of service on December 31st, 2021.
- Opensource Hadoop 3.2.2
- NFS-Ganesha support for 3.5 code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

Spectrum Scale

Data Access Services – Spectrum Scale on IBM Cloud! Similar to AWS experience - https://www.ibm.com/cloud/hpc



Data Caching Services

- Support of Google Cloud Storage(GCS) for AFM to Cloud Object Fileset.
- Support of creating and upload objects for empty directories in AFM to cloud object storage.
- Support of marking files and directories as local in AFM to cloud object storage fileset.

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

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

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

- Enabled support of replicating more than 2K metadata in AFM to cloud object storage fileset.
- *Manual Update (MU)* mode to support manual replication of files using a file list or ILM

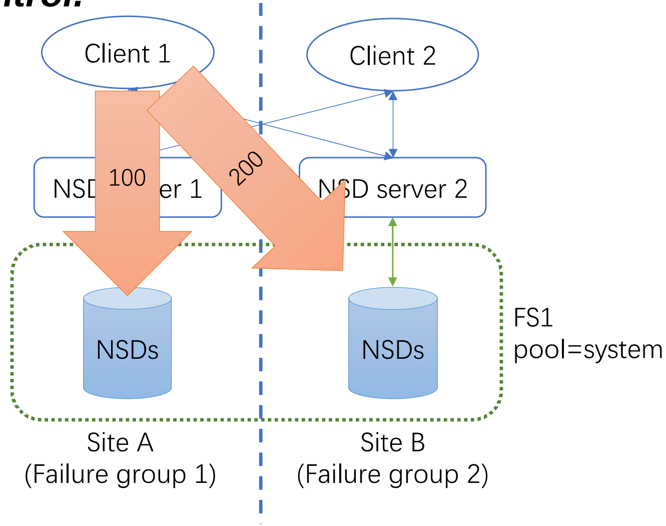
Data Management Services – Spectrum Scale Core Improvements

Spectrum Scale

Administration and reliability

Simpler and more flexible administration that allows better control.

- **mmxcp --copy-attrs** support for extra attributes – original compression is kept
- ACL garbage collection is aborted when it detects either mmunmount or mmchmgr is currently waiting to be run and will trigger a new GC run if needed
- Extend the rule of **readReplicaPolicy=local –** Add a new configuration parameter **nsdDiskAccessDistance** which allows distance for different disks by a unit of fs/pool/failuregroup.
- Online **mmfsck** since existing one requires downtime proportional to filesystem size
  - 5.1.4 – Reserved files, inodes and allocation map check



nsdDiskAccessDistance=
“FS1/system/1/100 FS1/system/2/200”
-N client1

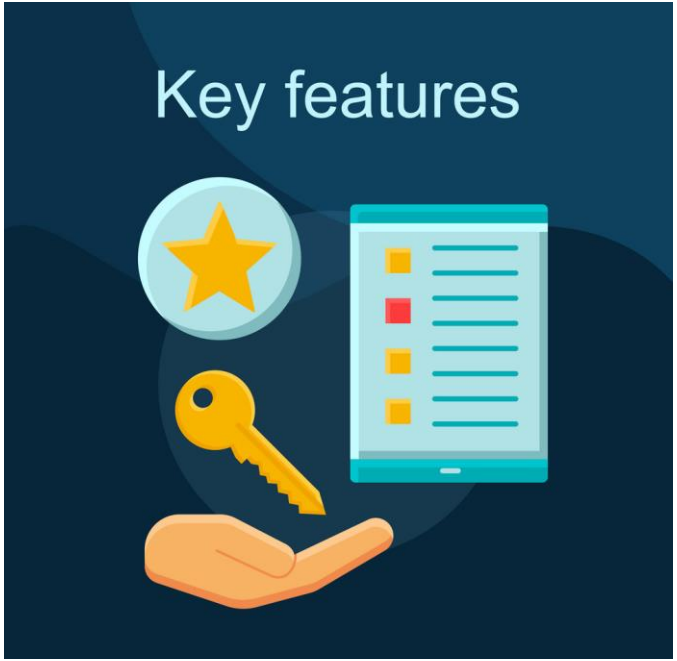
Fig.2 Different distances for different failure groups

Spectrum Scale

Data Caching Services – Performance - Spectrum Scale Core Improvements

Features that allow you to improve your resource performance.

- Allow **mmfsd** to dedicate specific TCP connections exclusively for ‘small message’ and ‘large message’ use.
    - # watch –n 5 “ls –ltr /fs1/lots_o_files_dir/”
- preferDesignatedMnode parameter – prefer metanode placement on manager node (that is usually the same node as token server for that file)
- New workload solutions
    - gpfsFineGrainReadSharing (FGRS) hint
    - gpfsFineGrainWriteSharing (FGWS) hint

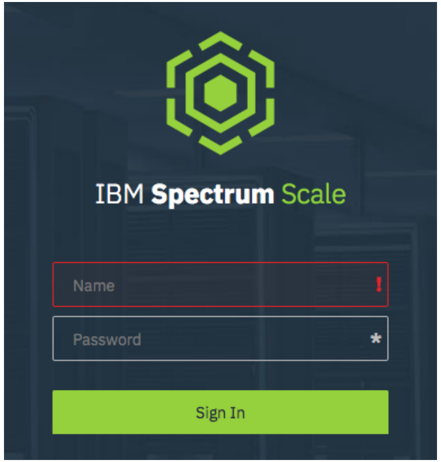


Data Management Services - GUI/API Changes

Administration and reliability

Simpler management.

- Updates to cache tables on AFM management pages
- Ensure High Availability for GUI/REST API
    - Replay logged jobs if failure occurs



Spectrum Scale

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

System Health & Monitoring

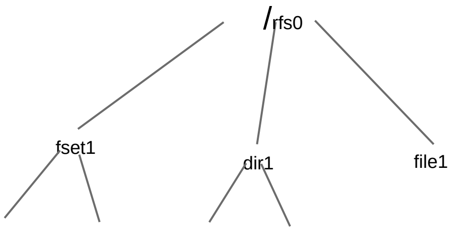
Enhanced awareness on the status of your system components

- New gpfs.snap options for HDFS and Hadoop information
- Enhanced stretch cluster monitoring via a new **STRETCHCLUSTER** component
- New **–server** option in the mmcallhome command allows call home servers to be specified explicitly
- Monitor AFM memory queue alerts in mmhealth.

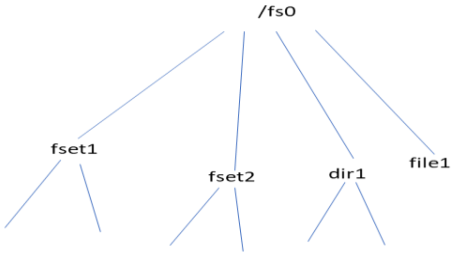
Data Security – Security – Remote Fileset Access Control (RFAC)

- 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.

Remote Cluster1



Home Cluster
Remote Cluster 1 – Allowed fsets : root, fset1



Data Security – Resiliency – Spectrum Scale Core Improvements

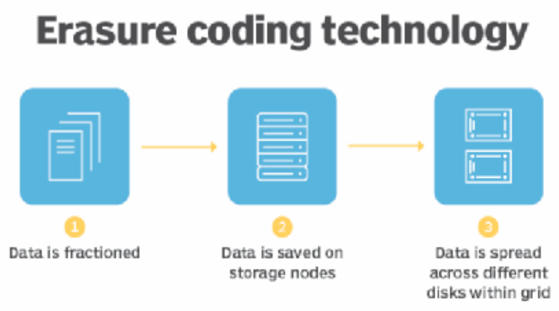
Scale on Z Systems

- **Updates to Erasure Code Edtion (ECE) guidance and usage for Spectrum Scale in Linux on Z**

- IBM z16 support
- RDMA/RoCE support for Linux on Z
- z/OS NFS client support

Data Security – Resiliency – Spectrum Scale Erasure Code Edition Changes

- 3 nodes 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
- Restricted support on background reclaim
  - User friendly automatic free space reclaim with trimming, instead of manually reclaim
  - **Need RPQ approval and customer test before using it**
- RAS improvements for kernel request hang
  - Cover both I/O path and SCSI tests
  - Reduce frequency of unnecessary disk slot location discovery unless configuration changes
  - Trigger kernel panic to reset the server node with proper setting, or trigger callback instead of kernel panic for user defined behavior
- Rebalance time and performance impact improvements


Erasure coding technology
1
Data is fractioned
2
Data is saved on storage nodes
3
Data is spread across different disks within grid

IBM ESS 3500: Next Generation

Spectrum Scale

NEXT GENERATION

Up to 12% better performance vs previous models

POWERED BY SPECTRUM SCALE

Supports Latest Global Data Platform Data Services

GREEN ENHANCED

Streamlined designed for better thermal results

ENHANCED HIGH AVAILABLITY

Enhanced non distruptive upgrades for scale-up

ESS 3500

Measured 91GB/s
300 μs latency
Measured 1+ million IOPs
(4k random reads)

* GA May 20, 2022

ESS 3500 & edge computing

Spectrum Scale

Optimized for entry configuration

Eliminate dedicated protocol node

Virtualized protocol services for 100s of clients

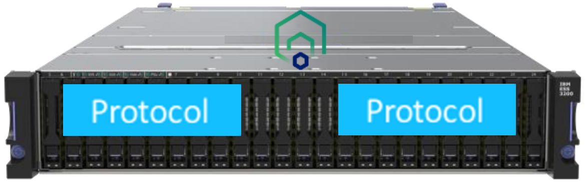
- NFS (1000)
- SMB (512)

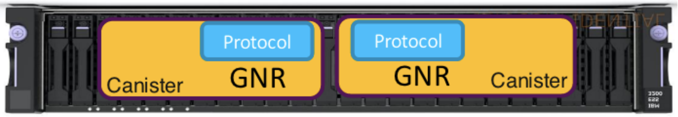
1 VM per canister

- 8 cores
- 64 GB RAM

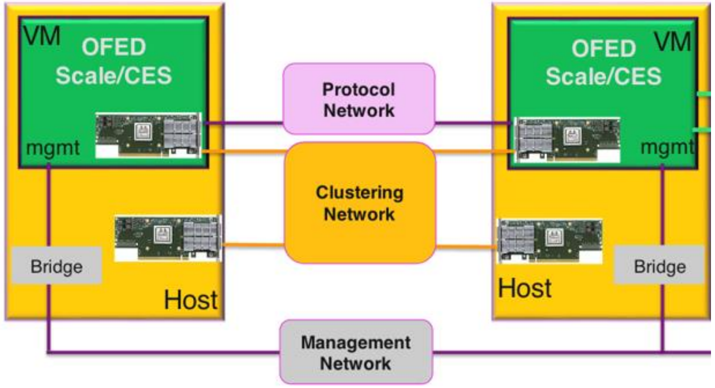
Adapters via PCIe-Passthrough

Don't forget about your EMS! ☺





ESS 3500



Spectrum Scale

Stabilized Features

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

Category	Stabilized functionality	Recommended Action	Since Version
cNFS	The use of TLS 1.0 and 1.1 for authorization within and between IBM Spectrum Scale clusters.	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

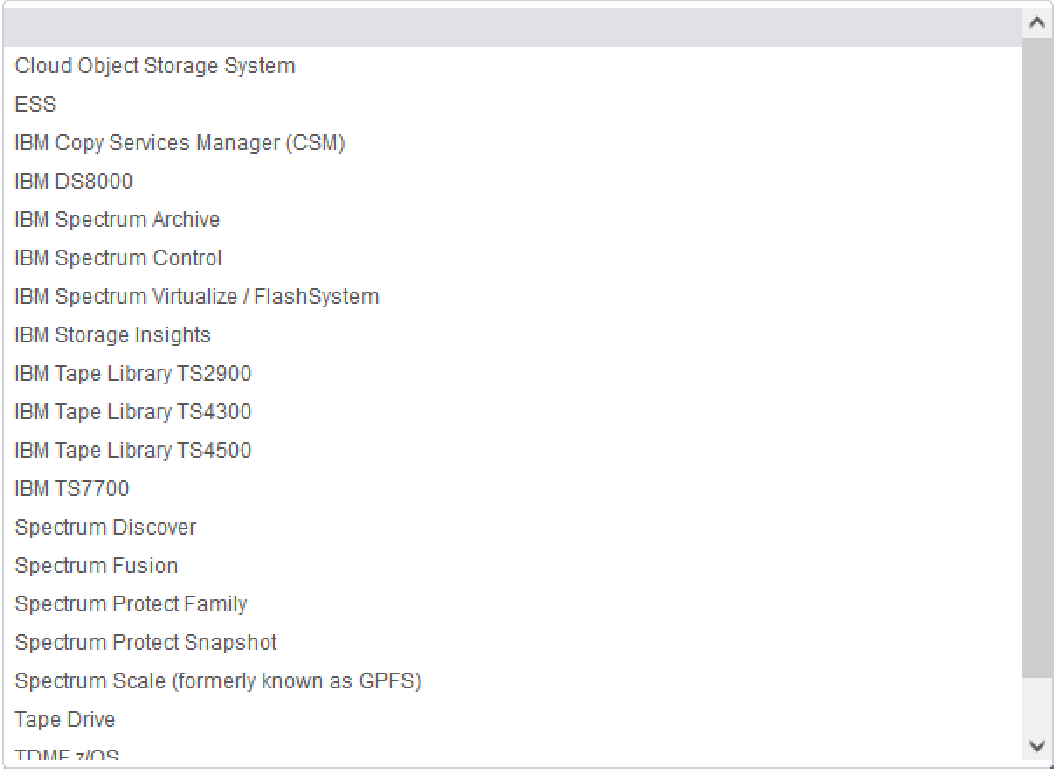
Spectrum Scale

Deprecated Features

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



Cloud Object Storage System
ESS
IBM Copy Services Manager (CSM)
IBM DS8000
IBM Spectrum Archive
IBM Spectrum Control
IBM Spectrum Virtualize / FlashSystem
IBM Storage Insights
IBM Tape Library TS2900
IBM Tape Library TS4300
IBM Tape Library TS4500
IBM TS7700
Spectrum Discover
Spectrum Fusion
Spectrum Protect Family
Spectrum Protect Snapshot
Spectrum Scale (formerly known as GPFS)
Tape Drive
TDMF z/OS

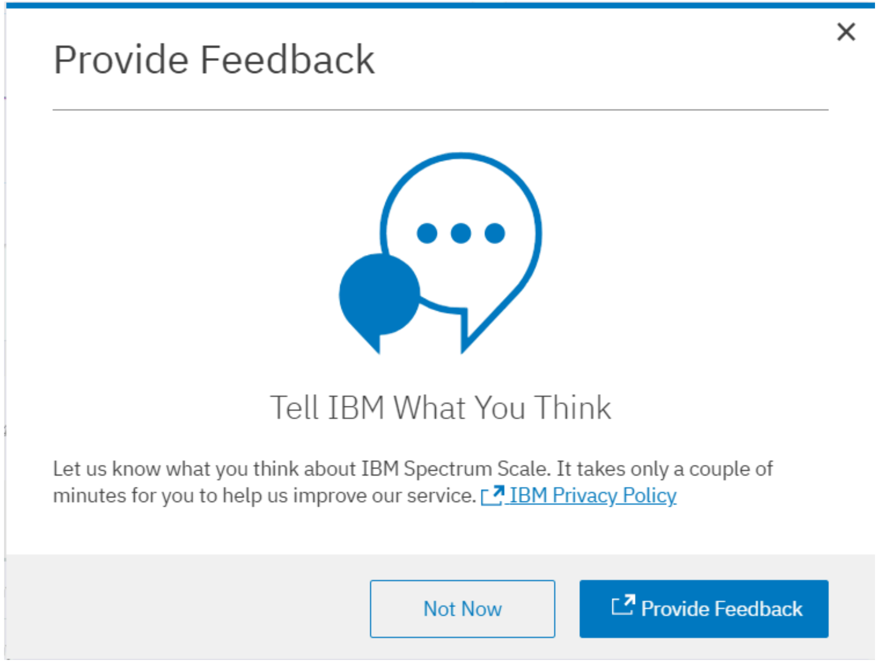
Check https://www.spectrumscaleug.org/experttalks for charts, show notes and upcoming talks

- 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

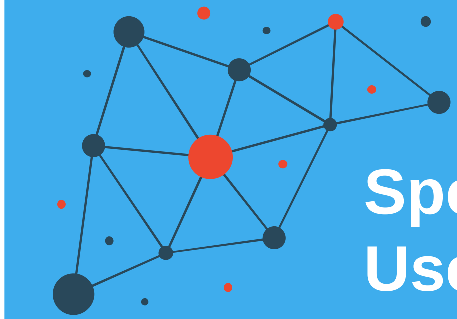
- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply



Provide Feedback

Tell IBM What You Think

Let us know what you think about IBM Spectrum Scale. It takes only a couple of minutes for you to help us improve our service. IBM Privacy Policy



Spectrum Scale User Group

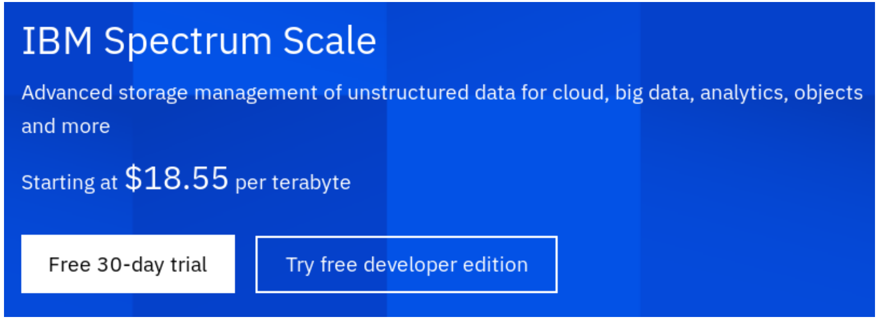
The Spectrum Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Spectrum Scale.

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!



IBM Spectrum Scale

Advanced storage management of unstructured data for cloud, big data, analytics, objects and more

Starting at $18.55 per terabyte

Fully functional!

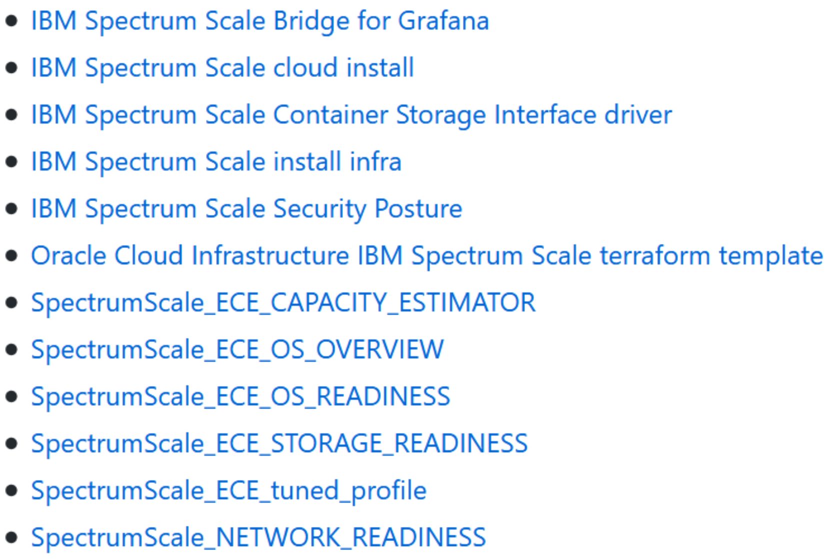
- – Based on first PTF of a release
- – Derived from **Data Management Edition (DME)**
- – Limited to 12 TBs:
  enough for a small test cluster
- – Available from the Scale ”try and buy” page on ibm.com

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!