Storage Scale

What's new in IBM Storage Scale (5.2.*) and Scale System (6.2.*)

The Global Data Platform for your best performing HPC and AI solution

Chris Maestas
IBM CTO, IBM Data and AI Storage Solutions
Chief Troublemaking Officer

IBM Storage Scale
1998 ~ 25+ Years ~ 2025

IBM

Disclaimer

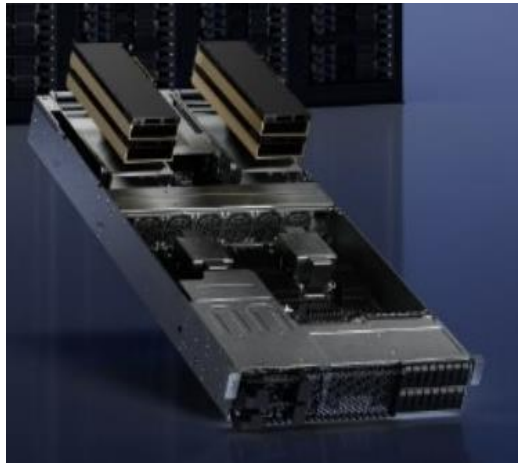
IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. 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's Storage Scale System works with NVIDIA solutions!

Storage Scale

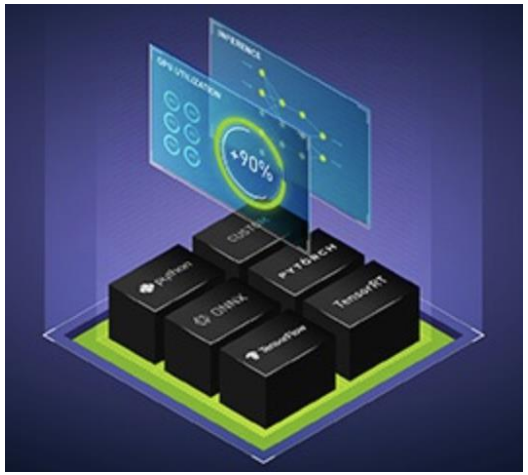
https://www.nvidia.com/en-us/data-center/dgx-superpod/



NVIDIA OVX

NVIDIA
DGX BasePOD

NVIDIA
DGX SuperPOD



NVIDIA
Cloud Partner
(NCP)

And wait there's more! NVIDIA certified system vendor based on HGX

https://www.nvidia.com/en-us/data-center/products/certified-systems/

IBM Storage Scale System for AI NVIDIA GPU Solutions

It really is this easy!

Storage Scale

Start small and scale predictably in response to business demand with the same IBM Storage Software

AI Entrant

1 DGX
or
1x HGX

- Half Populated 3500
- Up to 60 GB/s

or

- Half populated 6000
- Up to 150 GB/s read

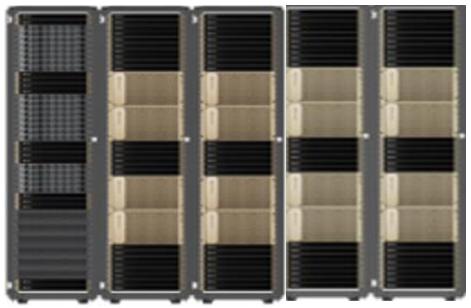
AI Medium

8 x DGX
or
8 x HGX

- 1 x 3500
- Up to 125 GB/s read

or

AI Master



16 x DGX
or
16 x HGX

- 2 x 3500
- Up to 250 GB/s read

or

- 1 x 6000
- Up to 310 GB/s read

AI Scaler

DGX SuperPOD

- 4 x 3500
- Up to 500 GB/s read

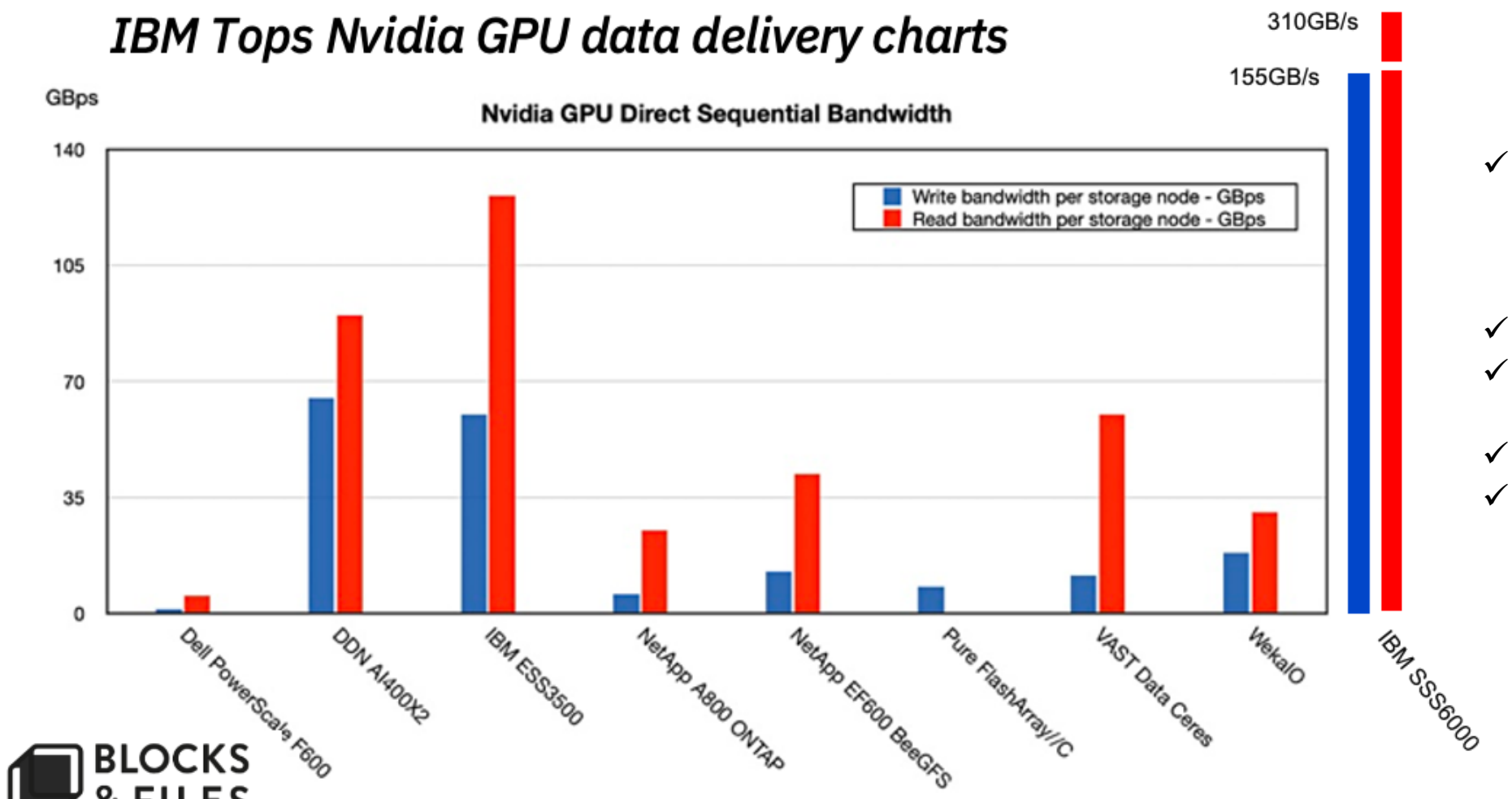
or

- 2 x 6000
- Up to 620 GB/s read

A simple, scalable upgrade path

IBM Storage Scale System 6000 sets new marks for performance

Storage Scale



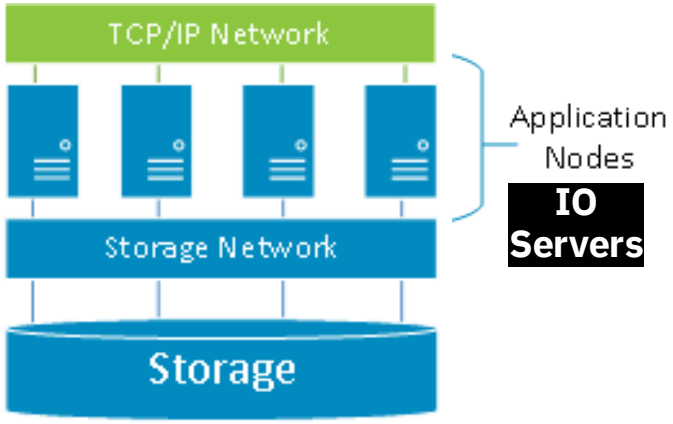
BLOCKS & FILES.

https://blocksandfiles.com/2023/08/15/ibm-nvidia-gpu-data-delivery/

- ✓ IBM 6000 is more than 2x more performant than the current 3500
- ✓ Read: 310+ GB/s
- ✓ Write: 155 GB/s
- ✓ Latest in networking
- ✓ Ready to support GB200 platforms

Scale Deployment model comparison

Storage Area Network (SAN) (NVMeoF,Fiber Channel,iSCSI)



Block based

SAN based filesystem

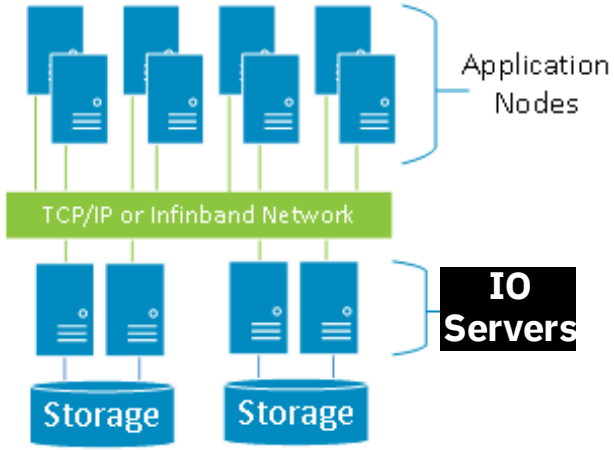
PureScale

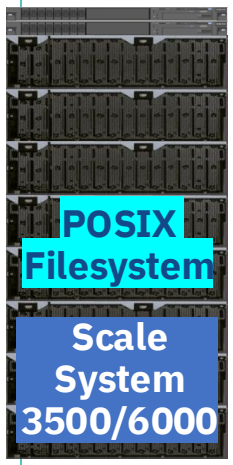
Sailfish

DS8k SDS Scale

Unify and parallelize storage silos

Twin tailed storage with erasure coding



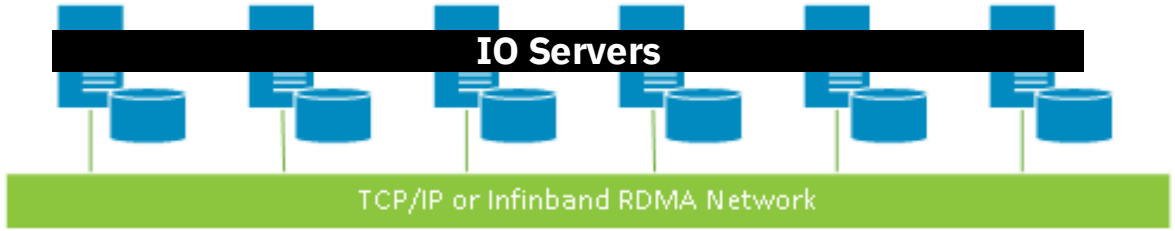


POSIX Filesystem

Scale System 3500/6000

Modular High-Performance Scaling

Shared Nothing Cluster (SNC) Model (Storage Rich Servers (replication, erasure code))



IBM Storage Ceph or OSS Ceph

IBM Cloud Object Storage (COS)

Scale (FPO)

Scale Erasure Code Edition (ECE)

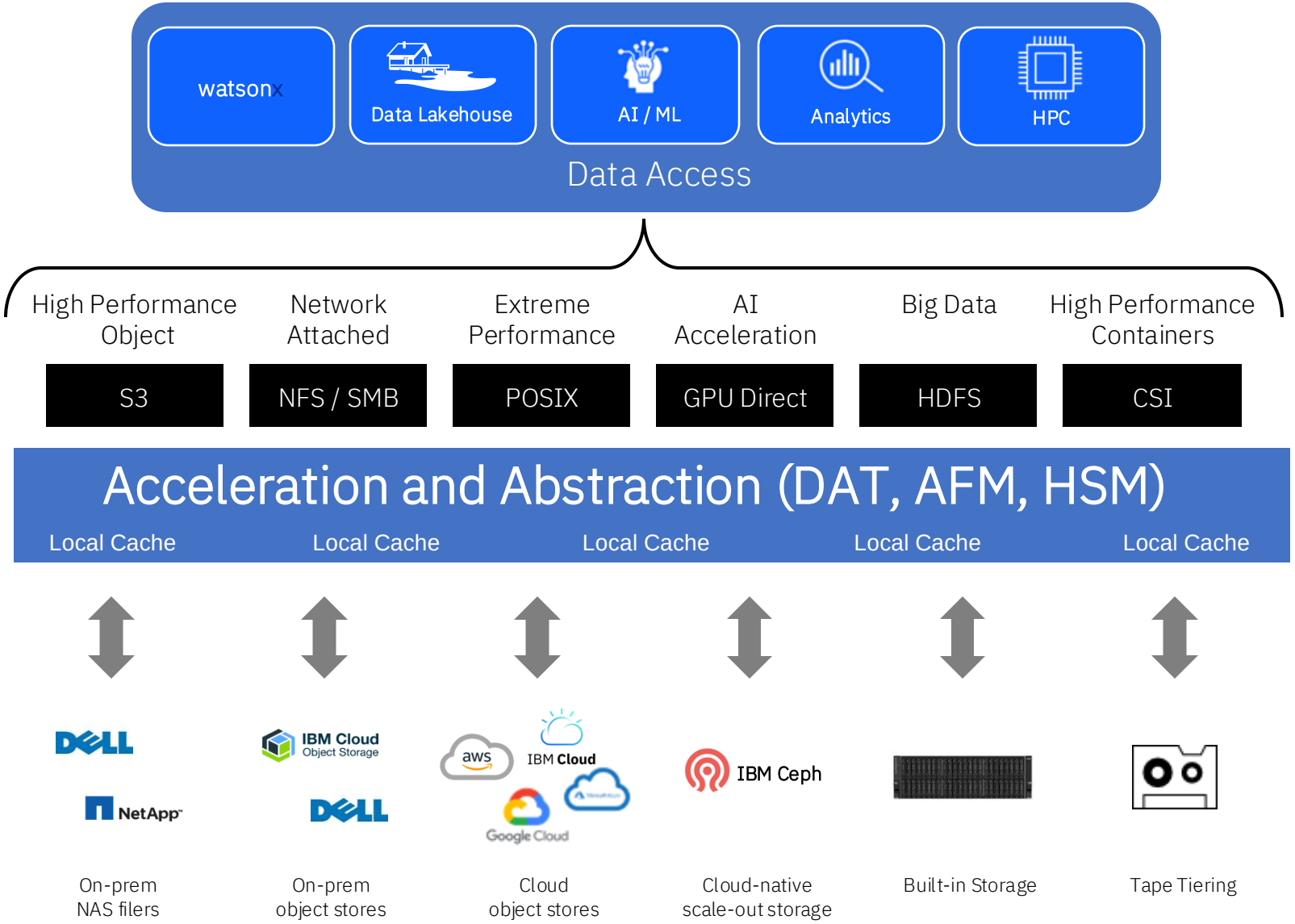
Fusion HCI or SDS

Span storage rich servers for converged architecture or HDFS deployment

IBM Storage Scale a Global Data Platform

Global Data {Access, Acceleration, Abstraction and Assurance}

Storage Scale



Multi-Protocol Data Access

Simultaneous multi-protocol access including GPU Direct support

Outcome: *Enable globally dispersed teams to collaborate on data regardless of protocol, location or format*

Storage Acceleration

Automatic, transparent caching of back-end storage systems

Outcome: *Accelerates data queries and improves economics by fronting lower performance storage*

Storage Abstraction

Single global namespace delivers a consistent, seamless experience for new or existing storage

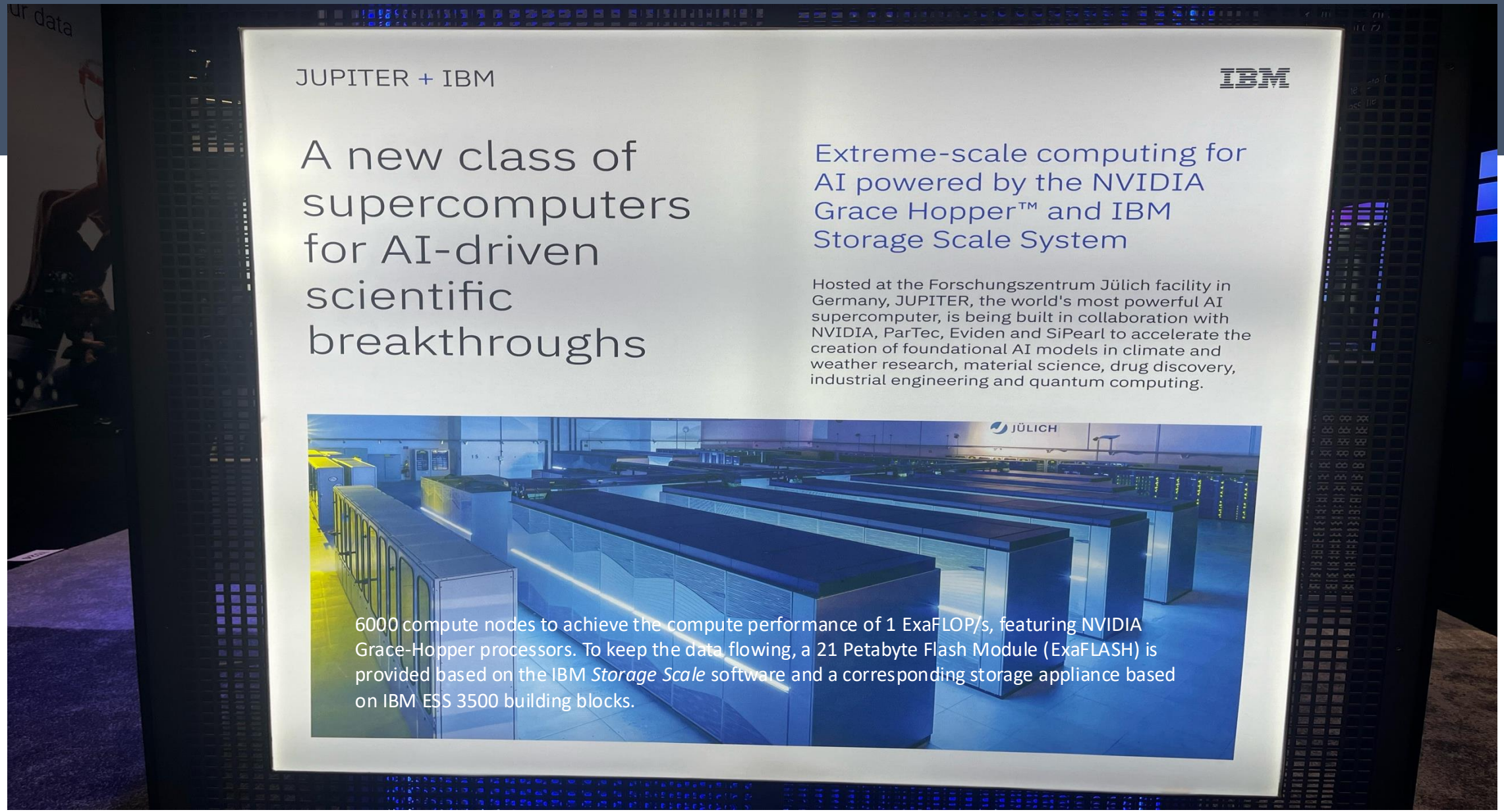
Outcome: *Reduce unnecessary data copies and improve efficiency, security and governance*

Storage Assurance

Date security from source to destination with governance and

Outcome: *Data accountability and integrity ensuring business continuity under any circumstance*

Julich Lab Jupiter Exascale AI: IBM Storage, NVIDIA GPU and ARM



More information @ https://www.fz-juelich.de/en/ias/jsc/jupiter/tech

IBM Storage Recent News

IBM Storage Scale System 6000 Now a Certified NVIDIA Cloud Partner

https://community.ibm.com/community/user/storage/blogs/mike-kieran/2025/01/10/ibm-storage-scale-system-6000-now-a-certified-nvid

IBM Storage Scale System 6000 is now a certified NVIDIA Cloud Partner (NCP) for HGX H100/H200/B200 systems. As a certified high performance storage partner for NCP, IBM Storage Scale System 6000 has demonstrated that it can deliver scalable high-performance IO to the most demanding AI training and inferencing workloads deployed on NVIDIA HGX GPUs in the cloud.

CoreWeave Partners with IBM to Deliver New AI Supercomputer for IBM Granite Models

NEWS PROVIDED BY
CoreWeave →
Jan 15, 2025, 08:00 ET

https://www.prnewswire.com/news-releases/coreweave-partners-with-ibm-to-deliver-new-ai-supercomputer-for-ibm-granite-models-302351465.html

- *One of the first deployments of NVIDIA GB200 NVL72 at supercomputing scale*
- *Supercomputer will leverage IBM Storage Scale System to power AI research and development*

"The supercomputer will leverage IBM Storage Scale System 6000 technology to deliver high-performance storage for AI, data analytics, and other demanding workloads.

As part of this agreement, CoreWeave customers can access the IBM Storage platform within CoreWeave's dedicated environments and AI cloud platform."

https://www.ibm.com/think/news/deepseek-r1-ai

NVIDIA GTC presentation for Content Aware STorage (CAST)!

In-Person | Talks & Panels

Enable Intelligent Storage to Process Data for AI Applications [S71937]

Vincent Hsu, VP, IBM Fellow, CTO for IBM Storage, IBM

Rob Davis, VP Storage Technology, NVIDIA

The common implementation of AI pipelines today is to bring data to AI. This works well when your dataset is relatively small and co-located. When we look at the next step of AI journey, we know one thing for sure: there will be a lot more data in a lot more locations. The effective way to address this challenge is to push AI processing closer to where the data is. This concept is “AI Content-Aware Storage (AI CAST).” The vision of content-aware storage is to enable intelligent storage to process data for AI applications. We'll demonstrate the architecture of AI CAST by leveraging NVIDIA Blueprints and NIMs to accelerate the retrieval-augmented generation (RAG) pipeline by incorporating storage and storage metadata in the Continuous Data Ingest and vector DB management.

Suggested Audience Level: Technical, All

Add to Schedule | **Monday, Mar 17 | 1:00 PM - 1:40 PM PDT**

IBM Storage Scale Workloads

Storage Scale

NVIDIA Solutions

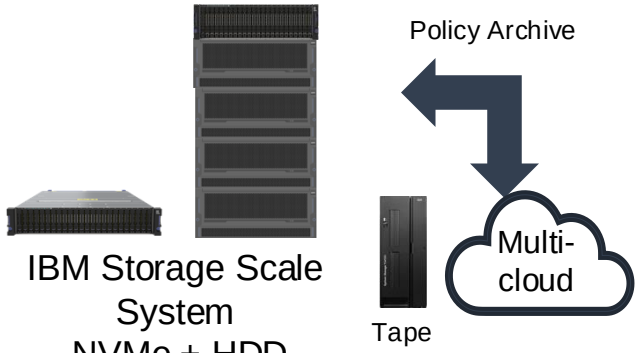
Analytics

SAS Viya require I/O throughput of 125 MB/s per physical core

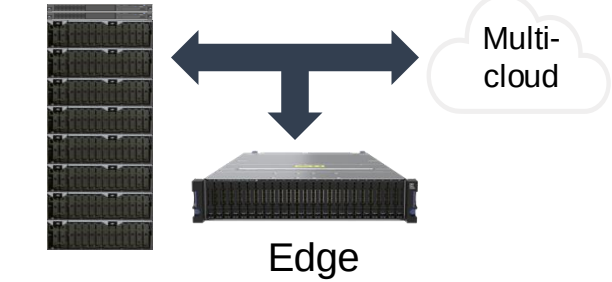
Platform modernization

IBM Cloud Pak for Data

Fast Backup and Archive

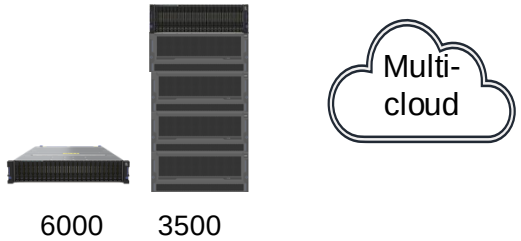


Data Lakes



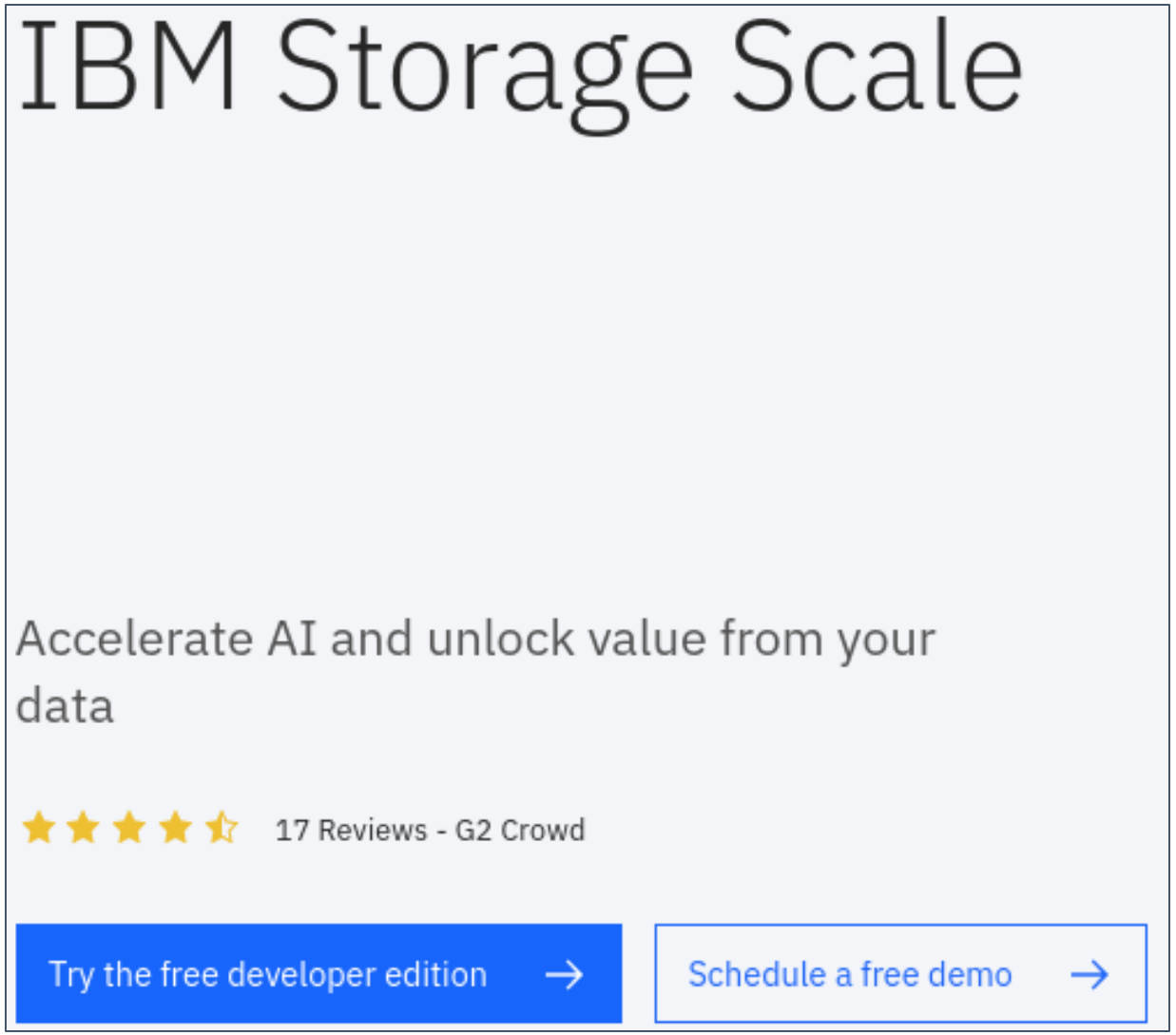
High Performance Computing

IoT/Video/Images/Genomes



IBM Storage Scale Developer Edition

https://www.ibm.com/products/storage-scale



Scale User Group

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

Techzone – https://techzone.ibm.com/collection/ibm-storage-scale-developer-edition-labs

Storage Scale

IBM Storage Scale Developer Edition Labs

Resources

(1) Rate this resource

Gold

Edit

Aug 25, 2024
Ibmcloud 2: us-east, us-south, ca-tor, eu-gb, eu-de, jp-tok, jp-osa, eu-es

IBM Storage Scale Developer Edition - Installation Experience

IBM Storage Scale Developer Edition - Installation Lab

Visibility
IBMers, Business Partners

Reserve

Aug 25, 2024
Ibmcloud 2: us-south, us-east, ca-tor, eu-de, eu-gb, jp-tok, jp-osa, eu-es

IBM Storage Scale Developer Edition Experience

IBM Storage Scale Developer Edition Installed on a 5 node system consisting of a GUI, 2 clients and 2 storage servers.

Visibility
IBMers, Business Partners

Reserve

Aug 25, 2024
Ibmcloud 2: us-south, us-east, ca-tor, eu-de, eu-gb, jp-tok, jp-osa, eu-es

IBM Storage Scale Developer Edition Lab - Cyber Security Experience with IBM QRadar

IBM Storage Scale Developer Edition Installed on a 5 node system consisting of a GUI, 2 clients and 2 storage servers along with IBM QRadar.

Visibility
IBMers, Business Partners

Reserve

Aug 25, 2024
Ibmcloud 2: us-south, us-east, ca-tor, eu-gb, eu-de, eu-es, jp-tok, jp-osa

IBM Storage Scale High Availability Experiences

Setup clusters for:

1. Erasure Coding
2. Active File Management Disaster Recovery
3. an RPO =0 Active/Active Stretch Cluster
4. a multi-cluster remote mount with AFM-POSIX or NSD remote mount

Visibility
IBMers, Business Partners

Reserve

Storage Scale Editions and Licensing

Editions have various function levels:

- Data Access Edition (DAE) – standard level often used for HPC
- Data Management Edition (DME) - adds advanced functions, valuable in commercial environments
    - Free Developer Edition (DE)
- Erasure Code Edition (ECE) - aimed at hyperscale, web-scale service providers

Capacity licensing: built for simplicity

- Easy to purchase, expand, budget, renew
- Entitled to unlimited number of IBM Storage Scale client and server licenses

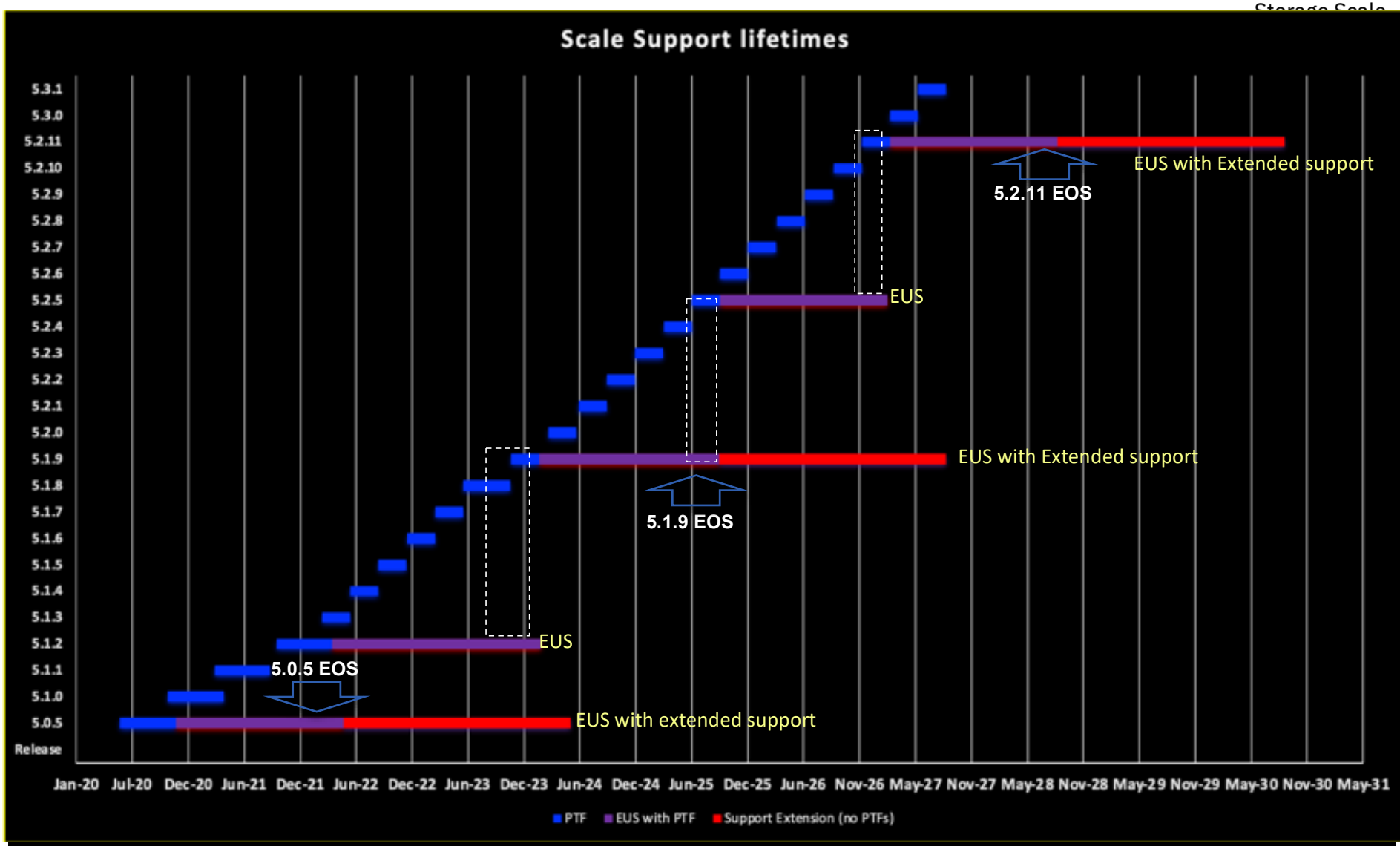
Feature	Data Access Edition	Data Management or Developer Edition	Erasure Code Edition
Multi-protocol scalable file service with simultaneous access to a common set of data	Yes	Yes	Yes
Facilitate data access with a global namespace, massively scalable file system, quotas and snapshots, data integrity and availability and filesets	Yes	Yes	Yes
Simplify management with GUI	Yes	Yes	Yes
Improved efficiency with QoS and compression	Yes	Yes	Yes
Create optimized tiered storage pools based on performance, locality, or cost	Yes	Yes	Yes
Simplify data management with Information Lifecycle Management (ILM) tools that include policy-based data placement and migration	Yes	Yes	Yes
Enable worldwide data access using AFM asynchronous replication	Yes	Yes	Yes
Immutability (WORM / Write Once Read Many)	Yes	Yes	Yes
Container Native Storage Access (CNSA)	Yes	Yes	Yes
Storage Scale Back-up Leverage	Yes	Yes	Yes
Asynchronous multi-site Disaster Recovery		Yes	Yes
Protect data with native software Encryption and secure erase, NIST compliant and FIPS certified		Yes	Yes
File audit logging		Yes	Yes
Watch folder		Yes	Yes
Fusion Data Catalog Entitlement (Discover)		Yes	Yes
Erasure coding	Scale System only	Scale System only	Yes

Release Cadence Goals

Extended Update Support goals:

- EUS with PTFs every 18 months
- Extended support on last EUS within a release
- Increase the number of Modification levels with new function
- Scale's Extended Update Support (EUS) approach is outlined in product FAQ

- **EUS release approach applies to non-containerized scale**
- CNSA currently doesn't have an EUS



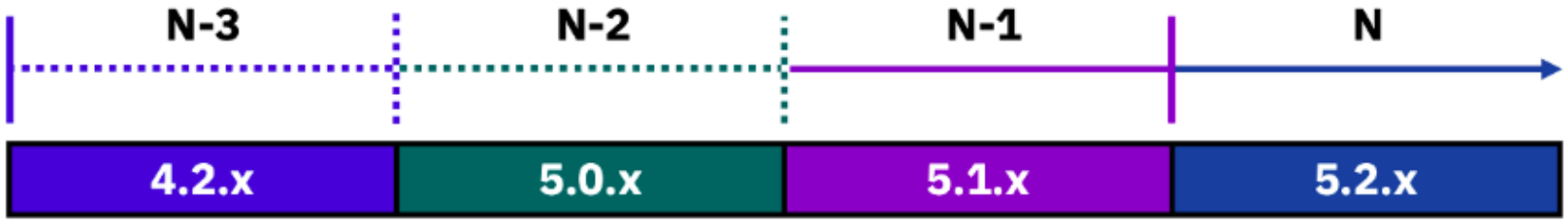
Note: Version numbers and release timing are for example purposes to demonstrate the goal of EUS every 18 months and do **not** represent a commitment to deliver a specific version or on a specific timeline

Release Cadence Goals

Can different IBM Storage Scale maintenance levels coexist?

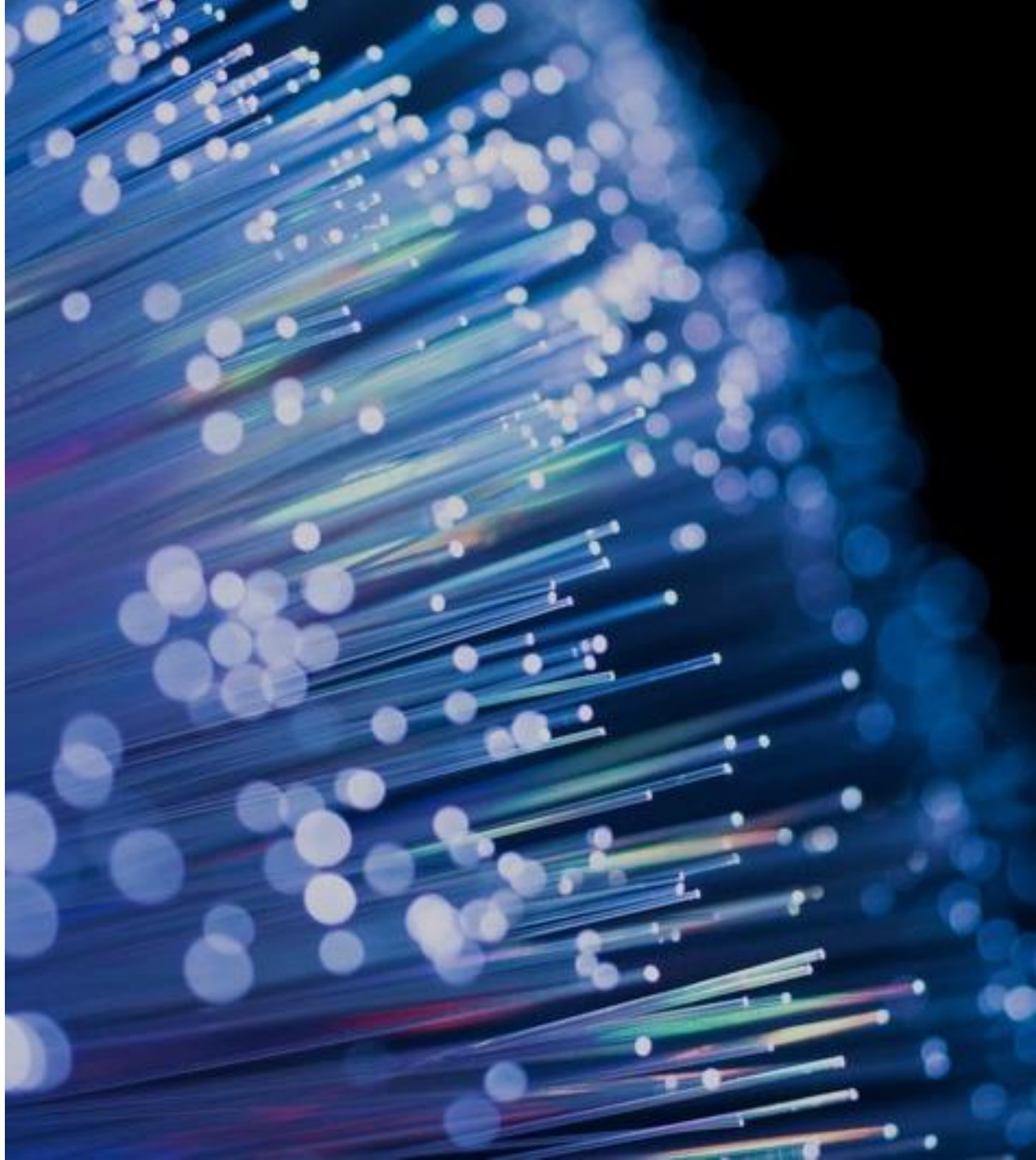
A2.8:

Different releases of IBM Storage Scale can coexist, that is, be active in the same cluster and simultaneously access the same file system. For release co-existence, IBM Storage Scale follows the N-1 rule. According to this rule, a particular IBM Storage Scale release (N) can co-exist with the prior release of IBM Storage Scale (N-1). This allows IBM Storage Scale to support an online (rolling) upgrade, that is a node by node upgrade. As expected, any given release of IBM Storage Scale can coexist with the same release. To clarify, the term release here refers to an IBM Storage Scale release stream and the release streams are currently defined as 4.2.x > 5.0.x > 5.1.x > 5.2.x.



These coexistence rules also apply for remote cluster access (multi-cluster remote mount). A node running release N-2 cannot perform a remote mount from a cluster which has nodes running release N, and vice versa.

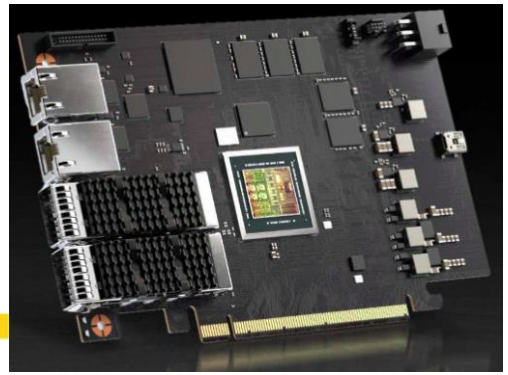
Access Services



Access Services – ARM

GA! -The official Architecture name is aaarch64

Wider support to use ARM functionality Data Processing Units (DPU)



Current goal: ARM client
compute nodes (Grace Hopper)

Make it a platform for Scale like any other

DPU (Blue Field-3) for exploitation
research spike

- BF-3 can be used as NIC with Scale as any other supported NIC given OS and MOFED supports it
- Work in Progress for further exploitation
- IOR results from Grace Hopper
  - Max Write: 46437.87 MiB/sec (48693.63 MB/sec)
  - Max Read: 47281.41 MiB/sec (49578.15 MB/sec)

- ***Included***
  - SE package / install toolkit / rpm based install
  - NSD client
  - Scale base functionality (IO, policies, remote mounts, snapshots, quotas, etc.)
  - Manager roles: file system manager / token manager / cluster manager
  - RDMA (IB or RoCE) including GDS
  - Health Monitoring
  - Target OS: RHEL 9.3 and Ubuntu 22.04 (ask to open RFE for customers askign for RHEL 8)
  - File audit logging, watch folders folders
  - Call home
  - GUI (can display ARM node, but cannot run on ARM)

- The NSD server functionality is now supported on arm64 platforms.

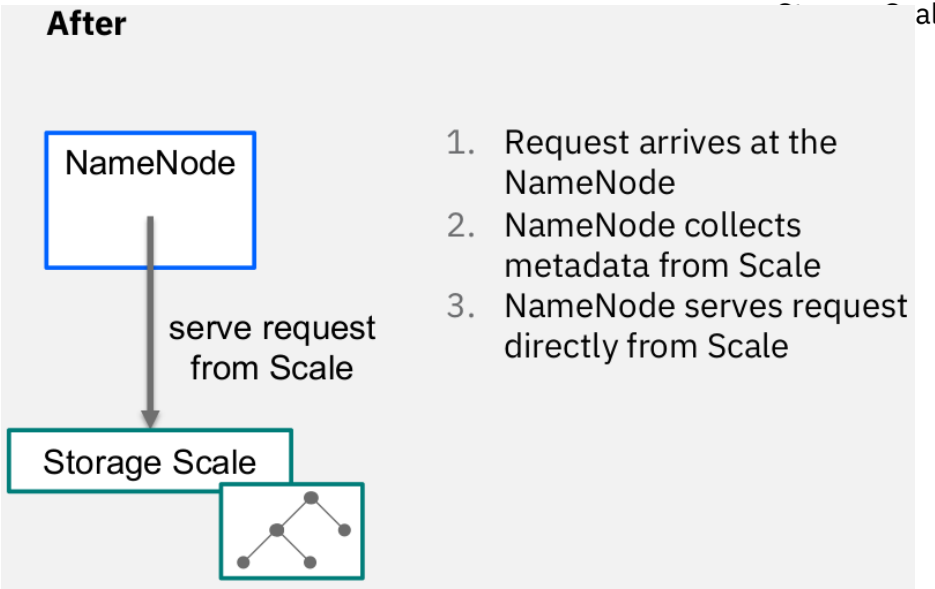
- ***Excluded***
  - SNC
  - Protocols
  - BDA / HDFS
  - CNSA
  - TCT (discontinued)
  - HSM

- We need to learn whether there are ARM designs that need code changes
  - so far the only one has been Raspberry Pie ;-)
  - … and that has been fixed but is still not supported

Access Services –NFS, SMB, HDFS

Support and Currency:

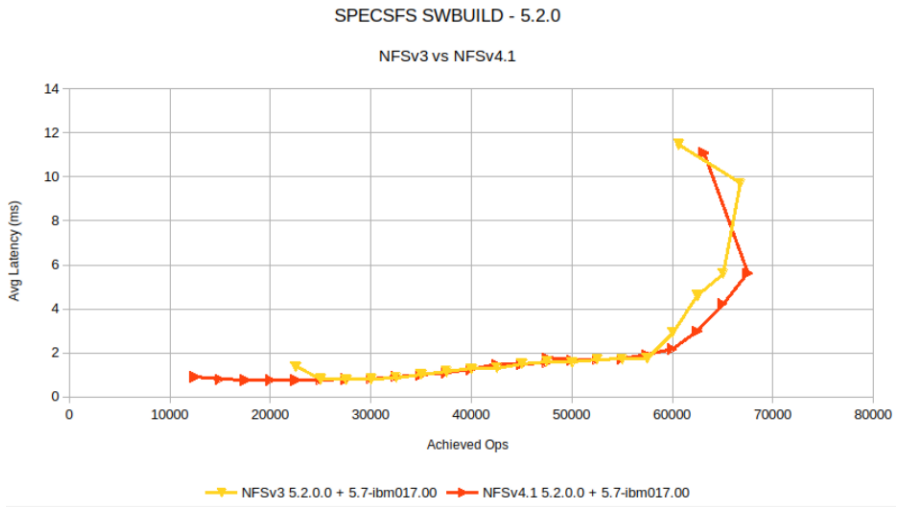
- Samba 4.19 release
  - The security improvements in recent releases (4.13, 4.14, 4.15, 4.16), mainly as protection against symlink races, caused performance regressions for metadata heavy workloads. While 4.17 already improved the situation quite a lot, with 4.18 the locking overhead for contended path based operations is reduced by an additional factor of ~ 3 compared to 4.17. It means the throughput of open/close operations reached the level of 4.12 again.
- NFS-Ganesha support for 5.7 code base
  - In presence of NFS IO the health check “rpc null check” may fail, and second check “performance counters” with it – leading to useless IP failover and failback, causing NFS Grace period and adding extra impact to NFS clients



Improved performance:

- NFS “meta data cache” component was revised resulting in significant performance improvements

  https://community.ibm.com/community/user/storage/blogs/mara-miranda-bautista/2024/05/20/ibm-storage-scale-ces-nfs-520-performance-eval



- HDFS transparency metadata redesign
  - Full parallelism for RPC calls (GPFSNamesystem)
  - No more lock contention in NameNode
- **Continued partnership with Tuxera for high-performance SMB**
- **Finishing MoSMB Evaluation as well**

Access Services – High Performance Object 2.0!

Storage Scale

Support and Currency:

- Swift is being Discontinued
  - You can use 5.1.8 Swift code in CES of 5.1.9
- New CES S3 is here!
  - https://www.ibm.com/support/pages/node/7145681

Multi-protocol data access support with POSIX, S3, NFS, SMB and CSI

ILM support including Tiering to Tape support via RPQ

2 billion objects in a single bucket ! - https://github.com/ghcoelhopsa/scale_s3_benchmark

IBM Technology Expert Labs can provide billable migration services (Swift to CES S3 and DAS S3 (HPO 1.0) to CES S3 (HPO 2.0))

Improved performance:

- IBM Storage Scale CES S3 (Tech preview) Performance evaluation of large and small objects using COSBench: https://community.ibm.com/community/user/storage/blogs/rogelio-rivera-gutierrez/2024/04/25/ibm-storage-scale-performance-ces-s3-tech-preview

Scaling limits for S3:

- Up to 10TB single object size
- Up to 5000 S3 accounts
- Up to 5000 S3 buckets
- Up to 100M objects per bucket (tested limit)
- Up to 3K client connections per CES node

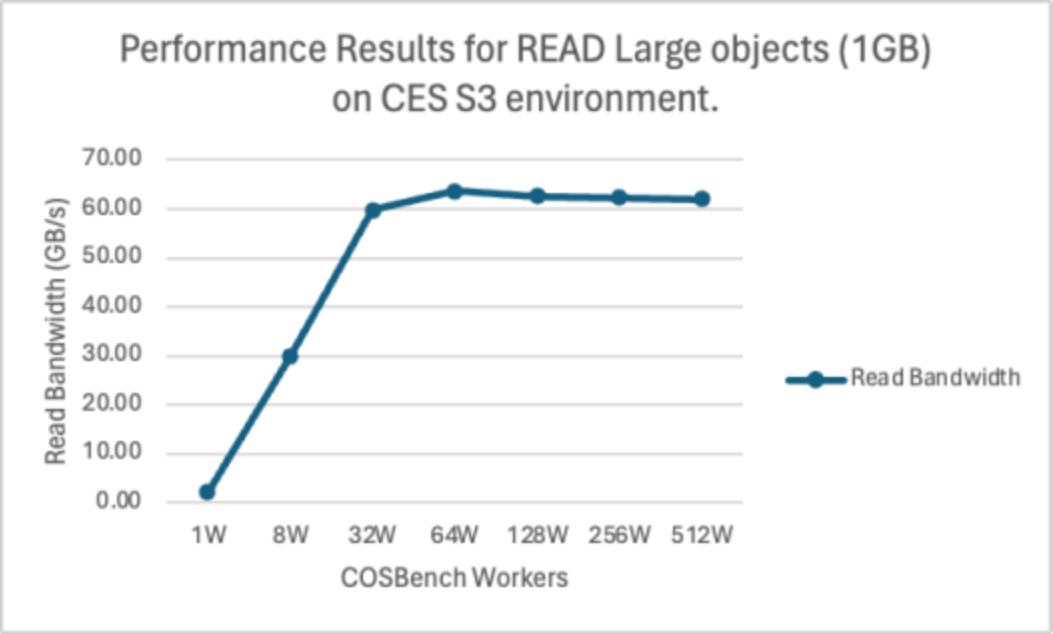
Higher scaling limits as compared to HPO 1.0 !

Deployment Requirements:	
Storage Scale Cluster:	Storage Scale 5.2.1
Operating System:	RHEL8.x or RHEL9.x
Architecture:	x86_64, Power(ppc64le), Z(s390x)
Storage Scale CES Cluster Size:	Up to 10-node CES cluster (tested limit)
*No support for upgrade from CES S3 Tech Preview to CES S3 MVP GA	

NO Openshift cluster required !

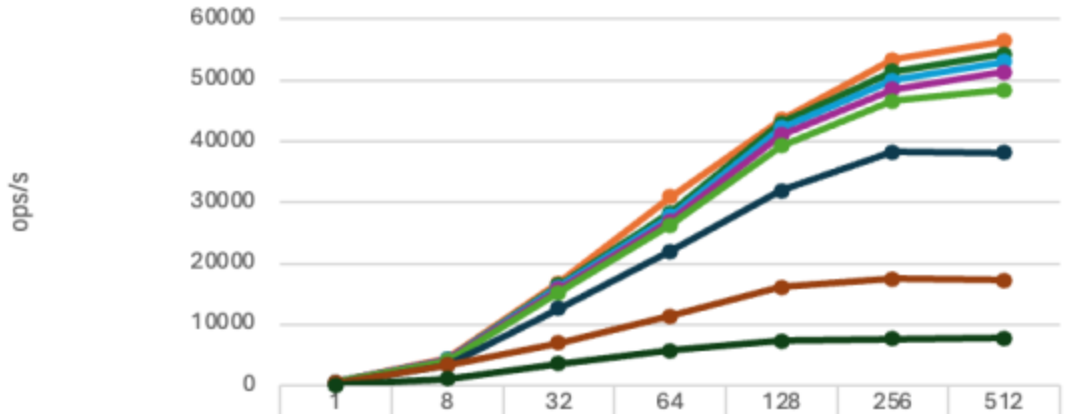
Access Services – Object Performance

Storage Scale



Op-Type	Obj Size	Workers	Op-Count	Byte-Count	Avg-ResTime	Avg-ProcTime	Throughput	Bandwidth	Succ-Ratio
READ	1GB	1	611 ops	625.66 GB	490.86 ms	4.83 ms	2.04 op/s	2.09 GB/S	100%
		8	8.78 kops	8.99 TB	273.26 ms	5.13 ms	29.27 op/s	29.98 GB/S	100%
		32	17.49 kops	17.91 TB	548.53 ms	7.67 ms	58.33 op/s	59.73 GB/S	100%
		64	18.61 kops	19.06 TB	1029.76 ms	15.79 ms	62.15 op/s	63.64 GB/S	100%
		128	18.28 kops	18.72 TB	2093.59 ms	28.6 ms	61.13 op/s	62.6 GB/S	100%
		256	18.12 kops	18.55 TB	4210.39 ms	60.39 ms	60.79 op/s	62.25 GB/S	100%
		512	17.91 kops	18.34 TB	8453.23 ms	106.39 ms	60.55 op/s	62.01 GB/S	100%

Table 2. Performance Results for READ Large objects (1GB) on CES S3 environment.

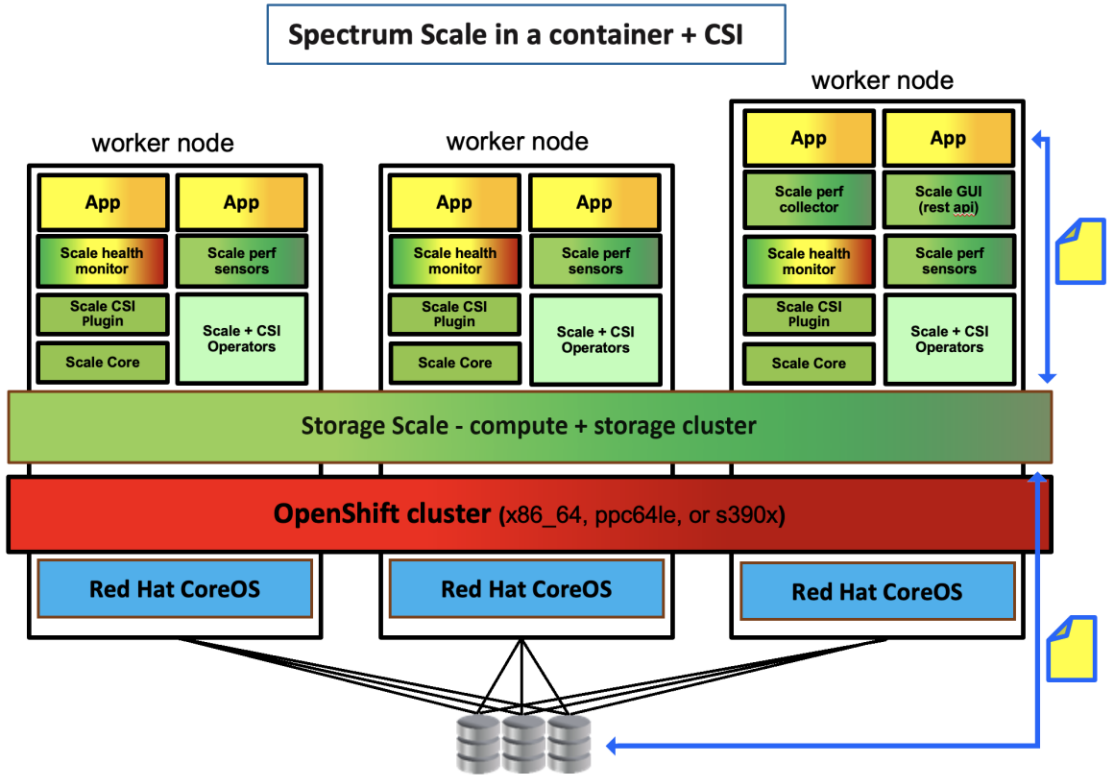


	1	8	32	64	128	256	512
4KB Object Size	557.45	4404.72	16838.96	30873.57	43559.88	53364.39	56416.84
32KBObject Size	552.89	4295.77	16439.61	28307.39	42945.14	51395.8	54249.48
64KB Object Size	540.26	4235.06	16161.52	27777.88	42140.26	50030.34	53075.99
128KB Object Size	536.22	4163.35	15764.64	26976.29	41039.65	48512.89	51357.97
256KB Object Size	509.1	4022.61	15200.64	26243.51	39258.35	46620.9	48416.68
1MB Object Size	421.53	3348.29	12619.59	21943.27	31959.01	38207.98	38115.95
4MB Object Size	441.76	3350.49	6965.76	11388.8	16161.77	17500.94	17282.13
8MB Object Size	135.47	1044.22	3607.19	5765.67	7334.11	7698.82	7735.21

https://community.ibm.com/community/user/storage/blogs/rogelio-rivera-gutierrez/2024/04/25/ibm-storage-scale-performance-ces-s3-tech-preview

Access Services – Container Native Storage Access (CNSA)

Improvements introduced in CNSA 5.2.2.0



Wider support to use the latest CNSA functionality.

Support for Red Hat OpenShift 4.15, 4.16, 4.17 with IBM Storage Scale container na

Support for parallel core pod upgrade

Avoid node reboots during upgrade

Multiple GUI hosts can be specified for CSI.
CNSA 5.2.1 will use multiple hosts in operator

Configure Resource limits of core pods

Internal GUI user password rotation
Starting with 5.2.0, the passwords of the internal REST users is changed every 90 days

Tech Preview!
Vela only: AFM caching via StorageClass addition of volumeType: “cache” as well as cacheMode options

Tech Preview of local disk attachment utilizing a direct disk attachment configuration, replacing prior technology preview of a shared nothing local disk configuration.

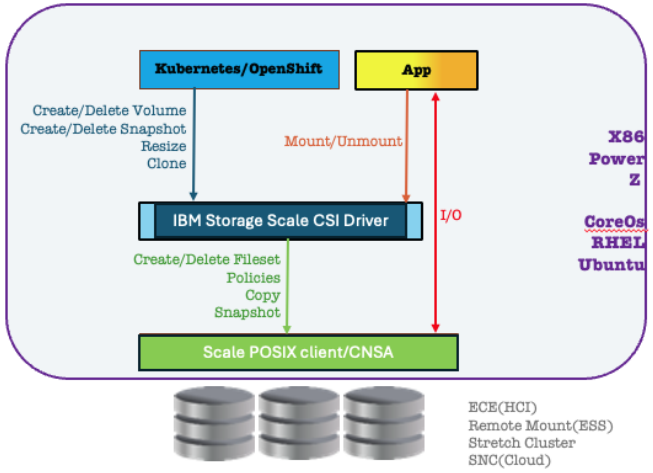
Infiniband RDMA support (previously technology preview in 5.2.1.0)

Access Services – Container Storage Interface

Improvements introduced in CSI 2.13.0

Storage Scale

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



Support for CSI specification 1.9

Support to customize the volume name prefix

Support for Kubernetes 1.31 and Red Hat® OpenShift® 4.17

Improvements in the script for debug data collection

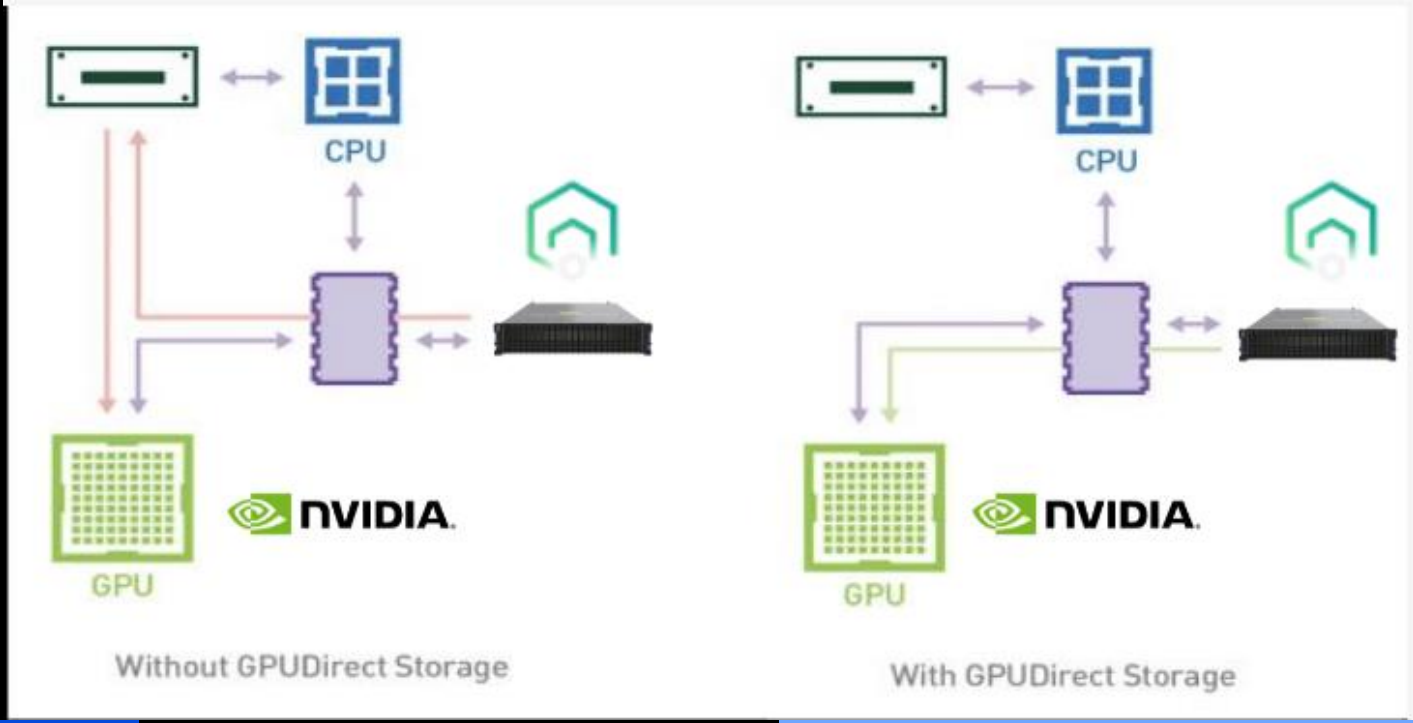
storage-scale-driver-snap.sh [-l | -n | -o | -p | -s | -v | -h]

Important: Starting with IBM Storage Scale Container Storage Interface 2.13.x, the support for OpenShift with RHEL worker nodes is discontinued.

Dynamic Provisioning - Create/Delete Volume
Static Provisioning
Volume Snapshot
Volume Expansion
Shallow Copy
Volume Cloning
Compression
Tiering
ConsistencyGroup
Remote Mount
Multiple Filesystem
FsGroup
GUI HA
Lightweight Volume/Fileset Based Volume

GPUDirect Storage

enables an explicit direct memory access (DMA) between GPU memory and storage when used in the application code



NVIDIA Magnum IO

Family of I/O Optimizations for GPU accelerated data centers

GPU Direct RDMA: Access peer node's memory without copying to host memory

GPU Direct Storage: Transfer data to/from GPU directly from storage without involving CPU and CPU memory

CUDA Toolkit

GDS will be in the CUDA toolkit

Development environment for GPU accelerated applications

Libraries, compilers, debuggers, optimizers, and tools

Leading GPU compute platform since 2006

GDS for Applications

Invoked using the CUDA Toolkit (cuFile) API

APIs must be explicitly called by the applications

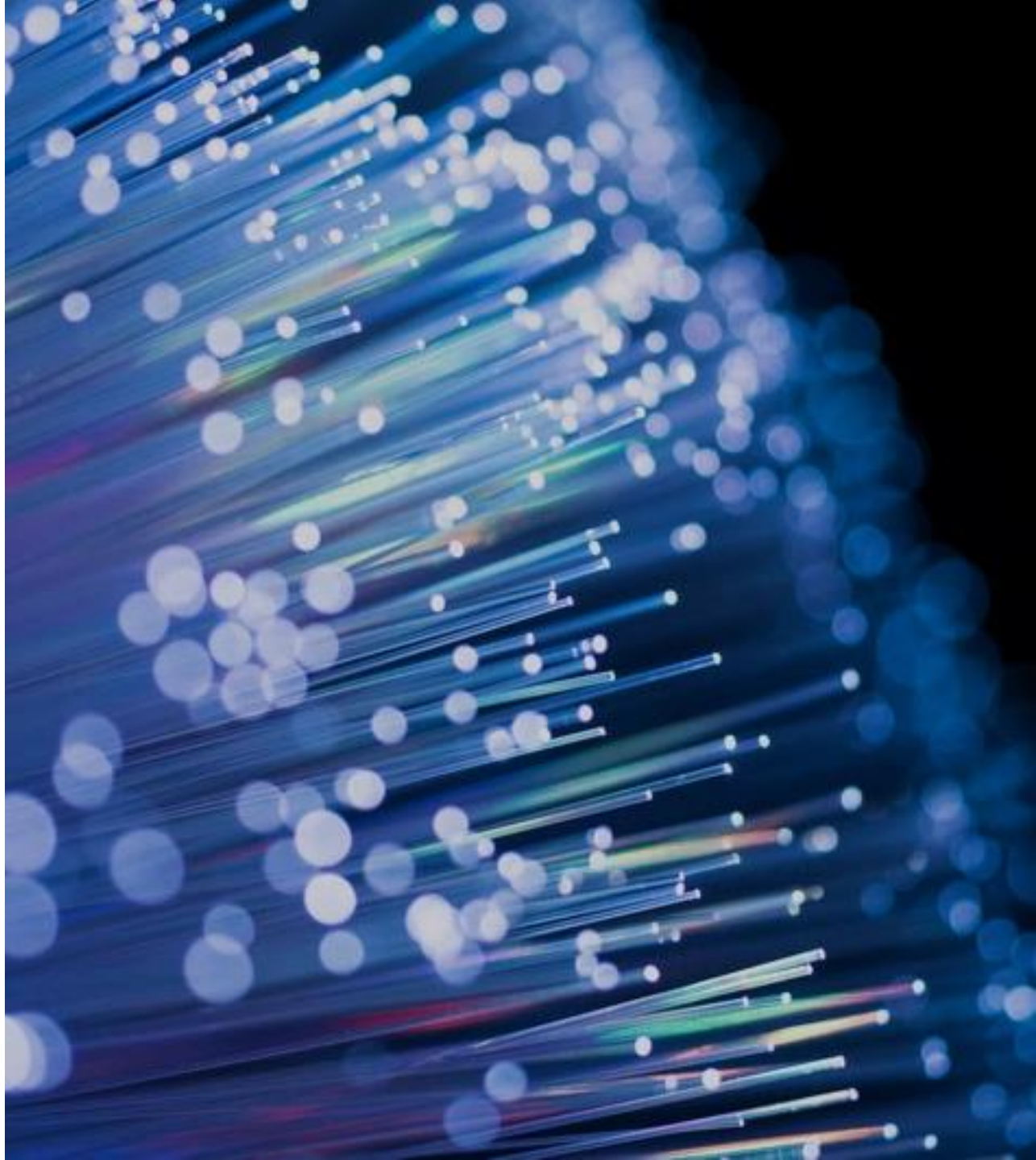
Storage must be GDS enabled. If not, GDS call falls back to regular data movement.

Why it matters

AI, HPC, Analytics are data hungry and require a very high data throughput.

GPUs are starved by slow I/O (and NFS is particularly slow)

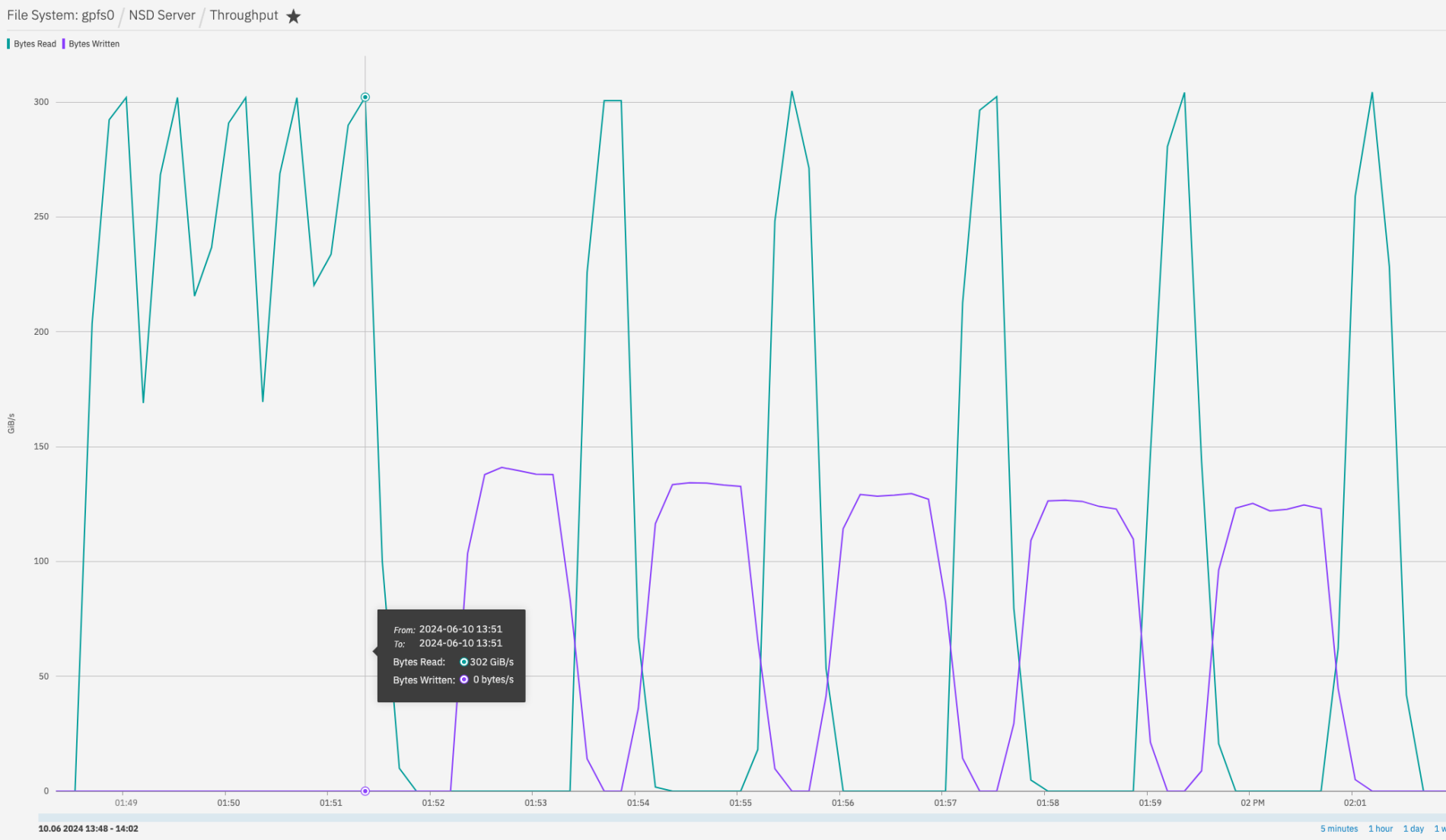
Acceleration and Abstraction



Performance update from 6000

Storage Scale

- **5x iterations:**
- Mean Write:
  - 156.35 GB/s
- Mean Read:
  - 320.49 GB/s



ECE - New EC code 16+2P/3P – 1

more efficient use of capacity and some improvement with full track writes

Storage Scale

Number of nodes	3WayReplication	4WayReplication	4+2P	4+3P	8+2P	8+3P	16+2P	16+3P
3	1 Node + 1 Device *	1 Node + 1 Device *	Not recommended 1 Node	Not recommended 1 Node	Not recommended 2 Devices	Not recommended 3 Devices	Not recommended 2 Devices	Not recommended 3 Devies
4	1 Node + 1 Device *	1 Node + 1 Device *	Not recommended 1 Node	1 Node + 1 Device #	Not recommended 2 Devices	Not recommended 1 Node	Not recommendd 2 Devices	Not recommended 2 Devies
5	2 Nodes	2 Nodes *	Not recommended 1 Node	1 Node + 1 Device	Not recommended 1 Node	Not recommended 1 Node	Not recommended 2 Devices	Not recommended 2 Devies
6	2 Nodes	2 Nodes *	2 Nodes #	2 Nodes	Not Recommended 1 Node	1 Node + 1 Device #	Not recommended 2 Devices	Not recommended 2 Devies
7	2 Nodes	2 Nodes *	2 Nodes	2 Nodes*	Not Recommended 1 Node	1 Node + 1 Device	Not recommended 2 Devices	Not recommended 2 Devies
8	2 Nodes	2 Nodes *	2 Nodes	2 Nodes*	Not Recommended 1 Node	1 Node + 1 Device	Not recommended 2 Devices	Not recommended 2 Devies
9	2 Nodes	3 Nodes	2 Nodes	3 Nodes	Not Recommended 1 Node	1 Node + 1 Device	Not Recommended 1 Node	Not Recommended 1 Node
10	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes #	2 Nodes	Not Recommended 1 Node	1 Node + 1 Device #
11+	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes	3 Nodes	Not Recommended 1 Node	1 Node + 1 Device
18	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes #	3 Nodes
19	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes	3 Nodes	2 Nodes	3 Nodes

- **To protect data from disk failure, all failure tolerances that are marked with # in the table need to be paid attention to for the spare disk space other than the erasure code. You can change the number of spare disk space to the same or bigger number than the node number before creating vdisks. For example, for 6 nodes with 4+2p erasure code, you can change all DA's spare disk space to 6 before creating vdisks.**
- **All failure tolerances that are marked with * are limited by recovery group descriptors rather than by the RAID code.**

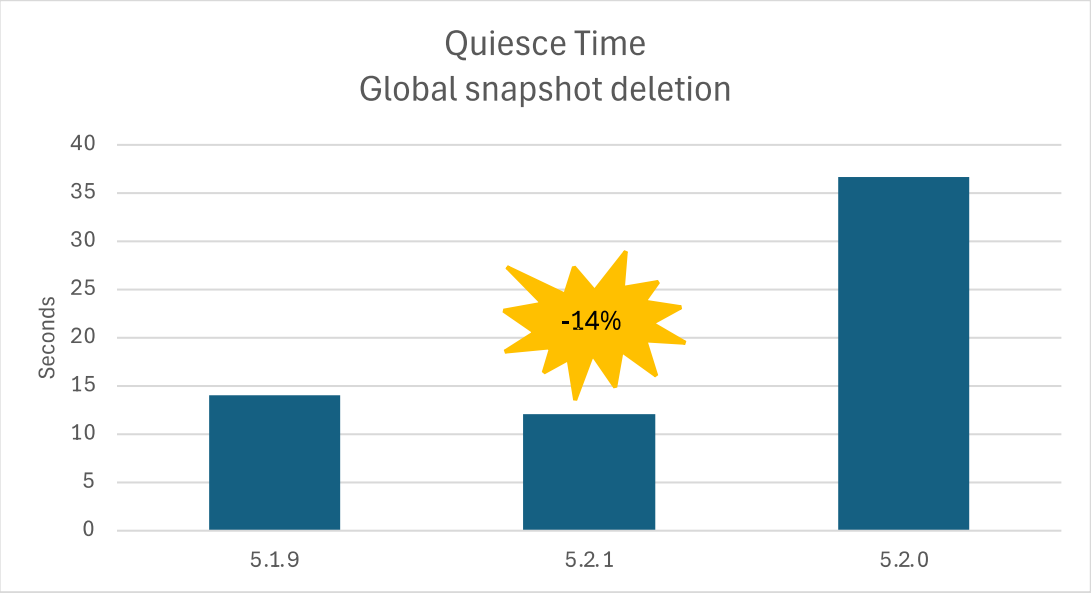
Read Append
Performance Improvement for small IO transfers

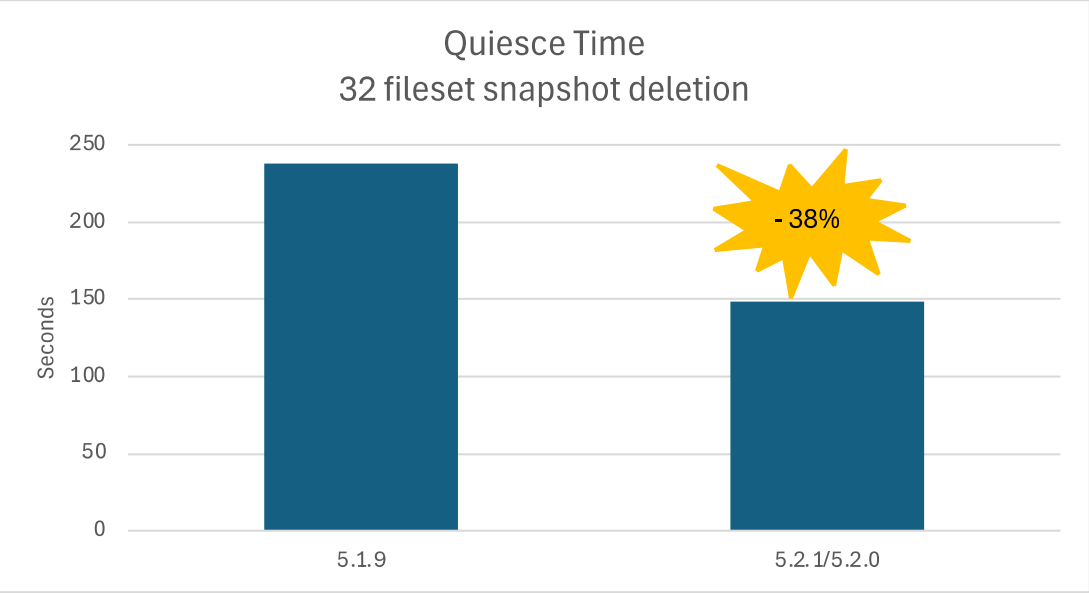
Time to read 20GB File – IO Size 8KB

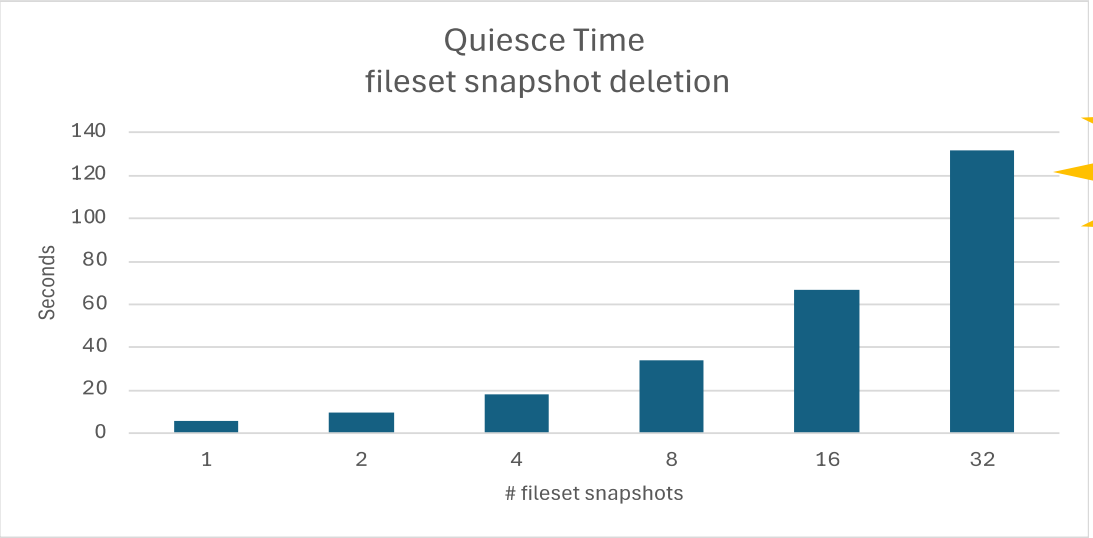
GPFS v5.1.9	GPFS v5.2.0
05min 55sec	00min 6.18sec

Performance improvement for cached objects cleanup for snapshot delete

Storage Scale







Linear batched deletions of mmdelsnapshot

Online Filesystem Check Updates!
mmfsckx improvement
Mitigate inode update slowness during online fsckx scanning on a large inode 0 files

Storage Scale

Improvement Summary:
Reduce impact to inode update performance during mmfsckx on a filesystem with large inode 0 files

Performance evaluation:
Evaluate file create/delete rate during mmfsckx
Monitor mmfsckx time

Configuration:
- x10 x86 client nodes
- ESS3500 performance model with 24 NVMe Drives
- 200Gib IB with RDMA

- 1B inode allocated
  fileset with 1M 1024KB files in 10 subdirectories, 4000 10GB files

Performance Results

	mmfsckx (min)	File create KIOPs	File delete KIOPs
R5.2.0	13		
		65	78
	13	IO aborted after being paused > 3m	
R5.2.1	14		
		66	78
	14	51	72

- mmfsckx time remains equivalent
- File create/delete was impacted slightly with mmfsckx

Abstraction - Data caching and tiering

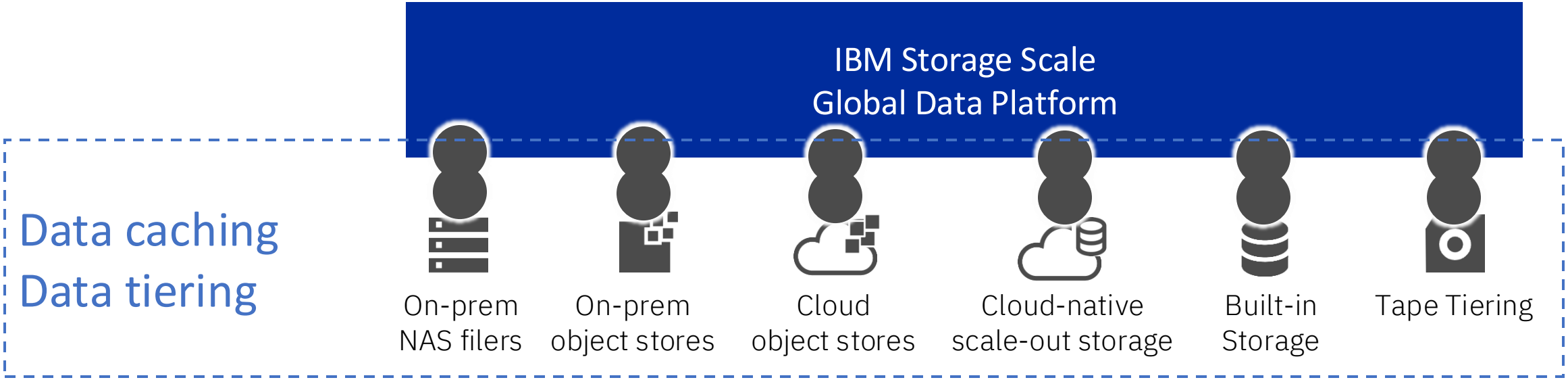
Storage Scale

IBM Storage Scale enables data caching and tiering with the following features.

Hierarchical Storage Management (HSM)

Active File Management (AFM)

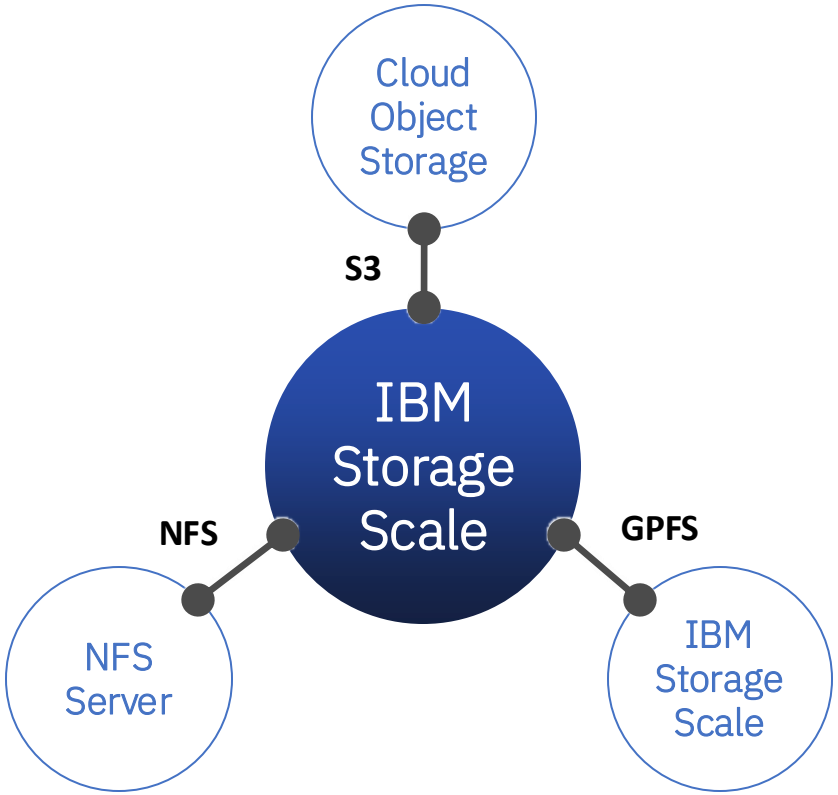
Policy-based Information Lifecycle Management (ILM)



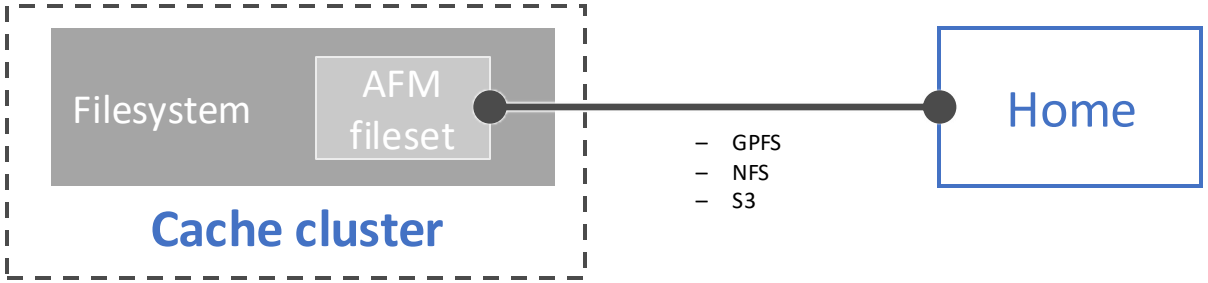
Active File Management (AFM) overview

Storage Scale

Active File Management enables caching data across other data sources.



- Each AFM fileset has a distinct set of AFM attributes.
- An IBM Storage Scale cluster that contains AFM filesets is called a cache cluster.
- A cache cluster has a relationship with another remote site called the home, where either the cache or the home can be the data source or destination.
- A cache cluster must be an IBM Storage Scale Cluster.
- A home can be an IBM Storage Scale, NFS server and Object Storage.

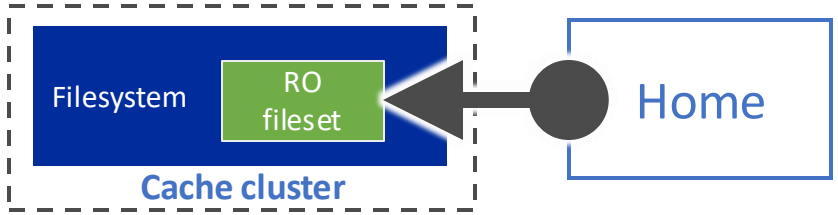


AFM Caching Mode

AFM has four caching modes.

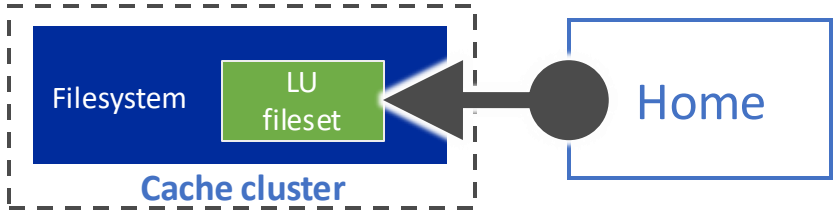
Read only (RO) mode

- Data in the cache is read only.
- Data source is the home and data destination is the cache.



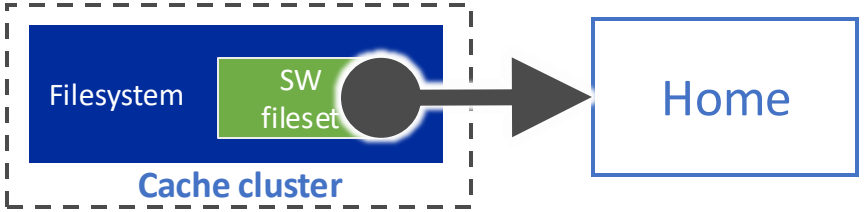
Local update (LU) mode

- Data in the cache can be read and written.
- Data which is created or modified in the cache is never updated by home.
- Data source is the home and data destination is the cache.



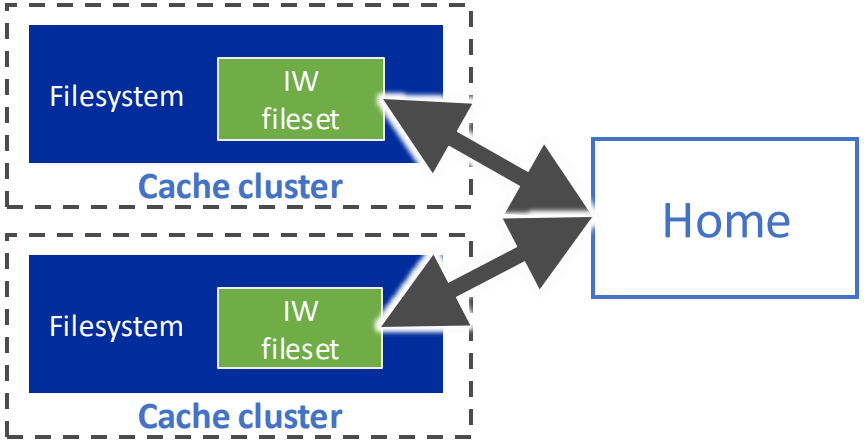
Single writer (SW) mode

- Data in the home should be read only.
- Data source is the cache and data destination is the home.



Independent writer (IW) mode

- Each cache reads from home and updates to the home independently of each other.
- Updates are propagated to the home in an asynchronous and can be delayed due to network.



AFM propagation

Storage Scale

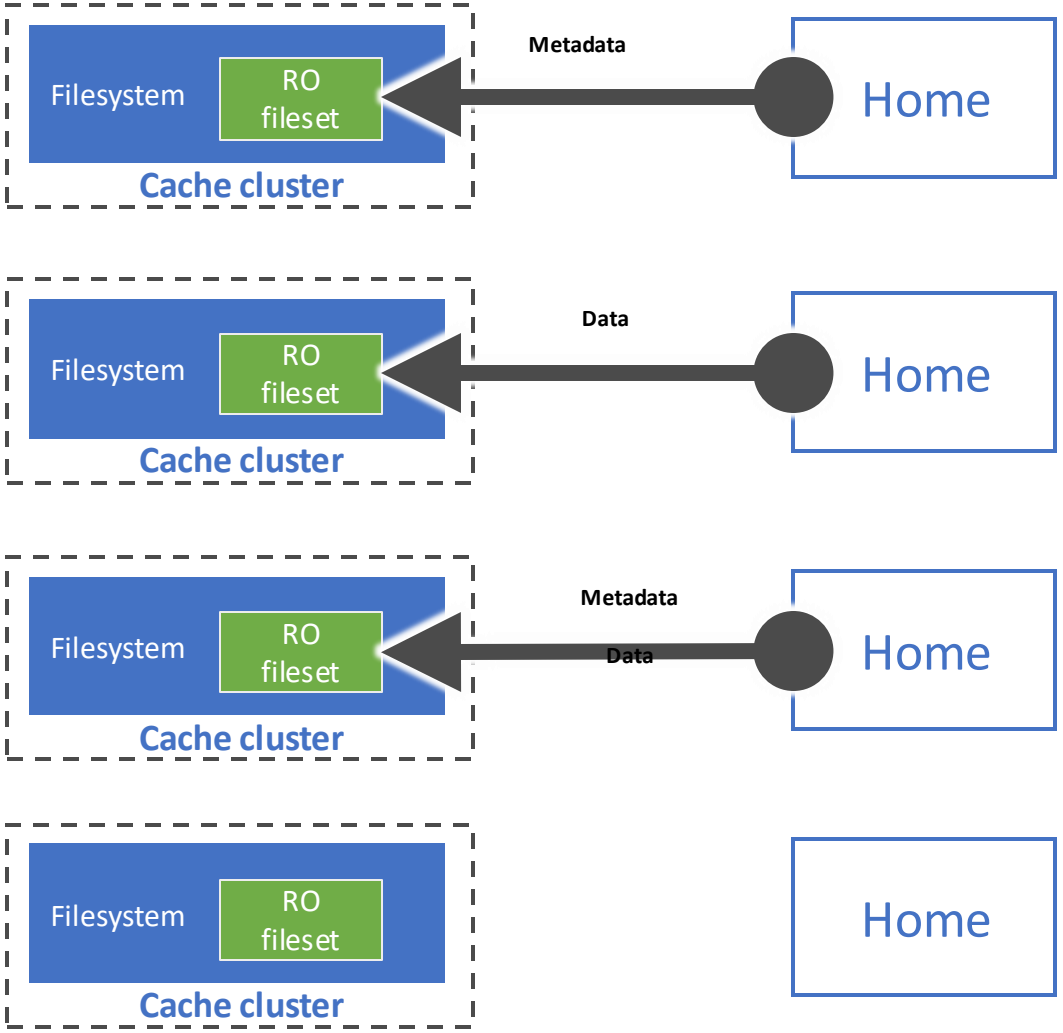
The metadata and data of the file is propagated when needed.

– Run the command, "ls" on the cache.
  Metadata is propagated from the home to the cache.

– Run the command, "cat" on the cache,
  Data is propagated from the home to the cache.

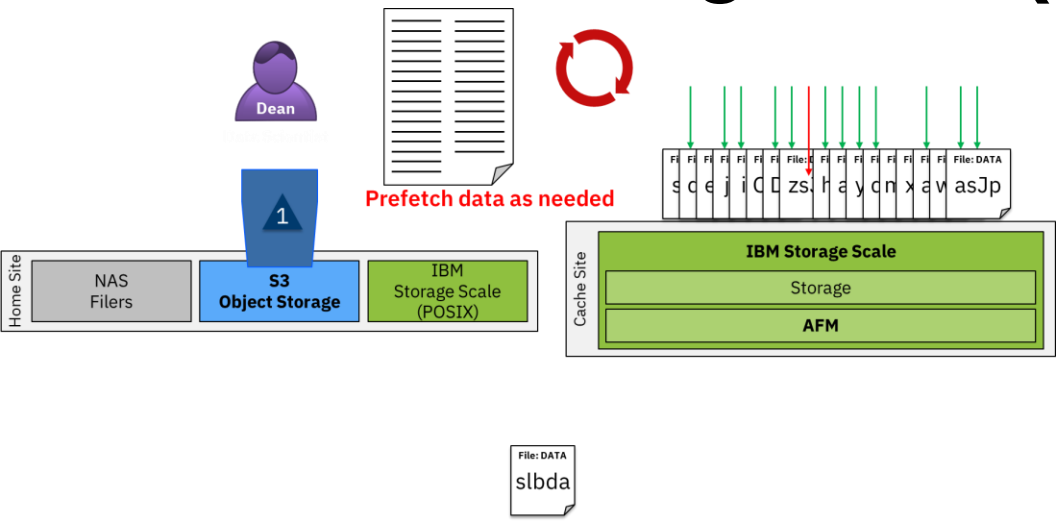
– Run the Scale command, "prefetch" on the cache,
  All metadata and data is propagated to the cache.

– Run the Scale command, "evict" on the cache
  Cached data is cleared in the cache.



Abstraction and Acceleration Services – Active File Management (AFM)

Storage Scale



Provide option to delete objects using Non-MDS gateway

Inode eviction from AFM cache

Simplification of migration commands improves user experience. Once command will ensure all critical steps are completed internally.

Validation Tool using REST API and Certification Tool using: mmafmtransfer

Migrate TCT enabled fileset to AFM-S3 MU (Tiering only)

New Command – reduced from 6 to 1 command!

#mmafmctl gpfs11 startCutover -j ro1

Check Status

#mmafmctl fs1 checkUncached -j ro1 --check-unmigrated [--dirpath /gpfs/fs1/ro1/data1]

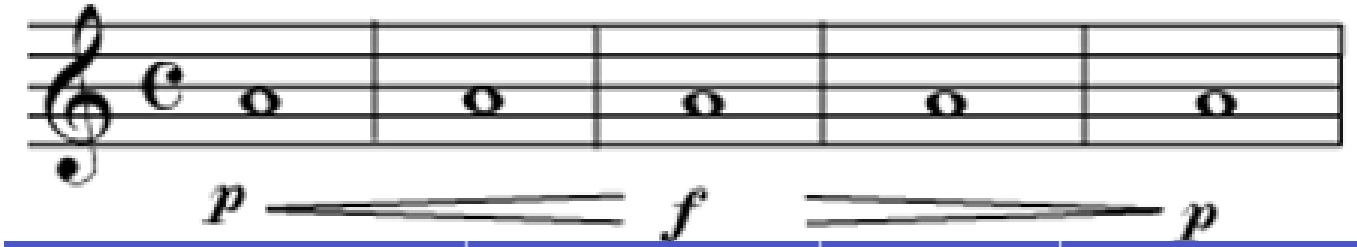
Abstraction and Acceleration Services – Dynamic Page Pool

Storage Scale

Dynamic workload management!

Scale detects a shortage of the pagepool memory, then attempts to increase the pagepool size.

When the Linux kernel detects the memory pressure, it requests Scale to shrink the size of the pagepool.



Config parameter	Allowed values	Default	Description
dynamicPagepoolEnabled	yes/no	no	Enable dynamic pagepool vs. static pagepool
pagepoolMinPhysMemPct	1-50	5	Minimum size of dynamic pagepool as percentage of physical memory.
PagepoolMaxPhysMemPct	10-90	75	Maximum size of dynamic pagepool as percentage of physical memory.
pagepoolChangeGracePeriod	1-86400	10	The grace period for growing the dynamic pagepool, in seconds. The dynamic pagepool grows only once every grace period

Configuration:

mmchconfig dynamicPagepoolEnabled=yes –N node1

mmchconfig pagepool=default –N node1

mmshutdown –N node1

mmstartup –N node1

mmdiag –pagepool

GPFSBufMgr monitor pagepool size via zimon

Default configuration changes with 5.2.N

Storage Scale

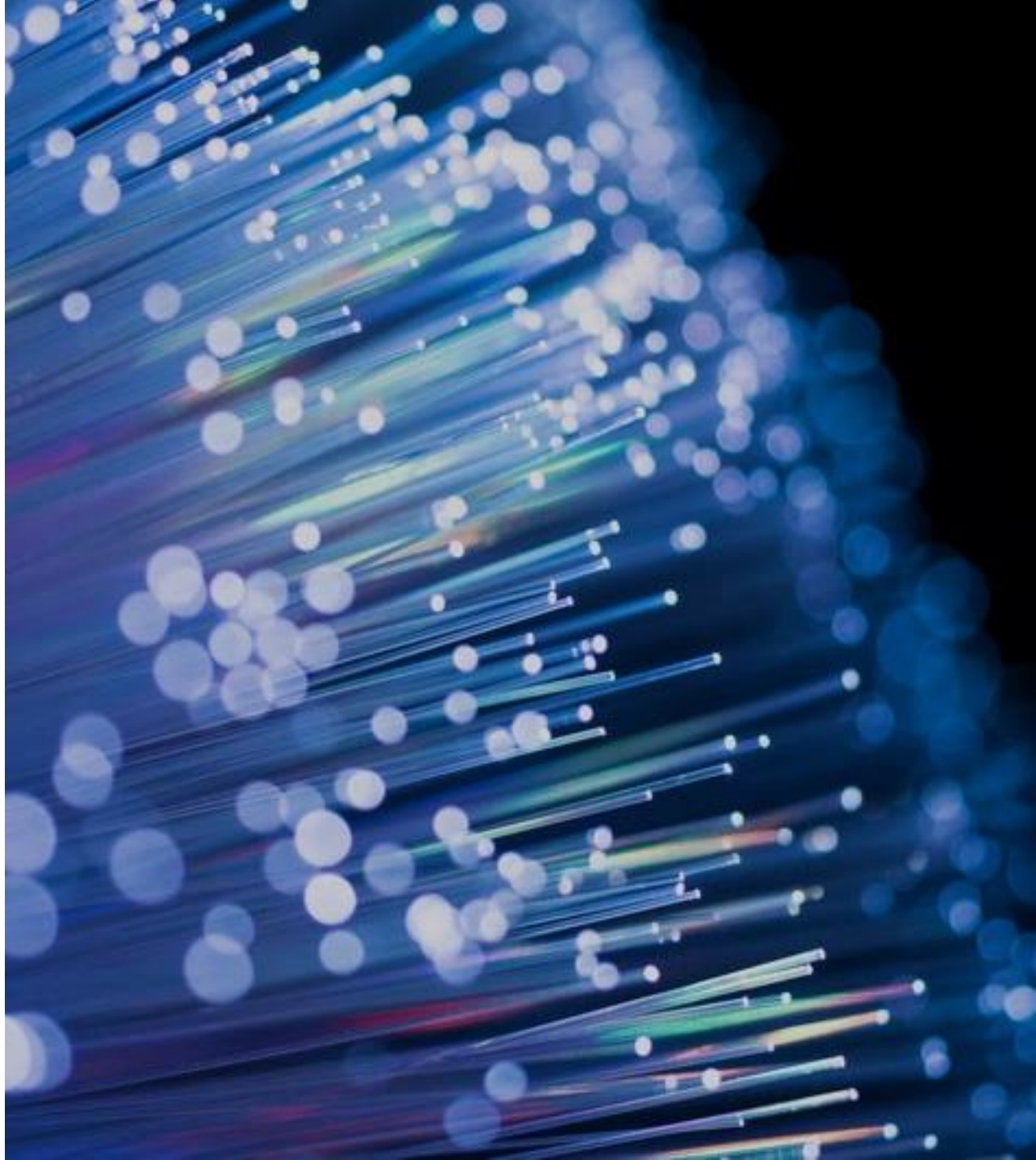
Provide better out-of-the-box performance for a wide variety of workloads.

Apply only for new 5.2.0 clusters. Do not apply for existing clusters, even with a 5.2.0 upgrade.

The new defaults are described in the mmchconfig man page!

config option	old default	new default
numaMemoryInterleave	no	yes
workerThreads	48	256
page pool	min(1G, 1/3 system mem)	min(4G, 1/3 system mem)
ignorePrefetchLUNCount	no	yes
dioRentryThreshold (undocumented)	0	1

Abstraction - Management and Orchestration



mmpstat - Live Performance Monitoring Tool

Storage Scale

- **Goal**: Faster identification of infrastructure bottlenecks
- **Findings**:
  - Existing command line tools (e.g. mmperfmon query) are designed to view historical data but they are clumsy when inspecting the current system performance
  - System administrators use other system tools and do not benefit from Zimon performance data
- **Solution**:
  - Inspired by Linux system monitoring tools (like iostat, vmstat, dstat) the new mmpstat command has been developed
    - show the current values of any Zimon metric in a table format
    - refresh/append new measurements to the table on a given interval
    - makes it easy to spot any changes to values to find slow downs and bottlenecks
    - Easy to compare values of different nodes, filesystems, nsds, etc.
- **CDM favorite command:**
  - Find some sort of equivalent command for:
  - **# dstat --noupdate --time --top-cpu --top-mem --top-io --top-bio --gpfs --gpfs-ops**

mmpstat - Live Performance Monitoring Tool

Storage Scale

Examples:

```
#> mmpstat cpu_system -N all
-----------------------------------
        |    testvm2|     testvm|
        |           |           |
Timestamp |  cpu_system|  cpu_system|
-----------------------------------
18:19:05        73.40|       0.00|
18:19:06        50.50|       1.00|
18:19:07        84.00|       0.51|
18:19:08        60.51|       1.00|
```

```
# mmpstat mem_memfree -N all
-----------------------------------
        |    testvm2|     testvm|
        |           |           |
Timestamp |  mem_memfree|  mem_memfree|
-----------------------------------
18:31:37       475812|      445032|
18:31:38       355220|      445032|
18:31:39       235819|      445032|
18:31:40       275812|      445032|
```

```
#> mmpstat netErrors --filter=netdev_name=ens3
-----------------------------------------------------------------------------------------------------
        |         testvm2|         testvm2|         testvm2|         testvm2|         testvm2|
        |            ens3|            ens3|            ens3|            ens3|            ens3|
Timestamp |  netdev_collisions|    netdev_drops_r|    netdev_drops_s|    netdev_errors_r|    netdev_errors_s|
-----------------------------------------------------------------------------------------------------
18:20:16                0|              0|              0|              0|              0|
18:20:17                0|             20|              0|              0|              0|
18:20:19                0|           2230|              0|              0|              0|
18:20:20                0|             10|              0|              0|              0|
```

```
#> mmpstat  "gpfs_pdds_bytes_written" --filter "gpfs_disk_name=.*e1s13.*" -b 10
-----------------------------------------------------------
        |             ess5kio2|             ess5kio2|
        |  ess5kio2::e1s13/path000|  ess5kio2::e1s13/path001|
Timestamp |  gpfs_pdds_bytes_written|  gpfs_pdds_bytes_written|
-----------------------------------------------------------
08:46:30             1212220|             4434344|
08:46:40                   0|                   0|
08:46:50             4444330|               23320|
08:46:50             4947330|               35420|
```

<u>Hint</u>:

Run „**mmpstat +**“ to print out a list of all known metrics

mmptop: top entities for a perfmon metric

Storage Scale

New command: mmptop

- Shows the top entities with the highest values (like Linux top command)
- Works for any perfmon metric!

- E.g.: show the nodes with the highest CPU usage

```
#> mmptop cpu -N all
2024-09-13 13:56:34.284255

                                  ---------------------------------
Key                          |        cpu_system |          cpu_user |       cpu_contexts |
-----------------------------------------------------------------
ess5kio1 | ess5kio2 |                     0.27 |              0.55 |               1943 |
ems5k.mmfsd.net |                        0.08 |               0.3 |               1501 |
                                           0.0 |               0.1 |               1450 |
```

- E.g.: show pdisks with the highest IO wait time

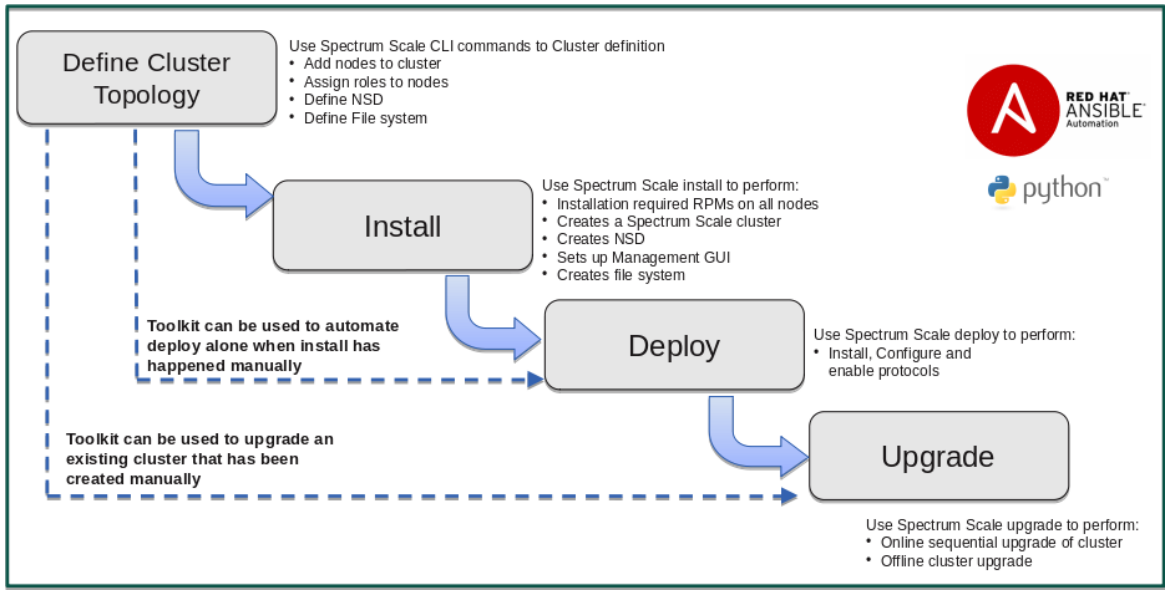
```
#> mmptop gpfs_nsdds_max_disk_wait_wr,gpfs_nsdds_max_disk_wait_rd 2024-07-10 10:40:50

Key ------------------------------------------------------------------------ | gpfs_nsdds_max_disk_wait_wr | gpfs_nsdds_max_disk_wait_rd | ----------------------

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
just6nsd09b.just | GPFSNSDDisk | RG002LG002VS007                                                                                                                 24.14 |                                0.0 |
just6nsd09a.just | GPFSNSDDisk | RG002LG001VS002                                                                                                                  1.38 |                                0.0 |
just6nsd09a.just | GPFSNSDDisk | RG002LG001VS012                                                                                                                   0.0 |                                0.0 |
just6nsd09a.just | GPFSNSDDisk | RG002LG001VS022                                                                                                                   0.0 |                                0.0 |
just6nsd09a.just | GPFSNSDDisk | RG002LG001VS023                                                                                                                   0.0 |                                0.0 |
just6nsd09a.just | GPFSNSDDisk | RG002LG001VS025                                                                                                                   0.0 |                                0.0 |
```

Storage Scale Deployment Toolkit

Storage Scale



[**CES S3**] CES S3 based Object protocol toolkit support for X86.

[**CES S3**] CES S3 based Object protocol toolkit support for PPC64LE.

[**CES S3**] CES S3 based Object protocol toolkit support for S390X.

[**Native Rest API Tech Preview**] Toolkit supported features for Native Rest API Tech Preview Deployment

[**Toolkit Arm**] Toolkit extended features support on ARM .

[**Python**] Smart Installer Revolutionises Python management, Automatically utilises Latest python installed version without user configuration.

[**Currency**] Extended OS currency

[**ESS Protocol Node**] ESS Protocol node certification with 5.2.1 Toolkit.

[**ECE**] ECE install toolkit enhancement to support in config populate with vdiskset in multi-DA and file system with multiple vdisk sets

[**Cloud**] NFS & SMB support for Cloud-Kit.

[**Cloud**] CES S3 Support for Cloud-Kit.

[**Open Source**] Open Source Ansible Role certification with 5.2.1.

[**Documentation**] Ansible tuning config for deployment consideration

Abstracting Cloud Service Deployment – Cloudkit!

Storage Scale

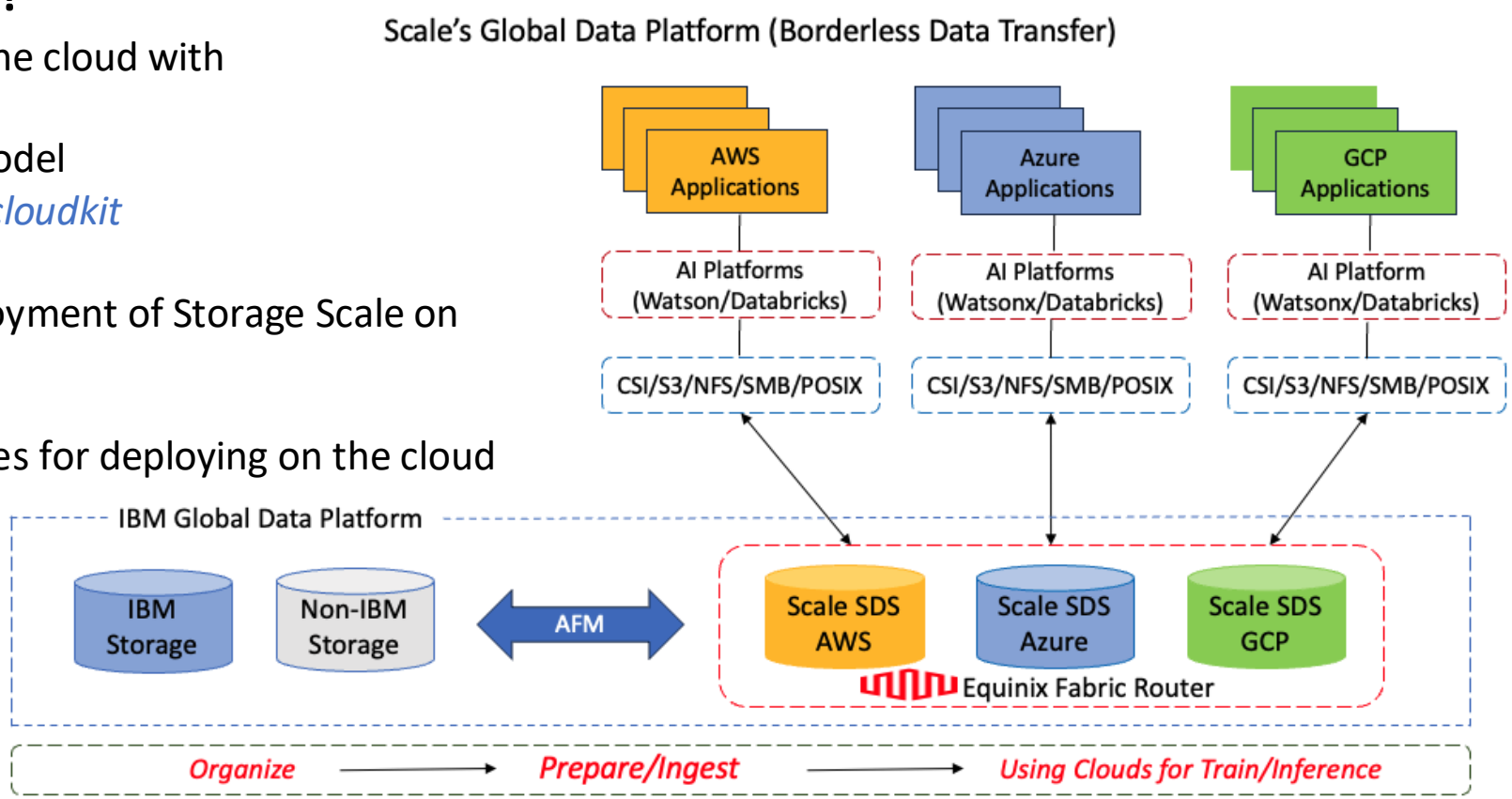
What is Storage Scale Cloudkit?

Create Storage Scale clusters on the cloud with

Bring Your Own License (BYOL) Model
Look in */usr/lpp/mmfs/VERSION/cloudkit*

Automates provisioning and deployment of Storage Scale on the cloud

Applies Storage Scale best practices for deploying on the cloud

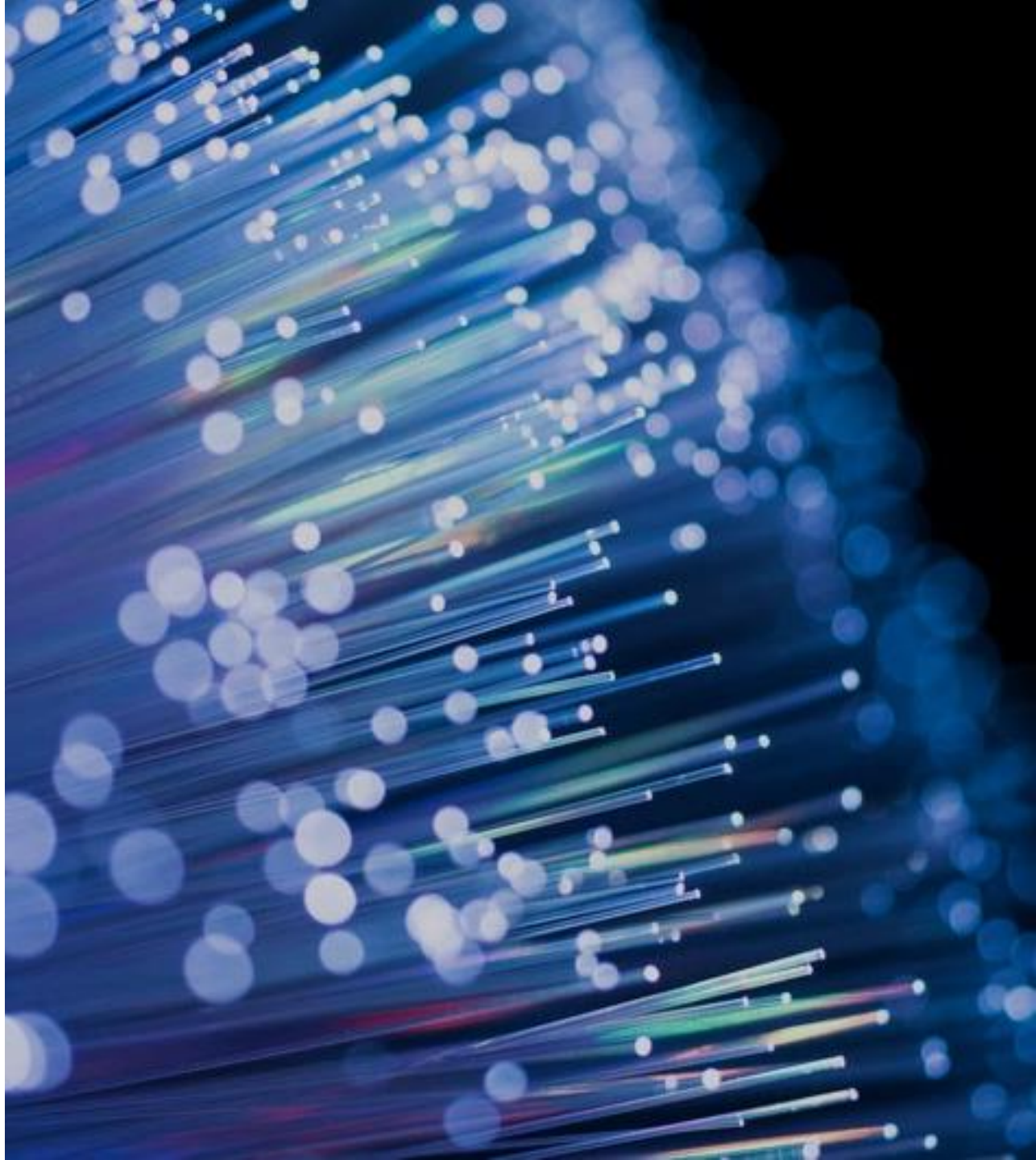


Advantages

Support for major public clouds Amazon (AWS) and Google (GCP)

- AFM-COS, Upgrades
- Tech-preview support for fleet support on AWS and GCP cluster instance

Assurance Services



Scale control plane and security architecture modernization

Security Improvements

Removal of SSH dependency

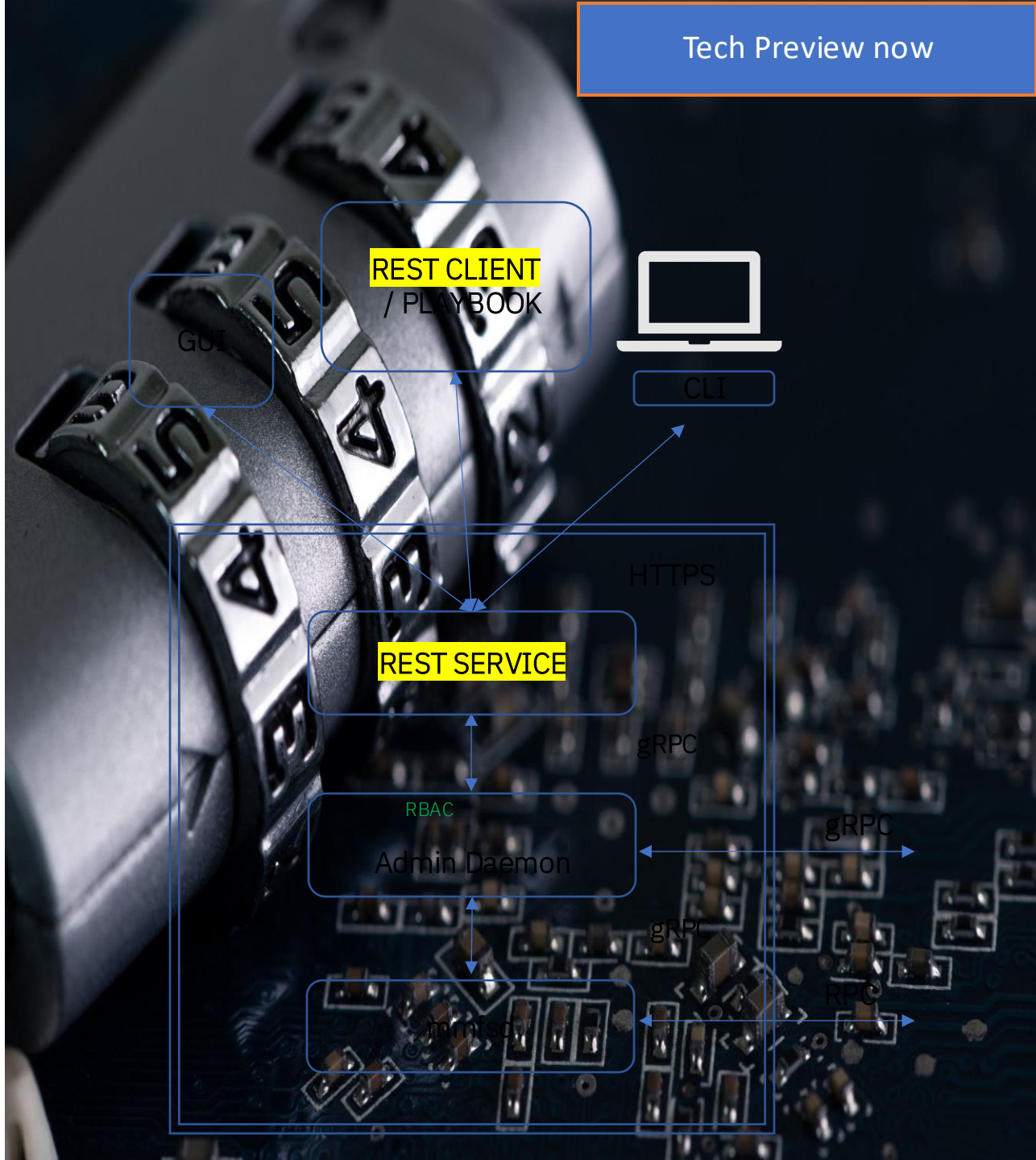
Removal of root requirement for control plane

Remote Administration

Fine-Grained Role Based Access Control
Declarative policy rules based on
Open Policy Agent

Control Plane Designed For Applications / Operators

Retain CLI for human management



File Audit Logging

- Lightweig
- All filesyst
- Audit logs
- Events for analysis o
- Fully-conf

Event Name	Description	Examples
CLOSEWRITE	Open for write access then closed	easier to determine when files have been modified on a file system or fileset
ACCESS_DENIED	A user was denied access to operate on a file.	open() with O_WRONLY where user has no write permission.
ACLCHANGE	A file's or directory's ACL permissions were modified.	mmputacl, chown, chgrp, chmod
CLOSE	A file was closed.	close(), cp, touch, echo, policy MIGRATE rule.
CREATE	A file or directory was created.	open(create flag), vi, ln, dd, mkdir
GPFSATTRCHANGE	A file's or directory's IBM Storage Scale attributes were changed.	mmchattr -i -e --indefinite-retention
OPEN	A file or directory was opened for reading, writing, or creation.	open(), mmlsattr, cat, cksum, ls (only for directories), policy LIST rule
RENAME	A file or directory was renamed.	rename(), mv
RMDIR	A directory was removed.	rmdir(), rm, rmdir
UNLINK	A file or directory was unlinked from its parent directory. When the linkcount = 0, the file is deleted.	unlink(), rm hardlink/softlink
XATTRCHANGE	A file's or directory's extended attributes were changed.	mmchattr --set-attr --delete-attr

Storage Scale
Storage Scale

Nodes
file IO
client IO

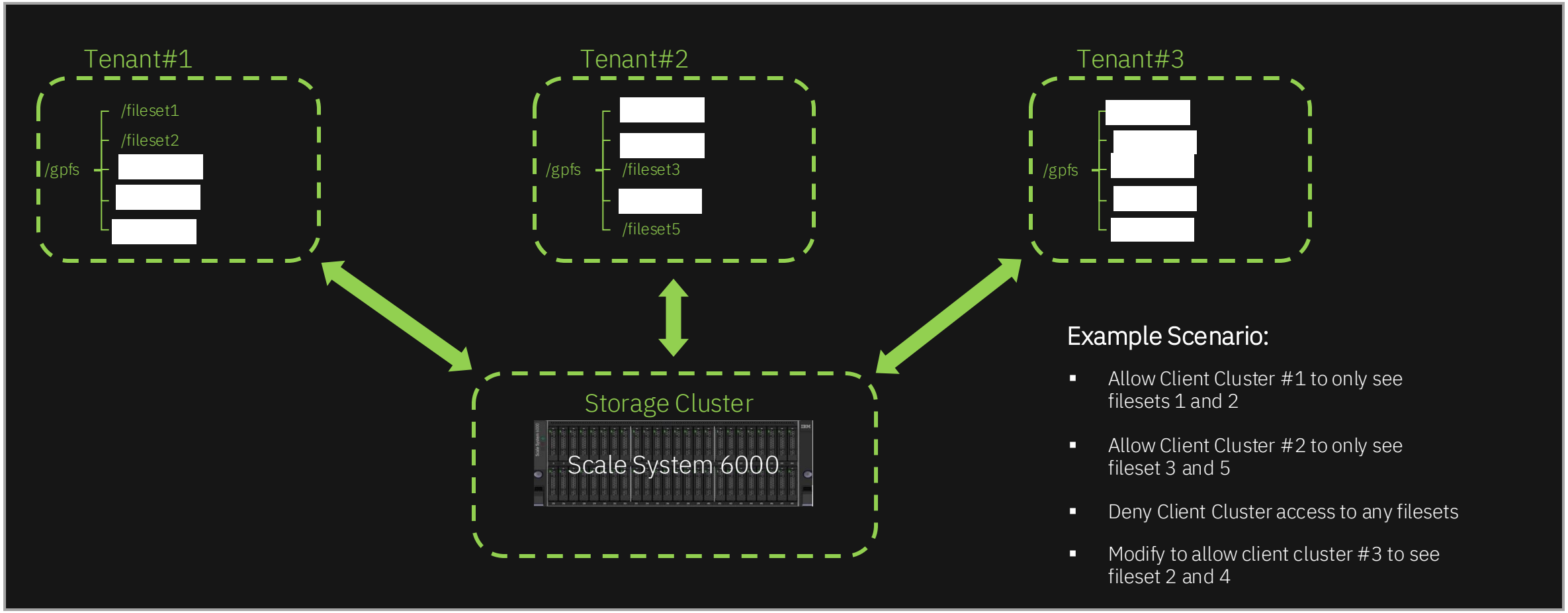
unk>

Radar®

orwarded to
gram

Remote Fileset Access Control

- Provides multi-tenancy capabilities for remote client clusters
- Define which remote clusters can see which filesets within a single filesystem namespace
- Dynamic ability to grant or deny fileset access to a remote cluster using *mmauth* allow or deny command
- Quotas and snapshots will only be visible for the authorized filesets, not all filesets within a filesystem



Example Scenario:

- Allow Client Cluster #1 to only see filesets 1 and 2
- Allow Client Cluster #2 to only see fileset 3 and 5
- Deny Client Cluster access to any filesets
- Modify to allow client cluster #3 to see fileset 2 and 4

Storage Scale

What’s new in

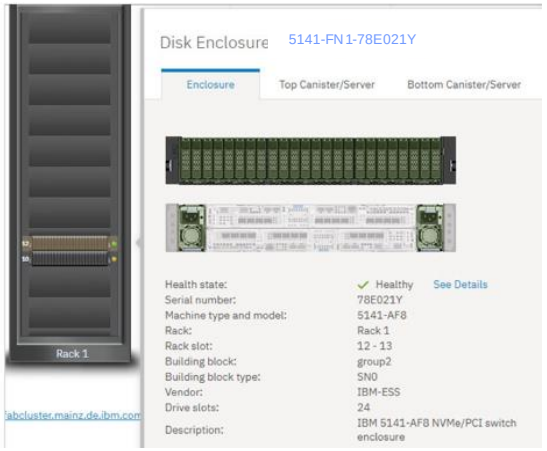
IBM Storage Scale System 6.2.*

Chris Maestas
IBM CTO, IBM Data and AI Storage Solutions
Chief Troublemaking Officer

IBM

IBM Storage Scale System

Integrated scale-out data management
for file and object



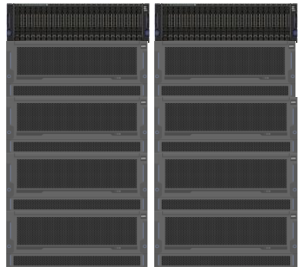
Simple GUI and wizards

Optimal building block for high-performance, scalable, reliable enterprise Storage Scale storage

- Faster data access with the choice to scale-up or scale-out
- Easy to deploy clusters with unified system GUI
- Simplified storage administration with IBM Storage Control integration

6000

3500



3500 or 6000
Hybrid/Capactity

One solution for all Storage Scale data needs

- Single repository of data with unified file and object support
- Anywhere access with multi-protocol support using protocol nodes: NFS 3/4.0/4.1, SMB, Object, and HDFS
- Ideal for big data analytics including full Hadoop transparency

Ready for business-critical data

- Disaster recovery with synchronous or asynchronous replication
- Ensure reliability and fast rebuild times using Storage Scale RAID’s dispersed data and erasure code
- Six 9s (99.9999%) of availability and online scalability and upgrades

Scale System models are built for speed and capacity

Storage Scale

Speed

Capacity

Hybrid

IBM Storage Scale System 6000

4U48 Enclosure
24 or 48 NVMe drives

Performance
Machine type 5149-F48

Up to 310+GB/s
IOR 100% read, InfiniBand

IBM Storage Scale System 3500

2U24 Enclosure
12 or 24 NVMe drives

Performance
Machine type 5141-FN2

Up to 125 GB/s
IOR 100% read, InfiniBand

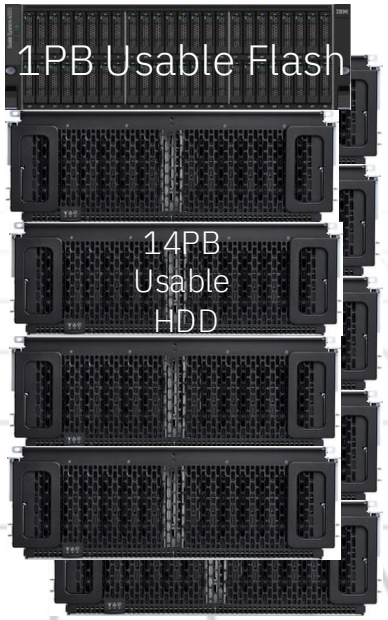
IBM Storage Scale System 3500 Hybrid (NVMe + HDD)

Up to 8x 4U102 JBOD

3500 Hx
Machine type 5141-FN2

Up to 91 GB/s – NVMe
IOR 100% read InfiniBand
Up to 48 GB/s – HDD
IOR 100% read InfiniBand

IBM Storage Scale System 6000 Hybrid (NVMe + HDD)



Up to 9x 4U91 JBOD

6000 Hx
Machine type 5149-F48

Up to 280 GB/s – NVMe
IOR 100% read InfiniBand
Up to 100 GB/s – HDD
IOR 100% read InfiniBand

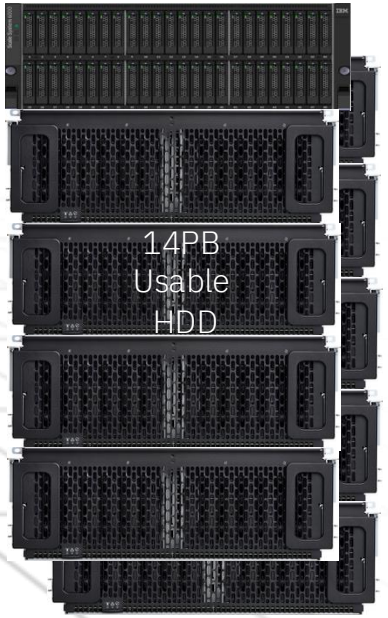
IBM Storage Scale System 3500 Capacity (HDD-Only)

Up to 8x 4U102 JBOD

3500 Cx
Machine types
5141-FN2

Up to 48 GB/s
IOR 100% read C4+ model, InfiniBand

IBM Storage Scale System 6000 Capacity (HDD-Only)



Up to 9x 4U91 JBOD

6000 Cx
Machine types
5141-FN2

Up to 90 GB/s
IOR 100% read C9 model, InfiniBand

New Scale System Software Features

IBM Storage & SDI
Storage Scale

Scale 5220
RH9.4 (UN / 6000)
Gen5 Samsung drives
Red fish script support for ESS 6000/UN (read-only)
mmptop (live CPU/memory info)
Improved call home ticket lifecycle management
S3 support on utility node (non-protocol VM (3500/6000)
Firmware updates for 6K/4u91
MES for Falcon/HBA (1-9 enclosures) ESS 6000 4u91
MES 24 to 48 FCM drive support
Protocol VM ESS 6K (POC)
SED support with TPM (no key server) ESS 6K
ESS HW metrics in Zimon

ESS 6000 All Supported Configurations

Storage Scale

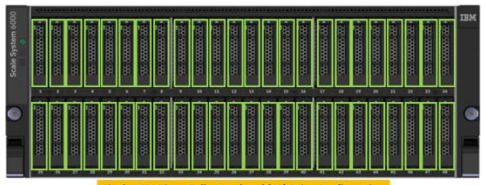
Performance Model : NVMe

Single NVMe Tier + Half Populated (24) Drive Configuration

Single NVMe Tier + Fully Populated (48) Drive Configuration

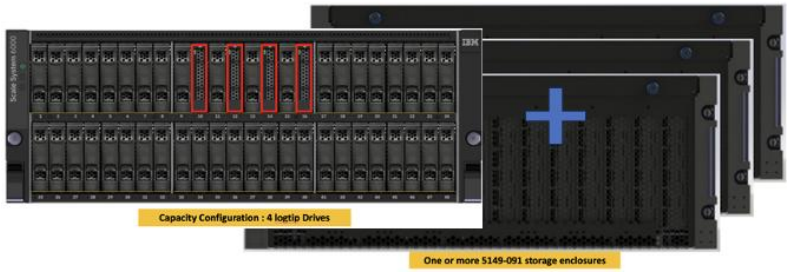
Performance Model : FCM

Single FCM Tier + Half Populated (24) Drive Configuration



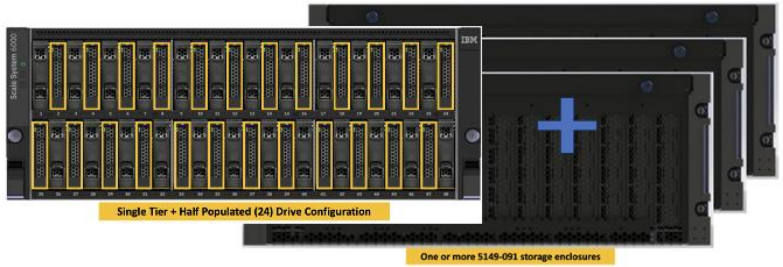
Single FCM Tier + Fully Populated (48) Drive Configuration

Capacity Configuration 4 logtip Drives – No Tiering

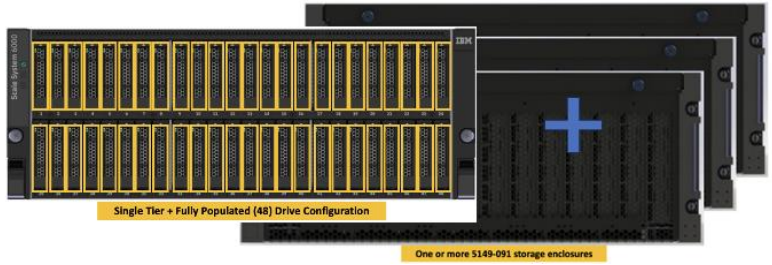


Hybrid Model : NVMe

H+S+24: Hybrid + Single NVMe Tier + Half Populated (24) Drive Configuration

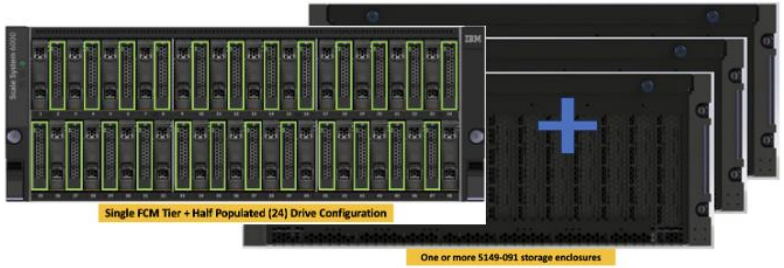


H+S+48: Hybrid + Single NVMe Tier + Fully Populated (48) Drive Configuration

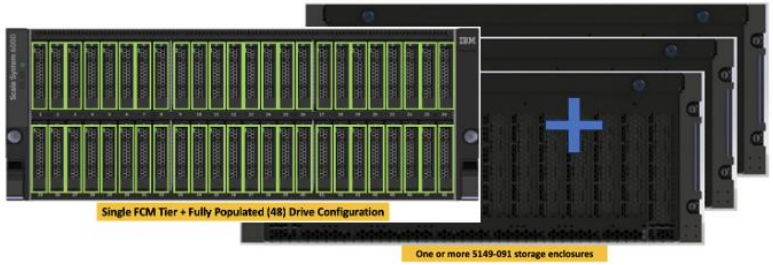


Hybrid Model : FCM

H+S+24: Hybrid + Single FCM Tier + Half Populated (24) Drive Configuration



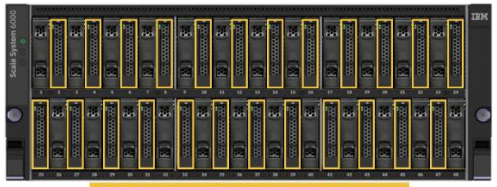
H+S+48: Hybrid + Single FCM Tier + Fully Populated (48) Drive Configuration



6K NVME MES (24 to 48 drives) 5149-F48 Flash Storage MES

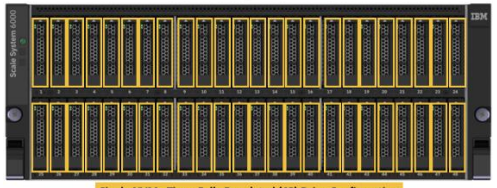
Storage Scale

Performance Model: NVMe MES



Single NVMe Tier + Half Populated (24) Drive Configuration

Installing single drive type — NVMe MES: 24 -> 48



Single NVMe Tier + Fully Populated (48) Drive Configuration

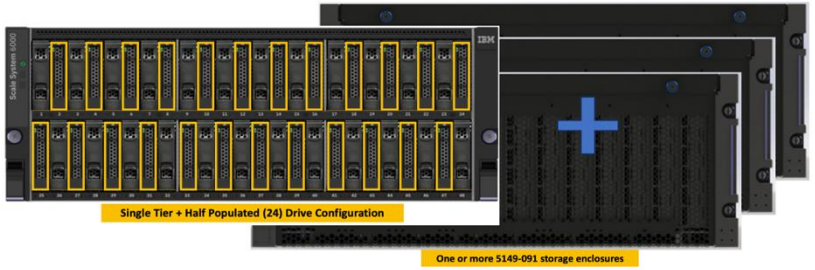
Performance Model: FCM MES

Single FCM Tier + Half Populated (24) Drive Configuration

Installing single drive type — FCM MES: 24 -> 48

Single FCM Tier + Fully Populated (48) Drive Configuration

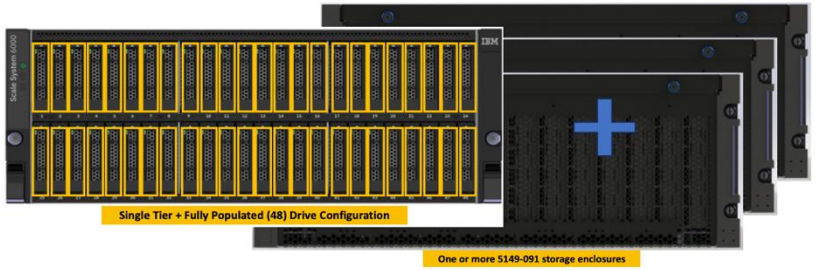
Hybrid Mode: NVMe MES



Single Tier + Half Populated (24) Drive Configuration

One or more 5149-091 storage enclosures

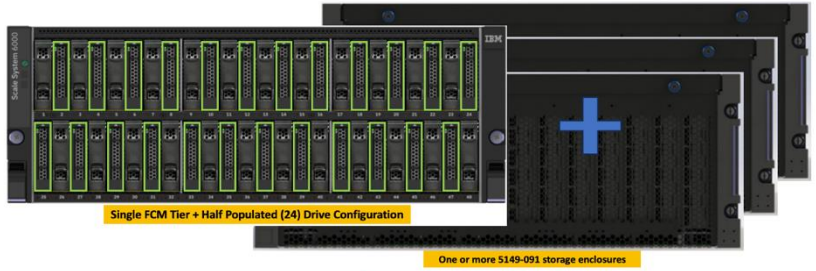
Installing single drive type — NVMe MES: 24 -> 48



Single Tier + Fully Populated (48) Drive Configuration

One or more 5149-091 storage enclosures

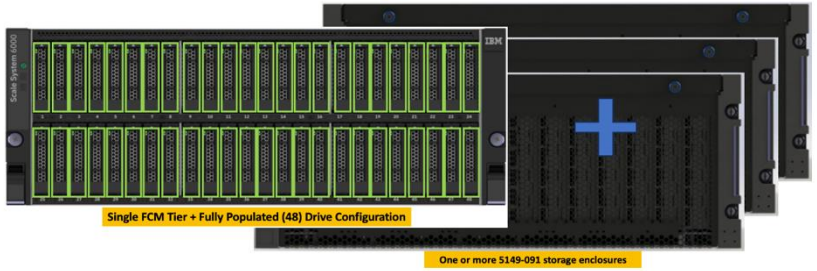
Hybrid Model: FCM MES



Single FCM Tier + Half Populated (24) Drive Configuration

One or more 5149-091 storage enclosures

Installing single drive type — FCM MES: 24 -> 48



Single FCM Tier + Fully Populated (48) Drive Configuration

One or more 5149-091 storage enclosures

Key considerations for storage enclosure MES

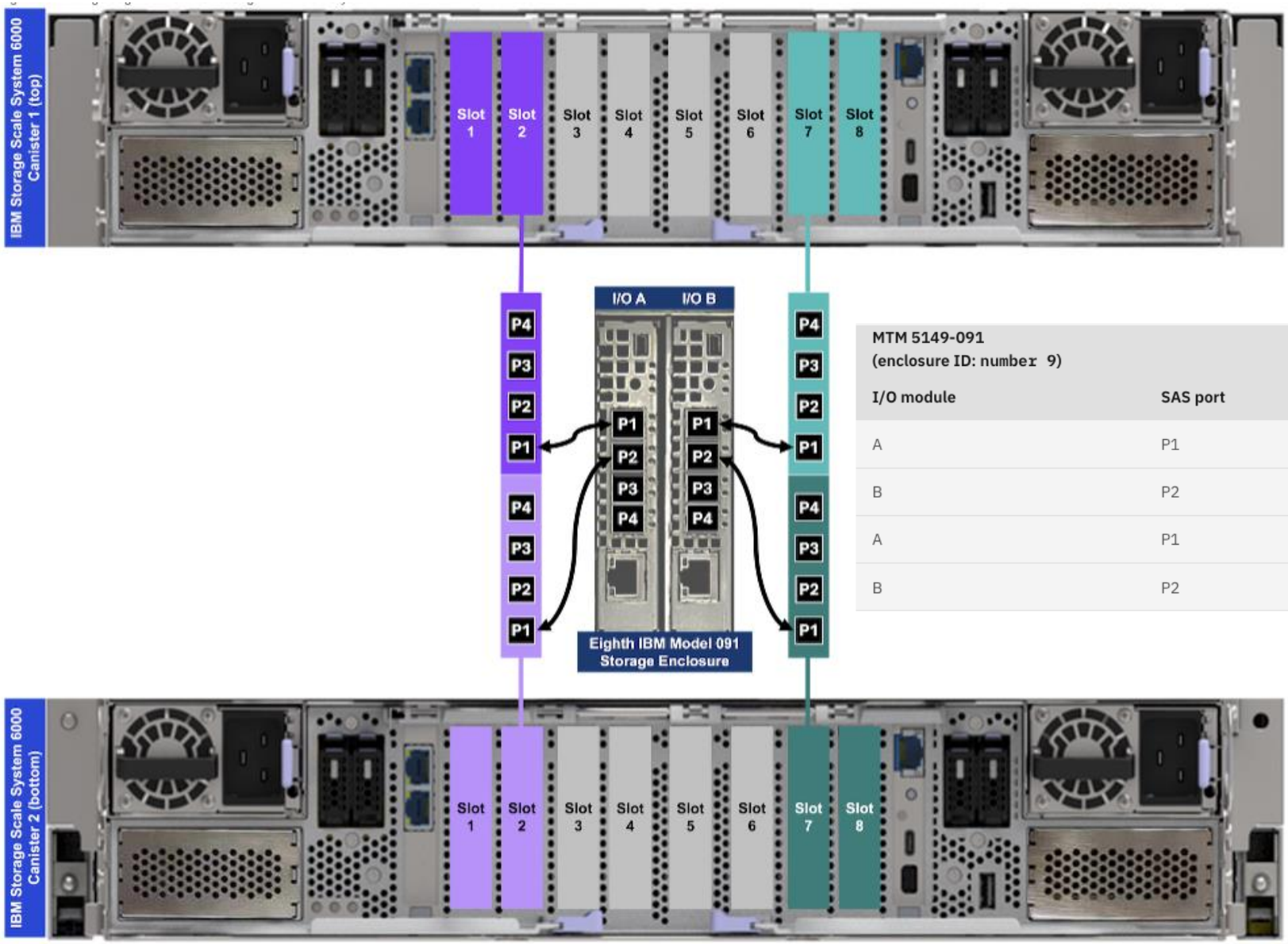
Storage Scale

- Storage Enclosure MES MAY require the following MES to be performed ***before*** adding storage enclosures
  - Server Memory MES
  - Server Adapter MES

- What are the right order to complete different kinds of MES as mentioned above?
  1. Complete the MES steps required in Server Node (ISS 6000) first
  2. Follow the Incremental Upgrade map to add storage enclosures **IN INCREMENTS**

- How do IBM Service handle the Incremental MES requirements in practice?

- For a customer with an IBM Storage Scale System 6000 building blocking with 3 4U91s to upgrade to 8 4U91s, how many incremental MES steps are required to complete the entire MES?



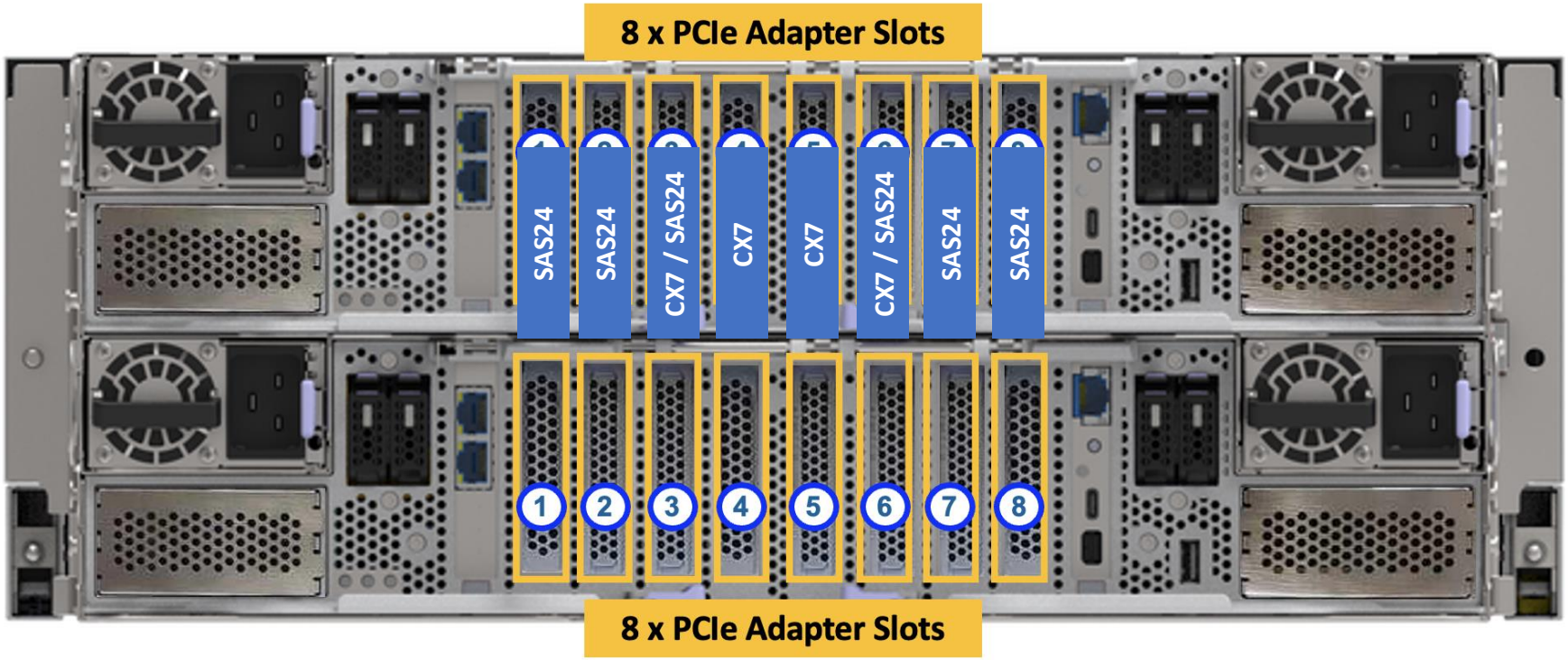
MTM 5149-091 (enclosure ID: number 9)		IBM Storage Scale System 6000 (enclosure ID: number 1)		
I/O module	SAS port	Canister	PCIe slot	SAS port
A	P1	1 (top)	Slot 2	P1 (C0)
B	P2		Slot 7	P1 (C0)
A	P1	2 (bottom)	Slot 2	P1 (C0)
B	P2		Slot 7	P1 (C0)

Storage Scale

MFG Built Config	Drive MES (Server Node)	Memory MES (Server Node)	Adapter MES (Server Node)	Storage Enclosure MES
Performance (24)	Performance (48)	optional	Host Attachment: optional	(P->H conversion)
Performance (48)	N/A	optional	Host Attachment: optional	(P->H conversion)
Capacity	N/A*	3-4 encls: optional 5-9 encls: mandatory	Host Attachment: optional 3-4 encls: 2 SAS mandatory 5-8 encls: 4 SAS mandatory 9 encls: 6 SAS mandatory	optional
Hybrid (24)	Hybrid (48)	3-4 encls: optional 5-9 encls: mandatory	Host Attachment: optional 3-4 encls: 2 SAS mandatory 5-8 encls: 4 SAS mandatory 9 encls: 6 SAS mandatory	optional
Hybrid (48)	N/A	3-4 encls: optional 5-9 encls: mandatory	Host Attachment: optional 3-4 encls: 2 SAS mandatory 5-8 encls: 4 SAS mandatory 9 encls: 6 SAS mandatory	optional

Adapter Placement

Storage Scale

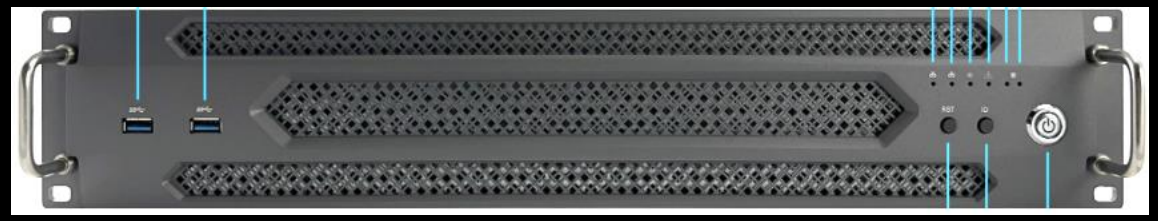


Adapter placement rules

Placement Priority	Feature Code	Description	1st pair	2nd pair	3rd pair	4th pair	5th pair	6th pair	7th pair	8th pair
1	AK18	Broadcom 9600-16e 24Gb SAS	Slot 1	Slot 8	Slot 2	Slot 7	Slot 6	Slot 3	N/A	N/A
2	AJQQ	CX-7 1-port 400Gb	Slot 4	Slot 5	Slot 3	Slot 6	N/A	N/A	N/A	N/A
3	AJQS	CX-7 2-port 200Gb	Slot 4	Slot 5	Slot 3	Slot 6	N/A	N/A	N/A	N/A

2U X86 Utility Node

All-purpose, powerful and fully integrated utility node, supporting multiple use cases and compatible with existing building blocks



Replaces existing power-based EMS and Protocol node and adds support for additional storage use cases

System Config

Processor: AMD EPYC (single/dual docket)

Memory: 128GB – 512GB

2x internal boot drives

High-Speed Network: 1-4 CX-6 adapters

1Gb/10Gb network

Versatility, Flexibility and Support for:

Management Server (EMS), GUI and Callhome

Protocol node functions

AFM gateway

GKLM (orderable via AAS)

IBM Storage Protect

Data Cataloging Service

IBM FlashCore™ Module 4

Capacity and Performance

2.5” dual ported U.2 NVMe Gen 4 PCIe
Industry leading density at 38.4 TB per drive
Inline hardware FIPS 140-3 encryption
Inline hardware 3:1 compression = 116 TB!

Internally tiered storage
-> MRAM -> SLC -> 3D QLC

Industry leading QLC endurance
15K Program/Erase cycles
Compared to 1500 for enterprise QLC

IBM Unique QLC management (100+ patents)
read calibration, heat binning, heal[illegible]inning,
error correcting codes, optimized voltage

Continuous health monitoring
keeps wear across all cells within 5%

Effective capacity depends on data type.
May be 3:1 or 2:1 if it compresses

Still on average data if we achieve 1.2-1.3 it’s about 45-50 TB per drive

Estimate your data via gzip/zip/lz4 or scale software compression and get the best idea of actual compression. Another tool is being tested (compressimator)

Non-compression - 35 GB/s write and 90 GB/s read
Compression testing shows 50 GB/s writes and 150 GB/s reads.

SED Support with GKLM : Overview

GA in Dec 2022 for ESS 6.1.5.0 (ESS 3500)

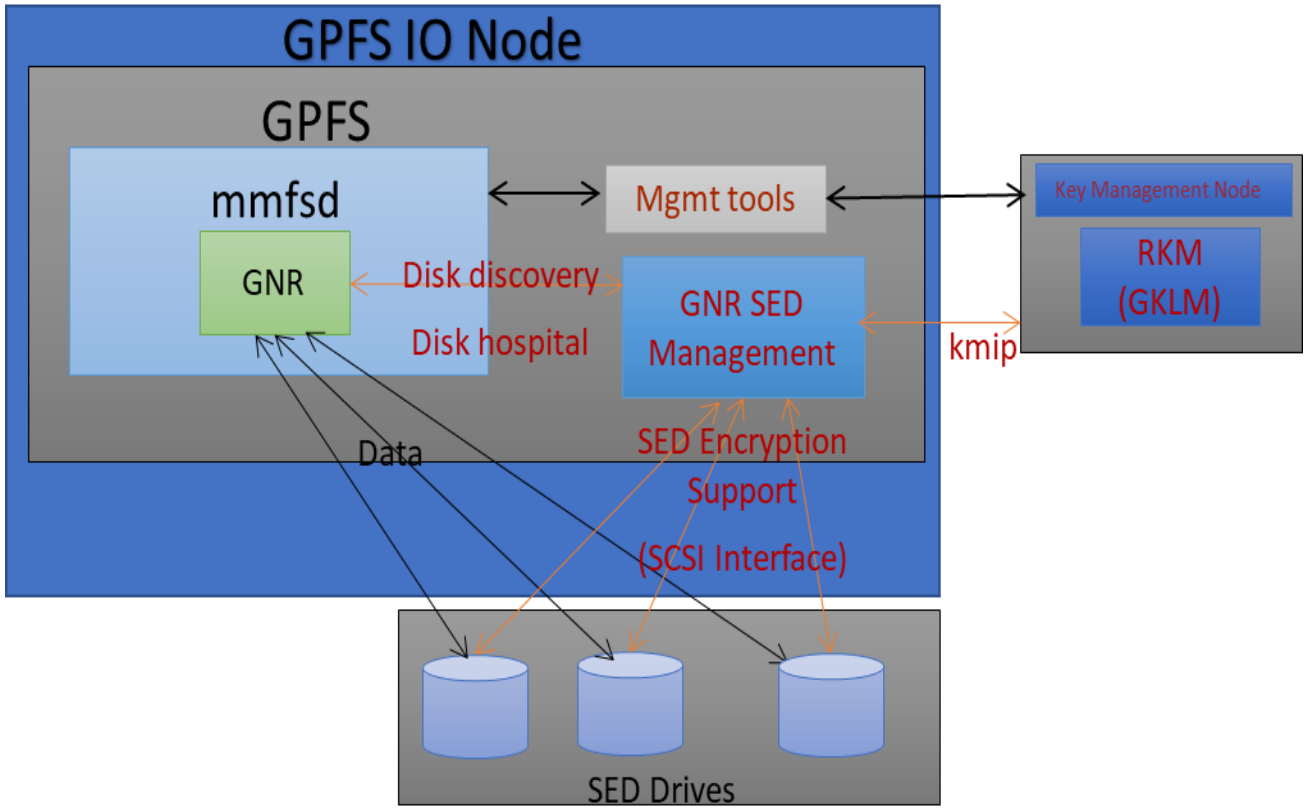
Storage Scale

Background:

- SED enabled by enrolling with MEK
- Auto lock on power off
- Data Security at Rest
- Need to unlock at Power ON using MEK
- Crypto erase by changing DEK

Challenges:

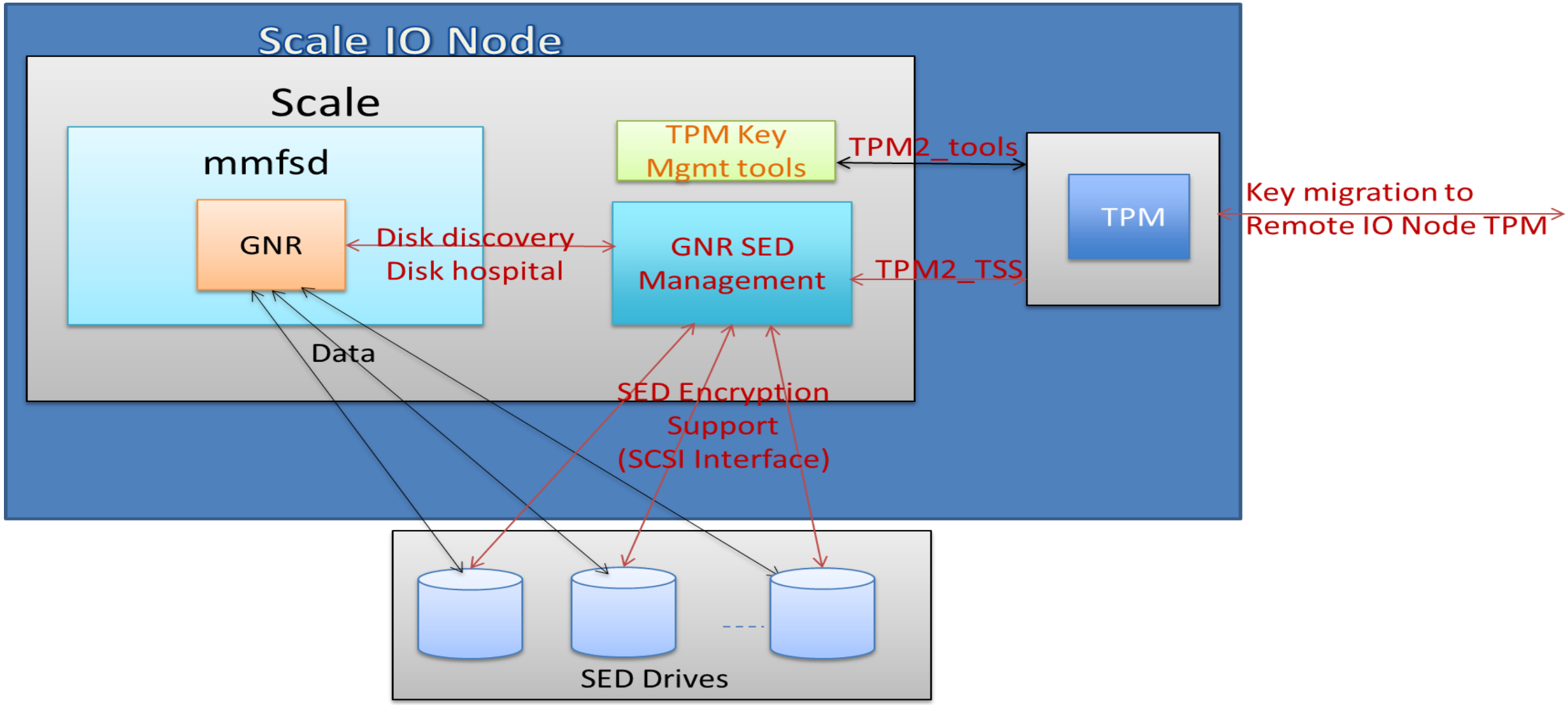
- External Key Managers are expensive
- Different Key Managers



What is TPM(Trusted Platform Module)

- A specialized hardware security chip.
- Provides secure cryptographic functions.
- Defined by TCG(Trusted Computing Group).
- TPM 2.0 is the latest standard of TPM.
- Secure key storage, encryption & decryption, platform integrity etc are some of its key features.
- secure boot ,disk encryption, device authentication etc are some of its applications.

SED Support using TPM : Overview

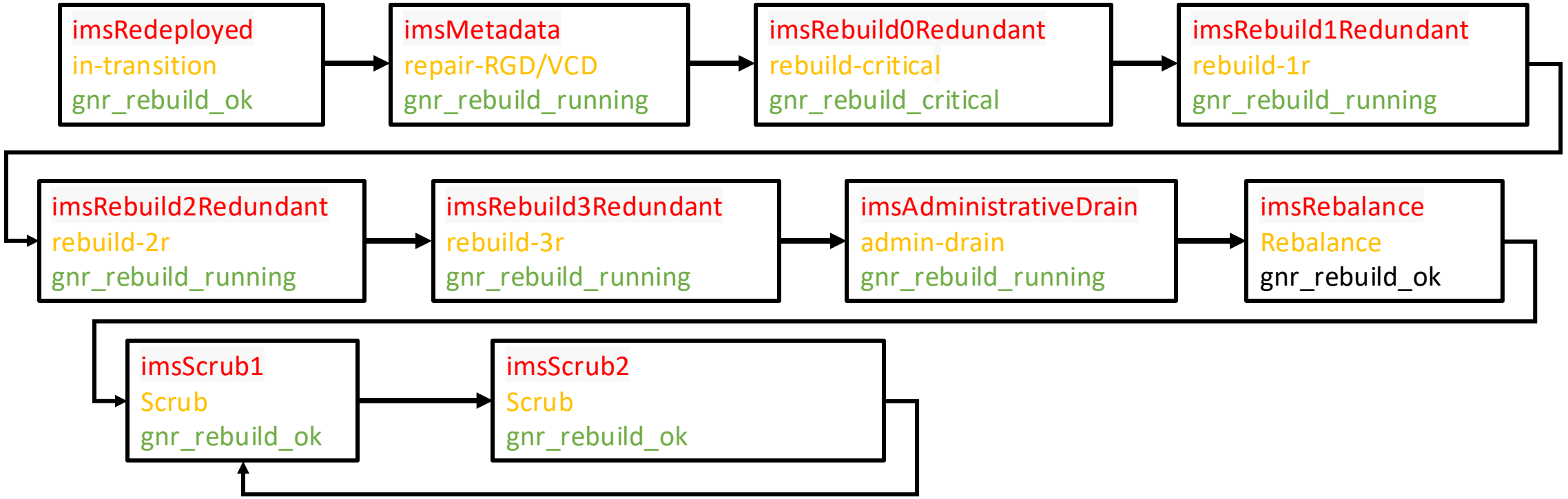


ECE / GNR Monitor Improvement

Storage Scale

- **Mapping of the different imStates to the bgTask field and health events:**

Legend: Ims state BgTask field health event



New ESS Hardware events to create Call Home tickets

Storage Scale

The following, hardware-related health events will automatically create a call-home ticket because the customer cannot fix them without the help of IBM support.

canister_failed
dimm_inspection_failed
dimm_size_wrong
dimm_module_size_wrong
dimm_speed_wrong
dimm_module_speed_wrong
cpu_inspection_failed
cpu_speed_wrong
cpu_unit_speed_wrong

This improves the reaction time of IBM support to a customer problem.

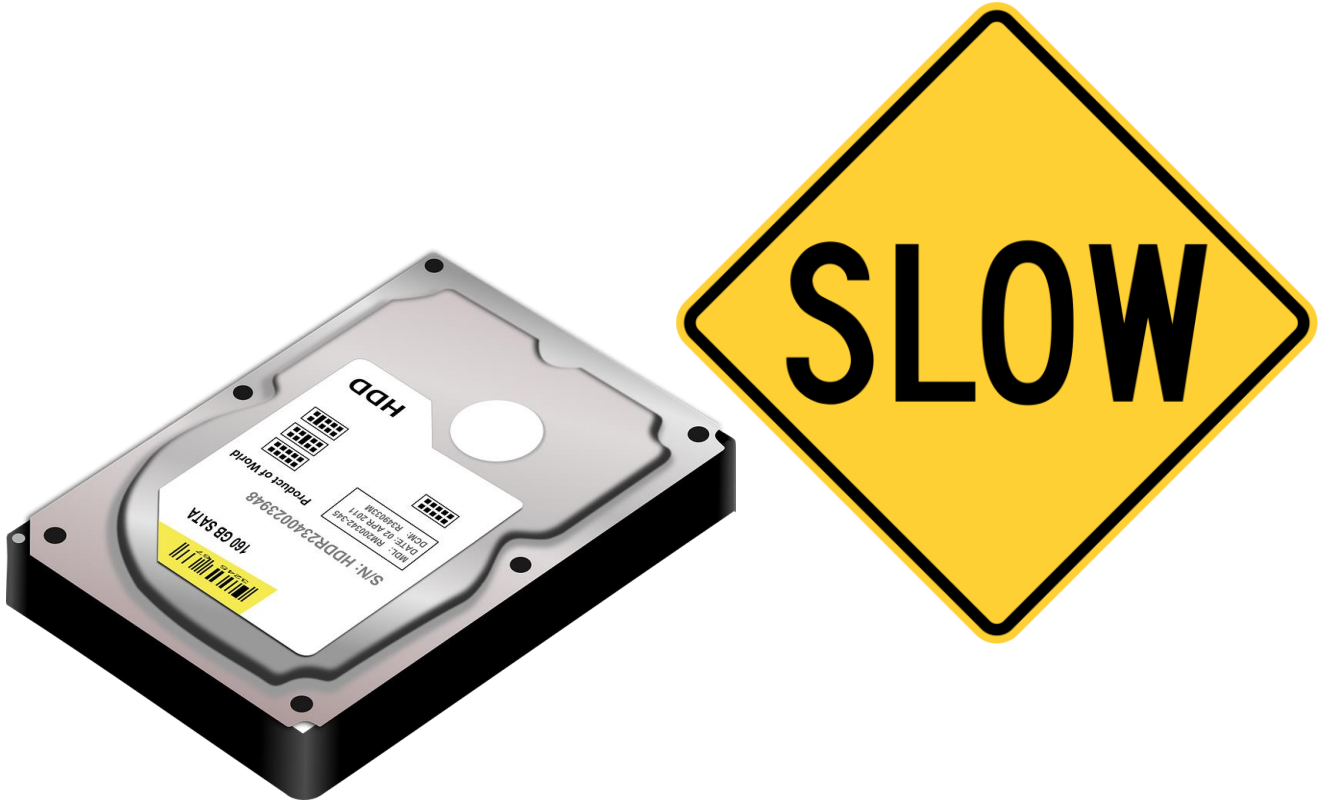
find slow pdisks on ESS

Storage Scale

Problem: Performance impact caused by some slow disks at a large ESS6000 customer installation in Germany with thousands of disks.

Solution: find the slow disks and replace them! DiskHospital knows them already, <u>but how to get the top 10 ?</u>

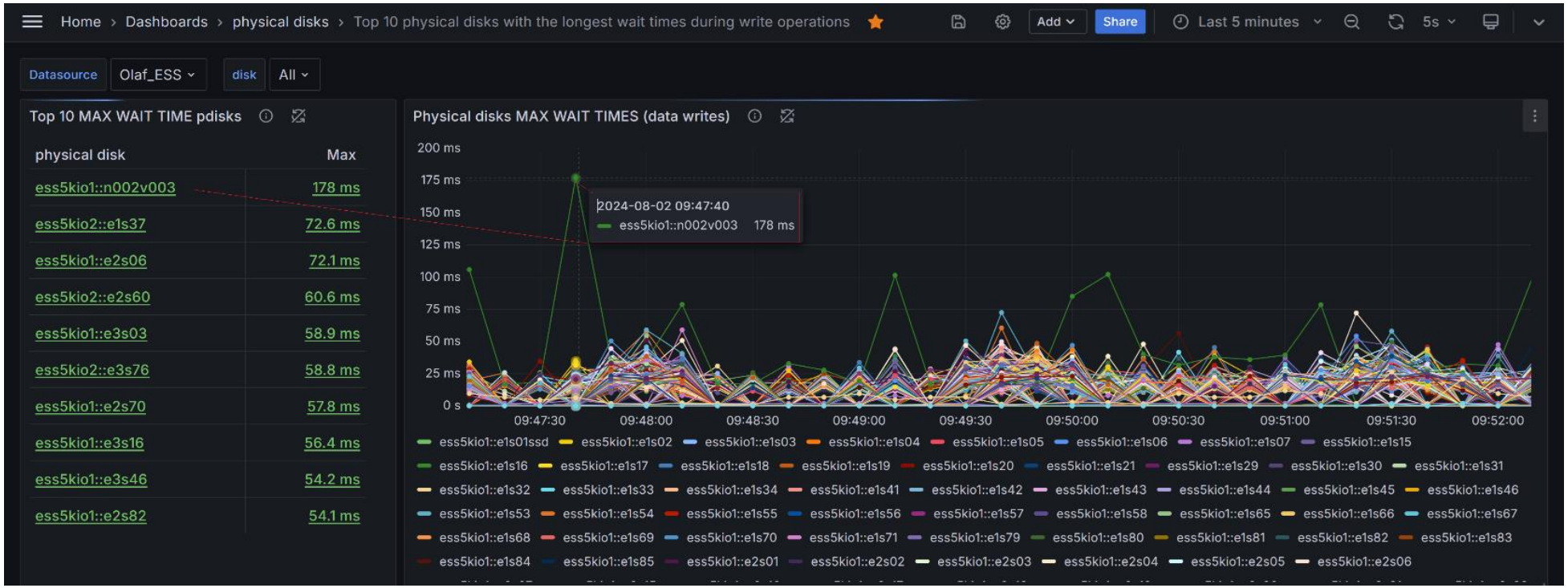
_____/usr/lpp/mmfs/samples/vdisk/findslowpdisks.sh




```
# /usr/lpp/mmfs/samples/vdisk/findslowpdisks.sh rg_nsd11
Get pdisk data for RG=rg_nsd11. Please wait !
Top 10 pDisks with lowest relativePerformance
-----------------------------------------------
0.756 e4s88 just6nsd11ad1.just
0.865 e4s72 just6nsd11ad1.just
0.970 e4s31 just6nsd11ad1.just
....
Top 10 pDisks with IOErrors
-----------------------------------------------
9 e4s85 just6nsd11ad1.just
4 e4s82 just6nsd11ad1.just
2 e4s74 just6nsd11ad1.just
...
Top 10 pDisks with IOTimeouts
-----------------------------------------------
16 e3s63 just6nsd11ad1.just
13 e5s53 just6nsd11ad1.just
...
Top 10 pDisks with pathErrors
-----------------------------------------------
2 e4s82 just6nsd11ad1.just
2 e4s80 just6nsd11ad1.just
...
Failed or disabled pdisks
-----------------------------------------------
```

Grafana pdisks dashboard (example)

Storage Scale



One of the most important measures of physical disk performance is the wait time for a disk write operation.

The **new** bundle of sample dashboards [physical disks]() allows you to identify ***the top 10 physical disks with the longest wait time*** for a write operation for the selected time period.

For more details on a particular disk, you can ***drill down*** from the table to the individual disk view.

Watch DEMO video *on the IBM Storage Scale bridge for Grafana [Wiki]() >>>*

Data Acceleration Tier (DAT): Ustore NVMEoF Monitoring

Storage Scale

Mmhealth enhancement for Ustore / NVMEoF

- Nvme component (ESS side)
  - show non-gnr nvme devices in mmhealth nvme component (server side). Existing nvme checks are done on exported nvmes too (e.g. nvme_temperature_warn)
  - Additional smart check through HAL (nvmeof_raw_disk_smart_failed)
- NVMeOF component (ESS side)
  - Detect if node exports nvmes (mmvdisk nvmeof list –Y) Noderole=NVMeoFTarget
  - Check packages, modules, multipath settings
- Disk Component (client side)
  - By default mmhealth shows NSDs which the node is NSD server only (not the case for ustore)
  - Evaluate nodeclass „nvmeofClients“ to show and monitor all NSDs incl. Ustore NSDs

```
# mmhealth node show nvmeof –v
Node name:      c145f11san06b.gpfs.net

Component     Status       Status Change          Reasons & Notices
-----------------------------------------------------------------------
NVMEOF        HEALTHY      2024-03-21 17:23:30     -

Event                    Parameter  Severity    Active Since        Event Message
-----------------------------------------------------------------------------------------------------------
nvmeof_modules_installed   NVMEOF     INFO       2024-03-21 17:23:30  NVMeoF modules are installed.
nvmeof_multipath_disabled  NVMEOF     INFO       2024-03-21 17:23:30  Native multipath is disabled for NVMeoF.
nvmeof_packages_installed  NVMEOF     INFO       2024-03-21 17:23:30  NVMeoF packages are installed.
```

Storage Scale

IBM