Storage Scale

# What's new in

# IBM Storage Scale

# the Global Data Platform for all data lakes

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

IBM

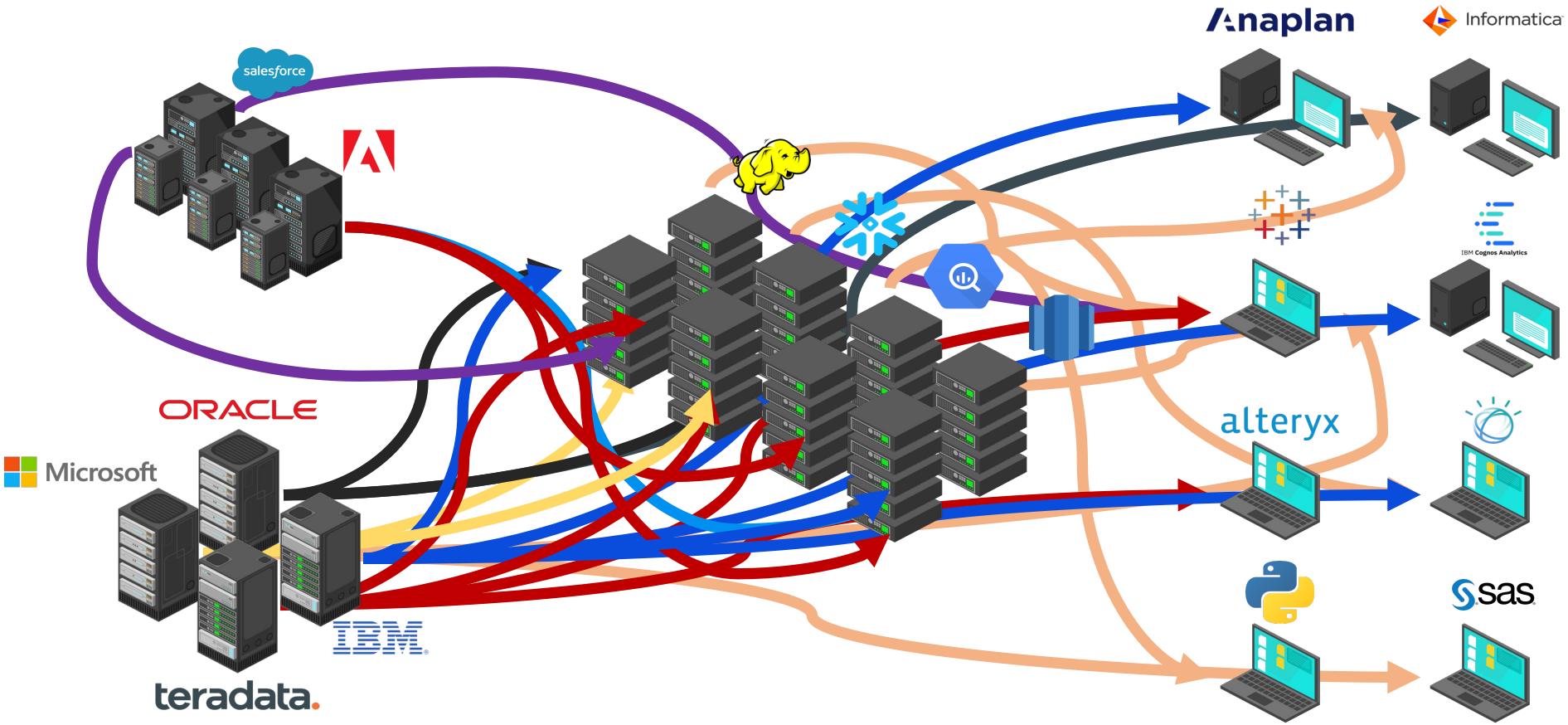
# Disclaimer

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.

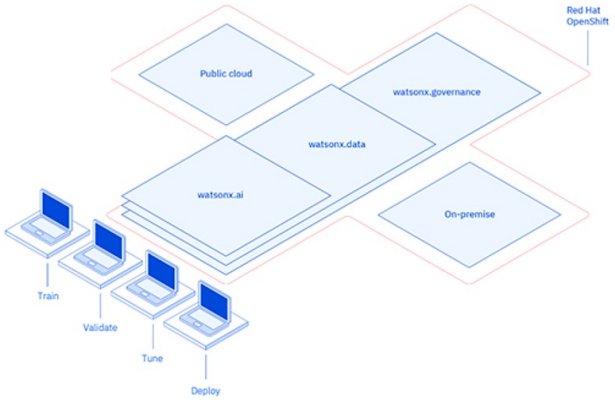
Storage Scale

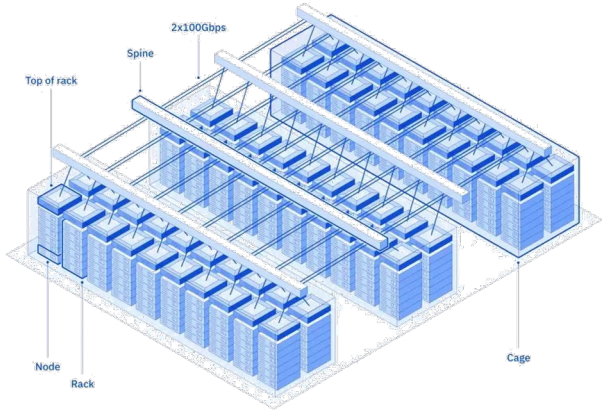
# Data is everywhere in a hybrid and multi-cloud world and the compute (GPU or CPU or DPU) wants data from remote to local!



# Storage requirements for AI and HPC

Storage Scale




## Tuning/Inferencing

| | | | |
|---|---|---|---|
| **watsonx .ai** | Storage Acceleration<br><br>Storage Abstraction | Efficient GPU support | Rapid deployment<br><br>HA/DR/Backup<br><br>Simplified Day-2 operations |
| **watsonx .data** | | Metadata catalog integration | |
| **watsonx .governance** | | | |

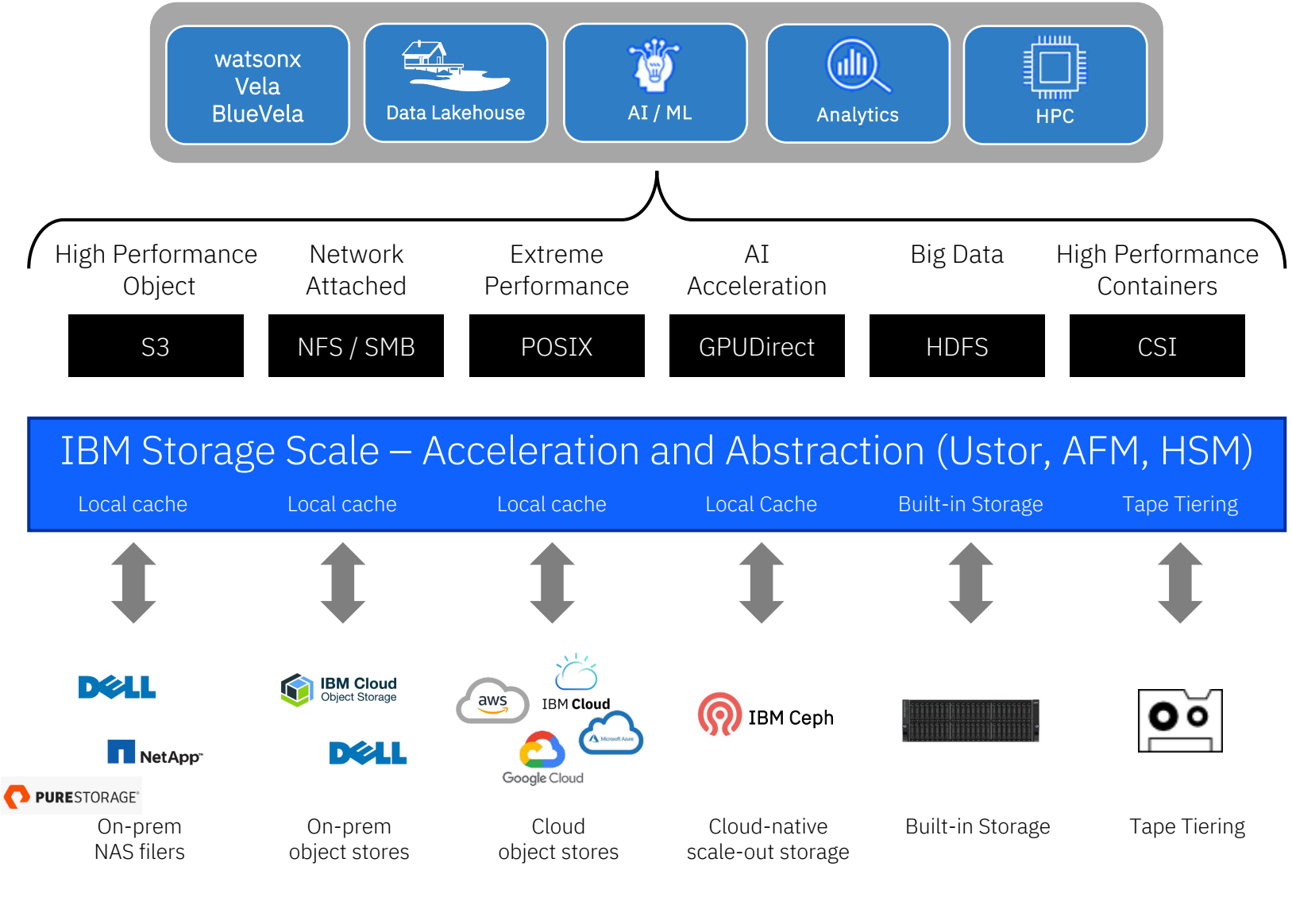
## Training and Modeling

| | |
|---|---|
| **Maximum Performance** | Efficient GPU support<br><br>High bandwidth<br><br>Low latency |
| **Scalability** | Linear capacity scaling<br><br>High density |

# IBM Storage Scale – A Global Data Platform

Storage Access, Abstraction and Acceleration to maximize CapEx investment and lower OpEx costs

Storage Scale




## Multi-Protocol Support Access

Simultaneous multi-protocol access including GPUDirect 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*

# MN5: IO Partition

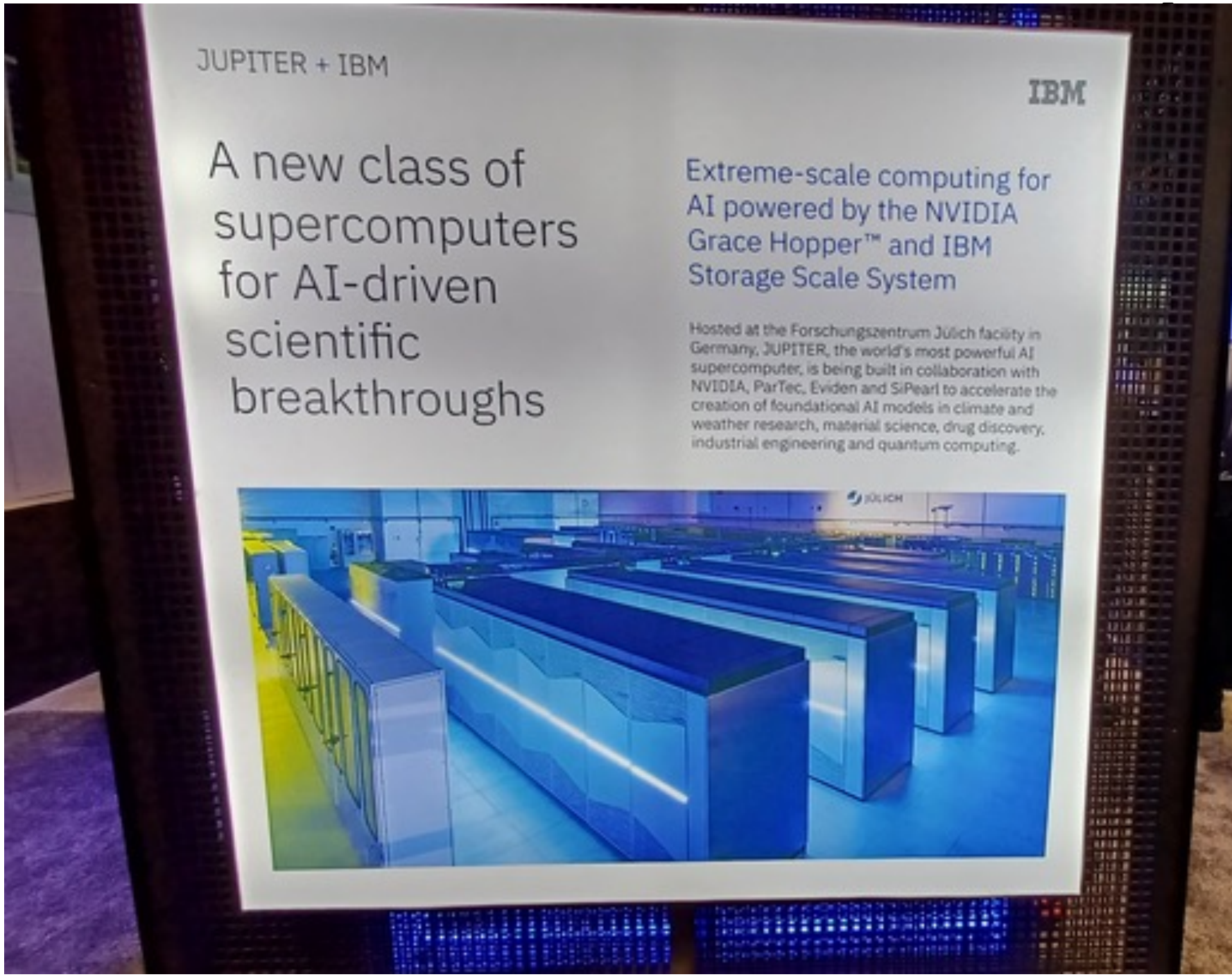
| ESS model | #ESS | Drive Capacity | Total # drives | Raw capacity | Net capacity | Read perf | Write perf |
|---|---|---|---|---|---|---|---|
| ESS 3500 Capacity model | 50 | NL-SAS 18TB | 20400 | 367PB | 248 PB (8+3P) | 1.6TB/s (IOR 100%read) | 1.2TB/s (IOR 100%read) |
| ESS 3500 Performance model | 13 | NVMe 15.36TB | 312 | 4.79PB | 2.81PB (8+2P) | 600GB/s 1Mio iops 4KB | 600GB/s 500K iops 4KB |

Total net storage capacity: **650** PB

| Element | Element | Size |
|---|---|---|
| IBM TS4500 | 2 | |
| Tape Enterprise | 20100 | 400 PB |
| Drives | 64 | |

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Julich Lab Jupiter Exascale HPC with IBM Storage



# IBM Storage Scale Security and Resiliency

## Active Protection for Cyber Resiliency built on the NIST framework

Storage Scale

### IDENTIFY

- Cyber Resiliency Assessment Tool, Probes 100s of different controls and best practices

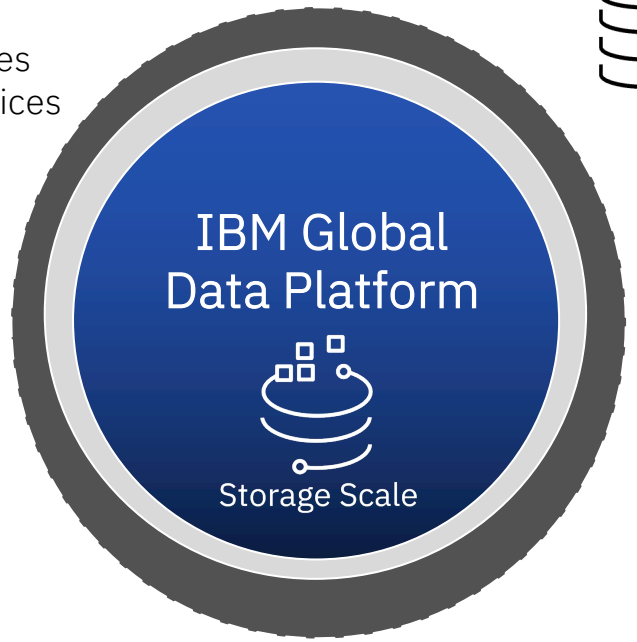
### Governance

- Data Catalog allowing for data orchestration and data migration control and accountability
- Watson Knowledge catalog

### RECOVER

Recover Operations and Data Quickly

- Instant Restore with Storage Scale AFM
- Storage Scale and Storage Protect – recover multi-petabyte filesystems in hours
- QRadar Incident Forensics




### PROTECT

Active Protection against cyber attacks

- Multifactor Auth, RBAC, Privileged Access Monitoring (IBM Security Verify)
- Safeguarded Copies via immutable snapshots, logical air gap
- Scan snapshots for signs of ransomware
- Log all Admin & user actions

### DETECT

Detect Suspicious Behavior

- QRadar and Splunk SIEM integration
- File Audit Logging, Watch Folders
- Analyze backup data for signs of ransomware (Spectrum Protect)
- Reporting: QRadar User behavior analytics
- IBM Flash Core Modules entropy detection

### RESPOND

Alert and take action

- Automated action upon threat detection (QRadar)
  - Snapshot, Block Session , Etc..
- Alerts automatically prioritized based severity of the threat and criticality of the assets involved

# Multiple ways to deploy IBM Storage Scale

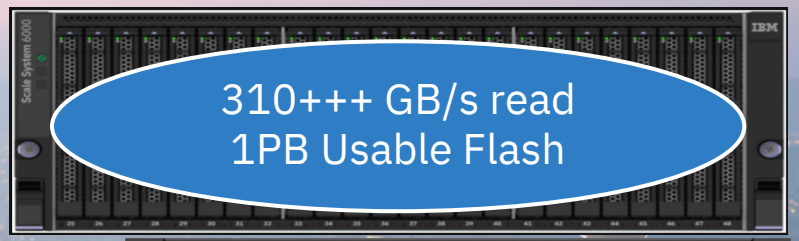
## IBM Storage Scale software

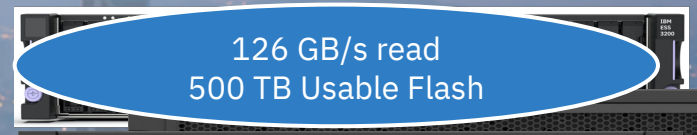
ARM, x86, IBM Power, IBM System z, Kubernetes or Virtual Machines

## IBM Storage Scale in Public Cloud

AWS Azure IBM Cloud Alibaba Oracle Cloud Google Cloud

## IBM Storage Scale System 6000 or 3500




# IBM Storage Scale Workloads

Storage Scale

## AI: NVIDIA GPU

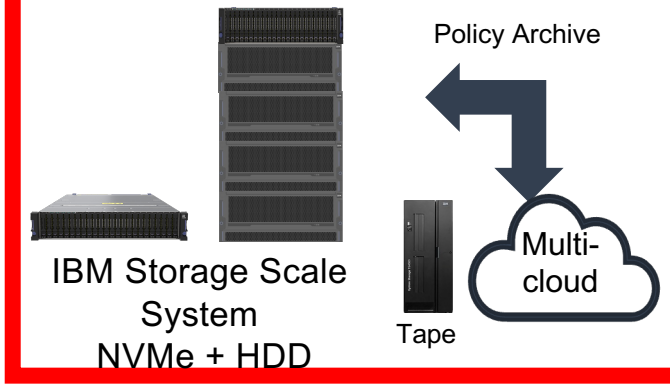

## AI: Analytics

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

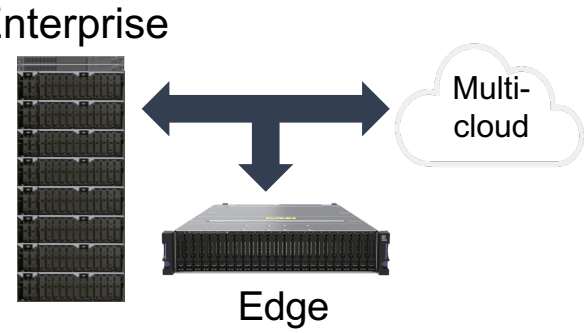

## AI: IBM Data Fabric

IBM Cloud Pak for Data


## Hybrid Cloud: Backup / Archive

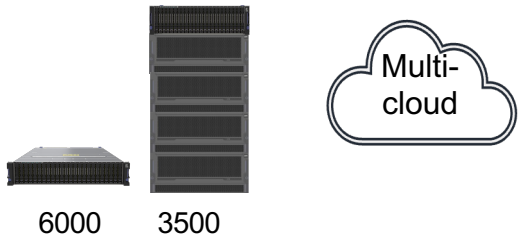



## Hybrid Cloud: Data Lakes




## Hybrid Cloud: HPC

IoT/Video/Images/Genomes



# The easiest way to deploy a Global Data Platform

## IBM Storage Scale System

All NVMe Flash 6000

- 48 TB to 1+ PB flash
- Up to 15+ PB capacity per 3500
- Scales from 1 to 10000s+ of clients

3500 Hybrid NVMe Flash + HDD

## Performance Optimized

**310+GB/s per 6000**

13M+ IOPS per 6000

Parallel access to data

**Locally cached global data with dynamic memory pools**

## Cost Savings

Storage Scale

**Compression-enabled NVMe QLC storage 2Q24**

Integrate existing non-IBM storage and cloud

**Turn off/turn on unused storage w/ tape**

Mix and match old/new

## Cyber Secure

End to end encryption with customer keys

Lower RTO with Safeguarded Copy quick recover option

**Ensure protection with CyberVault service**

## Global Connectivity

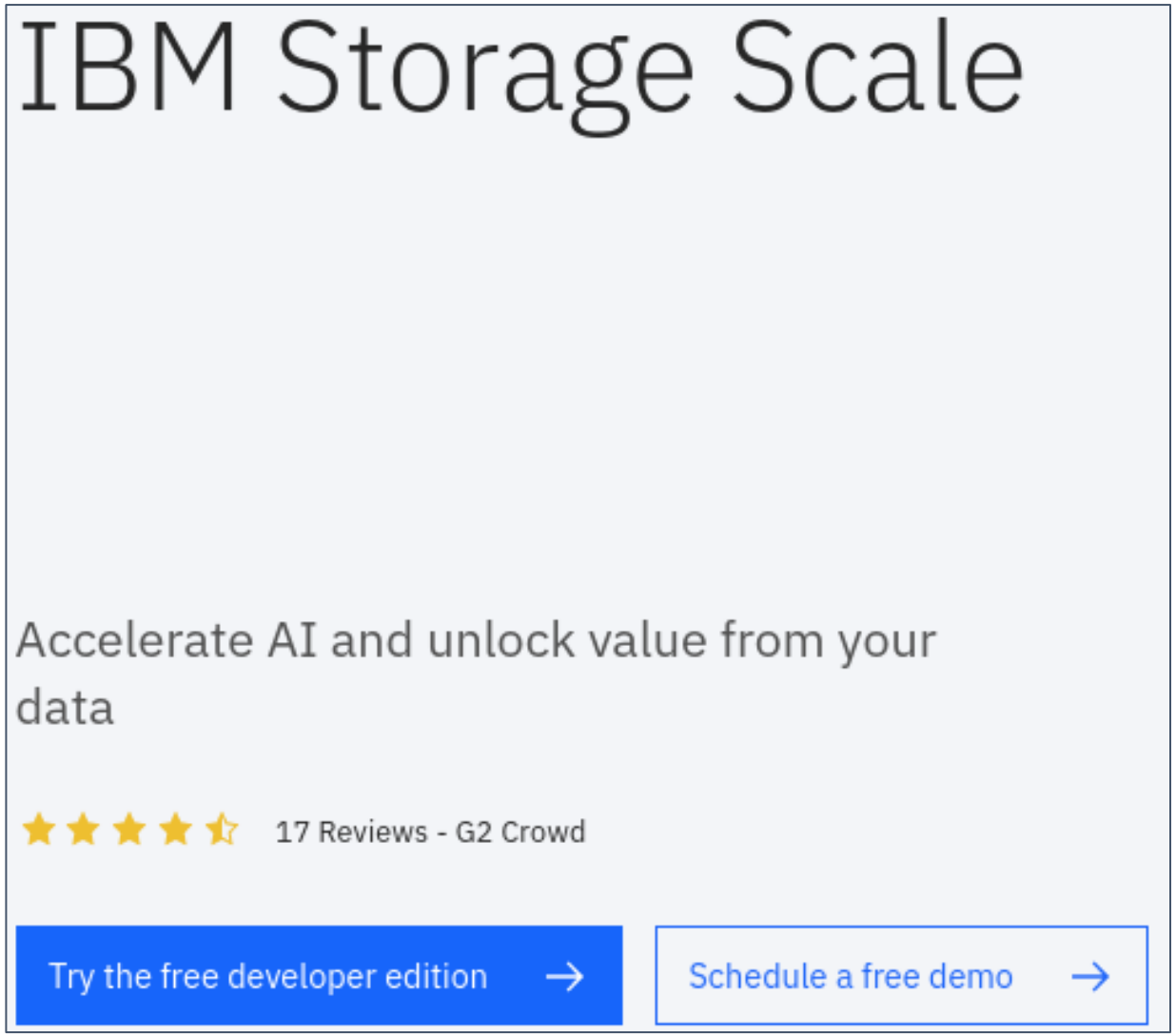
**Break down silos by connecting remote systems, cloud data and non-IBM storage**

Connect any node to a global data platform online when needed

1 Up to 2PB+ can be obtained with 3:1 compression using FCM disks

# 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.

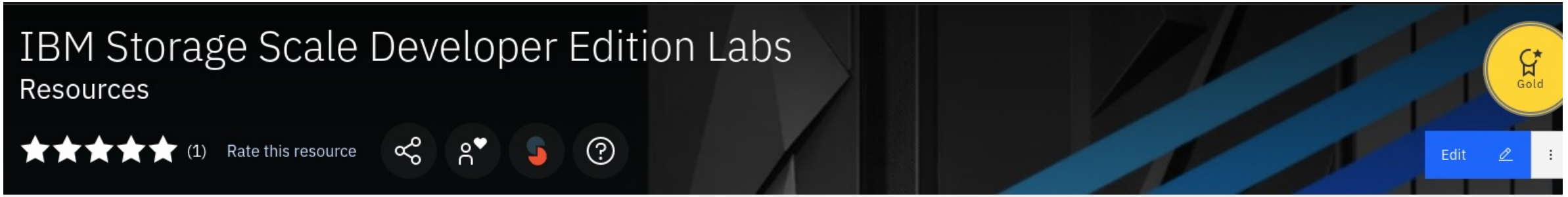
The format of the group is as a web community with events held during the year, hosted by our members or by IBM.

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




Mar 14, 2023
Login - to Stor
Edition Labs v

Powerpoint step
Storage Scale De
key method.

Visibility
IBMers, Business P

Mar 14, 2023
Video - Storag
Management

Video showing ex
of using Storage
Management (ILM
manage data.

Apr 28, 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

Apr 28, 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.

Apr 28, 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.

# 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

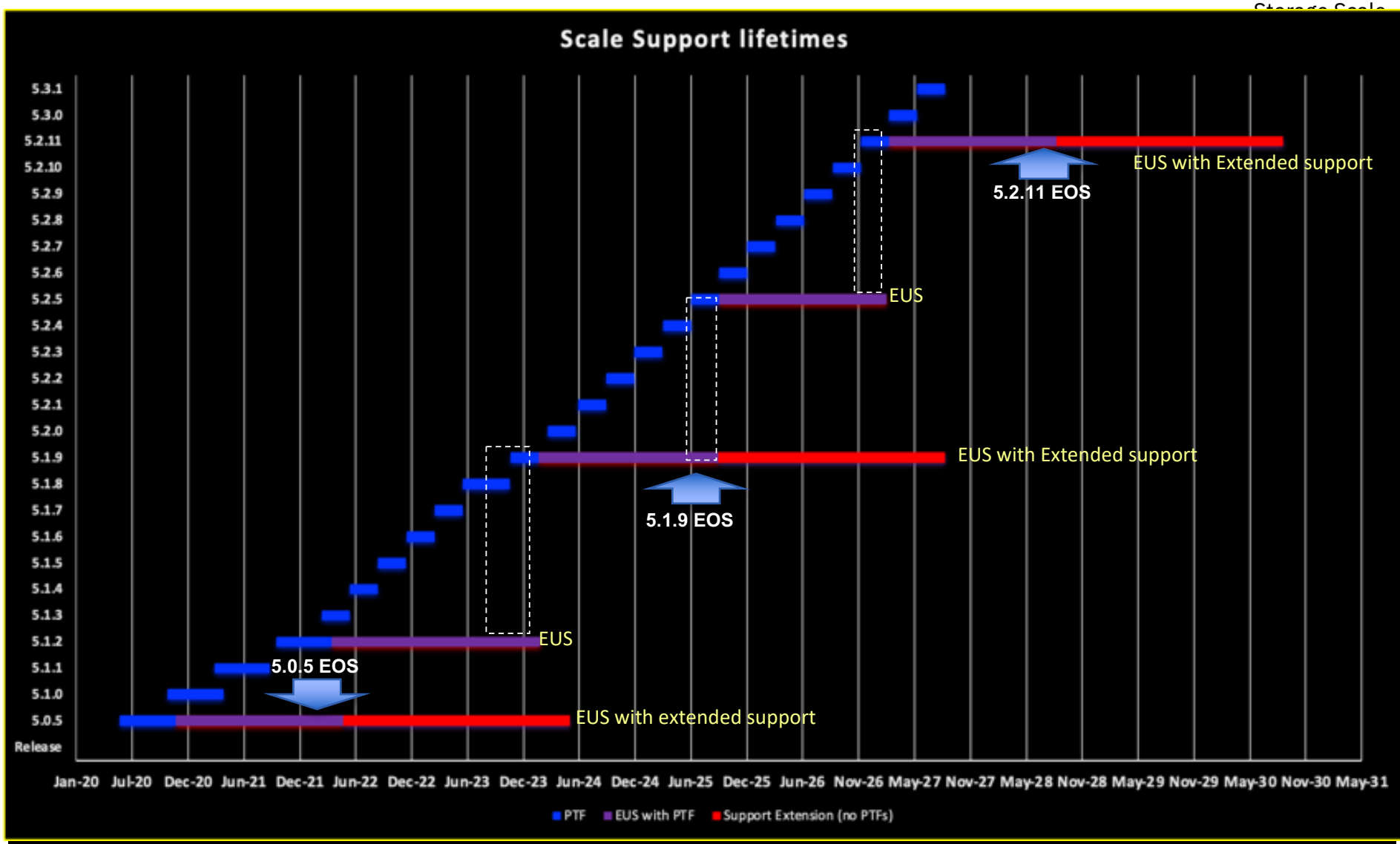
| Feature | Data Access Edition | Data Management or Developer Edition | Erasure Code Edition |
|---|---|---|---|
| Multi-protocol scalable file service with simultaneous access to a common set of data | Yes | Yes | Yes |
| Facilitate data access with a global namespace, massively scalable file system, 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



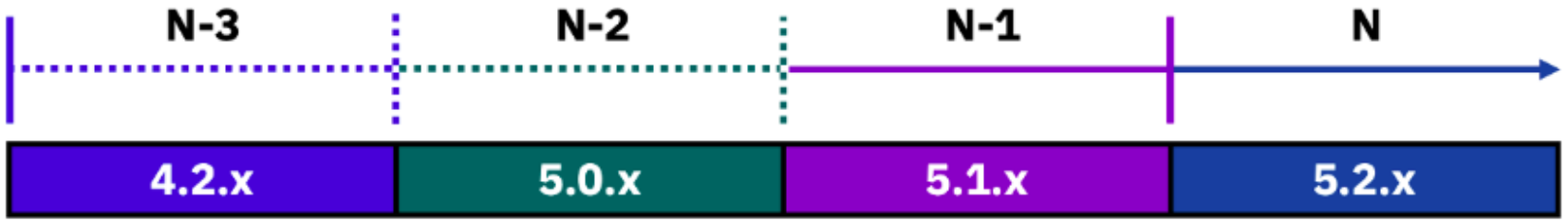

**Note:** Version numbers and release timing are for example purposes to demonstrate the goal of EUS every 18 months and do **not** represent a commitment to deliver a specific version or on a specific timeline

# 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.

# IBM Storage Scale - Highlights of 2022/23 releases

## Access Services

### Modernize protocol stack

- **High Performance Object:**
  - Containerized S3
  - Targeting AI/Analytics
  - Initial support ESS storage

### AI / ML/ GPU acceleration

- NVIDIA GDS over RoCE
- GDS Write acceleration
- GDS Perf enhancements

### Containerized Environments

- OpenShift integration for upgrade resiliency
- FSgroup support for RWO
- VMWare 7 support for CNSA

## Abstraction and Acceleration Services

### Data Abstraction

- AFM policy-based tiering to object storage - AWS, Azure, Google
- AFM DR 60 min RPO
- AFM Containerization
- AFM to S3 download and upload performance
- AFM Recovery improvements for rename and remove operations

### Core Services

- Multi-Rail Over TCP
- Improved Prefetch performance ~2x
- Fileset scaling
- IOPS and io500 improvements

## Orchestration Services

### Visibility, Control and Automation

- Remote Fileset Access Control
- Independent filesets scalability
- Scale GUI HA, GUI HA in ESS Environment – Stretch Cluster (Active/Active)
- Online mmfsck
- RHEL 9 support

### Monitoring, Availability & Proactive Services

- Enhanced stretch cluster monitoring
- Monitor AFM memory queue alerts in mmhealth
- Detection of hung/unresponsive nodes

## Security Services

### Security

- **Safeguarded Copy:** Prevents modification or deletion of copies due to user error, malicious destruction, or ransomware attack
- Encryption -Using Hashicorp/ KMIP

### Resiliency

Erasure Code Edition Enhancements:

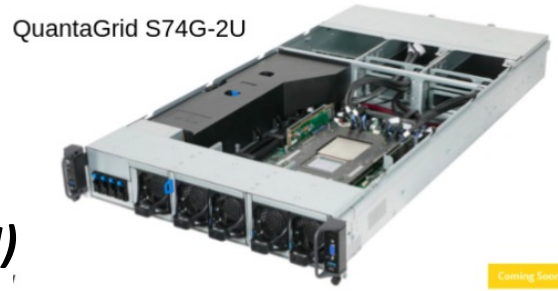
- High density server (60 disks)
- Small ECE - 3 nodes
- Trim - Automated reclamation

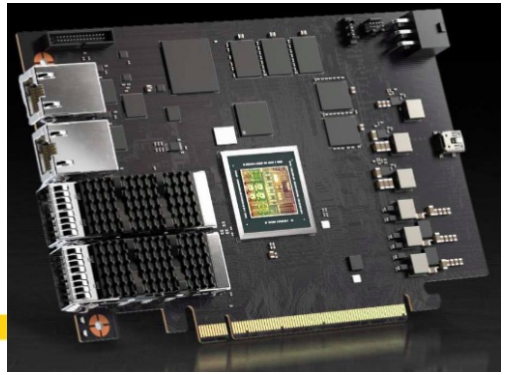
# Access Services

# 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)

DPU (Blue Field-3) for exploitation
research spike

Make it a platform for Scale like any other



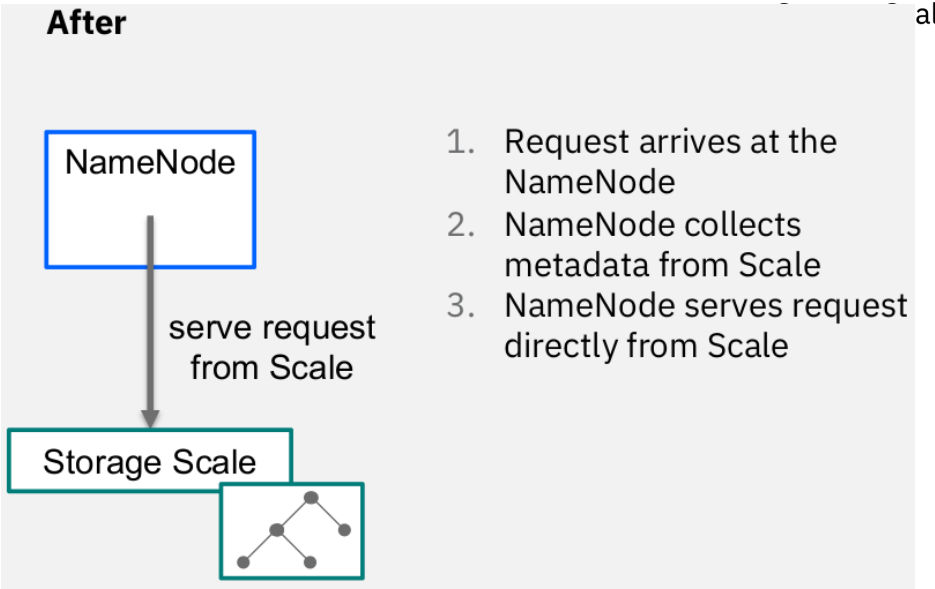
- ***Included***
  - SE package / install toolkit / rpm based install
  - NSD client
  - Scale base functionality (IO, policies, remote mounts, snapshots, quotas, etc.)
  - Manager roles: file system manager / token manager / cluster manager
  - RDMA (IB or RoCE) including GDS
  - Health Monitoring
  - 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)
- ***Excluded, but planned for future releases***
  - NSD servers (has been tested and used, requested Real World testing in 5.2.1)
  - GNR/ECE (successful sniff test done, tdb when this will become a product)
- ***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 – HDFS, NFS, SMB

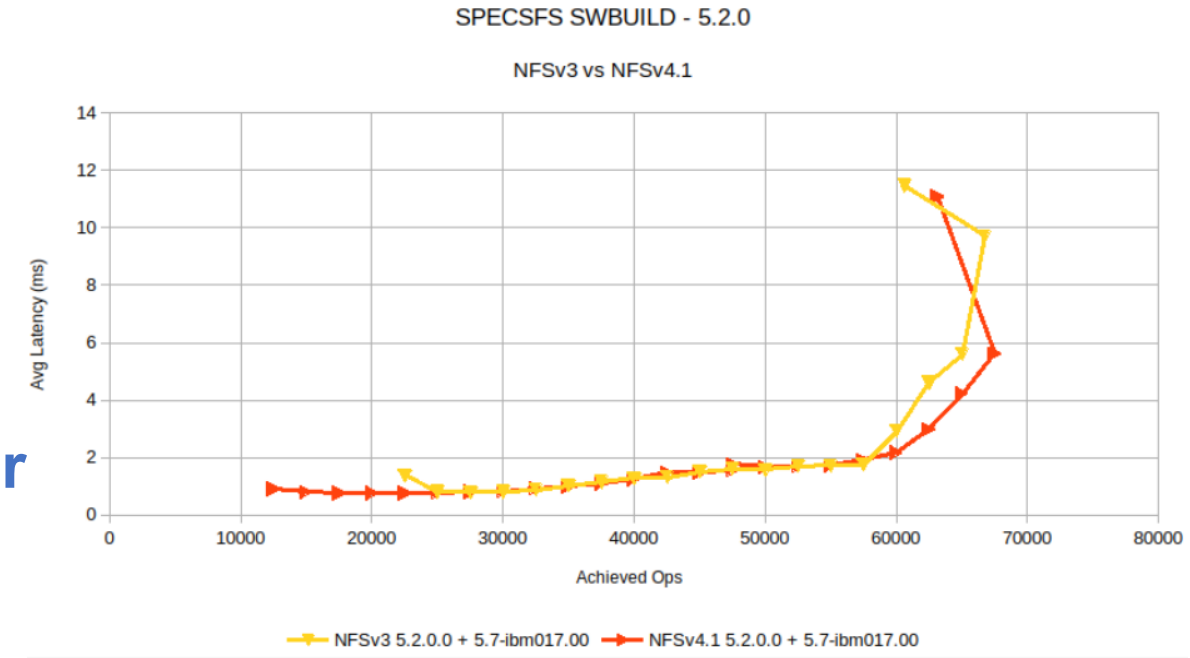
***Support and Currency:***

- Cloudera Data Platform (CDP) Private Cloud Base is certified with IBM Storage Scale on x86_64 and ppc64le since December 2020.
- Opensource Hadoop 3.1.3, 3.2.2, 3.3.0
- Includes HDFS Transparency 3.1.1-16/17, HDFS Transparency 3.2.2-7/8
- Support **mmhdfs config dump** – dumps in use configuration
- NFS-Ganesha support for 5.7 code base




***Improved performance:***

- NFS “meta data cache” component was revised resulting in significant performance improvements
- HDFS transparency metadata redesign
  - Full parallelism for RPC calls (GPFSNamesystem)
  - No more lock contention in NameNode




- **Continued partnership with Tuxera for high-performance SMB**

# Access Services – Object

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
- New: IBM Storage Scale CES S3 Technical Preview for S3 access for AI, data and analytics workloads: https://community.ibm.com/community/user/storage/blogs/madhu-punjabi/2024/04/26/ibm-storage-scale-ces-s3


***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

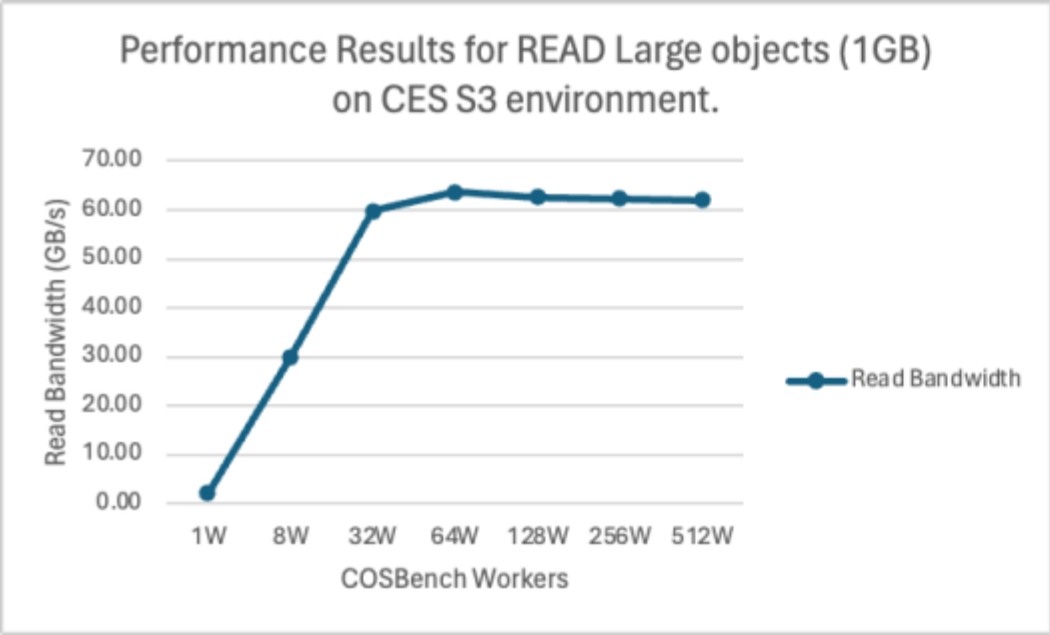
Scaling limits for S3:

- Up to 10TB single object size
- Up to 5000 S3 accounts
- Up to 5000 S3 buckets
- Up to 4,000,000 (4M) objects/bucket. 100M objects/bucket planned for MVP GA.

Higher scaling limits as compared to HPO !

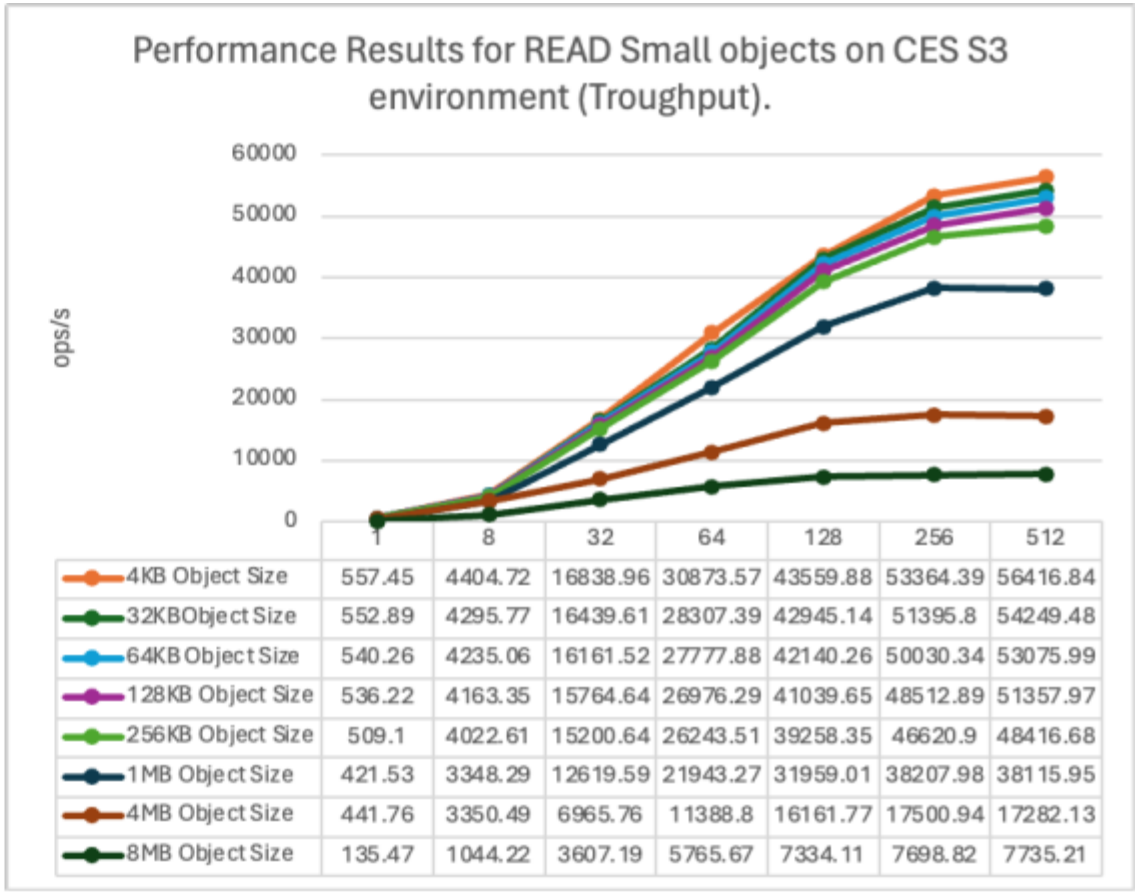
# Access Services – Object Performance

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% |

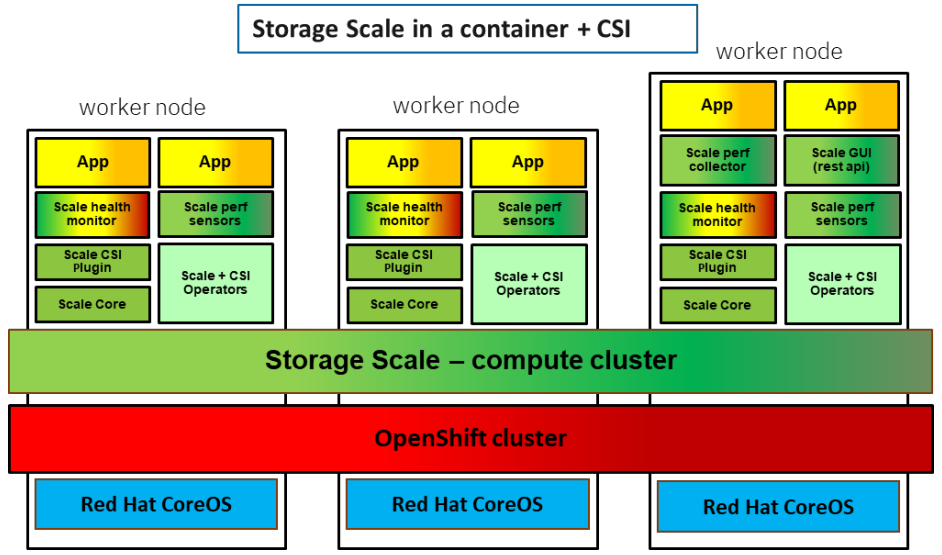
**Table 2. Performance Results for READ Large objects (1GB) on CES S3 environment.**



# Access Services – Container Native Storage Access (CNSA)

Storage Scale

## Improvements introduced in CNSA 5.2.0.0




***Wider support to use the latest CNSA functionality.***

Are you upgrading from a previous version of CNSA? < 5.1.5.0?
Don't skip the upgrade to 5.1.5.0!

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

Support for RedHat OpenShift Container Platform 4.[13-15]

Enhanced status reporting
Node Cordons - Which nodes? Did the operator do it? Did someone else?
Node Drains - Is the operator draining? Which pods is the operator currently evicting?
Did the operator fail to evict pods? Why did eviction of pods fail?
Pod Deletions – Operator planned?, Operator wants? What is stopping it?

Tech Preview!
Support for SNC Filesystem – 3 way replication
Infiniband Remote Direct Memory Access (RDMA)
Basic enablement of RDMA (within gpfs) via CR
Ability for CNSA + Scale storage cluster to utilize IB RDMA in goodpath environments

```
apiVersion: scale.spectrum.ibm.com/v1beta1
kind: Cluster
spec:
  ...
  daemon:
    roles:
    - name: client
      resources:
        memory: 4Gi
        cpu: "2"
      limits:
        memory: 10Gi
        cpu: "5"
```

# Access Services – Container Storage Interface

## Improvements introduced in CSI 2.11.0

Storage Scale

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

Planned support for Red Hat OpenShift 4.[13-15] and Kubernetes 1.28/1.29

Support for Shallow Copy Volume

CSI spec do not have concept of mounting a snapshot. The only way to access content of a snapshot is to create new volume by copying content of snapshot and then mount that volume for workloads. – It's for backup!

Support to configure resource limits of IBM Storage Scale Container Storage Interface driver

can be configured with the higher limits if the user notices that the pods are being stopped due to an OOM

Upgraded Kubernetes CSI sidecars

Improvements in the script for debug data collection

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

# IBM Storage for Data & AI Solutions with NVIDIA

Storage Scale

## *IBM Storage & NVIDIA Collaborations*

- *DGXH100 SuperPOD RA : 2023*
- *1st SuperPOD BCM installation: 2022*
- *DGX BasePOD validated storage partner : 2022*
- *SuperPOD installations: 2021*
- *DGX A100 SuperPOD RA : 2021*
- *GPU Direct to Storage (GDS): 2021*
- *DGX A-100 2/4/8 RA: 2021*
- *Red Hat OpenShift on DGX: 2020*
- *DGX-2H SuperPOD RA: 2019*
- *DGX-1 / DGX-2 POD RA: 2018/2019*
- *IBM Data Science Pipeline 2018*

## *US DOE Summit & Sierra*

- *Built in 2018*
- *#2 and #3 fastest supercomputers in the world*
- *Summit: 27,648 Tesla GPUs*
- *2.5 TB/s single stream IOR*
- *2.6 M 32k file creates*
- *16 GB/s r/w per node*


## *NVIDIA Circe*

- *Built in 2018 in 3 weeks*
- *#61 Top 500 in 2018*
- *36 NVIDIA DGX2 nodes*
- *Mellanox EDR*

## *NVIDIA DGX-2H SuperPOD*

- *Built in 2019*
- *#22 Top 500 in 2019*
- *96 NVIDIA DGX2 nodes*

## *NVIDIA Certified Systems (EGX/HGX system)*

- *X86 server business partners*
- *IBM Storage Fusion system*
- *IBM Storage Scale S/W*
- *IBM ESS3500*

## *NVIDIA DGX Compute*

- *NVIDIA Business partners*
- *IBM Storage Sale systems (ESS)*
- *BasePOD and SuperPOD Architecture*
- *Flywheel / XNAT BasePOD*

## *Service providers*

- *“AI/GPU as a Service” providers in the region*
- *Cloud Service Providers*
- *IBM Storage Scale or ESS*

IBM

# IBM Storage for Data and AI & NVIDIA GPU Solutions

*A full spectrum of scalable AI solutions*

Storage Scale

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

## AI Entrant

1 DGX H100
or
1x HGX H100

- Half Populated 3500
- Up to 60 GB/s

**or**

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

## AI Medium

8 x DGX H100
or
8 x HGX H100

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

**or**

## AI Master

16 x DGX H100
or
16 x HGX H100

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

**or**

- 6000
- Up to 300 GB/s read

## AI Scaler

DGX H100 SuperPOD
or
HGX H100 SuperPOD

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

**or**

- 6000
- Up to 300 GB/s read

*A simple, scalable upgrade path*

## IBM Storage:

Simple building blocks – scalable seamless storage upgrade path as needs grow from 1st DGX to AI CoE DGX SuperPOD

Global Data Platform – Data fidelity capabilities to automate AI workflows.

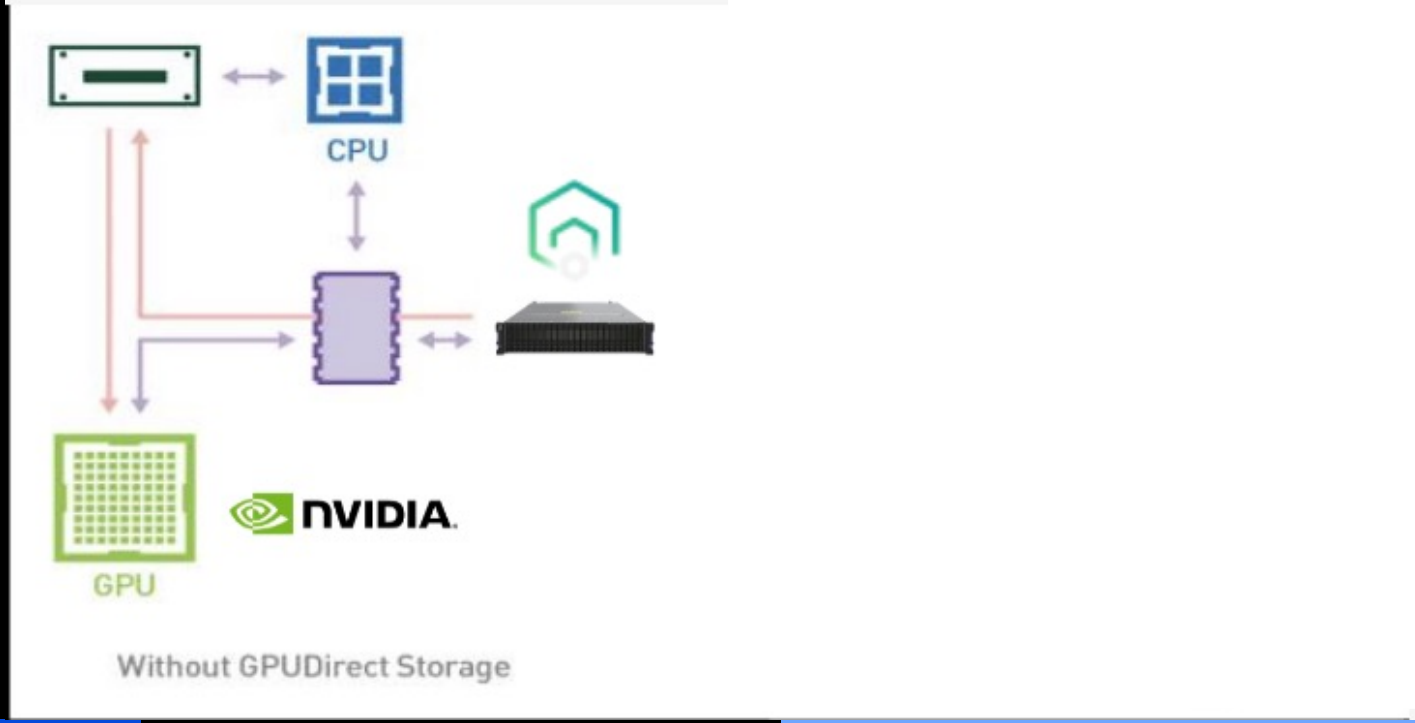
Data Economics – Eliminate copies and transparently tier

Trusted, global enterprise level support and services.

Successful deployments across the globe

# 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**

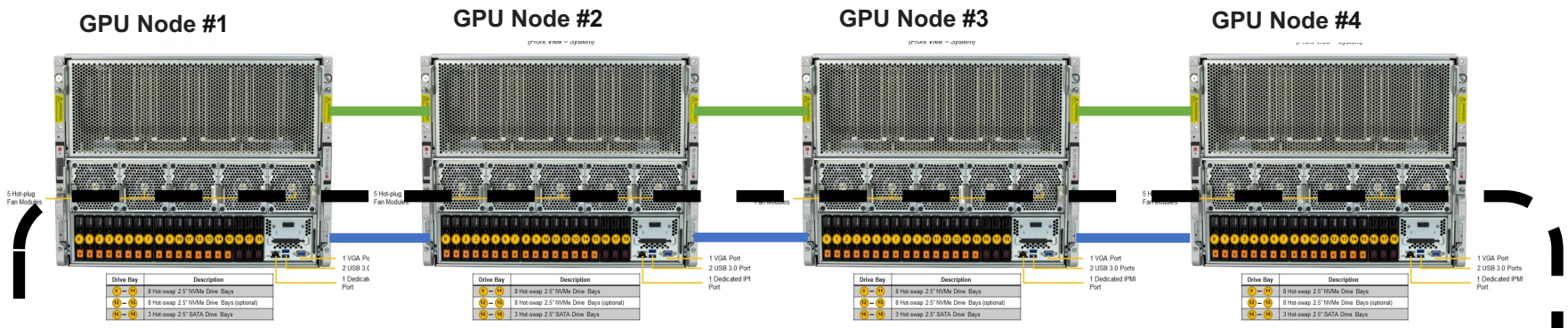
AI, HPC, Analytics are data hungry and require a very high data throughput.

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

# Abstraction and Acceleration

# Performance leadership:  Nothing comes close to IBM

# A new approach to Data at the compute – History - Converged Solution Architecture

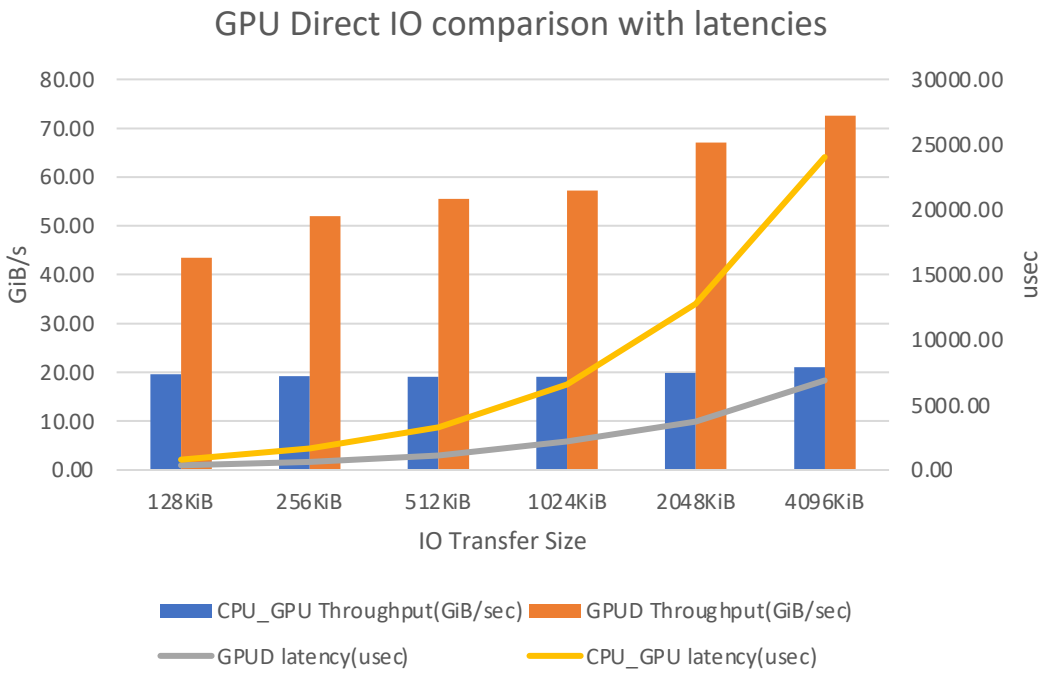


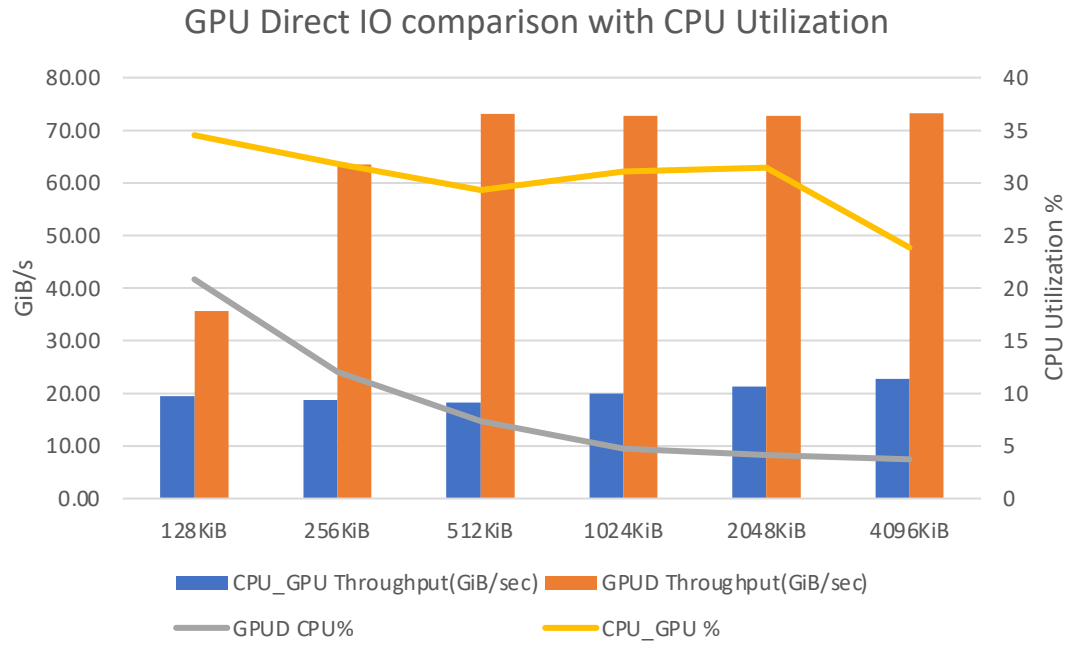

## IBM Storage Scale configuration

- Converged GPU and Storage solution for AI training workloads
- High Performance parallel storage using NVMe local drives from GPU nodes
  - Minimum 2 drives per node; 8 drives across 4 node cluster
  - Max 16 drives per node; 64 drivers across 4 node cluster
- 2 x 200 Gbps storage network per node

# Acceleration Services - GPU Direct Storage Performance – converged solution

Storage Scale

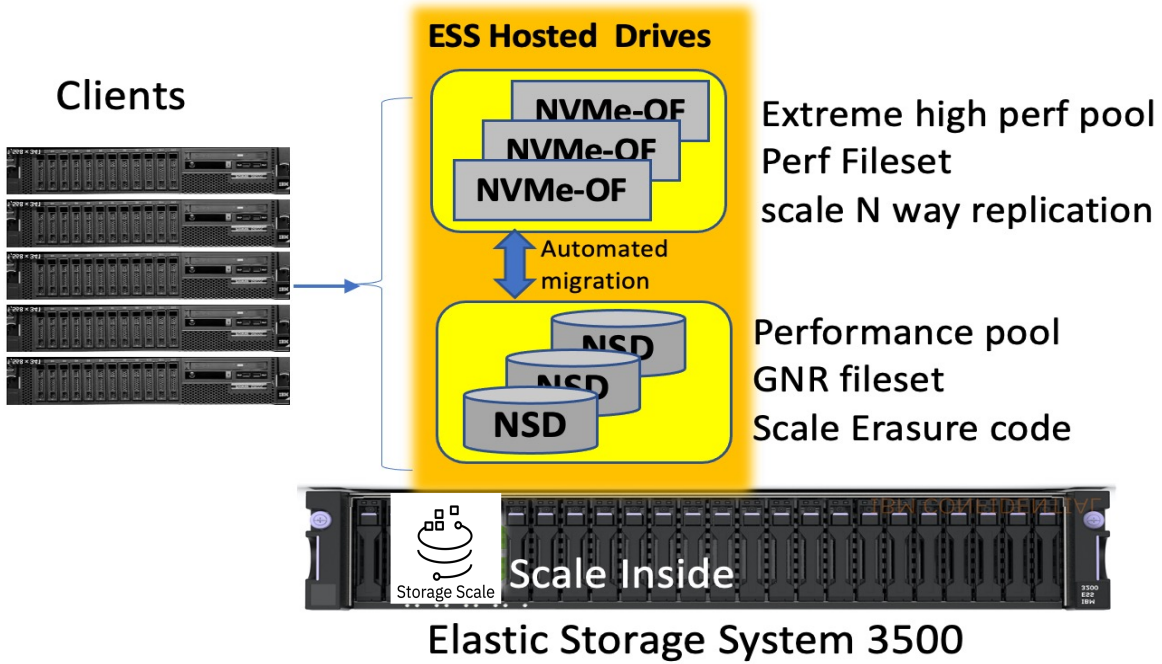



Note : 16 Threads per GPU; Total 128 threads for read

Up to 3.5x higher Bandwidth; 50% reduction in latencies;

# At SC22 demonstrated: Integrated NVMeoF for an extreme performance tier to the compute




Measured over 10+ Million IOPs and 00s GB/s

## Use Case

Data analytics (AI/ML) needing very high rand IOPS with high throughput

High performance Scratch / Shuffle space

- **System Config**
- 3.84 TB, 7.68 TB, 15.36 TB or 30.74 TB
- 4x CX6-VPI Adapters / canister

- **Performance and Features**
- **Integrated extreme high IOPs storage Pool**
- **Dedicated performance pool (12x drives)**
- **Easy configuration and setup**
- **Automatic data migration between pools**
- **Integrated RAS support**

# Introducing IBM Storage Scale Ustore feature! Caching and Acceleration to the compute

## Accelerate AI and Analytics by storing the data as close to the compute as possible

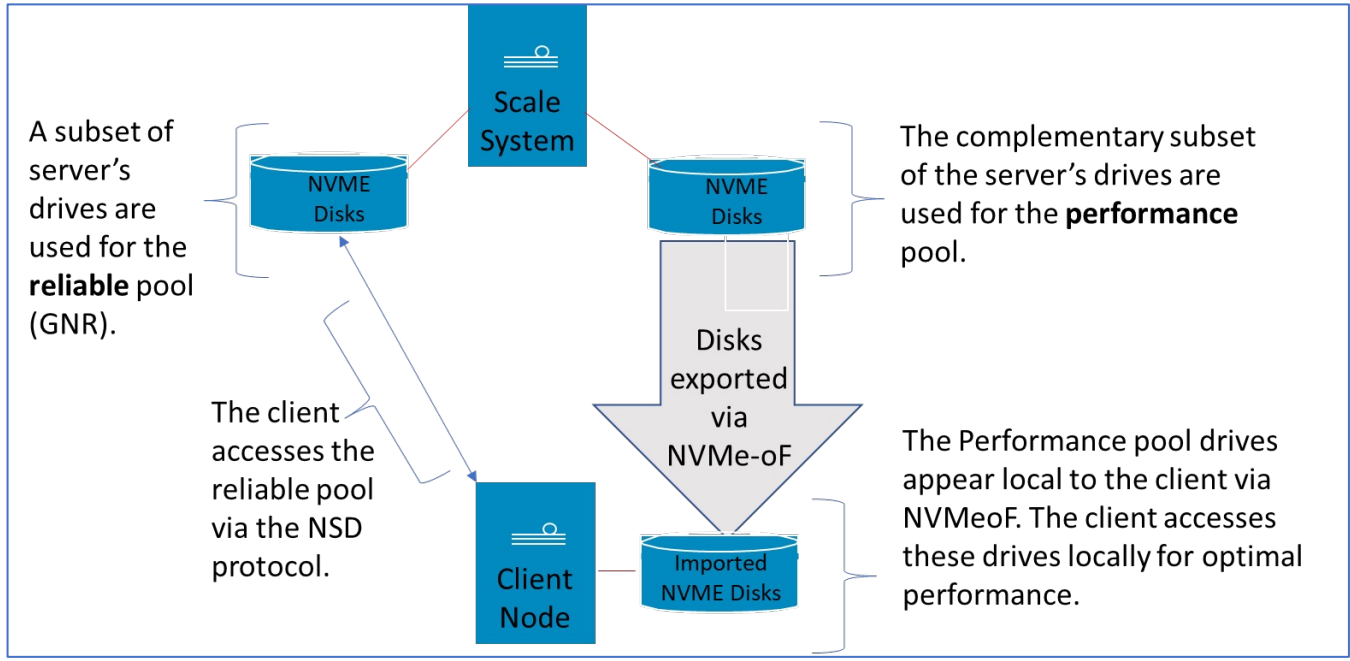
Storage Scale

**Leverage both shared storage (e.g. NVMeoF) and storage inside of the compute node**

Support **Asymmetric Replication** with one erasure-encoded copy of the data and one performance copy for high-speed access

Creates a **Shared Co-operative Cache** across all compute nodes. Any node can access all cached data, regardless of physical location.

The first release, writes update all copies. In a follow-on release, allow writes to performance copy only with **Eventual Reliability** (e.g. Burst Buffer)



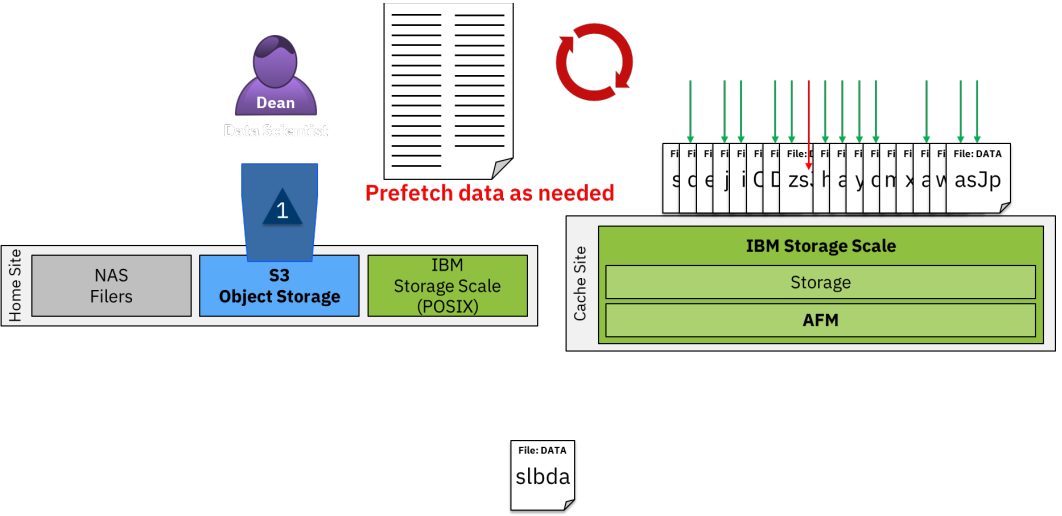
# Read Append
## Performance Improvement for small IO transfers

### Time to read 20GB File – IO Size 8KB

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

# Abstraction and Acceleration Services – Active File Management (AFM)

Storage Scale




Validation Tool using REST API and Certification Tool using: mmafmtransfer

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

Support of multi site replication in AFM to cloud object storage MU mode
- Support currently enabled with similar API providers. :
  - AWS S3, IBM COS, Ceph, Minio (using common S3 API)
  - Azure blob at multiple sites

New callback added for AFM DR RPO sync: afmDRRPOSync
- Acknowledges RPO peer snapshots are created successfully on both sites.

New cluster parameter : '**afmObjMUCheckFName**'
- Enable if AFM fileset is converted to MU mode fileset. This will reset old AFM attributes and update with new attributes.

New cluster parameters : '**afmFastLookup** & **afmLookupMapSize**'
- During lookup and readdir operations, AFM required to run a read-only mutex on the AFM fileset to collectthe information about last successful refresh operation.

CES S3 (Nooba) use-case with AFM Testing
- Using CES S3 Nooba as object store for AFM-S3 fileset.
- Configure CES S3 on top of AFM filesets

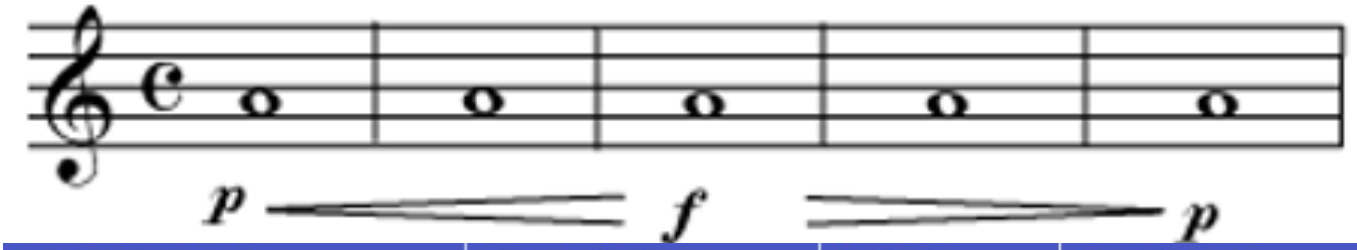
# Abstraction and Acceleration Services – Dynamic Page Pool

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.




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**

| 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 |

# Default configuration changes

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 |

# Management and Orchestration

# Data Orchestration

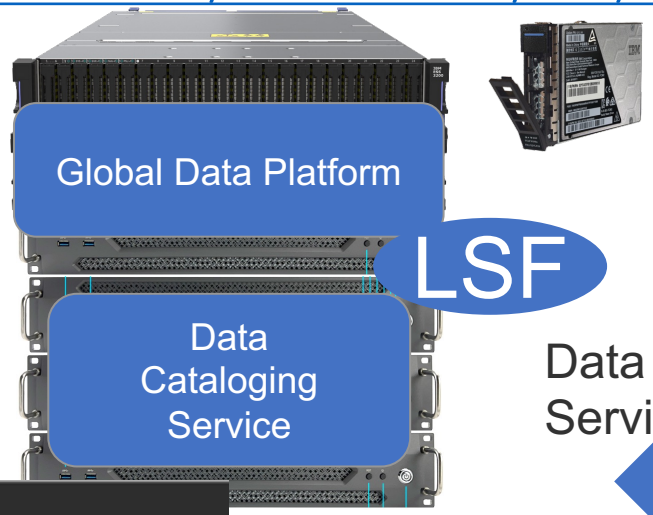
# Data Orchestration providing a future scalable architecture with Tape - IBM S3 Deep Archive

https://www.ibm.com/downloads/cas/84KXPK05

IBM Storage & SDI
Storage Scale

50TB

Access Services

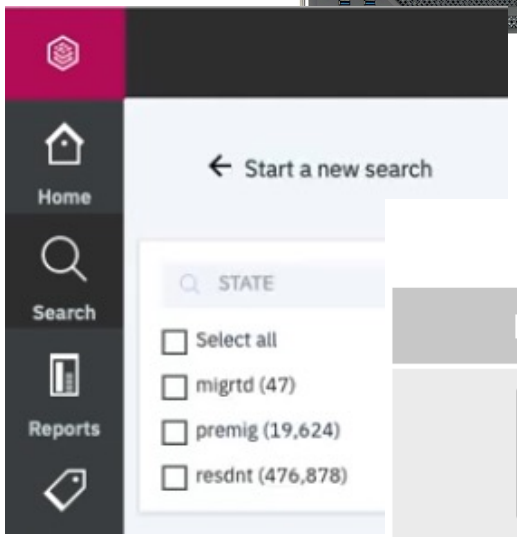


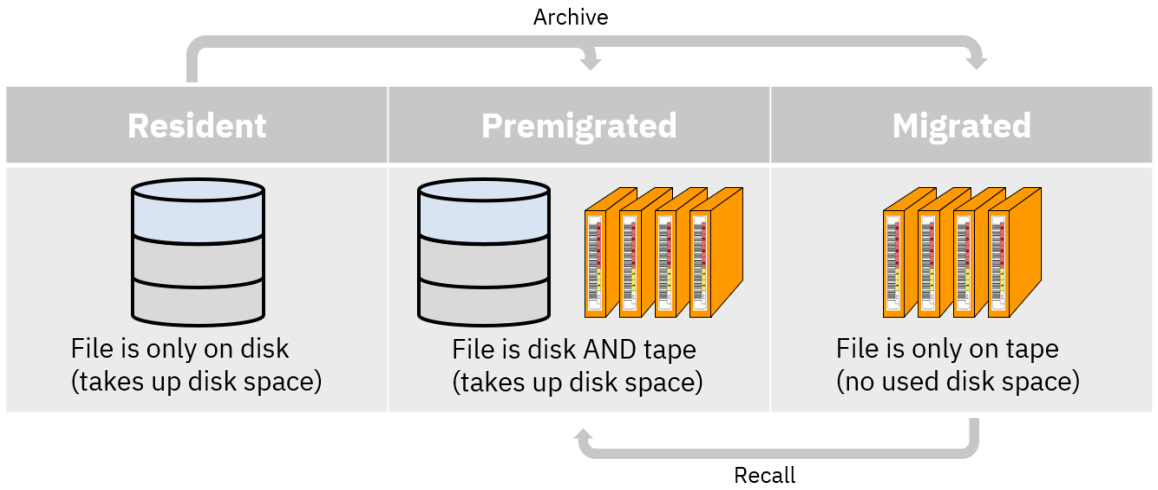



**Architecture to ingest and recall data to tape libraries**

Data Orchestration Services

Orchestration




TS4500 or Diamondback Racks

# Monitoring and Health

Storage Scale

mmhealth event list resolvable | tips

mmhealth node show --last-check

CES S3: monitoring, FTDC and call home coverage

GNR monitoring: failure detection speedup via callbacks

Check for lost NSD paths

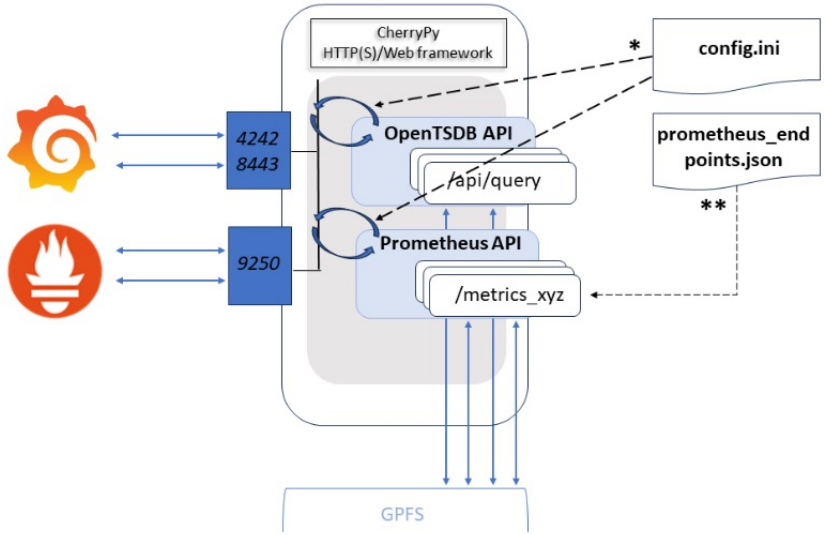
**mmhealth** TIPs for new **mmchconfig** defaults

Unified Call Home for all ESS (no need for ESA agent)

Upload to pmr with: **gpfs.snap --pmr XXXXXX**

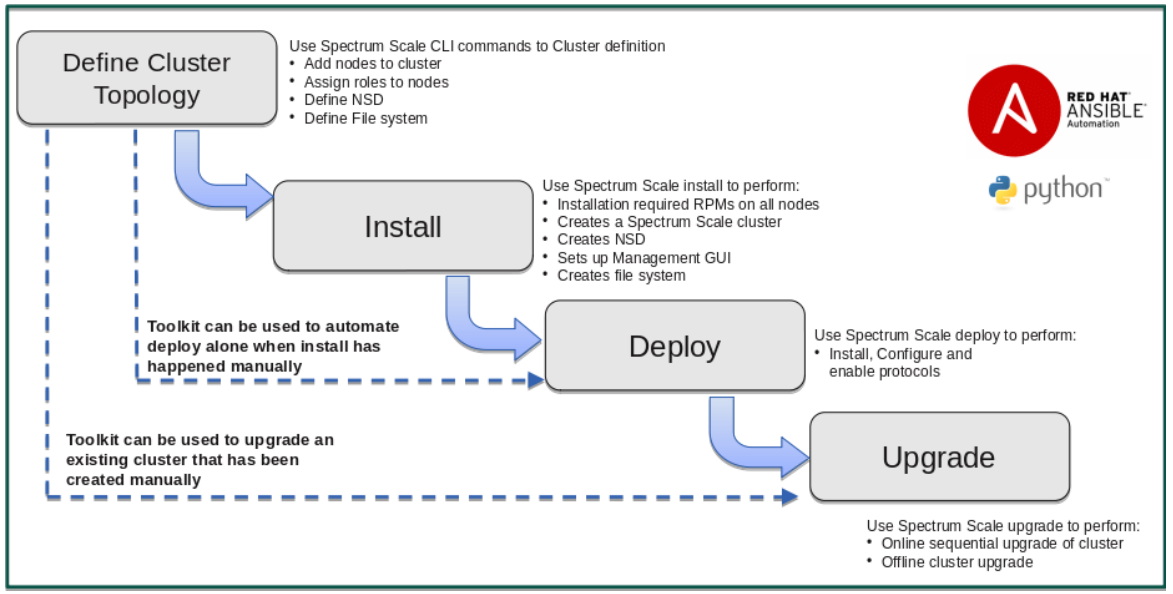
Prometheus Exporter for Zimon Data

| Callback | Resulting Events |
|---|---|
| postRGTakeover | gnr_rg_takeover<br>gnr_rg_takeover_warn |
| postRGRelinquish | gnr_rg_relinquish<br>gnr_rg_relinquish_warn |
| rgPanic | gnr_rg_server_panic |
| daRebuildFailed | gnr_da_rebuild_failed |



# Storage Scale Deployment Toolkit

Storage Scale




CES S3 based Object protocol toolkit support for complete workflow.

ARM tech preview support on UBUNTU.

Toolkit install and upgrade restrict and tolerance for RHEL 7 node deployment

Toolkit support for mmhealth cluster show migration from mmces state command

Extended OS currency

Scale System Utility Node Protocol Services node certification with 5.2.0 Toolkit.

ECE install toolkit enhancement to create file system with multi-vdiskset

# Orchestration Services – Cloudkit!

Storage Scale

## What is Storage Scale Cloudkit?

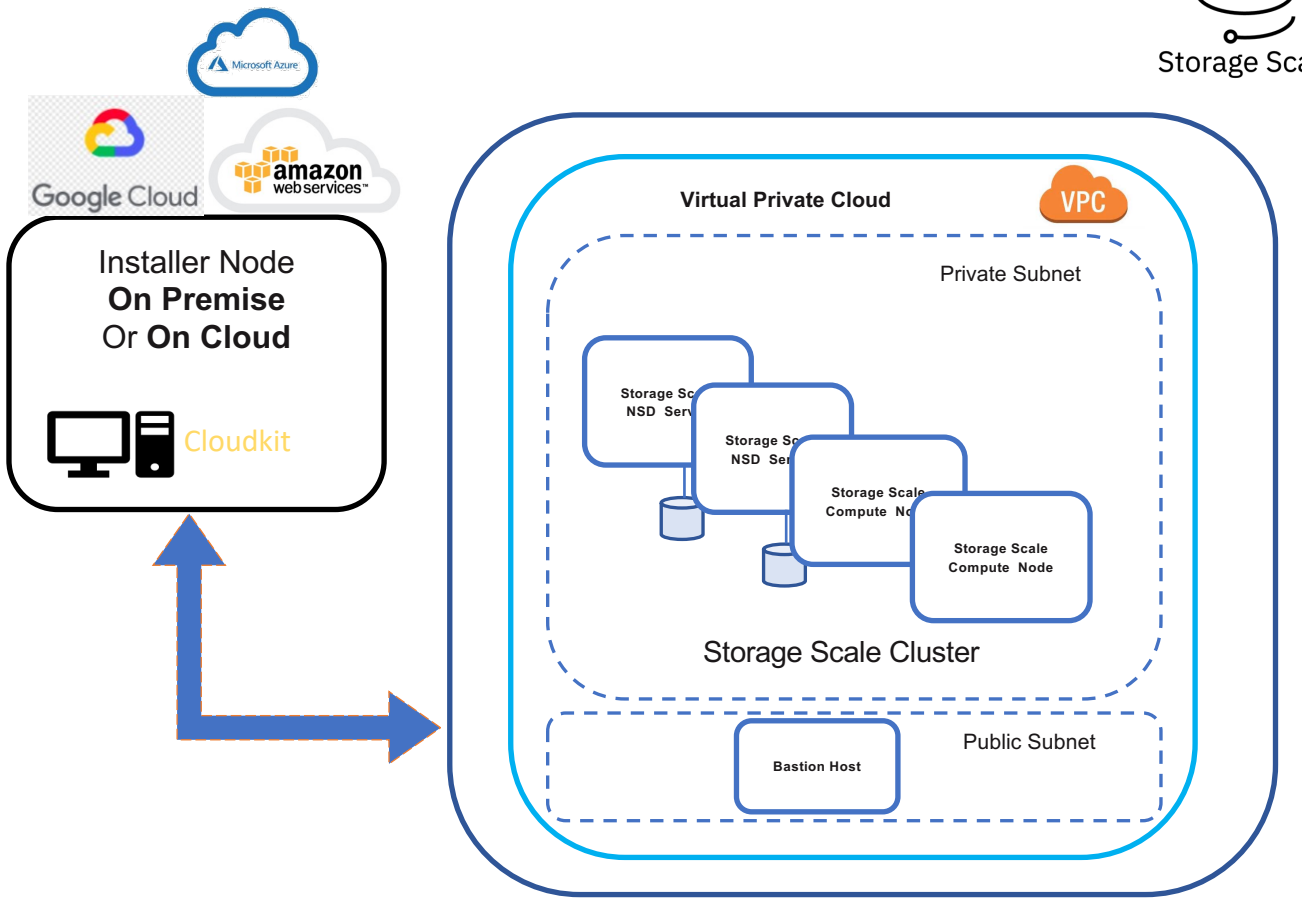
Create Storage Scale clusters on the cloud with

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 (Tech-preview on GCP, GA on AWS), Upgrades
- Tech-preview support for fleet support on AWS and GCP cluster instances
- Tech-preview Azure deployment

# Orchestration Services – Cloudkit!

## Cloudkit Fleet Performance Measurement

Google Cloud

amazon web services

Storage Scale

- Fleet (aka. Rapid expansion of compute/client nodes in a remote mount setup)
  - Traditional `mmaddnode` takes around `45sec` to add a single node (batching of 50 nodes is possible but it's is a sequential operation)
  - Quickly adding large nodes and deleting nodes before/after the burst situation is problematic.
  - Silent disappearance of nodes is a problem (as log recovery takes lot of time), which limited usage of spot instances (which comes at much cheaper price than an on-demand price).

| Fleet Size | Minimum Time | Maximum Time | Average Time |
|---|---|---|---|
| 1 | 42.69s | 42.69s | 42.69s |
| 100 | 21.86s | 32.33s | 26.94s |
| 200 | 20.98s | 28.85s | 23.74s |
| 300 | 10.74s | 1m7s | 22.84s |
| 400 | 19.38s | 1m19s | 24.91s |

# Orchestration Services Services – mmcrfs, mmbackup, mmxcp


## mmcrfs

- --inode-segment-mgr {yes | no} - feature useful for workloads of parallel file creation
- default value of the -n option to mmcrfs is now interpreted as 512; before was 256.
  - Scale now allocates at least 512 inode alloc regions, instead of at least 256 inode alloc regions.
    - Allows larger clusters (over 100 nodes) that are created in this manner to have better inode allocation performance.

## mmbackup

- --other-policy-options policyOptions - Allows to specify multiple mmapplypolicy options such as --sort-command, --sort-buffer etc. [--other-policy-options "—sort-buffer-size 3%"]
- MMBACKUP_PROGRESS_CONTENT to instruct mmbackup to display total data size transferred

## mmxcp --hardlinks and copy-attrs/verify

- Add options to override maxfiles and threadlevel for mmapplypolicy
- added that executes an additional pass through the source files searching and copying hardlinked files as a single batch
- appendonly and immutable, have been added which copies and verifies the appendonly and immutable attributes, if present.

# mmfsckx –repair

## May sacrifice speed for file system uptime

Storage Scale

**Usage:**

```
mmfsckx Device [--repair] [--check-reserved-files-only ]
               [--max-threads MaxThreads] [--max-pagepool-percent MaxPagepoolPercent]
               [--qos QosClass] [-N {all | mount | Node[,Node...] | NodeFile | NodeClass}]
```

- *mmfsck will continue to be supported*

- Directory related corruptions
  - Bad directory blocks
  - Bad directory entries
  - Incorrect file link counts
  - Orphaned inodes
  - Fileset identifier validation

- To repair these run the mmfsck command with the file system unmounted.

- *Work in progress* to include handling the above corruptions in future releases.

| Type of Error | Description |
|---|---|
| Reserved file corruption | Reserved files are files internal to scale and not user visible. They contain all file system metadata. |
| Duplicate disk addresses | An inconsistency caused when two or more disk addresses in the same or different files point to overlapping sectors on the disk |
| Improperly freed blocks | Blocks that are marked as available in the block allocation map system file but are referenced in some files. |
| User file metadata corruption | This includes the file inode, indirect blocks, and extended attribute blocks. |
| Lost blocks | Blocks that are marked as in-use in the block allocation map system file but do not belong to any file. |
| Inconsistent inode allocation status | An inconsistency where in-use files are shown as free in the inode allocation map or vice-versa. |

# Resiliency Services

# Modernization of Scale v2

**Multi-Tenancy Improvements**

**Resiliency**

- Failure Isolation / Blast Radius
- Network Fault handling
- Bad / Slow Nodes
- Rapidly changing configuration

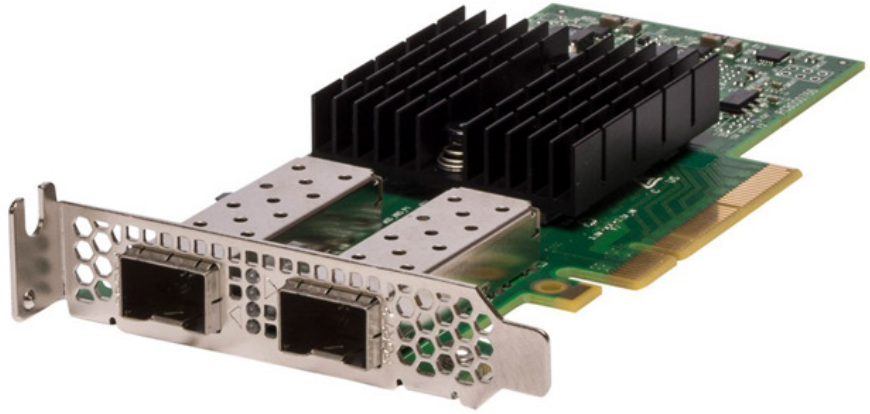
**Performance Isolation**

- QoS / SLA
- Data and Metadata Isolation
- Shared Metadata contention (Quotas & Locks)

**Manageability**

- Rapid deployment / shutdown
- Parallelism on all operations
- Management Isolation
- First Time Failure Data Capture
- Job level statistics & monitoring

orage Scale



# Network Resiliency

Scale is a clustered file system and depends on timely and reliable TCP/IP communication between all nodes in the cluster to ensure data integrity and good performance

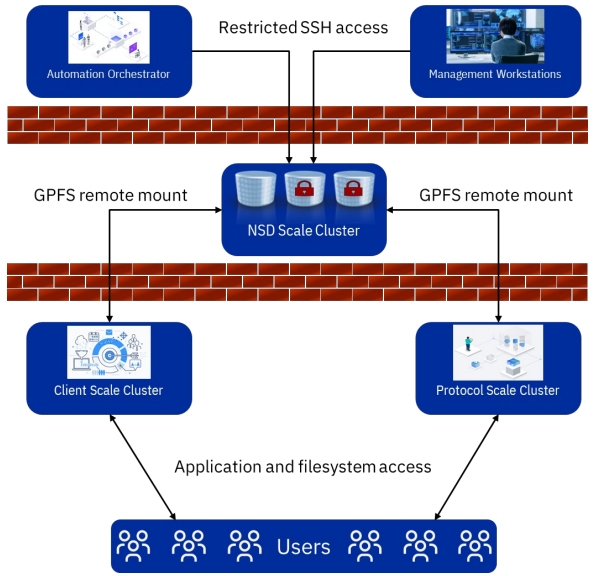
Proactive Reconnect, Prioritized critical RPC, improved mmnetverify, improved error logging and integration with System Health,

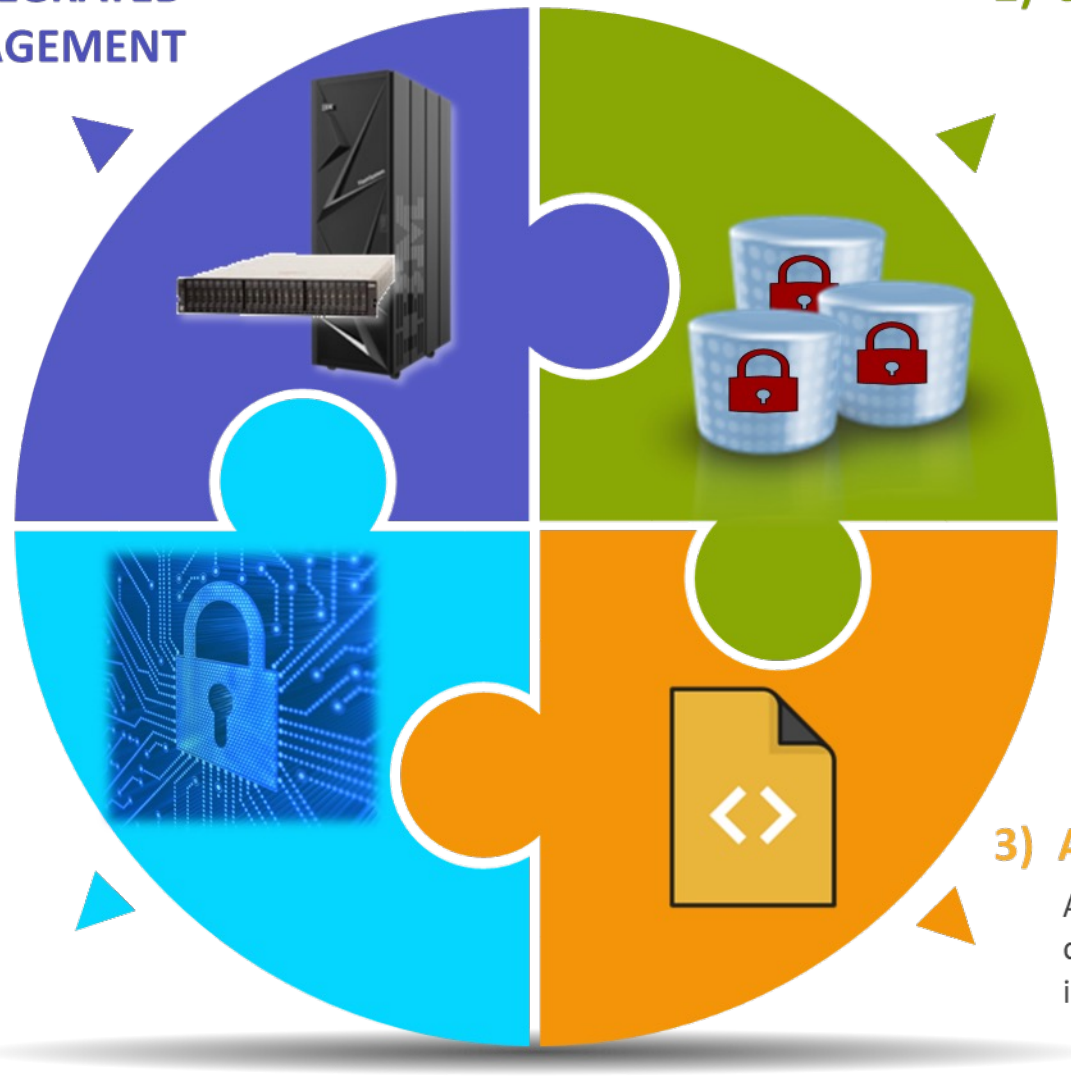
Efficiently detect and recover from the case of failed nodes to give up tokens more rapidly

# Resiliency Services: Cyber Vault Framework

Storage Scale

## 1) IBM STORAGE WITH INTEGRATED COPY MANAGEMENT



## 2) SAFEGUARDED COPIES

Protected PIT copies:
Immutable and Isolated with stringent RBAC's

Safeguarded copy
https://www.ibm.com/docs/en/Storage-scale/5.1.5?topic=administering-protecting-file-data-Storage-scale-safeguarded-copy

sudo-wrapper set up and configuration
https://community.ibm.com/community/user/storage/blogs/nils-haustein1/2020/12/17/Storage-scale-sudo-wrappers

## 3) AUTOMATION

Automated data validation, data recovery and application integration

## 4) SECURITY INTEGRATION

Detect and respond to threats real-time from a wide variety of data sources.

# Modernization of Scale (MOS): Security

Tech Preview

## Security Improvements

Removal of SSH dependency

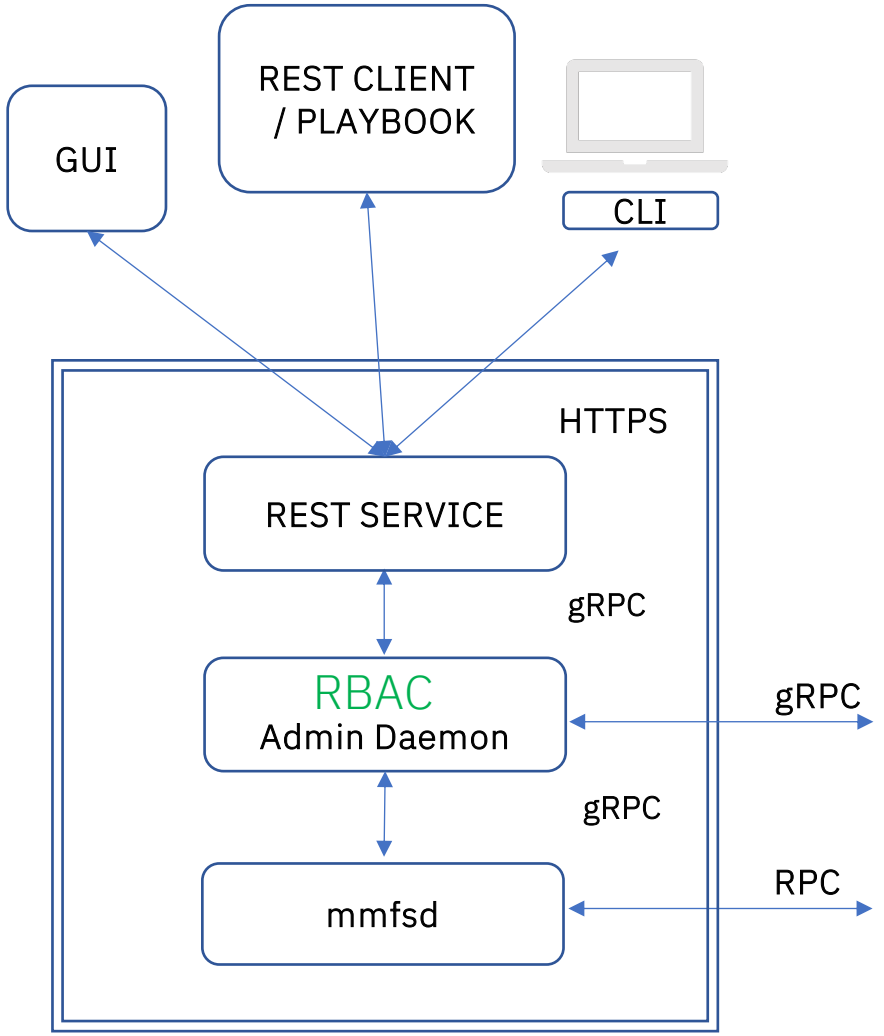
Removal of root requirement for control plane

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

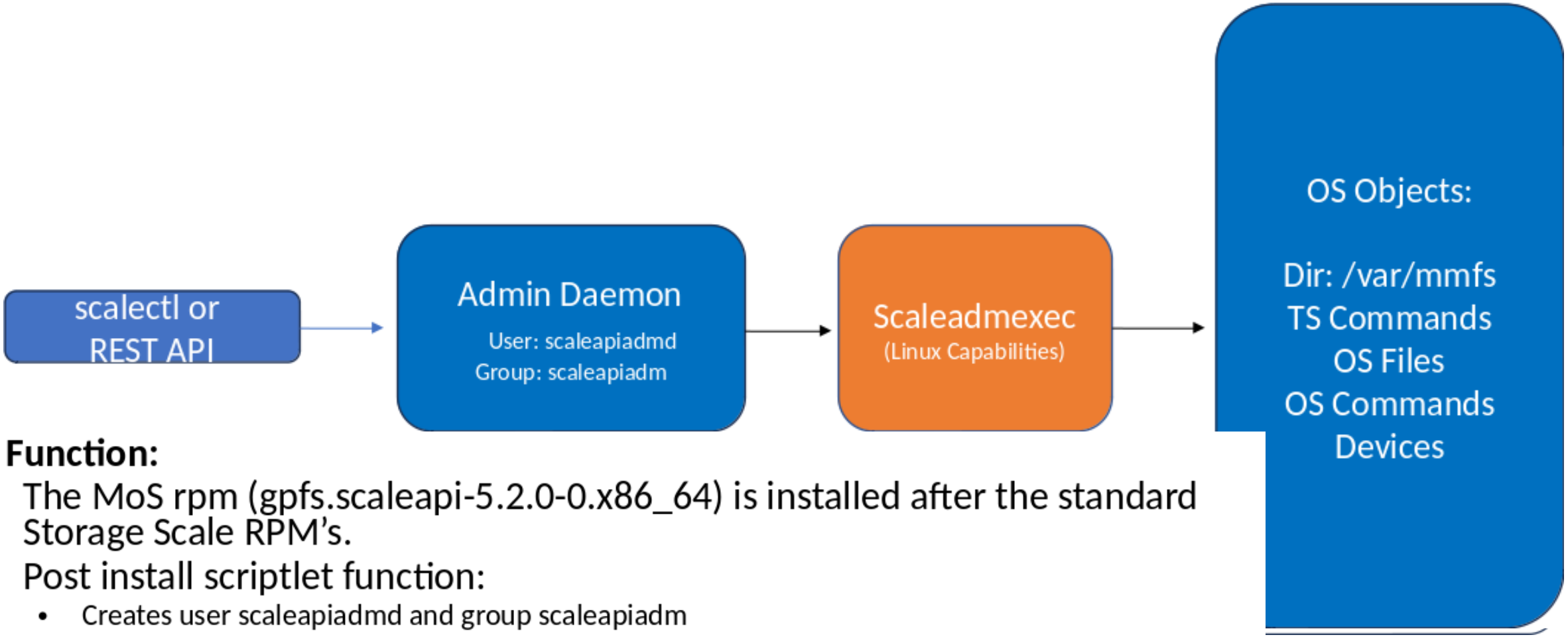


# Modernization of Scale

Storage Scale

Admin Daemon Client Interfaces: REST API's and scalectl (CLI client)

API Server Node: Listens and processes client requests




**RPM Function:**

- The MoS rpm (gpfs.scaleapi-5.2.0-0.x86_64) is installed after the standard Storage Scale RPM's.
- Post install scriptlet function:
  - Creates user scaleapiadmd and group scaleapiadm
  - Starts the scaleadmd systemd service as the non-interactive user scaleapiadmd
  - Creates or updates files under /var/mmfs to have a specific set of permissions as well as sets the MoS user and group where needed