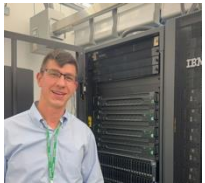
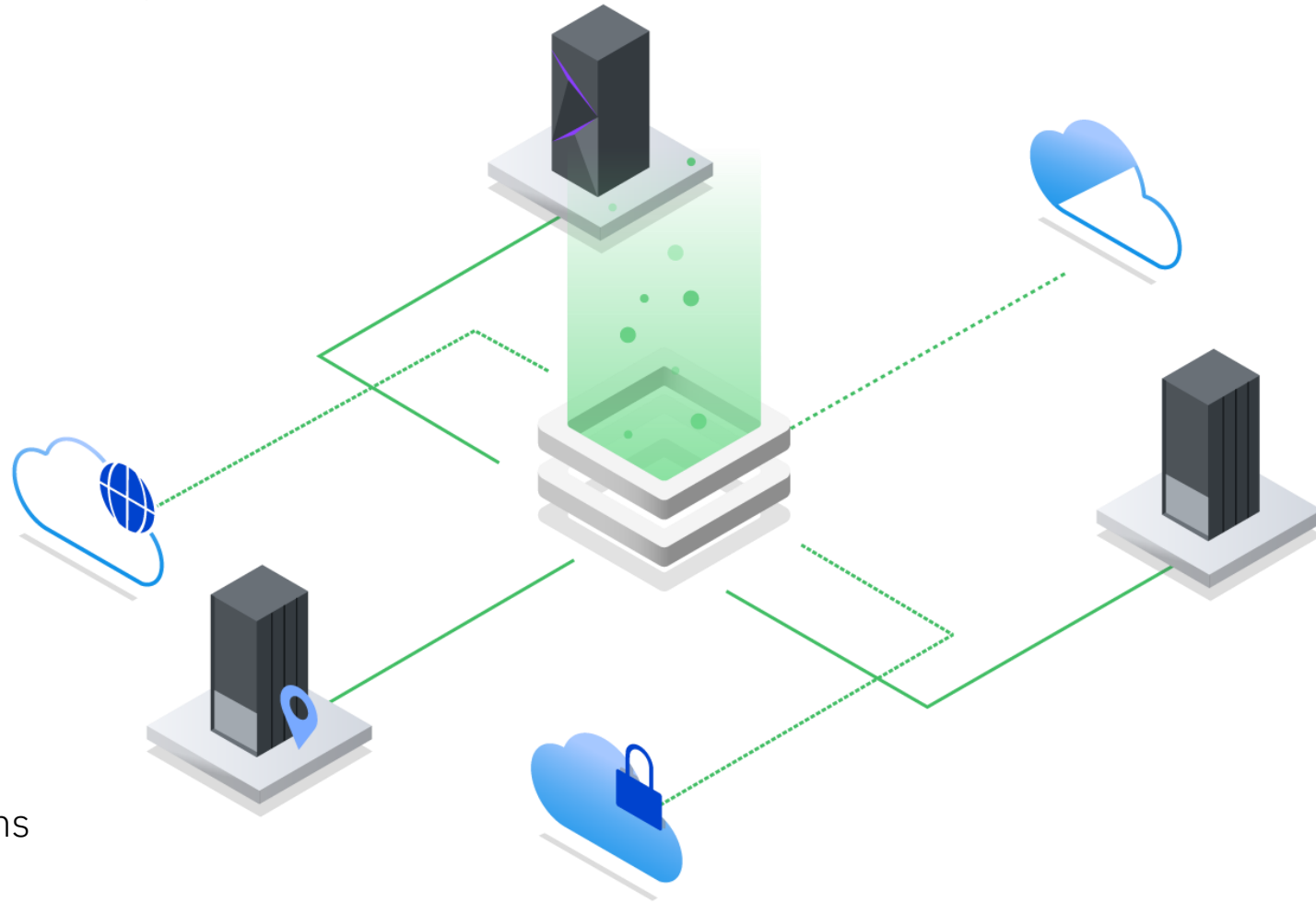


IBM Storage Scale

IBM's family of software defined storage, storage hardware, and storage management software



Matthew Klos
Senior Solutions Architect
Americas SWAT Team



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



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.

Leadership Matters!



For the **ninth consecutive year**, **IBM** is recognized **leader** in the Gartner Magic Quadrant for File and Object Storage Platforms.

Recognized for:

- ✓ AI Leadership
- ✓ Global File and Object Services
- ✓ Product Innovation



The IBM Storage Scale Journey and **Support reference guide**

Accelerate
AI & HPC

Scale System
for HPC & AI

Data
Acceleration
Tier

Disaggregated
Storage

NVIDIA AI

Storage
Certification

Content Aware
Storage

Accelerated
Data Path

Commercial
AI

Scale System
Usability

Consumability

Enhanced
File & Object

Foundation

API Driven
Control Plane

Containerized
Workloads

Scale as a
Service



Support reference guide for Storage Scale:

<https://www.ibm.com/support/pages/node/6252403>



Support reference guide for Storage Scale System:

<https://www.ibm.com/support/pages/node/6252477>



Note: In a Severity 1 case, please describe your business impact.



IBM Storage Scale Software Version

Recommendation Preventive Service Planning:

<https://www.ibm.com/support/pages/ibm-storage-scale-software-version-recommendation-preventive-service-planning>



IBM Storage Scale

Introduction

What is IBM Storage Scale?

IBM Storage Scale is an enterprise-grade parallel file system that provides superior resiliency, scalability, and control.

IBM Storage Scale delivers scalable capacity and performance to handle demanding data analytics, AI/ML, content repository and technical computing workloads.

Aka. Spectrum Scale and GPFS

Storage Scale Capabilities

Scalable performance
and Capacity

Data Caching and
Acceleration (AFM)

Multi-Protocol
Support

Update & Deployment
Automation

Data Resiliency
And Privacy

Intuitive Management
GUI, APIs and CLI

Container Native
Support and CSI

Quotas, QoS,
Snapshots, Tiering

Nvidia
GPU Direct

...and More!

IBM Storage Scale

Deployment Methods

IBM Storage Scale previously known as *Spectrum Scale* or *GPFS* is a software storage solution which can be deployed on-prem and in the cloud on existing SAN storage infrastructure, using IBM Storage Scale system, using Storage-Rich servers, or all the above!

1 Software Defined Storage

- Bring your own servers and storage
- Deploy On-Prem or in the cloud
- Supports variety of system types x86, Power, s390x, arm
- Highly Scalable

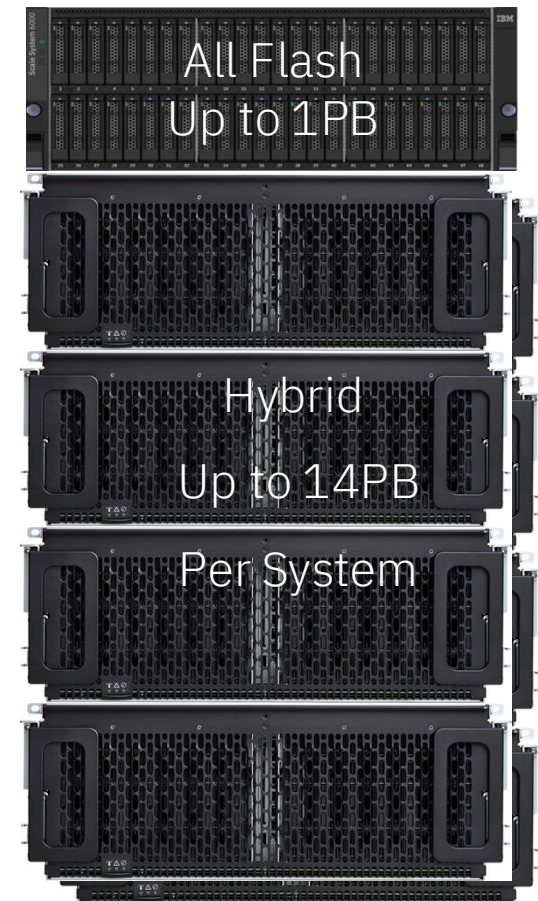
2 Storage Scale System

- Building block storage solution based on GNR
- NVMe, HDD, or Hybrid building blocks
- Highly Scalable – add more building blocks for more Capacity and Performance
- Six 9's of Availability

3 Erasure Code Edition

- Bring your own storage rich servers
- NVMe and HDD support
- Utilize GNR RAID
- Add more servers for more capacity and performance

IBM Storage Scale
System 6000



Scale Deployment model comparison

VAST -NVMeoF

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

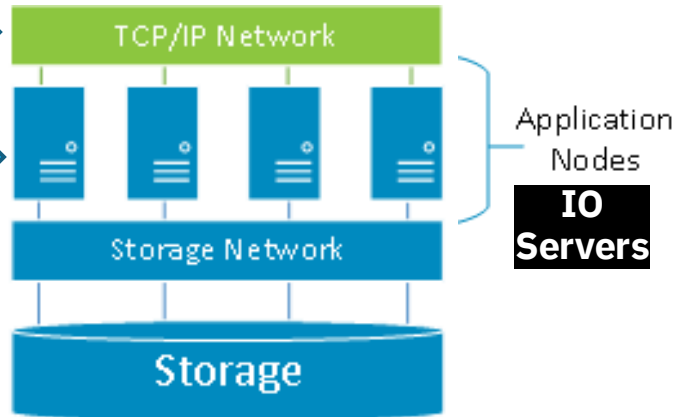
IBM Client-site GNR

RH OpenShift
Virtualization with
Scale

PureScale

Sailfish

DS8k SDS
Scale

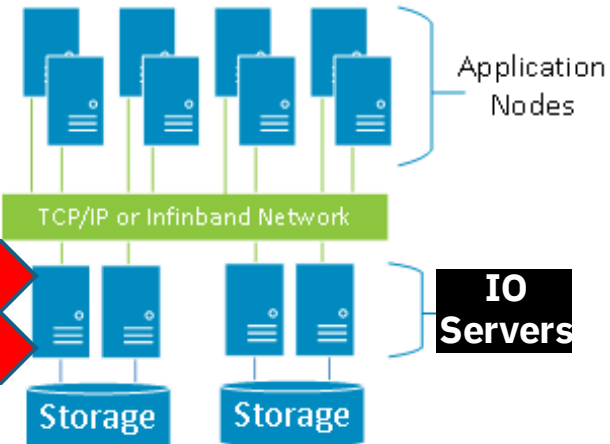


Unify and parallelize storage silos

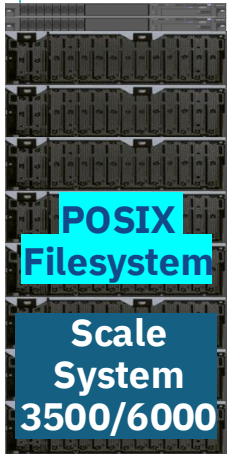
Twin tailed storage with erasure coding

Netapp

DDN



Modular High-Performance Scaling



POSIX
Filesystem

Scale
System
3500/6000

PowerScale

Weka

Pure

VAST -v2

Netapp v2

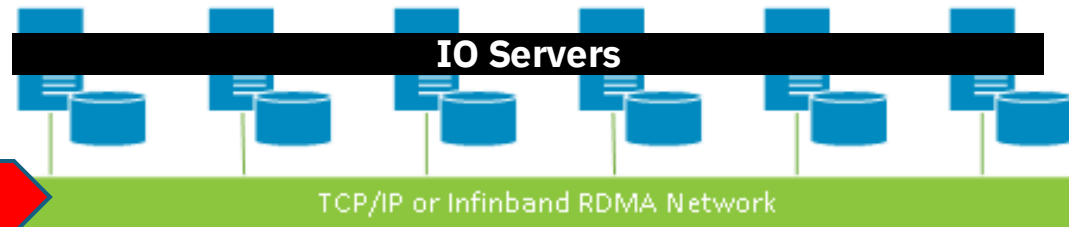
Scale (FPO)

Scale Erasure
Code Edition (ECE)

Fusion HCI

DDN Infinia

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



Span storage rich servers for converged architecture or HDFS deployment

IBM Storage Ceph

IBM Cloud
Object Storage
(COS)

Hadoop

IBM Storage Scale

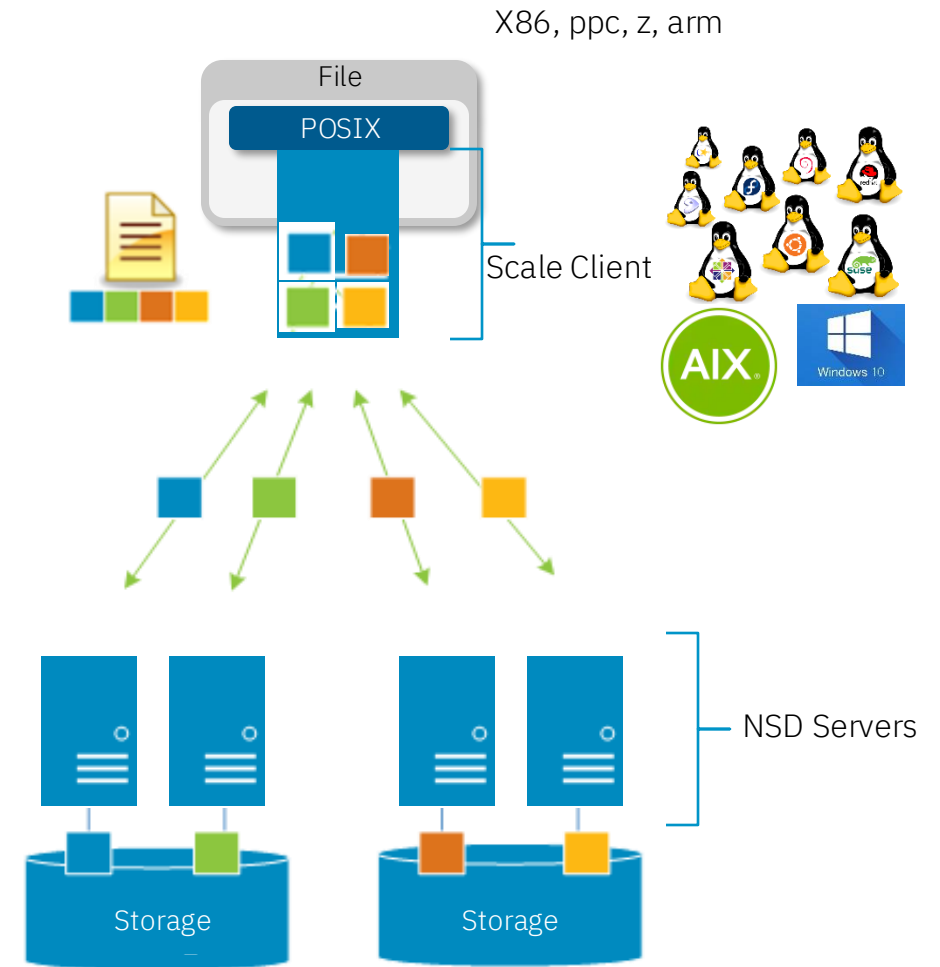
Parallel Architecture



Network Shared Disks (NSDs) - POSIX

- All NSD servers export to all clients in active-active mode
- IBM Storage Scale stripes files across NSD servers and NSDs in units of file-system block-size
- File-system load spread evenly
- Easy to scale file-system capacity and performance while keeping the architecture balanced

No Hot Spots!

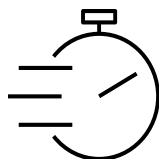


Data Service Requirements for Enterprise Storage Infrastructure



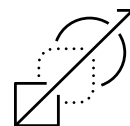
Data Access

- Improve data availability
- Support existing applications
- Ensure data access for power intensive, costly GPUs



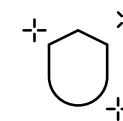
Data Acceleration and Abstraction

- Eliminate bottlenecks
- Preserve existing investment
- Hybrid cloud



Data Awareness

- Data orchestration that optimizes data feeding to the AI data pipeline
- Gain deeper insight into data

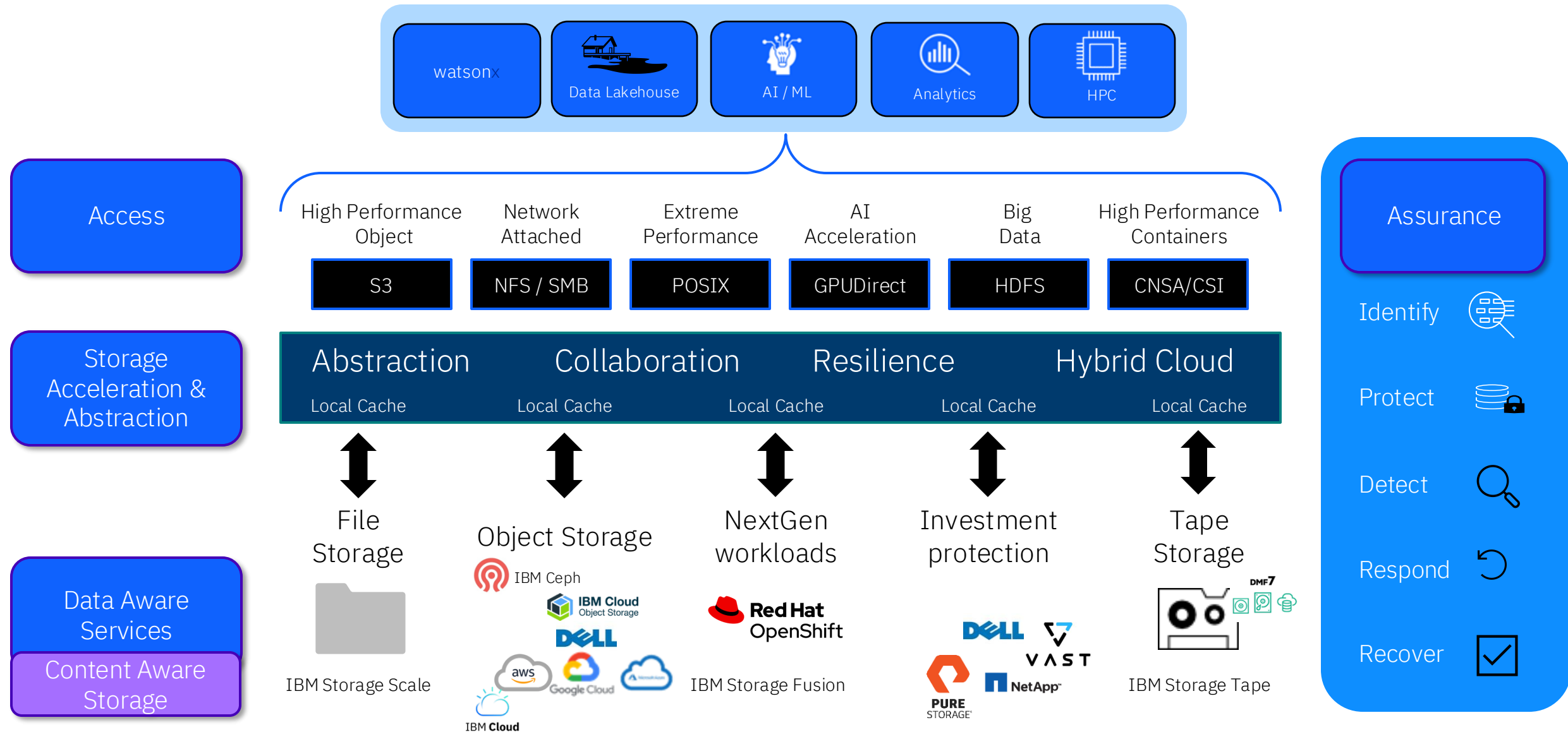


Data Assurance

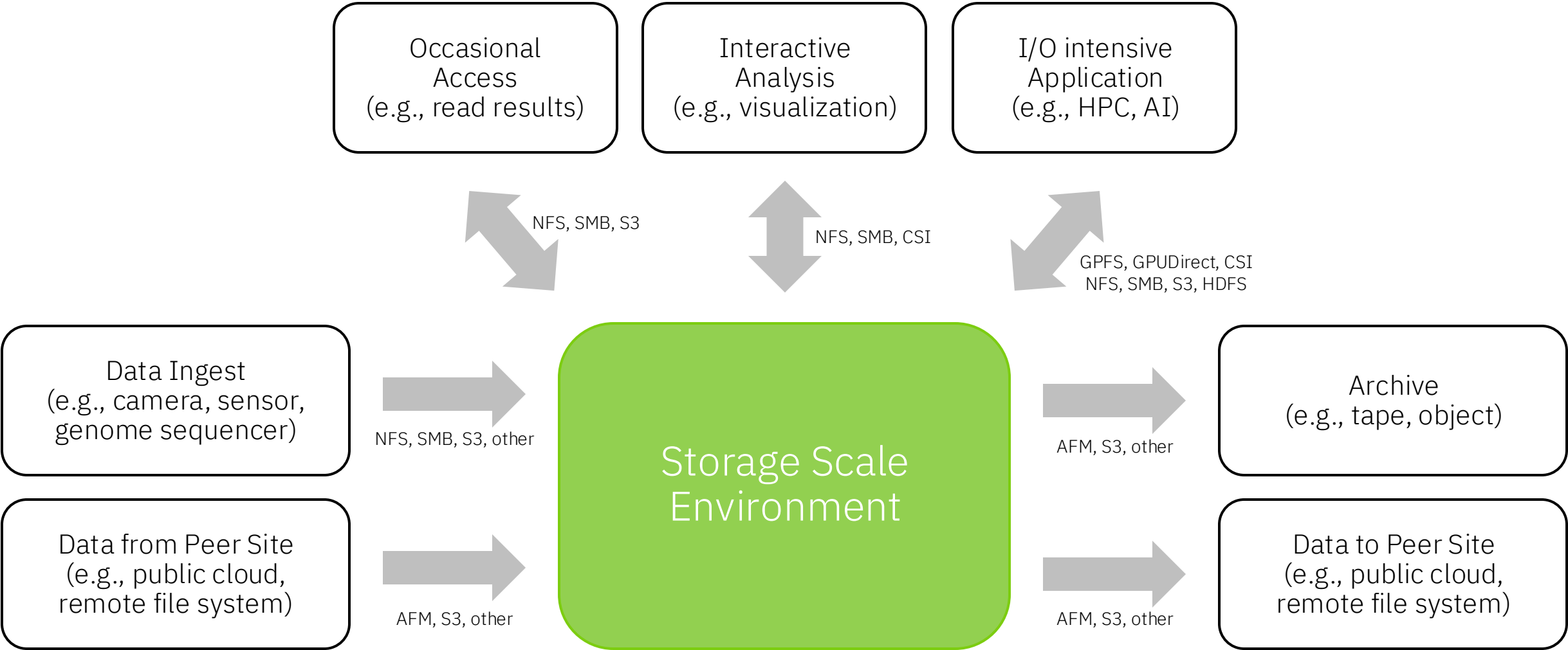
- Security of sensitive information from cyberthreats, especially in cloud computing
- Governance over data used, trained, turned, and inferred

IBM Storage Scale - global platform for storage (gpfs) services of unstructured data

1) Accessibility, 2) Storage Acceleration and Abstraction 3) Data Aware Services, and 4) Assurance



Data-Intensive Workflows for Unstructured Data



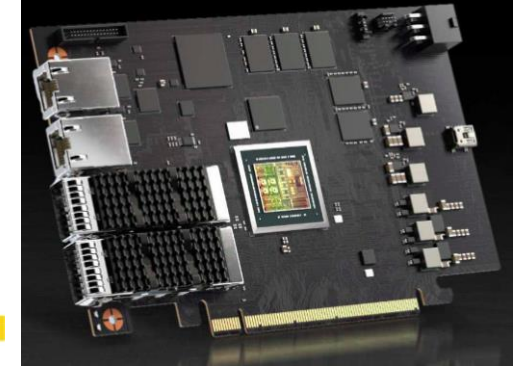
- Many unstructured data is generated and processed outside the data center.
- File and object protocols allows devices and applications to access and process data on remote servers and systems.
- Many of the remote applications and devices stick to one of the many file and object protocols (e.g., genome sequencers).

Access Services – ARM

GA! -The official Architecture name is aarch64

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

QuantaGrid S74G-2U



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

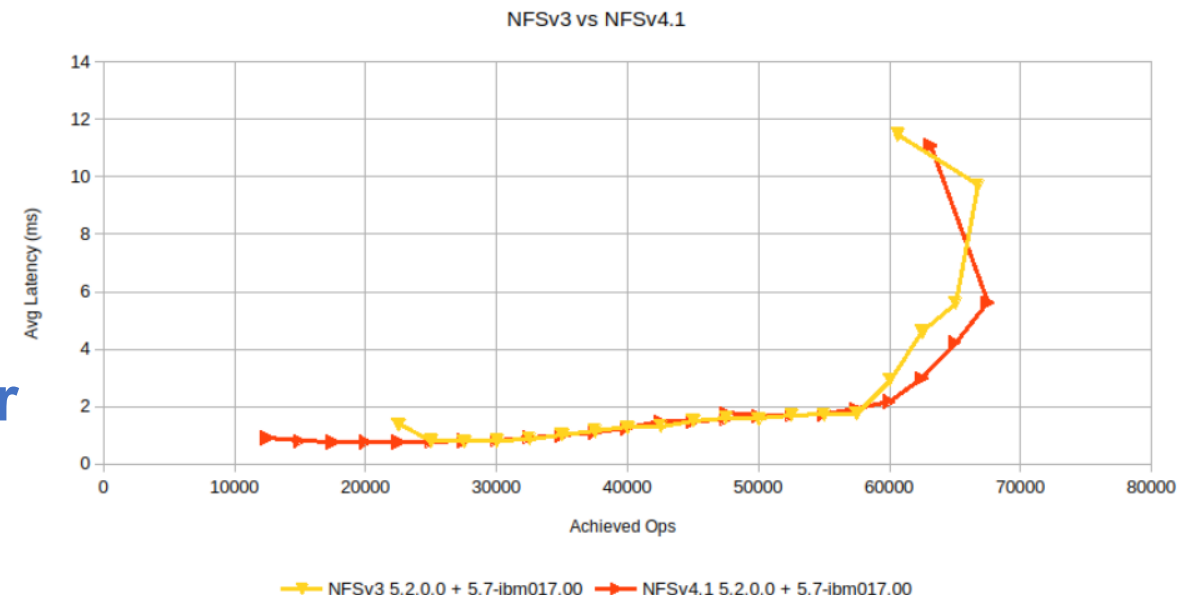
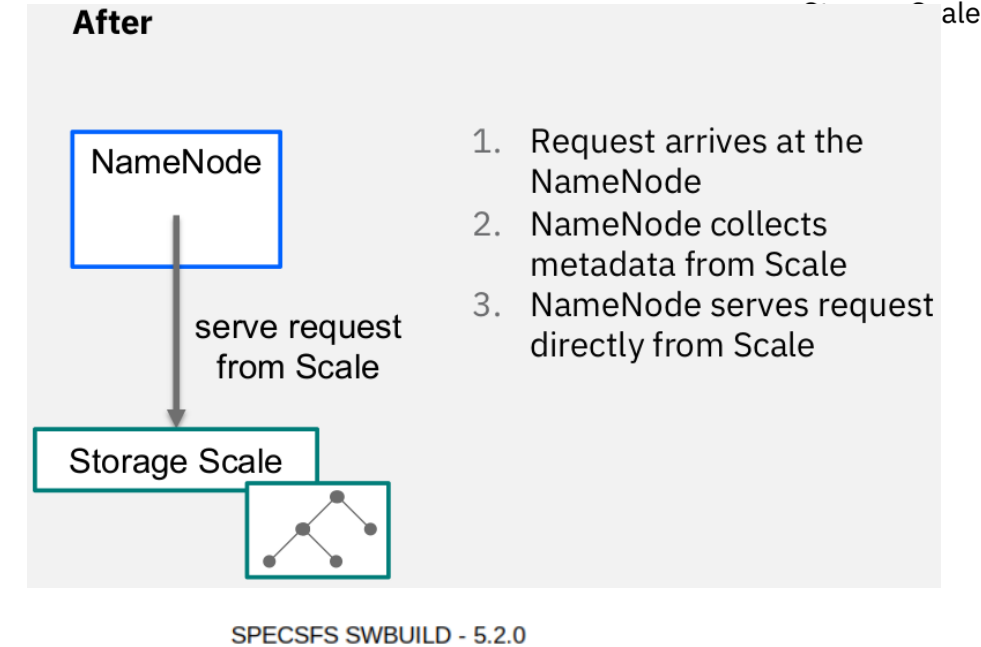
Access Services –NFS, SMB, HDFS

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:

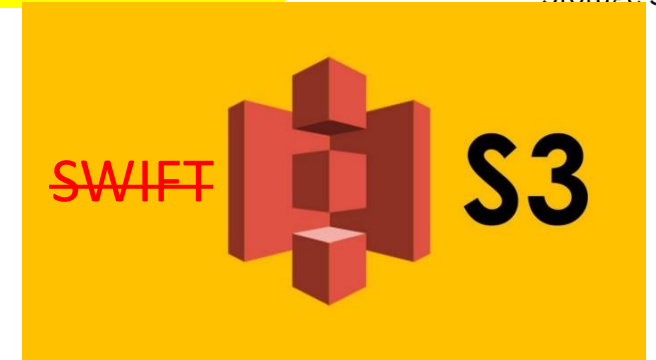
- NFS “meta data cache” component was revised resulting in significant performance improvements
- HDFS transparency metadata redesign
 - Full parallelism for RPC calls (GPFSNamesystem)
 - No more lock contention in NameNode
- **Continued partnership with Tuxera for high-performance SMB**



Access Services – High Performance Object 2.0!

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

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/rogerio-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 !

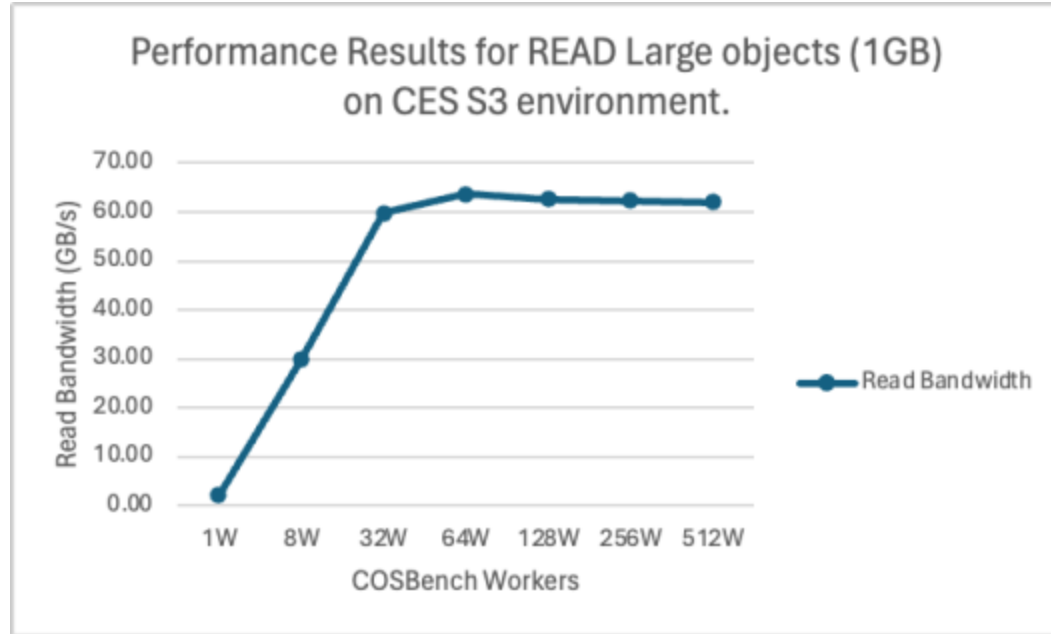
Deployment Requirements:

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

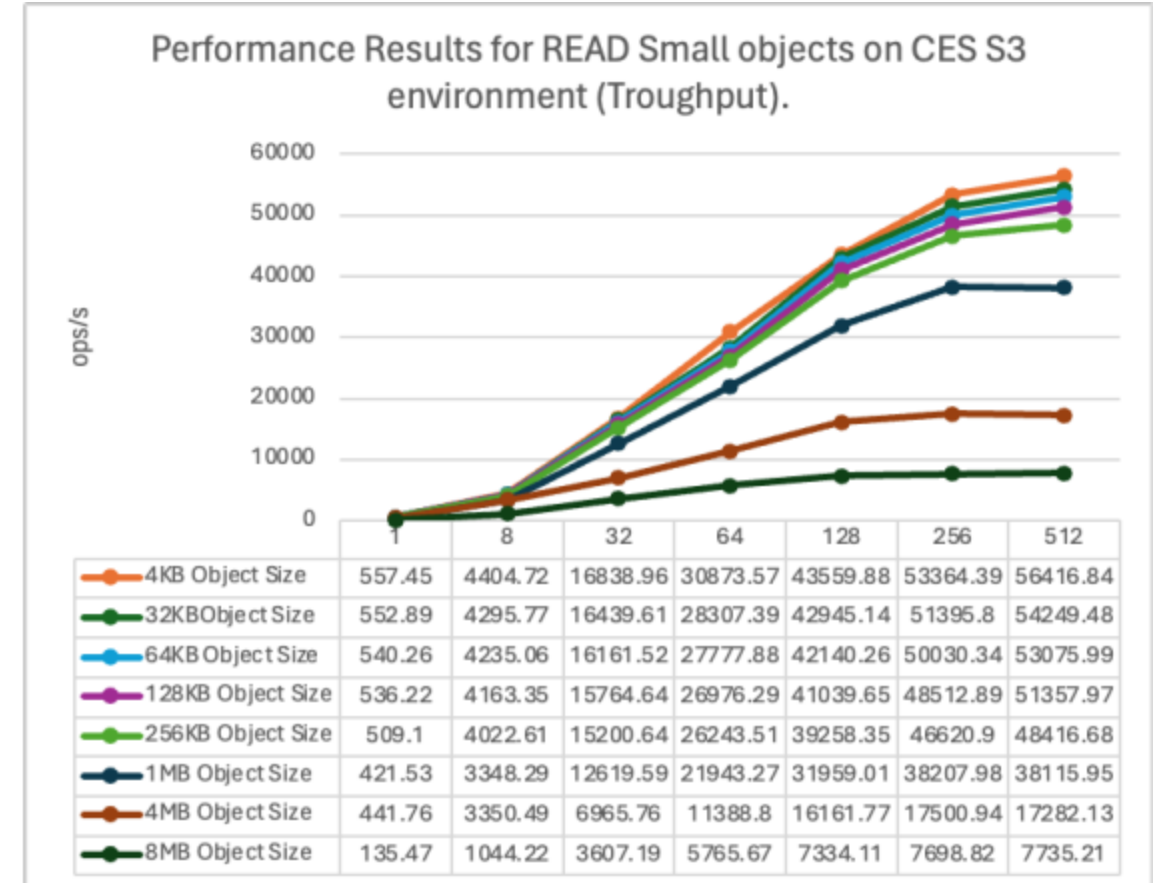
NO
Openshift
cluster
required !

Access Services – Object Performance



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.



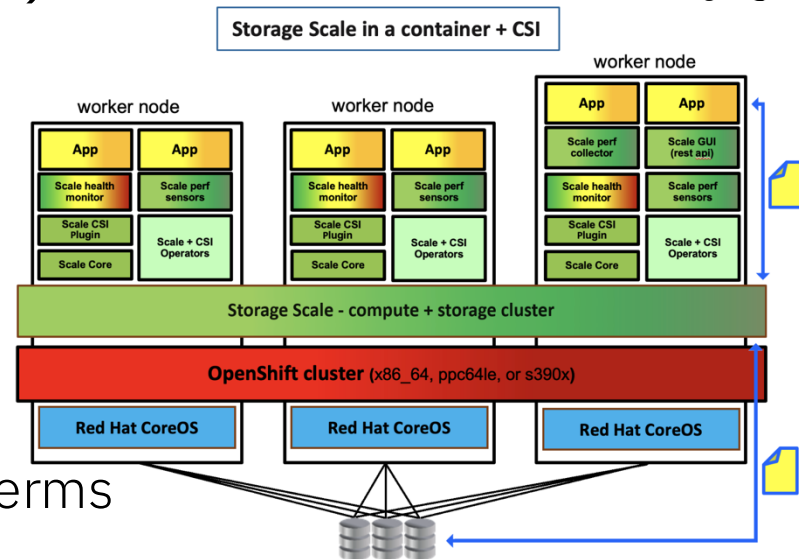
<https://community.ibm.com/community/user/storage/blogs/rogerio-rivera-gutierrez/2024/04/25/ibm-storage-scale-performance-ces-s3-tech-preview>

Container Native Storage Access (CNSA)

Improvements introduced in CNSA 5.2.3.0

OpenShift and
regular Kubernetes!

	Supported Versions
Kubernetes	1.29, 1.30, and 1.31
Architectures	x86_64
RHEL	8.10, 9.4 and 9.5
Ubuntu	22.04 and 24.04



- 5.2.3 GUI: ContainerOperator user role includes CsiAdmin user role perms
- Support for Google Kubernetes Engine (GKE) is technology preview
- AFM Cache Volumes Support
- RDMA support with IBM Storage Scale container native
 - RDMA over Converged Ethernet in Scale 5.2.3 (InfiniBand in 5.2.2)
- Shared disks can be used (have connection to at least two K8s nodes)
 - SNC disks are not supported (have connection to one K8s node)
- Red Hat OpenShift Virtualization (Fusion Access)
 - Goal is to gain VM use cases and customers
 - Uses CNSA local file system shared disk functionality: SAN attach FC / iSCSI
 - Announced @ Red Hat Summit

Demos are
available to show!

Works with
OpenShift
Virtualization
Engine!

Supporting multiple workloads with a single platform!

Speed AI results and accuracy

AI and HPC: NVIDIA GPU



Eliminate duplicate data

AI and HPC: Analytics



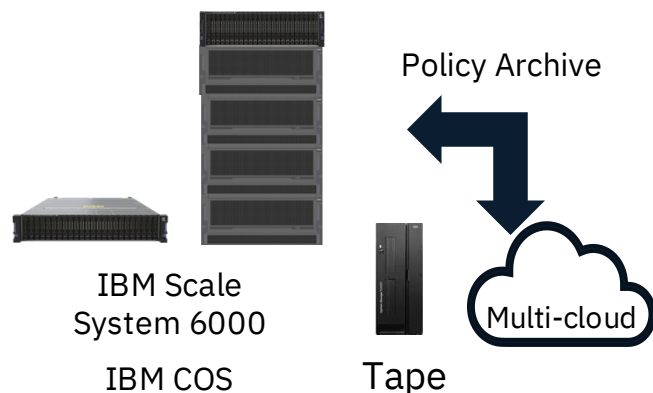
Speed deployment of NextGen apps

AI and HPC: IBM Data Fabric



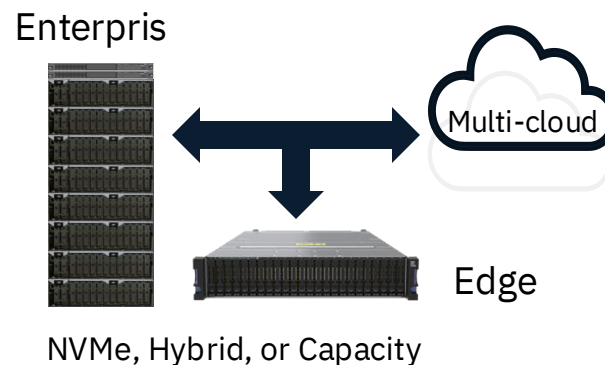
Improve application agility

Hybrid Cloud: Backup / Archive



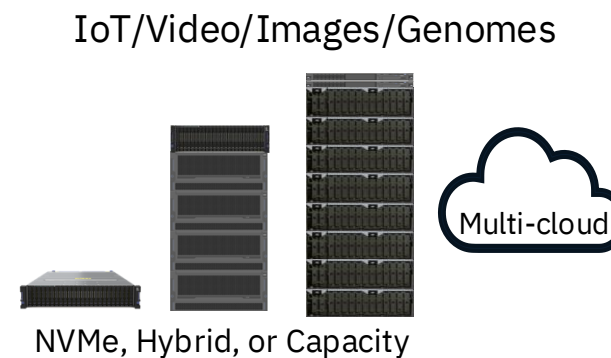
Support multiple concurrent projects

Hybrid Cloud: Data Lakes

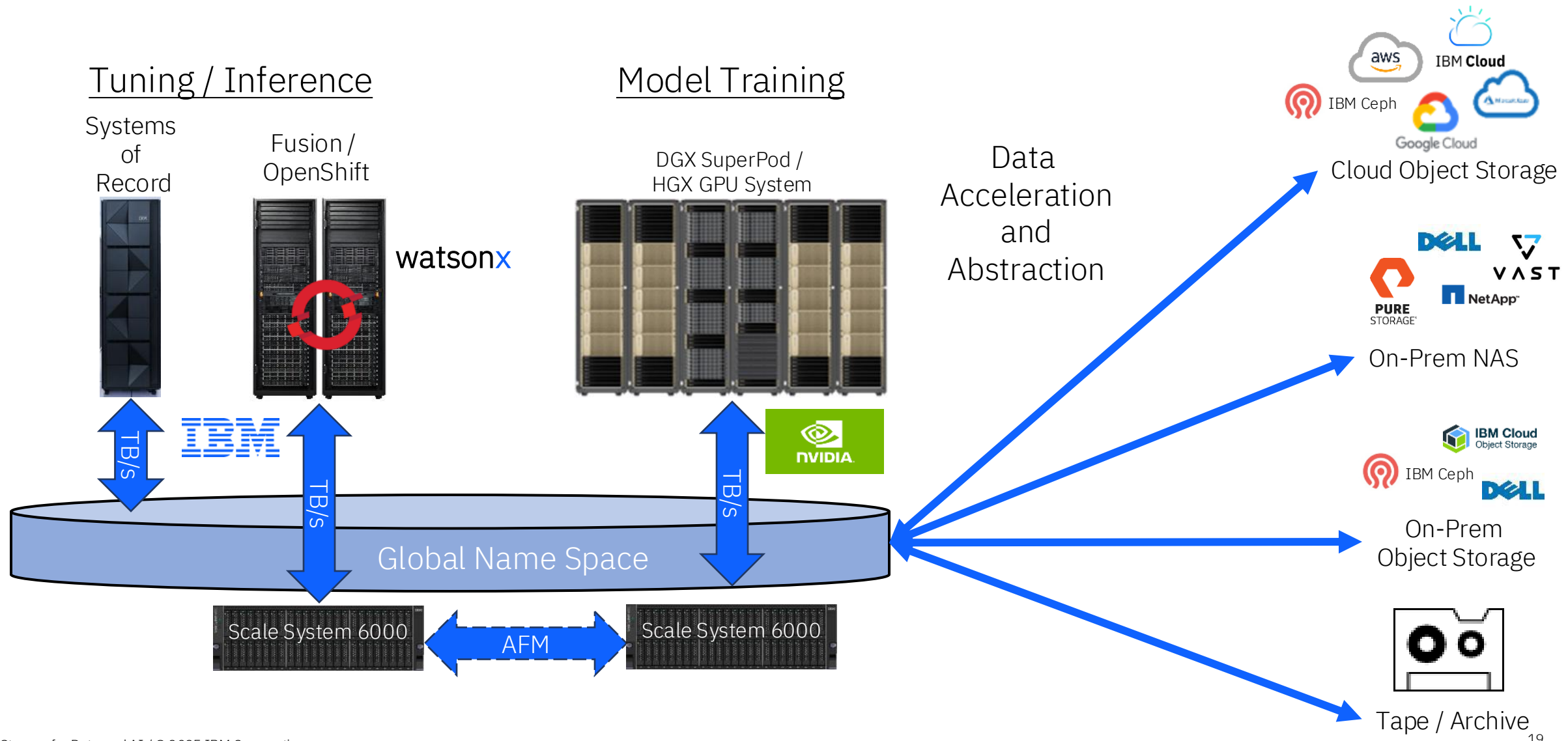


Control costs with real-time data

Hybrid Cloud: HPC



Storage Scale Global Data Platform



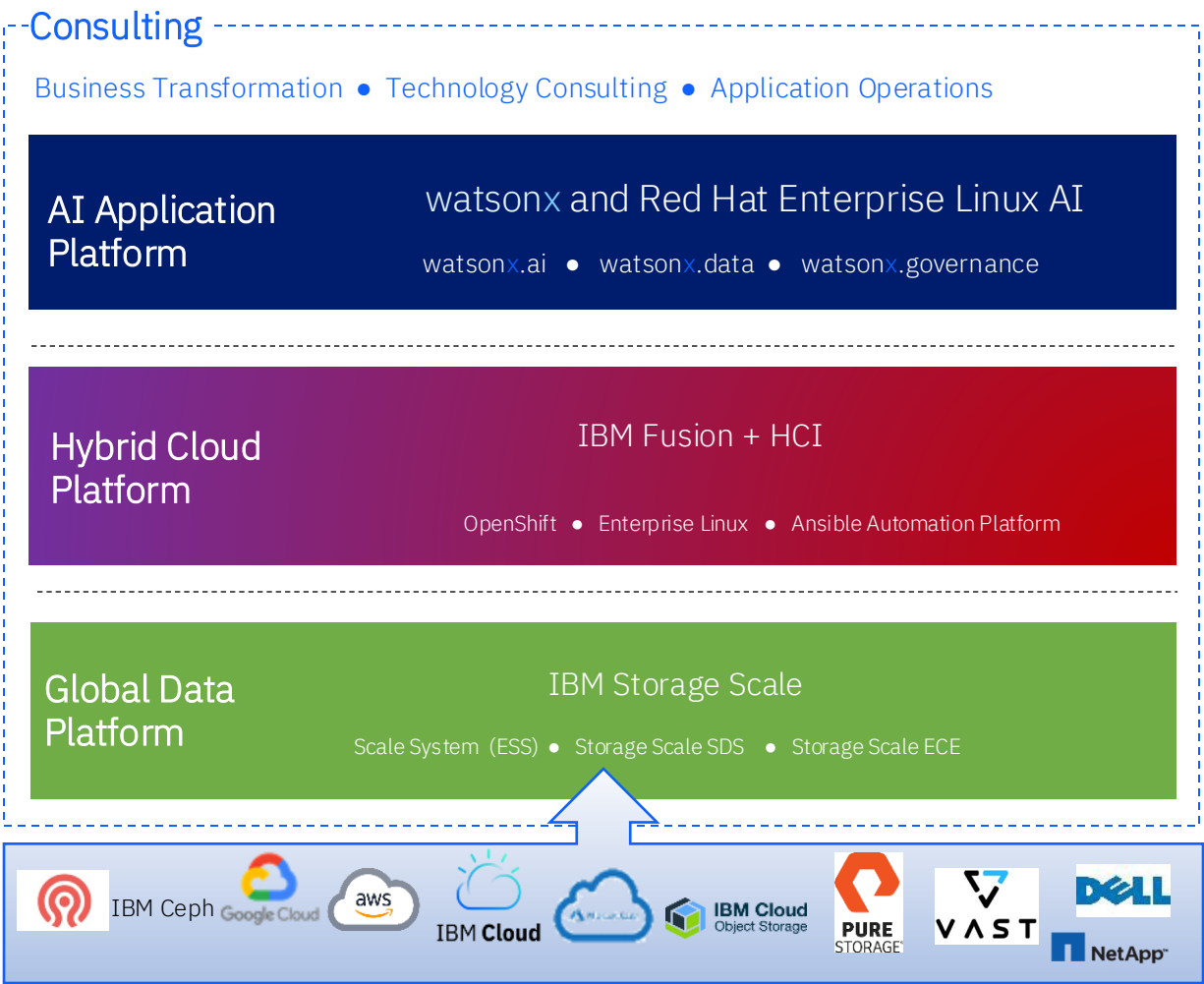
IBM's Enterprise Data and AI Solution

Our AI infrastructure stack spans between public cloud, private cloud, and on-prem compute systems to meet the distinct needs of current and future AI workloads.

Common storage and platform layers provide the foundation for AI and data platforms.

This hybrid AI infrastructure and platform enables business to scale, optimize, and deliver AI solutions.

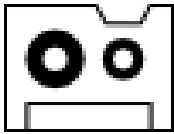
End-to-End Enterprise AI Solution



IBM Fusion HCI



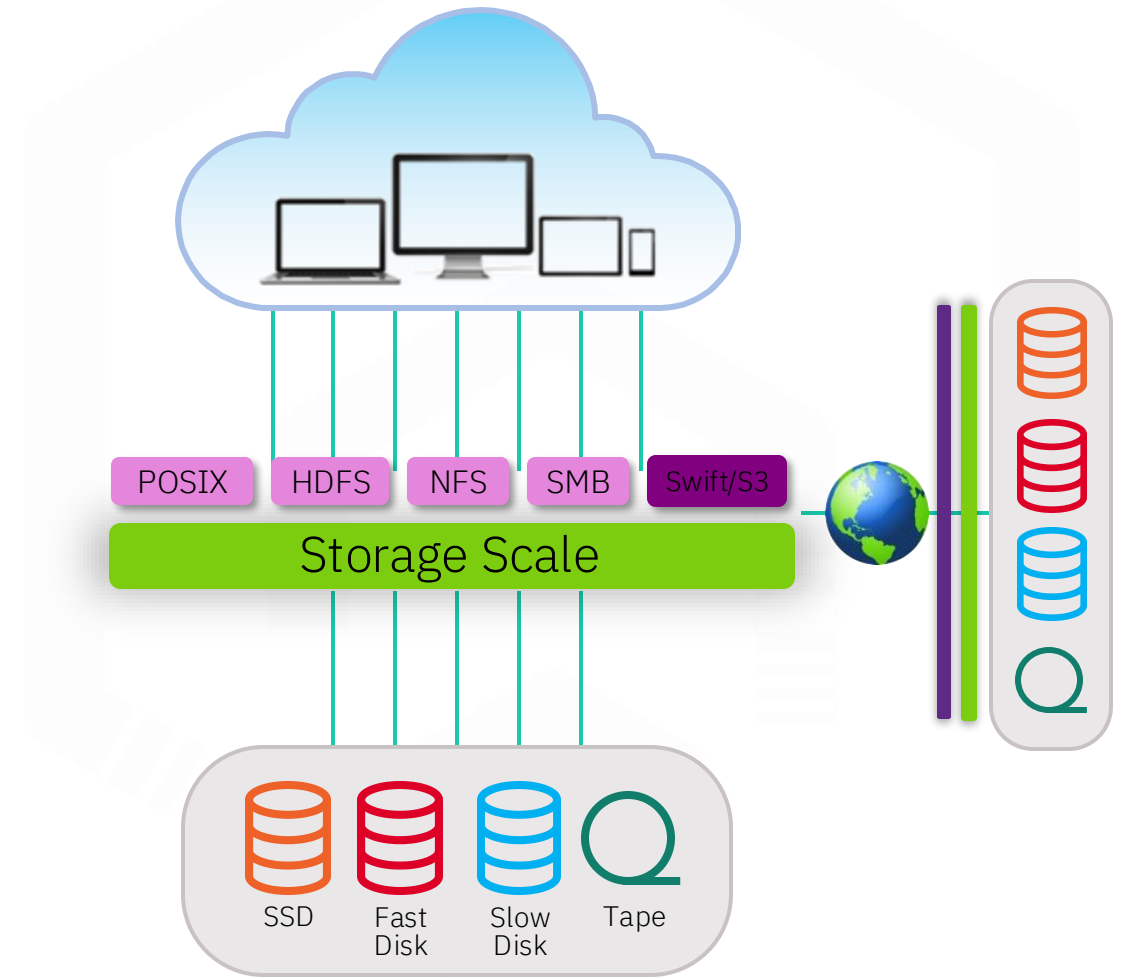
IBM Scale System 6000



IBM Storage Scale

Let's you store everywhere. Run anywhere.

- **Unified Scale-out Data Lake**
 - Access data using multiple protocols
 - High-performance concurrent access with integrity
 - Analytics on demand
 - Single management plane
 - Cluster replication and global namespace
 - Enterprise storage features across file, object, and HDFS
- **Global collaboration with Active File Management**
 - Filesystem caching and single namespace view across multiple geographically distributed remote sites
 - Extend collaborative workflows
 - Mitigate network bottleneck with advanced routing
 - Flexible configuration with writer and read-only sites
 - Disaster recovery for enterprise resiliency



IBM Storage Scale

Management and Monitoring GUI

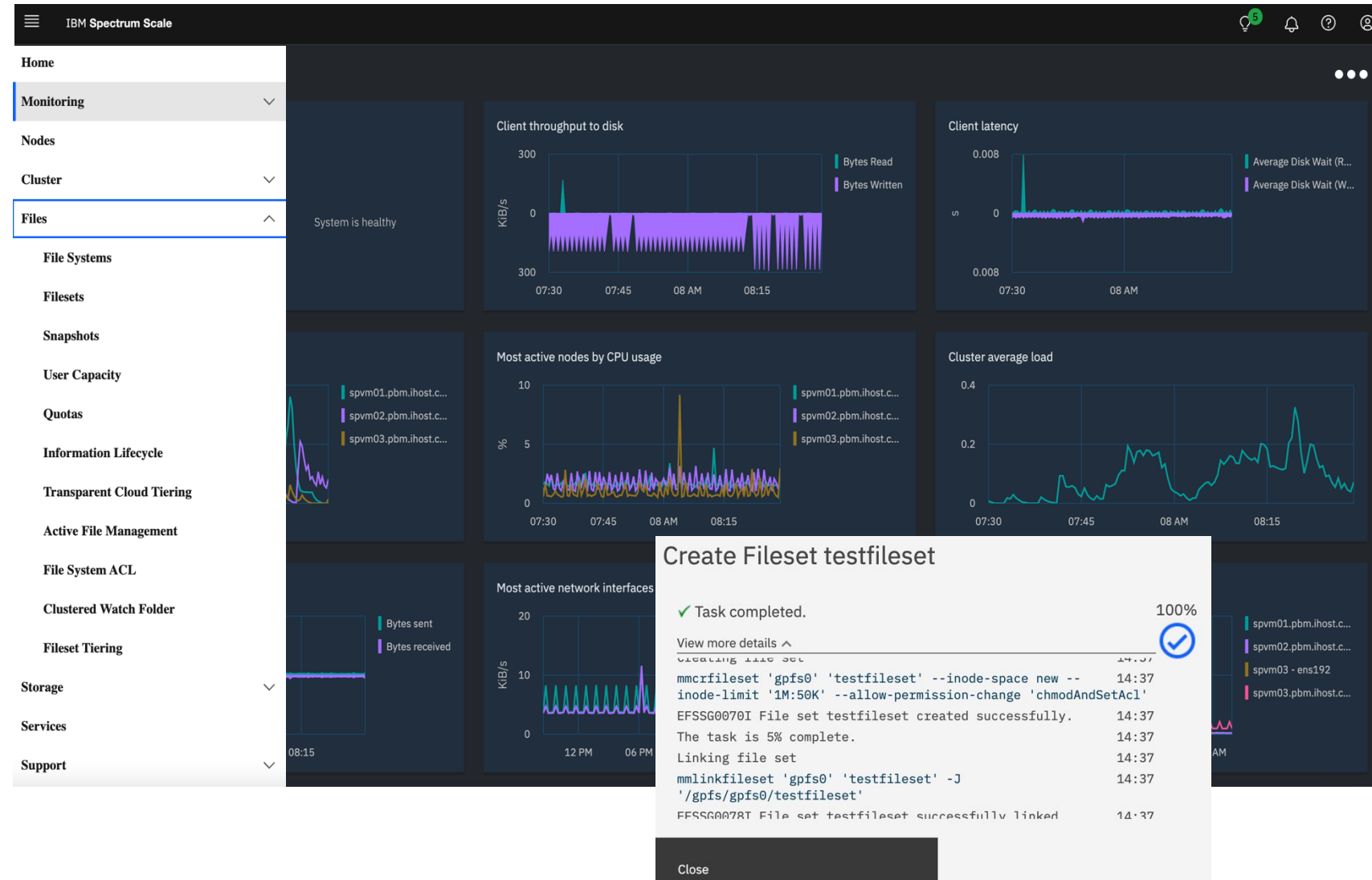
- System Health
- Node performance
- Network traffic
- Historical Trends
- Integrated into Spectrum Control
 - Storage portfolio visibility
 - Consolidated management
 - Multiple clusters

Cluster management tasks

Fileset management

Snapshots

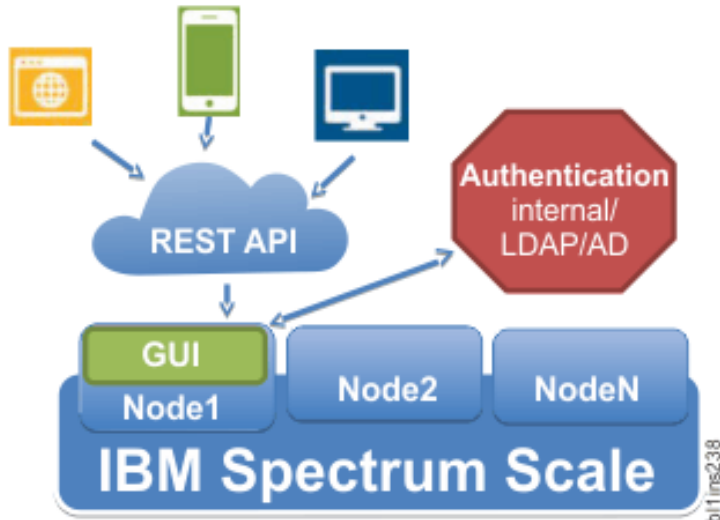
Quotas, ACLs, ILM, and more!



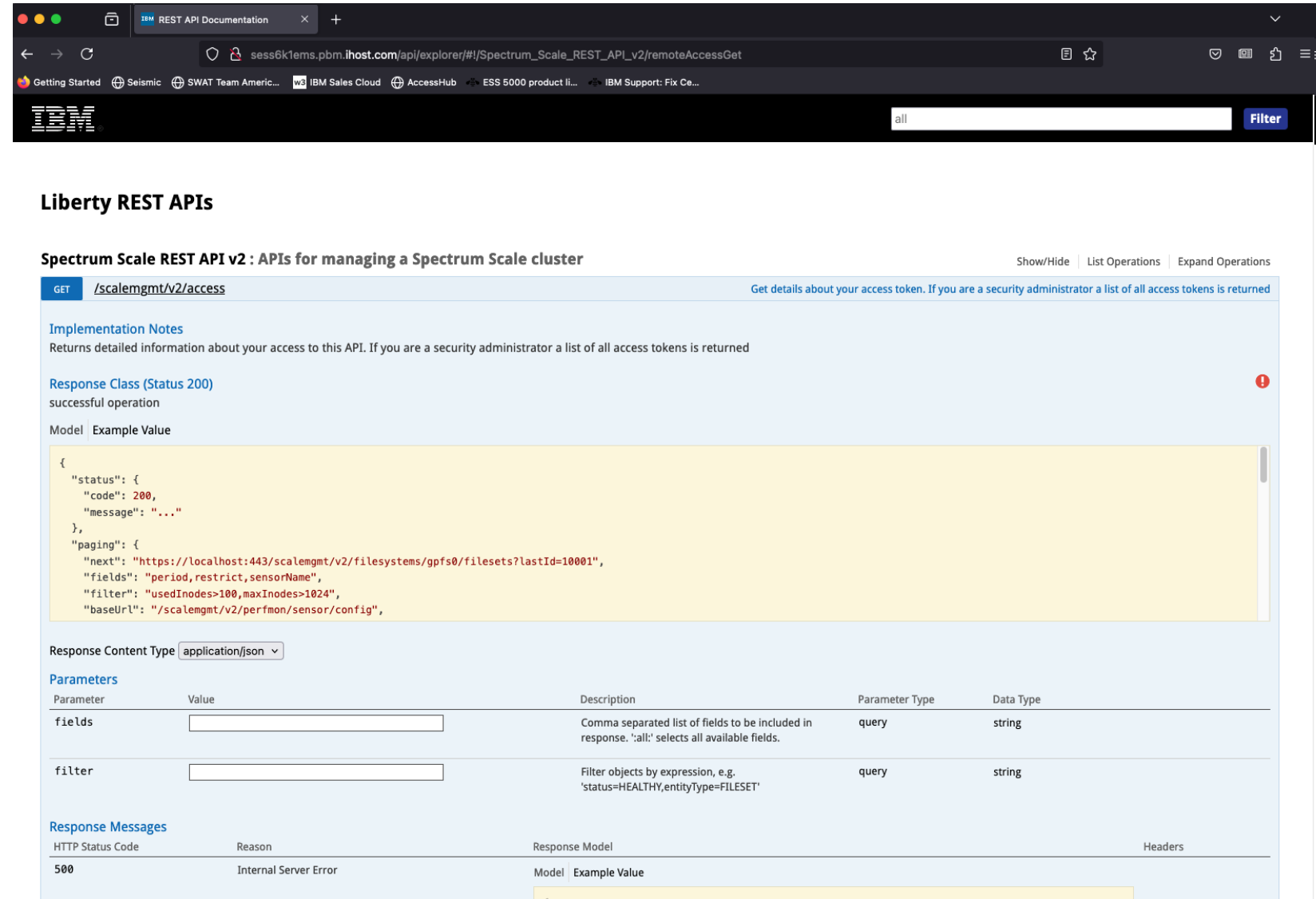
IBM Storage Scale

REST APIs

- Scale REST APIs run on the GUI node and re-use GUI capabilities
- Role-based access control
- Highly scalable
- Easier deployment and administration



https://www.ibm.com/docs/en/storage-scale/5.2.0?topic=overview-storage-scale-management-api#bl1adm_restapi_main.dita



The screenshot shows the IBM REST API Documentation page for Spectrum Scale REST API v2. The page title is 'Liberty REST APIs' and the specific endpoint is 'Spectrum Scale REST API v2 : APIs for managing a Spectrum Scale cluster'. The endpoint is '/scalemgmt/v2/access'. The page shows the 'GET' method and the 'Response Class (Status 200)' which is 'successful operation'. The 'Model' is 'Example Value' and the 'Response Content Type' is 'application/json'. The 'Parameters' section shows two parameters: 'fields' and 'filter'. The 'Response Messages' section shows a 500 status code with the reason 'Internal Server Error'.

Liberty REST APIs

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

Get details about your access token. If you are a security administrator a list of all access tokens is returned

GET /scalemgmt/v2/access

Implementation Notes
Returns detailed information about your access to this API. If you are a security administrator a list of all access tokens is returned

Response Class (Status 200)
successful operation

Model Example Value

```
{
  "status": {
    "code": 200,
    "message": "...",
  },
  "paging": {
    "next": "https://localhost:443/scalemgmt/v2/filesystems/gpfs0/filesets?lastId=10001",
    "fields": "period,restrict,sensorName",
    "filter": "usedInodes>100,maxInodes>1024",
    "baseUrl": "/scalemgmt/v2/perfmon/sensor/config",
  }
}
```

Parameters

Parameter	Value	Description	Parameter Type	Data Type
fields	<input type="text"/>	Comma separated list of fields to be included in response. 'all' selects all available fields.	query	string
filter	<input type="text"/>	Filter objects by expression, e.g. 'status=HEALTHY,entityType=FILESET'	query	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
500	Internal Server Error	Model Example Value	

Released!

Security Improvements



Remote Administration

Control Plane Designed For Applications / Operators

Released!

The diagram illustrates the system architecture:

- REST CLIENT / PLAYBOOK** (highlighted) interacts with the **REST SERVICE**.
- GUI** (laptop icon) and **CLI** (terminal icon) also interact with the **REST SERVICE**.
- The **REST SERVICE** interacts with the **Admin Daemon** via **gRPC**.
- The **Admin Daemon** (which includes **RBAC**) interacts with **miniscd** via **gRPC**.
- miniscd** interacts with external components via **RPC**.
- External components include a database (cylinder icon), a storage unit (floppy disk icon), and a server rack.
- Communication protocols shown are **HTTPS**, **gRPC**, and **RPC**.



REST SERVICE

RBAC

Admin Daemon

gRPC

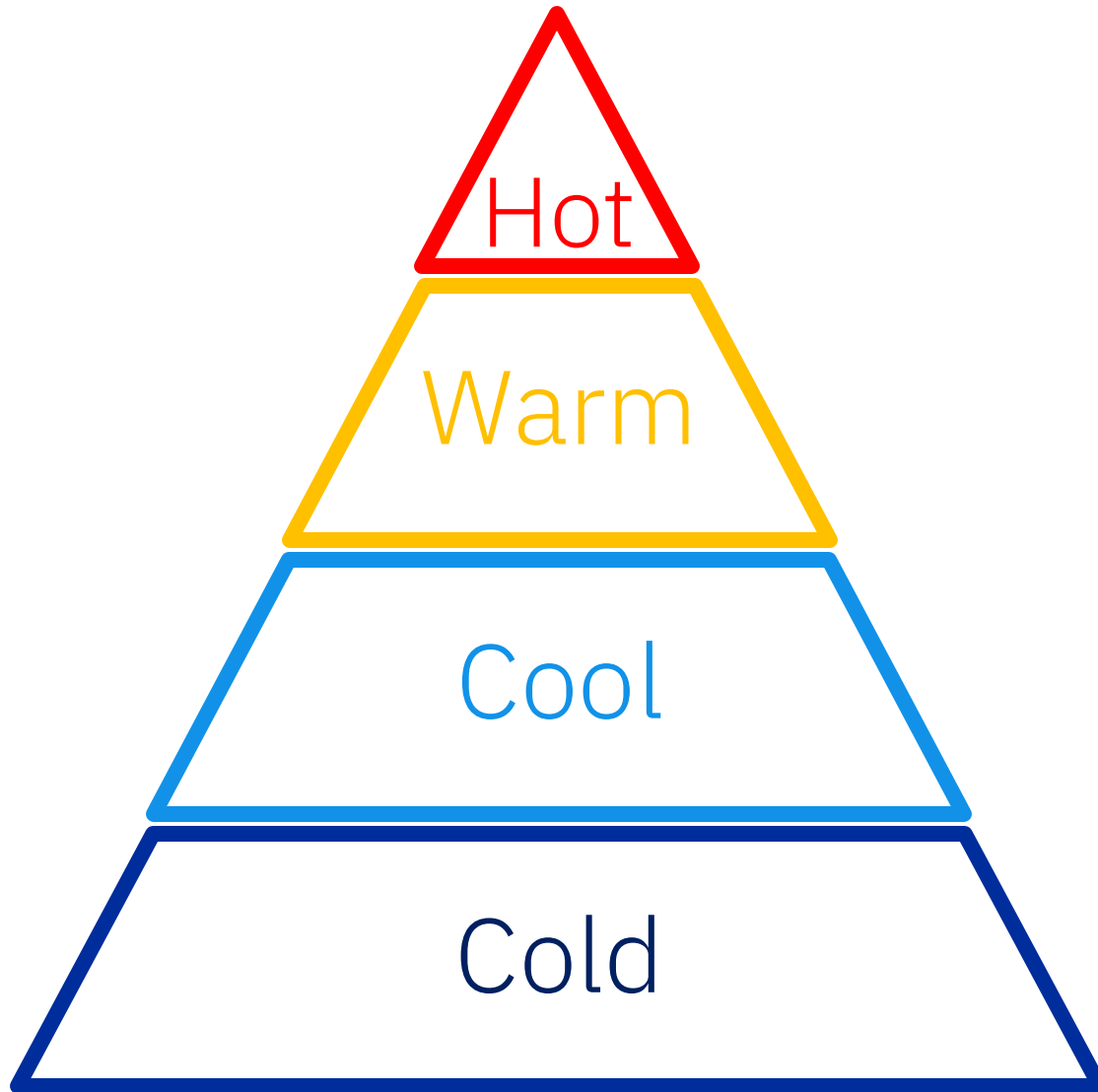
gRPC

gRP

m m f s d

RPC

Covering Every Tier with IBM Storage



High-performance SSS6000 delivering 1PB+ usable NVMe and 320GB/s+ per 4U for AI inferencing and training

Capacity model SSS6000 with HDDs delivering up to 90GB/s with automated tiering for tokenization and ready access

Storage Ceph providing scalable object storage for data extraction, collection, and aggregation

Diamondback securing 27PB per rack for cloud-like deep archive S3 storage ready to recall

IBM Storage Scale

Automated Deployments and Upgrades



Cloudkit: Resource provisioning and deployment of Storage Scale on public Clouds

- Command Line Interface to create Storage Scale clusters on public clouds (AWS, GCP, Azure (tech preview))
- Provides end to end automation to create and bring up an IBM Storage Scale cluster on public clouds (in minutes)
 - Automates infrastructure provisioning on the cloud
 - Automates the deployment of IBM Storage Scale on the cloud
 - Applies IBM Storage Scale best practises for deploying on the cloud
- Easy to use, guided interface

Install Toolkit: Installation, Deployment and Initial Configuration

- Fully automated CLI to install, Deploy and Configure Storage Scale on Bare Metal servers or Virtual Machines
- Provides ability to install, create and bring up a Storage Scale cluster
- Supports the automated creation of filesystems
- Supports the installation and initial configuration of advanced functionality such as Scale data access services (NFS, SMB, HDFS and Object protocols), AFM, File Audit Logging etc.

Upgrade

- “One button” rolling upgrade: Support for rolling upgrade of the Storage Scale cluster
- Offline parallel upgrade: Upgrade entire cluster parallelly when the cluster is shutdown

Ansible based Install Toolkit Overview

Install toolkit Workflow

Define Cluster Topology

Use Storage Scale CLI commands to Cluster definition

- Add nodes to cluster
- Assign roles to nodes
- Define NSD
- Define File system

Install

Use Storage Scale install to perform:

- Installation required RPMs on all nodes
- Creates a Storage Scale cluster
- Creates NSD
- Sets up Management GUI
- Creates file system

Deploy

Use Storage Scale deploy to perform:

- Install, Configure and enable protocols

Upgrade

Use Storage Scale upgrade to perform:

- Online sequential upgrade of cluster
- Offline cluster upgrade

Toolkit can be used to automate deploy alone when install has happened manually

Toolkit can be used to upgrade an existing cluster that has been created manually

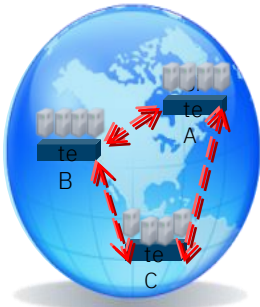
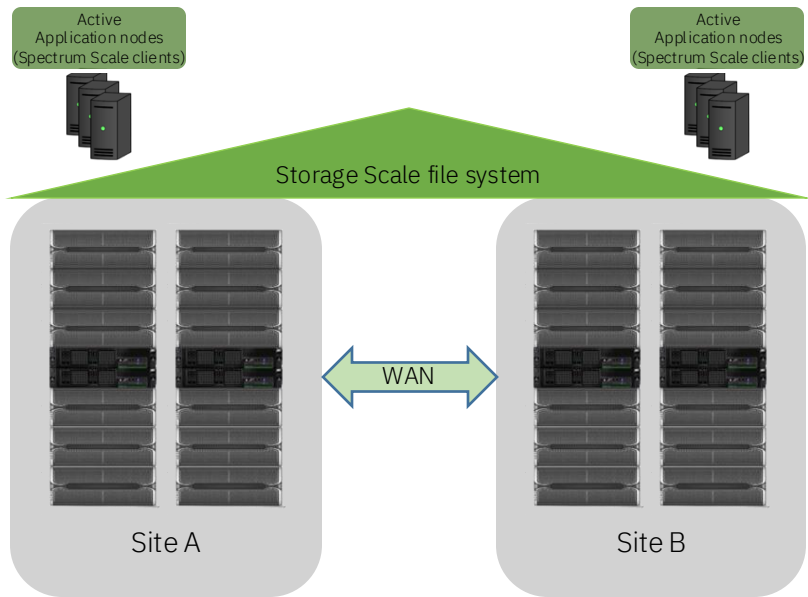


IBM Storage Scale

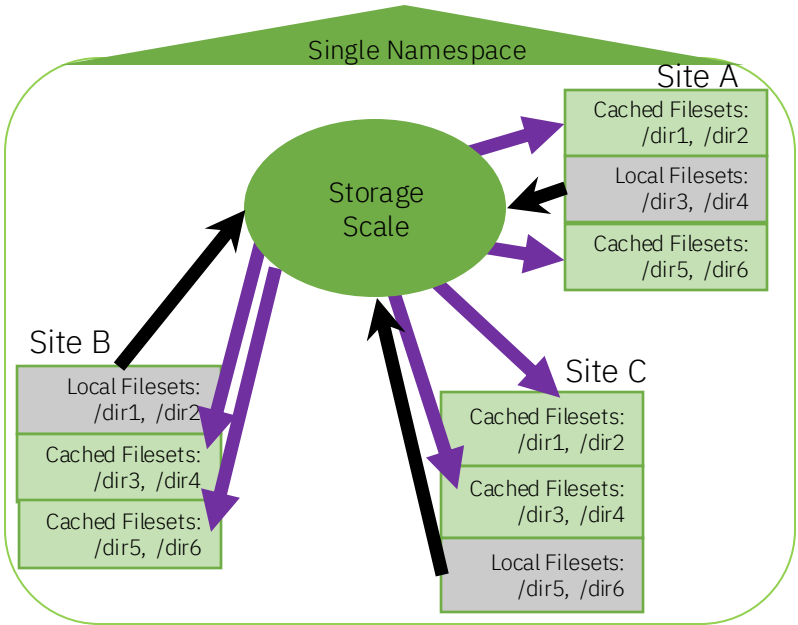
Multi-Site Collaboration/High-Availability



Stretch Cluster (synchronous)



Active File Management (asynchronous)



IBM Storage Scale

Active File Management (AFM)

Active File Management (AFM) lets Scale extend over geographic distances

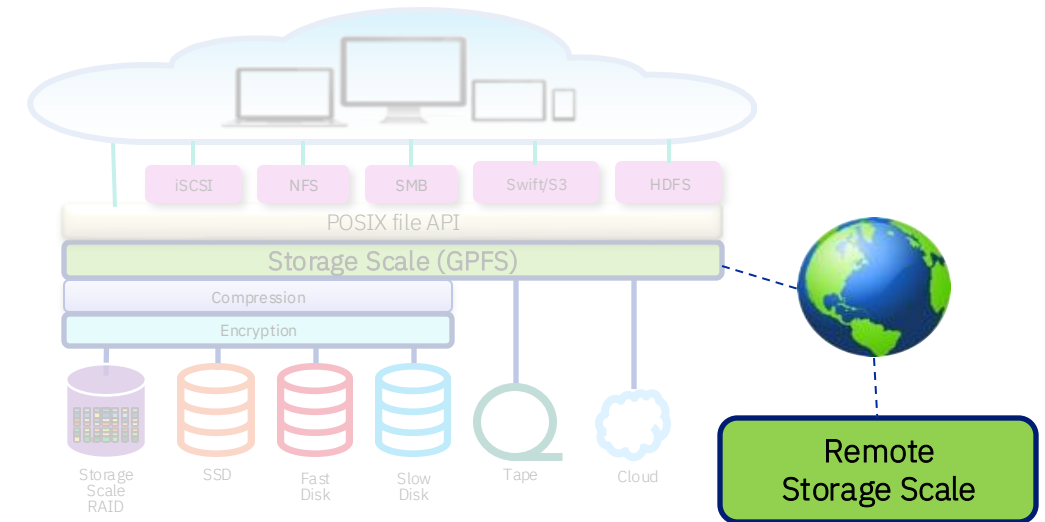
- Tolerates unreliable, high-latency networks (like a WAN).
- *Caches* copies of data from a remote file system into the local Storage Scale cluster.
- Cached files have the same read and write performance as other local files.
- However, there is no locking between different caches.

AFM enables efficient data access to collaborators and resources around the world

- Unifies heterogenous remote storage

Asynchronous DR (AFM-DR) is a special case of AFM

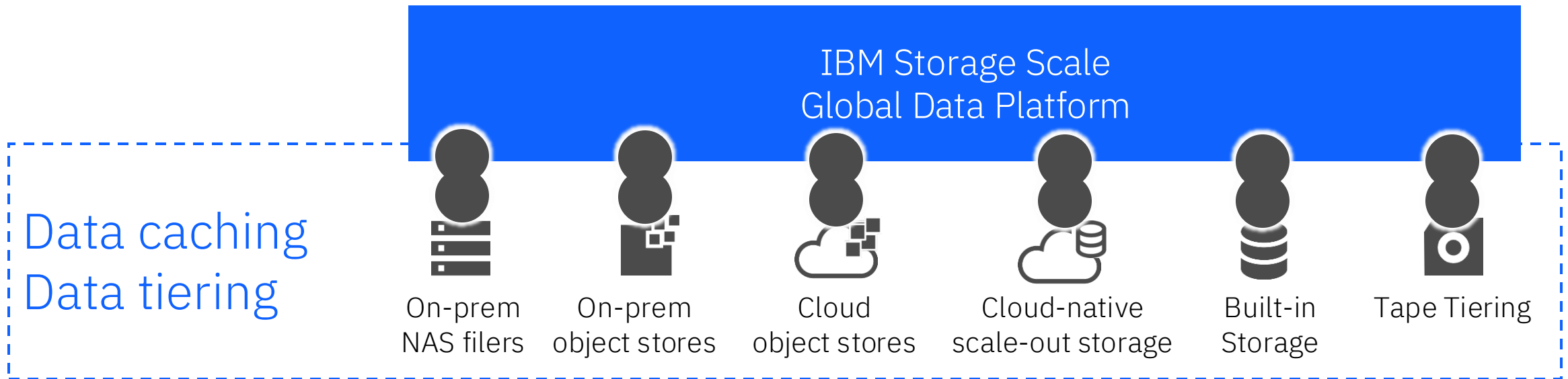
- Bidirectional awareness for failover and failback with data integrity
- Recovery Point Objectives (RPO) for application consistency points



Data caching and Tiering

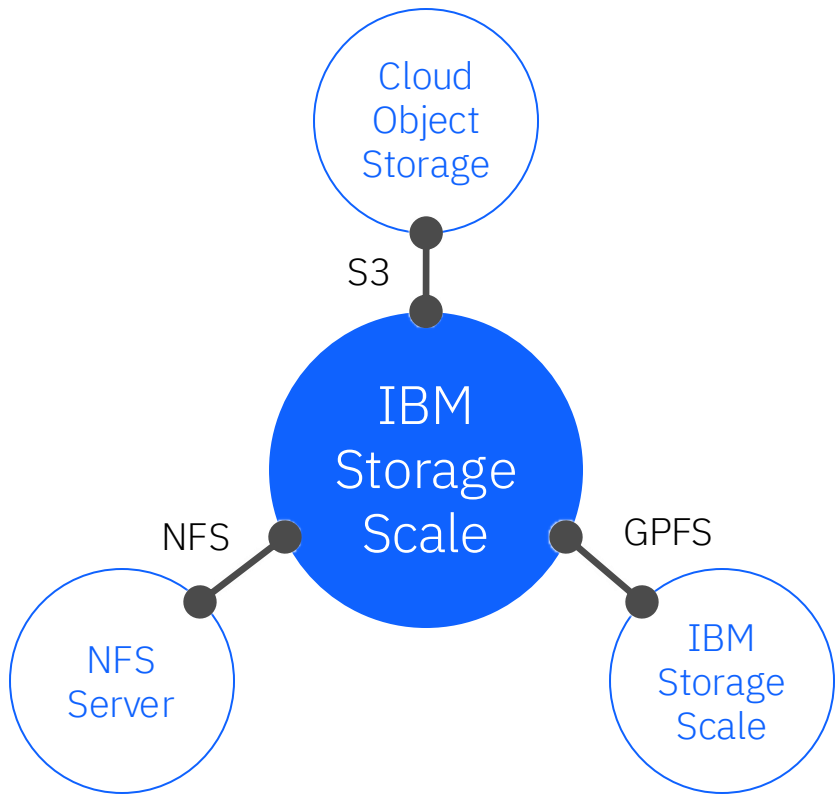
IBM Storage Scale enables data caching and tiering with the following features.

- Active File Management
- Policy-based Information Lifecycle Management (ILM)

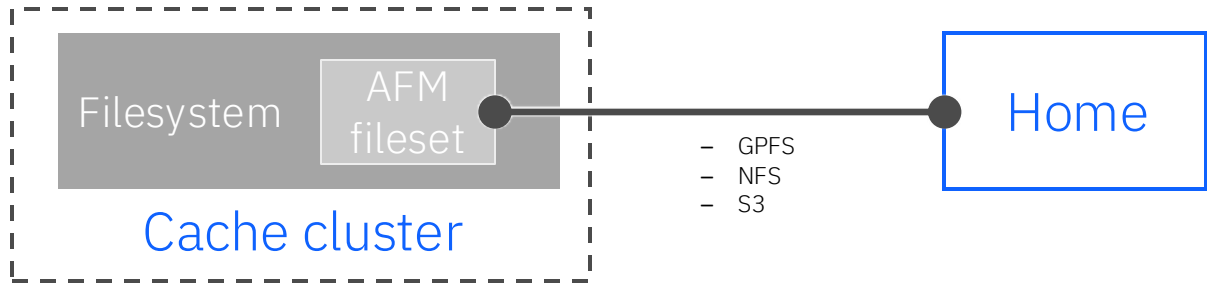


Active File Management (AFM) overview

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 (Five) caching modes.

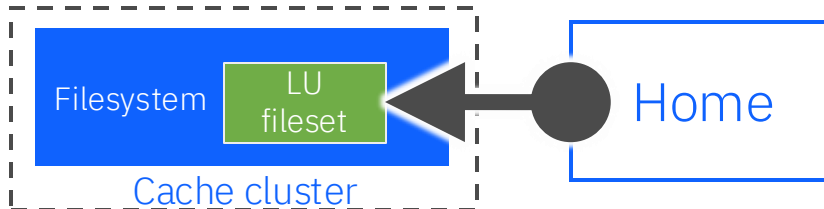
Read Only (RO) mode

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



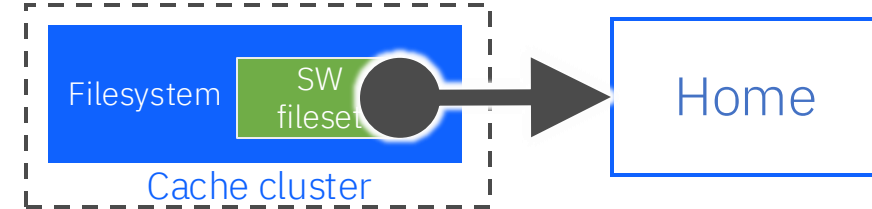
Local Updates (LU) mode

- 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.



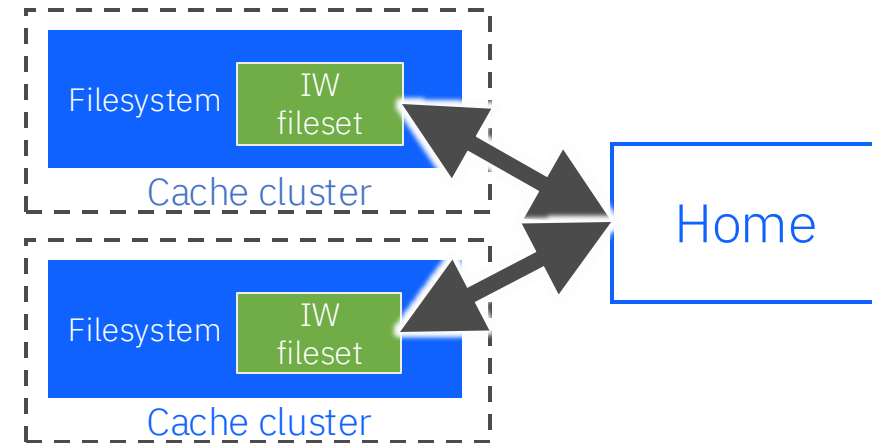
Single Writer (SW) mode

- 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.



Storage Scale

AFM S3 Caching Mode Additions

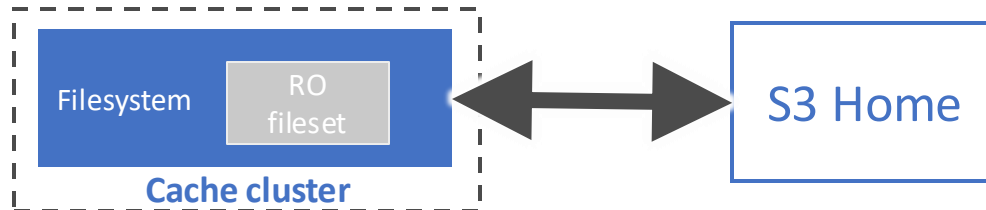


Storage Scale

AFM S3 has two other caching modes.

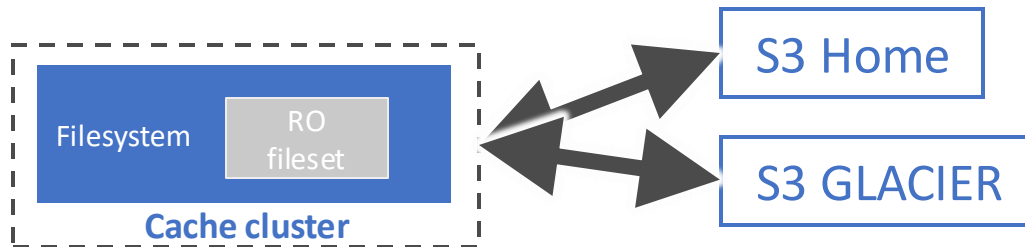
Manual Update (MU) mode

- Data in the cache or in the S3 object store is not automatically moved.
- Data can be moved bi-directional via Integrated Life Cycle Management (ILM).



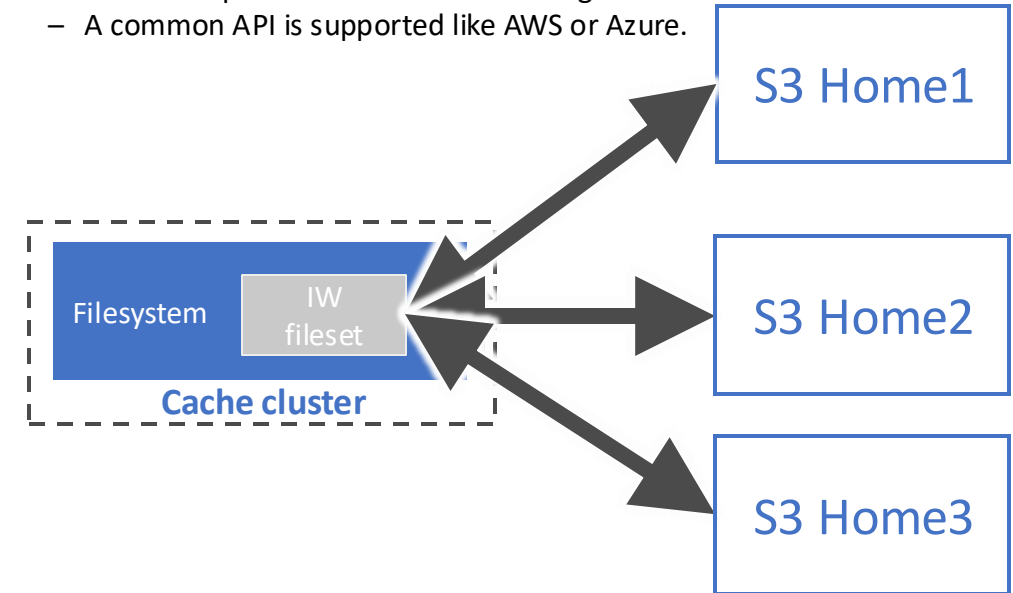
Deep Archive Mode

- Data in the cache or in the S3 object store is not automatically moved.
- Data can be moved bi-directional via Integrated Life Cycle Management (ILM).



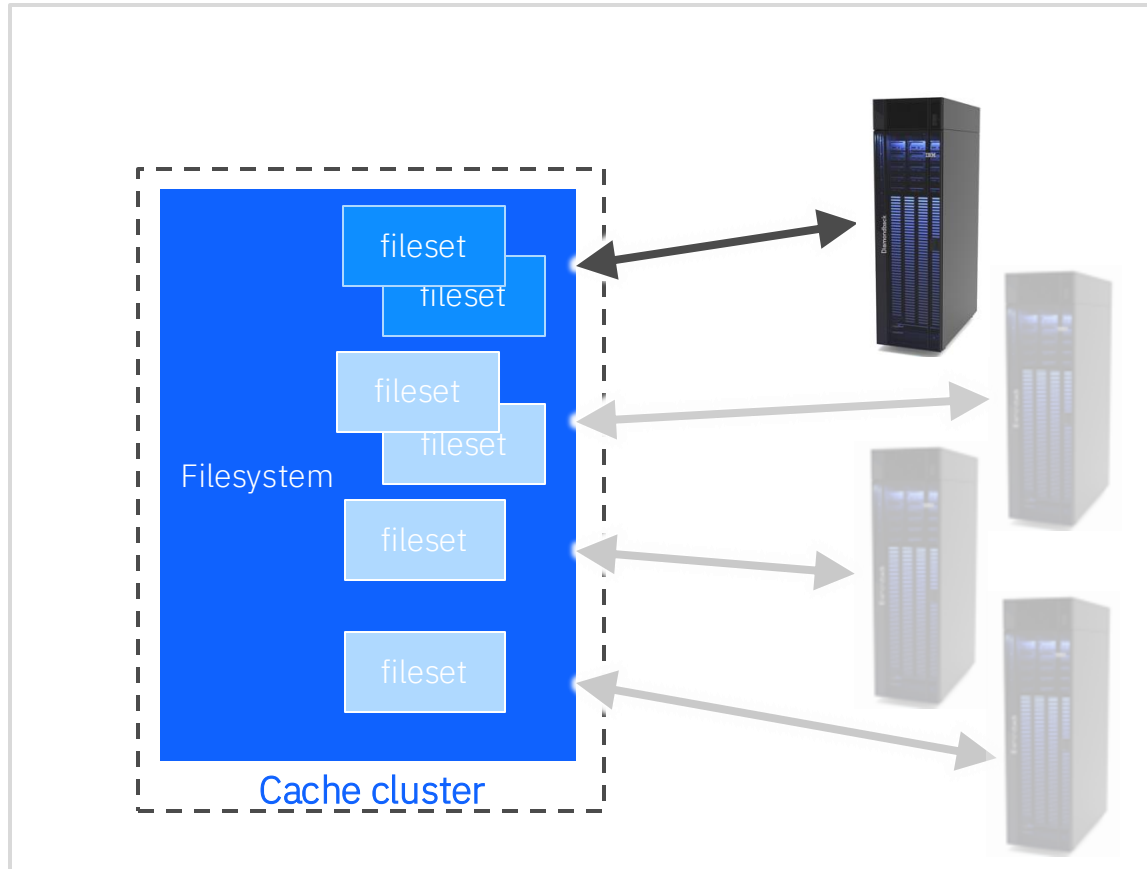
Multi-target replication

- Cache can push to two different S3 targets.
- A common API is supported like AWS or Azure.



- Ingest from any access method for export, share, S3 URI
- **ROADMAP:** Move from Scale to S3 to Tape (Glacier)
- **Not than multi-target**

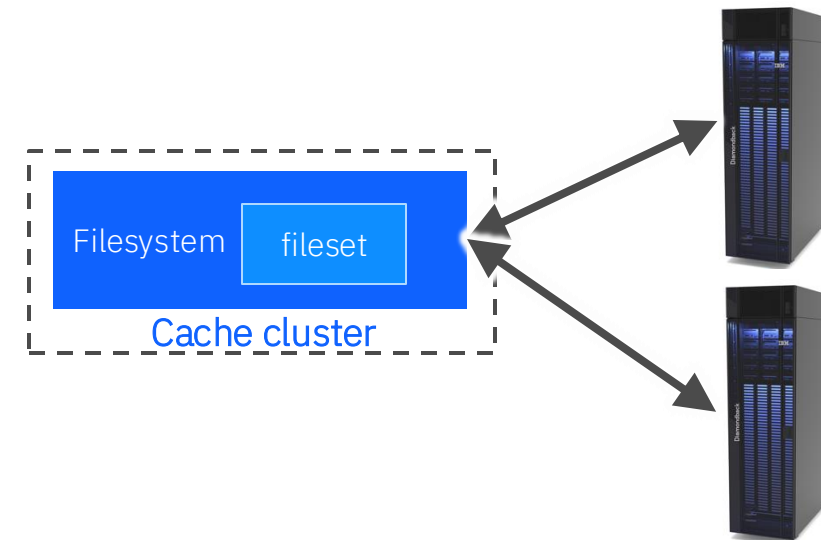
AFM S3 Support for Storage Deep Archive



Scale the number of buckets and/or Storage Deep Archives with additional filesets

Multi-target replication

- Cache can push to two different S3 targets (can be mixed)

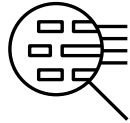


Eliminate all single-points-of-failure/repair with AFM-based replication

IBM Storage Scale

Assurance for Cyber Resiliency

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

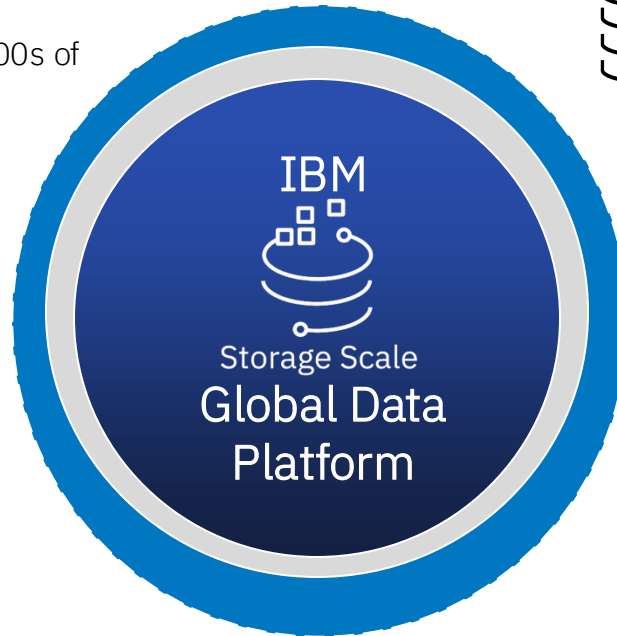
- QRadar and Splunk SIEM integration
- File Audit Logging, Watch Folders
- Analyze backup data for signs of ransomware (Spectrum Protect)
- Reporting: QRadar User behavior analytics
- IBM Flash Core Modules entropy detection

RESPOND



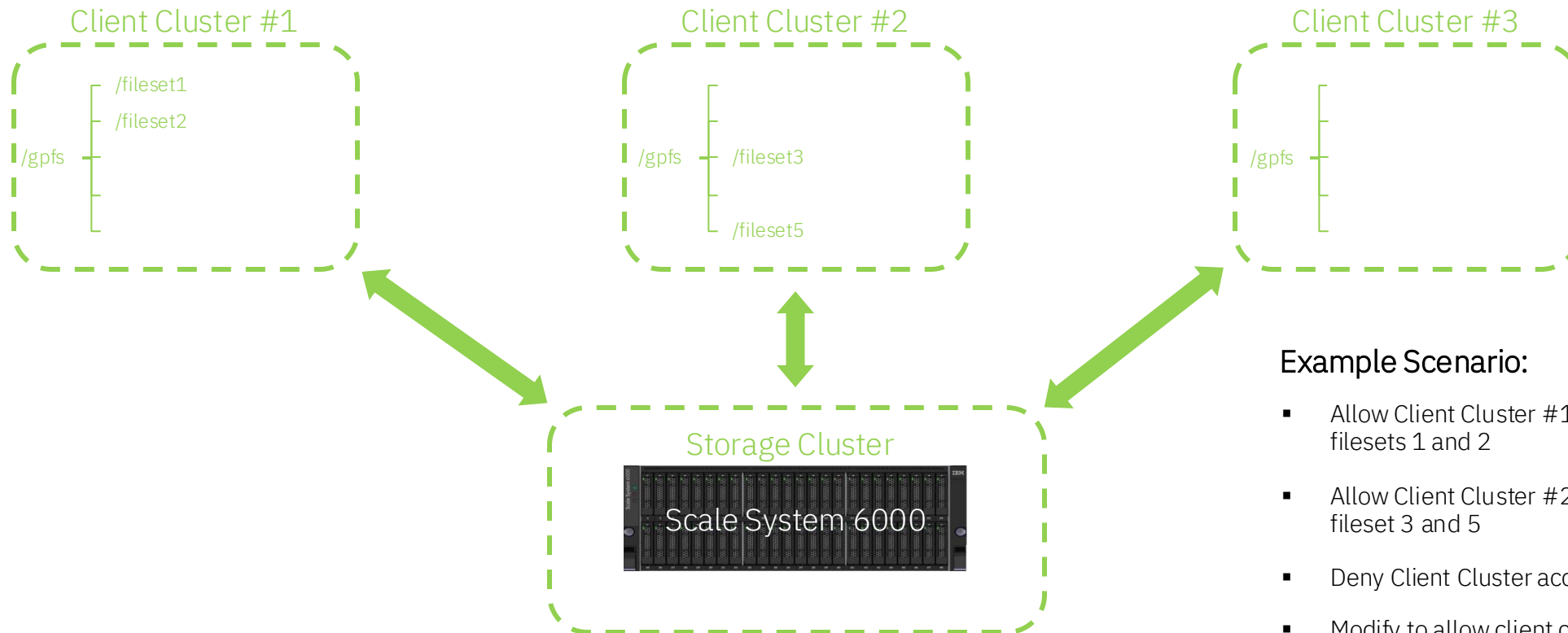
Alert and take action

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



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:

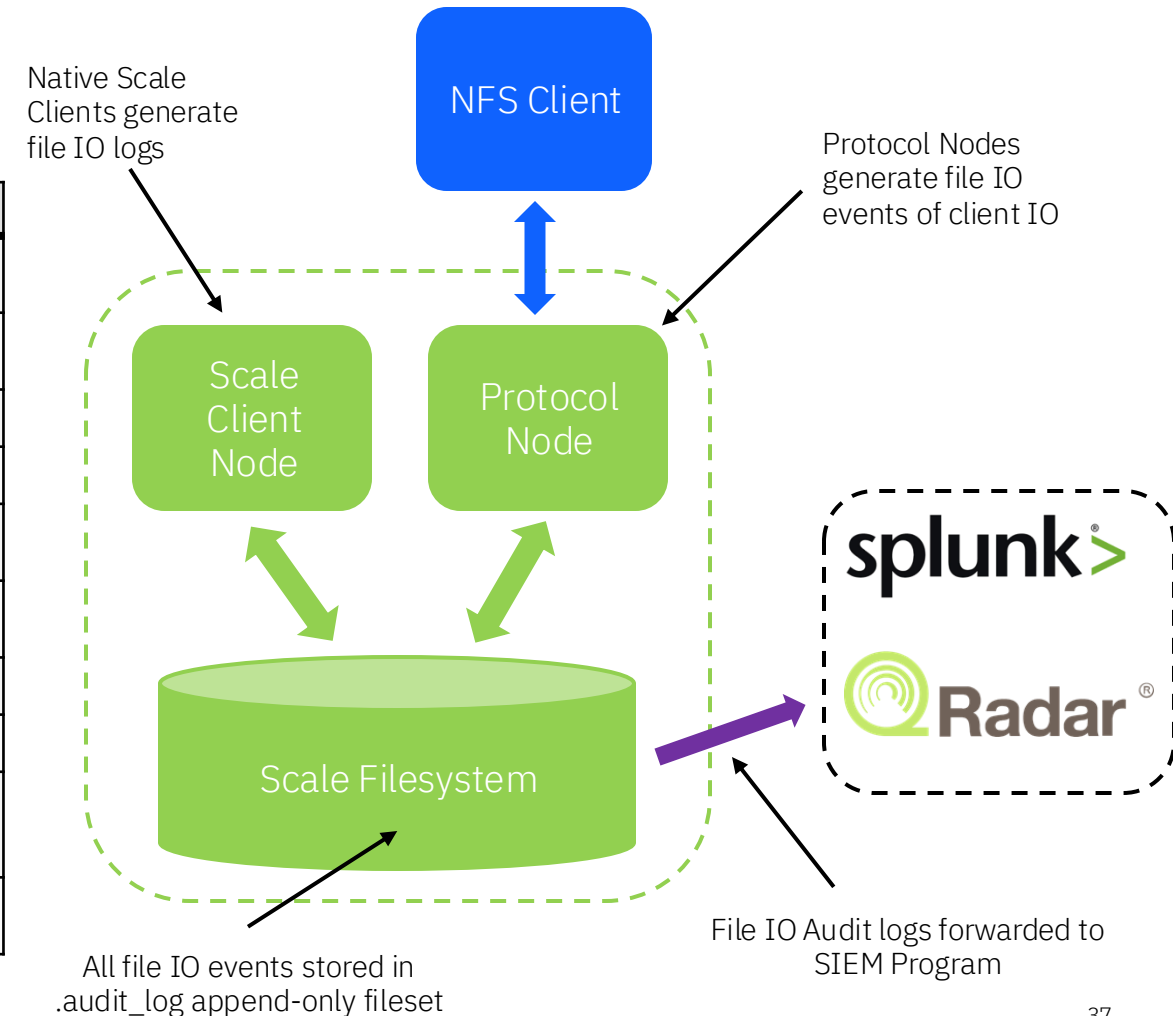
- Allow Client Cluster #1 to only see filesets 1 and 2
- Allow Client Cluster #2 to only see fileset 3 and 5
- Deny Client Cluster access to any filesets
- Modify to allow client cluster #3 to see fileset 2 and 4

File Audit Logging

- Lightweight File IO Event logs stored in JSON format
- All filesystem IO events captured from root, users, Protocols, etc...
- Audit logs stored in an append-Only fileset
- Events forwarded to SIEM program such as IBM Qradar or Splunk for analysis of known access patterns
- Fully-configurable based on needed file events

Video on how to forward to Splunk:
<https://www.youtube.com/watch?v=FGVsYysk1Q>

Event Name	Description	Examples
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



IBM Storage Scale

Native Filesystem Encryption

Data is encrypted while “at rest” on disk and decrypted on the way to reader – data not metadata

Encryption takes place on the node(s) from which the user drives the I/O

File content travels encrypted to the NSD server

MEKs can be accessed by nodes that have appropriate RKM credentials

Nodes that cannot access keys cannot access files, irrespective of file permissions

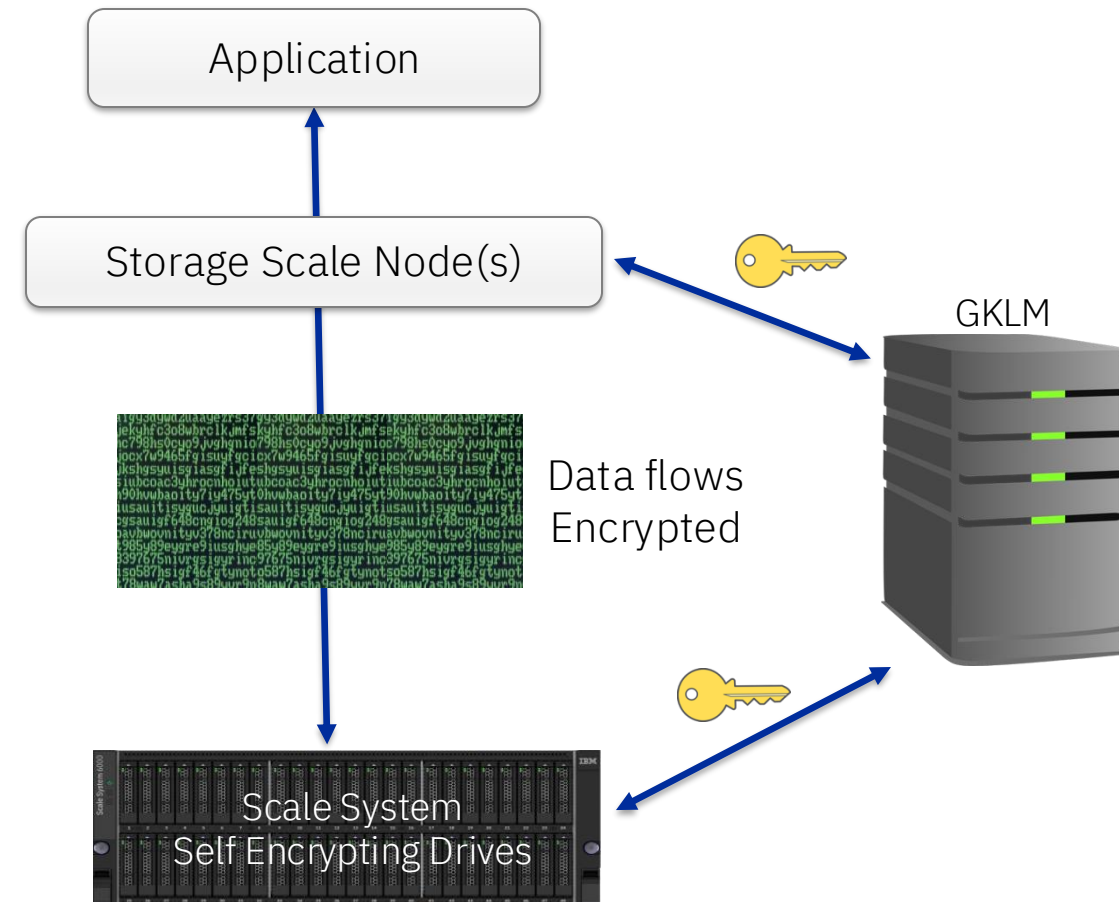
Granularity is per file or per fileset, as determined by encryption policies

Scale encryption can protect against attacks targeting disks (theft/acquisition of improperly discarded disks)

Secure data deletion using cryptographic erasure

IBM Storage Scale System supports Self Encrypting Drives (SED) for data at rest protection

Investment in Quantum Safe algorithm support



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 Cataloging Entitlement (Storage Discover)		Yes	Yes
Erasure coding	Scale System only	Scale System only	Yes

IBM Storage Scale Developer Edition

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

IBM Storage Scale

Fully functional!

Derived from the **most advanced Scale Technology**

Download it or schedule a demonstration!

Accelerate AI and unlock value from your data

★★★★☆ 17 Reviews - G2 Crowd

Try the free developer edition →

Schedule a free demo →



Storage Scale User Group

The Storage 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.Storagescaleug.org

IBM Storage
Resources

★★★★★ (1)

Rate

Reserve

Apr 16, 2023
IBM Storage Scale Developer Edition - Installation Lab

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

IBM Storage Scale Developer Edition - Installation Lab

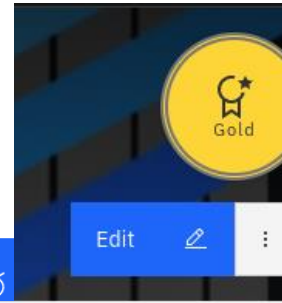
Visibility
IBMers, Business Partners

Apr 19, 2023
IBM Storage Scale Developer Edition Lab

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

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



Mar 14, 2023
Login - to Storage Scale Developer Edition Labs via SSH key method

Powerpoint step by step - login to Techzone Storage Scale Developer Edition Labs via SSH key method.

Visibility
IBMers, Business Partners



Mar 14, 2023
Lab Guide - Installation Toolkit Lab

README.md file step by step instructions, commands to do Storage Scale install using Storage Scale Install Toolkit.

Visibility
IBMers, Business Partners



Mar 14, 2023
Lab Guide - Storage Scale guided exercises for GUI and command line

README.md file step by step instructions to perform various Storage Scale operations on GUI and command line.

Visibility
IBMers, Business Partners



Mar 14, 2023
Video - Storage Scale GUI Overview

Video showing extensive step by step examples of using Storage Scale GUI.

Visibility
IBMers, Business Partners



Mar 14, 2023
Video - Storage Scale Data Management Services

Video showing extensive step by step examples of using Storage Scale Integrated Life Cycle Management (ILM) - place data, move and manage data.

Mar 14, 2023
Video - Storage Scale Data Caching Service

Video showing setup Storage Scale Active File Management (AFM) connections to NFS or S3 target.

Mar 28, 2023
IBM Storage Scale New User Group

This is intended for a New Scale User Group Session. The agenda is as follows:

- 5 minutes - Welcome
- 25 minute - Planning for Data-Intensive Science

Mar 28, 2023
2-3 day - IBM Storage Scale and Scale System Client Workshop

Day 1

- 15 minute - Welcome and introductions
- 45 min - Overview of the IBM Global Data Platform powered by Storage Scale

IBM Scale System

IBM Storage Scale System (SSS)



Fast time to value

Preconfigured, fully tuned

Easy to install or update by sysadmin or developer

Perfect for growing GPU workloads

Linear performance scaling

Operational efficiency

Containerized install and update software for a fast and easy out-of-the-box experience

High performance, high density: 310GB/sec and up to 1PB usable NVMe capacity per 4U system

High performance, high capacity: up to 55 GB/sec and 11 PB HDD usable capacity per ESS 5000 (SC9 model)

Reliability

IBM Storage Scale erasure coding

Fast, non-disruptive data rebuild

Automated monitoring of key hardware components

Disk Hospital

Deployment flexibility

High-performance tier for any Spectrum Scale / Elastic Storage System cluster

Start as small as 46 TB, scale out to exabyte capacity

Loosely-coupled edge data management as component of global unified data solution

Supported with AI solutions based on Intel, IBM Power, Z, and LinuxONE

IBM Scale System (SSS) Solution Packaging



Integrated solution

IBM Storage Scale is integrated, tested, and factory preloaded

Leverage the latest IBM Storage Scale releases

Data Management Edition and Data Access Edition

Optimal storage capacity & economy

SSS has various models providing NL-SAS, QLC, or TLC NVMe storage

Choose from various sizes of HDD, SSD, and NVMe

Rack-mountable solution

Non-disruptive upgrades

Capacity upgrades can be performed without application disruption

Software automatically rebalances data across all drives

High performance connectivity

Supports EDR/HDR/NDR InfiniBand or 100 GbE / 200 GbE / 400 GbE Ethernet high-performance networking

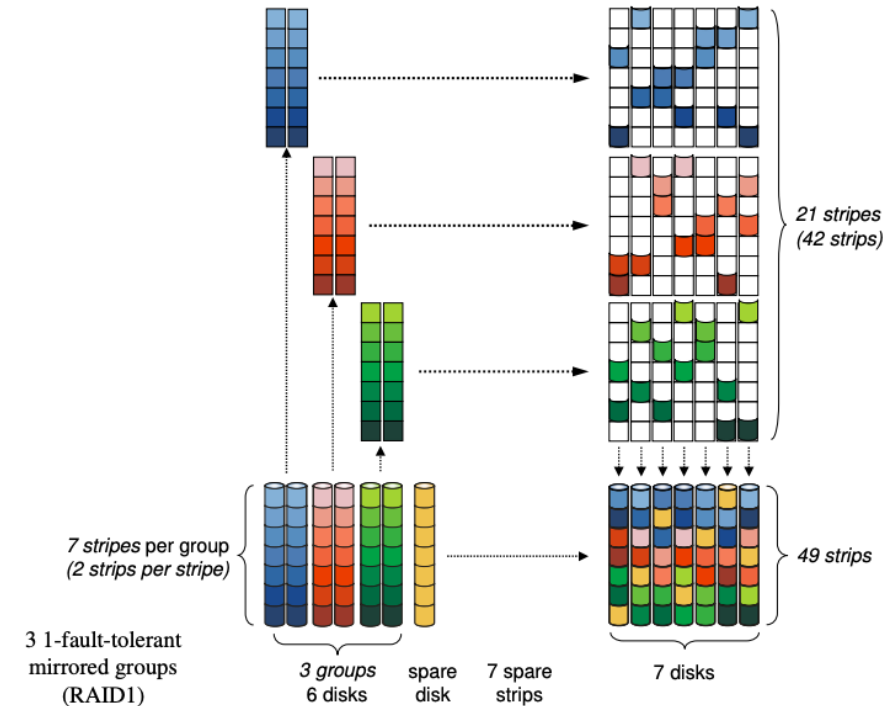
SSS is tested with NVIDIA network switches and firmware

Nvidia GPU-Direct support and certification

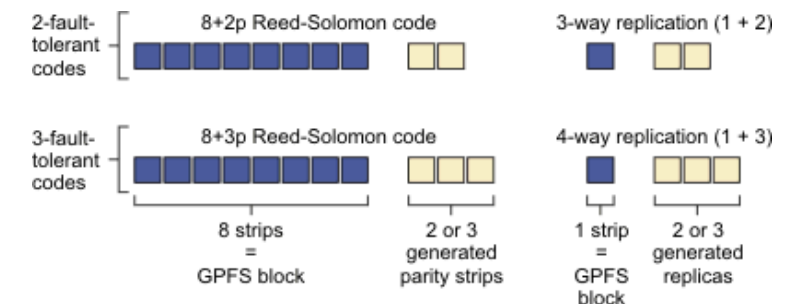
IBM Storage Scale RAID Overview

- Declustered erasure coding provides for data and parity to be distributed over all the disks and nodes in the de-clustered array for fastest performance out of the chosen media
 - Faster and more intelligent rebuild operations, using more drives in parallel
 - Prioritize normal vs critical conditions to better use node resources
 - Spare capacity is also distributed across all drives and nodes, so no dedicated spare disks are needed
- Improved storage efficiency and performance
 - 8+2P and 8+3P utilize less overhead vs 100% - 200% for 2X-3X replication
 - Patented algorithms optimize I/O data paths, read and multi-layer write caching

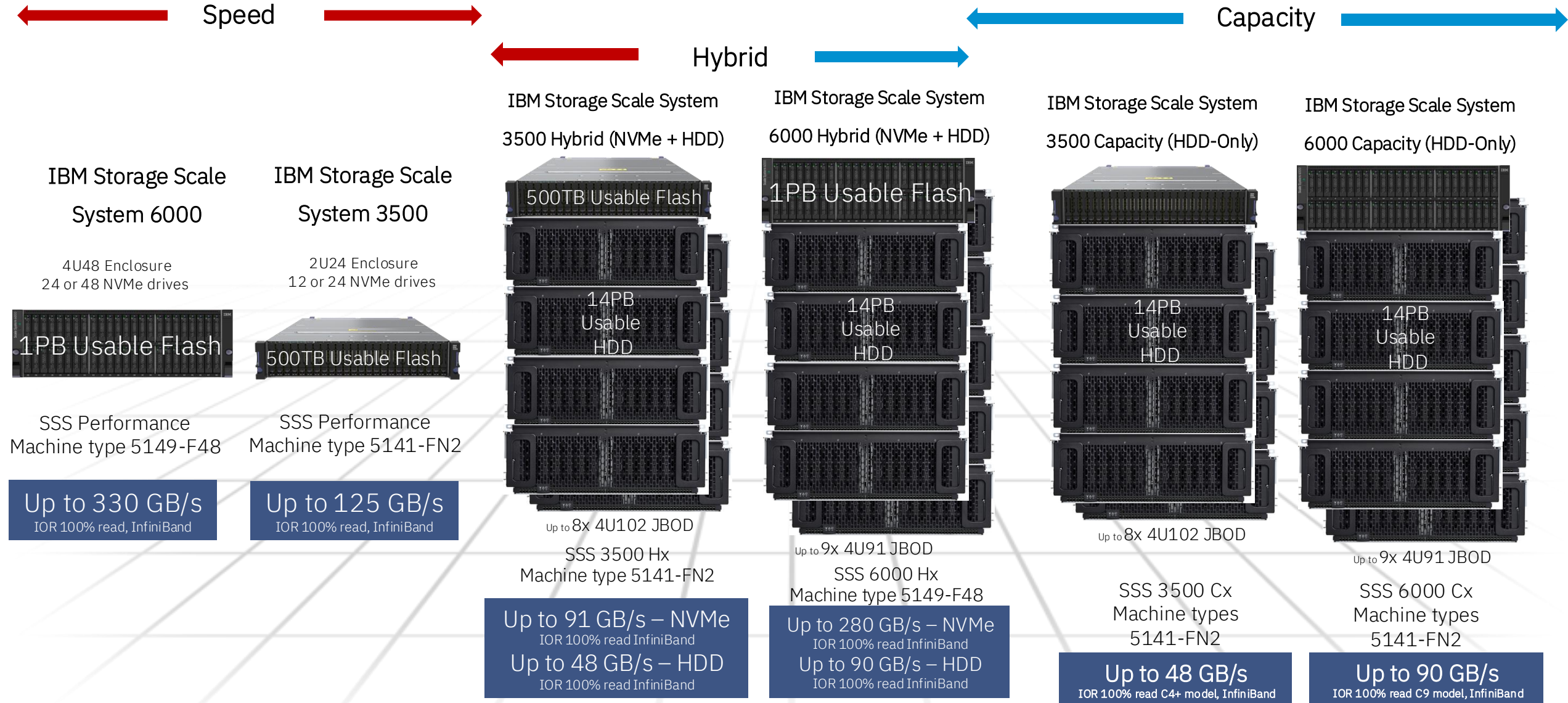
De-Clustered RAID Example



Erasure Encoding Examples



Scale System models are built for speed and capacity



What the Scale System 6000 looks like under the covers

A single 4U node with active-active controllers and redundant hardware to maximize always on data



Engineered for performance and efficiency

Processor per canister

Dual AMD EPYC Genoa 48C

Memory per Canister

24 x 32GB (768GB) – default base

24 x 64GB (1536GB) - option

Storage

48 U.3 G5 NVMe (24 and 48 drives)

NVMe (TLC): 3.84TB, 7.6TB, 15TB, 30TB

NVMe (QLC): 30TB, 60TB

FCM: 38TB

Networking

NVIDIA CX7 supported cards:

400Gb VPI Single port (IB/ETH) x16 Gen5 (OSFP)

200Gb VPI dual port (IB/ETH) x16 Gen5 (QSFP)

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, health binning,
error correcting codes, optimized voltage

Continuous health monitoring
keeps wear across all cells within 5%



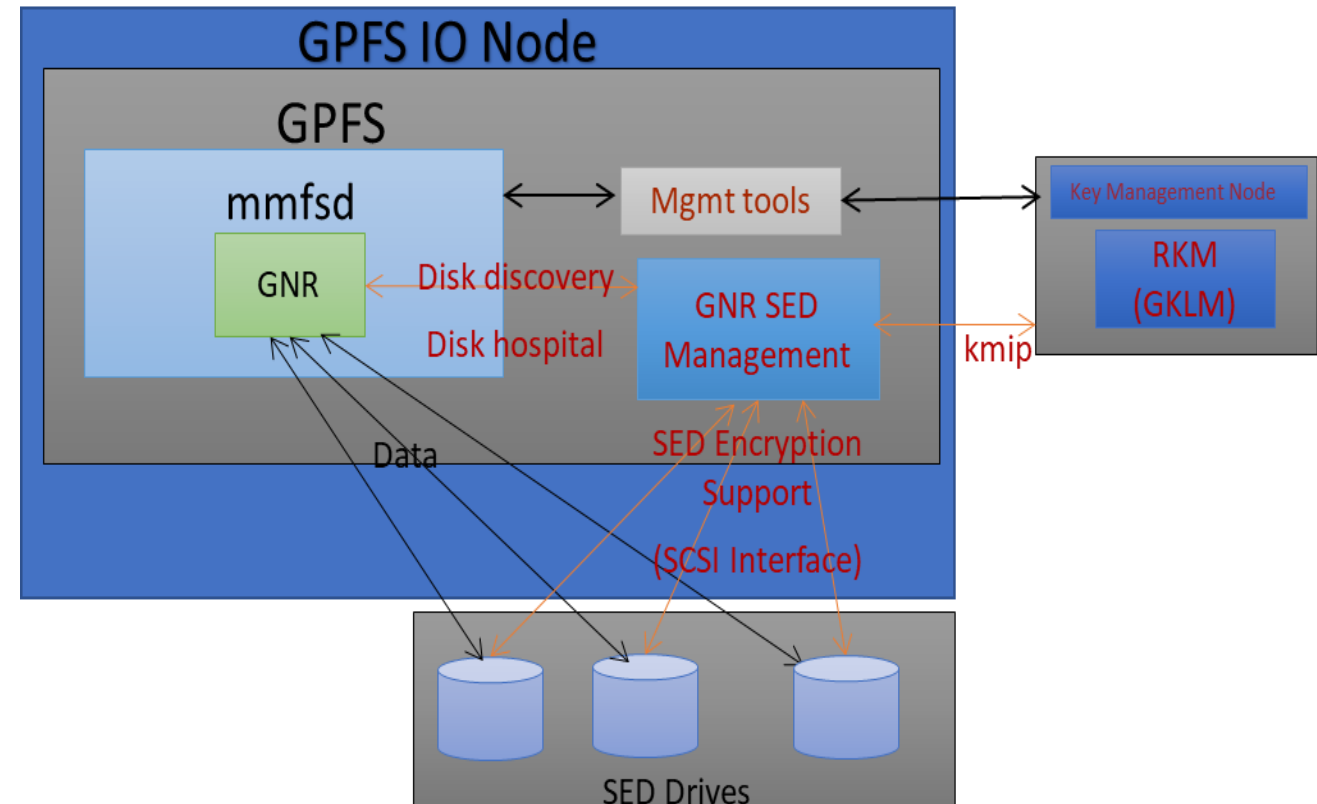
SED Support with GKLM : Overview

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



ESS 3500 Under the Hood

Faster Performance

- Two Active-Active Controllers
- AMD 7642 Processor with 125GB/s performance
- Four PCIe Gen4 slots for enhanced performance and options

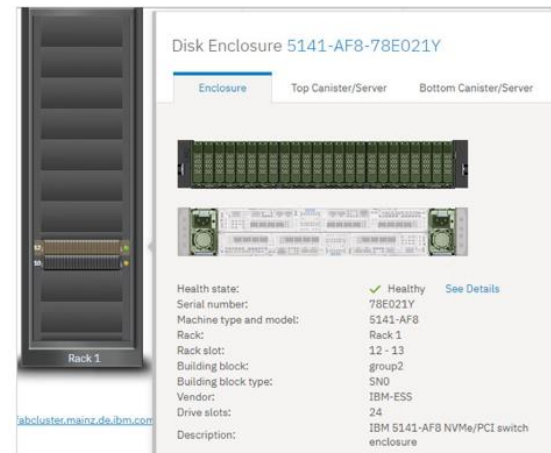
Easy to Manage

- Two integrated network ports and call home management
- Simplified service with client installable drives and parts
- Easy to manage with graphical user interface

Sustainable

- New power supply and streamlined design for better thermal results
- 4U Industry in industry standard 19" rack, depth 39.4" deep
- Combine HDD and NVMe disk drive in one node with transparent software lifecycle management with enhanced sustainability to turn off data (archive or tier) to tape or to cloud

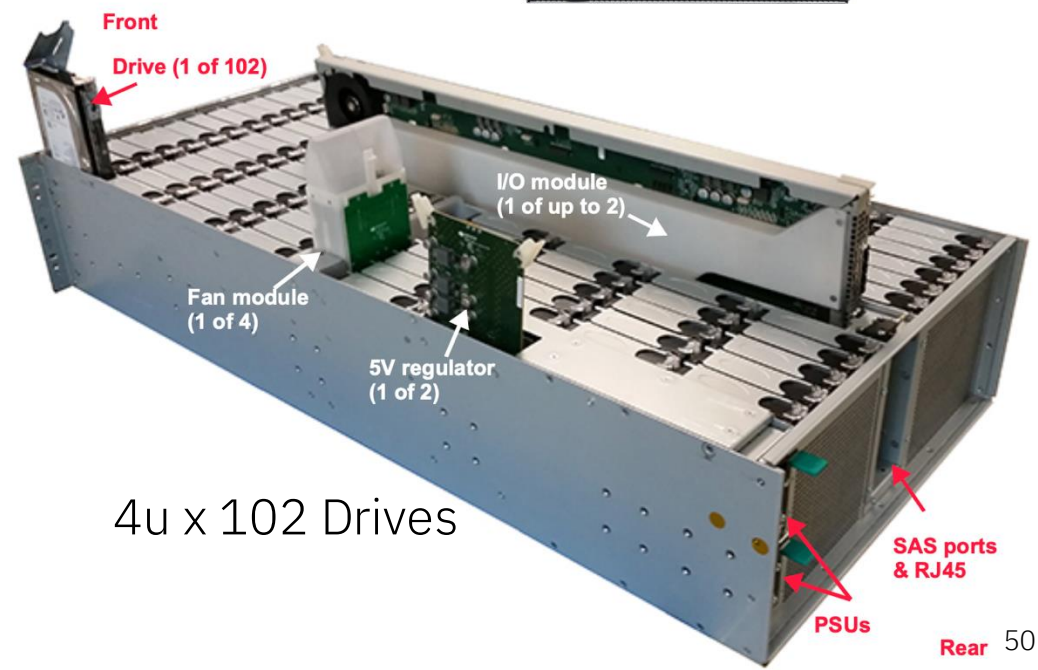
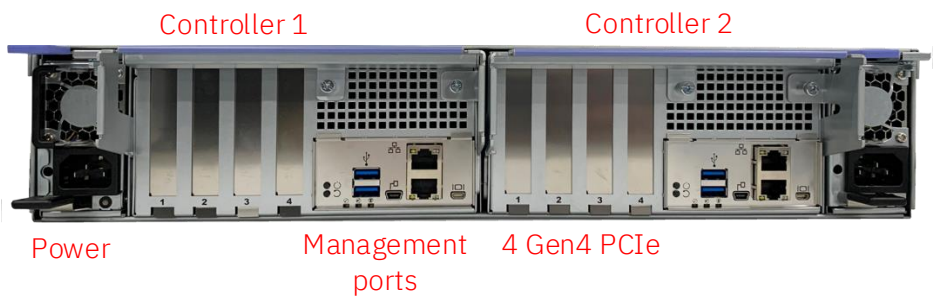
ESS GUI



ESS 3500



2u x 24 Drives



4u x 102 Drives

IBM's Purpose-Built Storage Solution for AI

The world's fastest systems need the world's best storage.
IBM has the best storage for NVIDIA GPUs

Highest Performance Platform

- Fastest performance for reads, writes, and density
- Linearly scalability for future growth

A Robust Enterprise Platform

- Up to Six 9's for all apps: AI, Analytics, HPC, Back-up, Archive, Cloud
- Cyber-resilient, encryption, WORM, and immutability

Collapse Layers & Simplify Workflows

- Eliminate extra copies and share data globally with all protocols
- Data cataloging and tiering for economics and data flexibility

High Efficiency and Low TCO

- Minimal power, cooling, and rack space without sacrificing performance
- Leading **GB/s-per-kW** and **GB/s-per-RackU**
- 5-Year TCO lower than leading competitors

IBM Storage Scale System 6000

A single 4U node with active-active controllers and redundant hardware to maximize always on data



Ultimate Performance and Scalability

up to 330 GB/s read performance per node

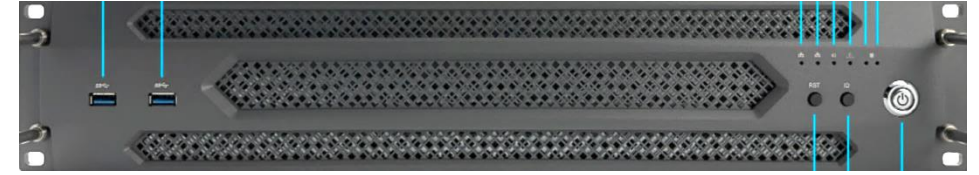
up to 155 GB/s write performance per node



Utility Node (5149-23E)

Features

- 2U (25" Depth) Rack Server, CRPS Redundant P/S
- Processor
 - Dual Socket SP3 (LGA4094)
 - Support AMD® EPYCTM 7003 and 7002 Processors
- Memory
 - Up to 8TB
 - 16 DIMM slots per CPU (8 channels per CPU)
 - Supports RDIMM, LRDIMM, RDIMM/LRDIMM-3DS
- PCIe Adapter Slot
 - 4 x PCIe 4.0 x16 + 1 PCIe 4.0 x8 RISER Slots (not used)
 - 1 x OCP NIC 3.0 Slot (PCIe 4.0 x16) (not used)
- Boot Drive
 - 2 x M.2 Micron 7450 PCIe 4.0 x4 for OS
- System Management and BMC
 - 1 x 1Gb Enet for BMC (IPMI 2.0) and SOL
 - 2 x 10Gbit LAN for management (Mgmt/SSR/Others)



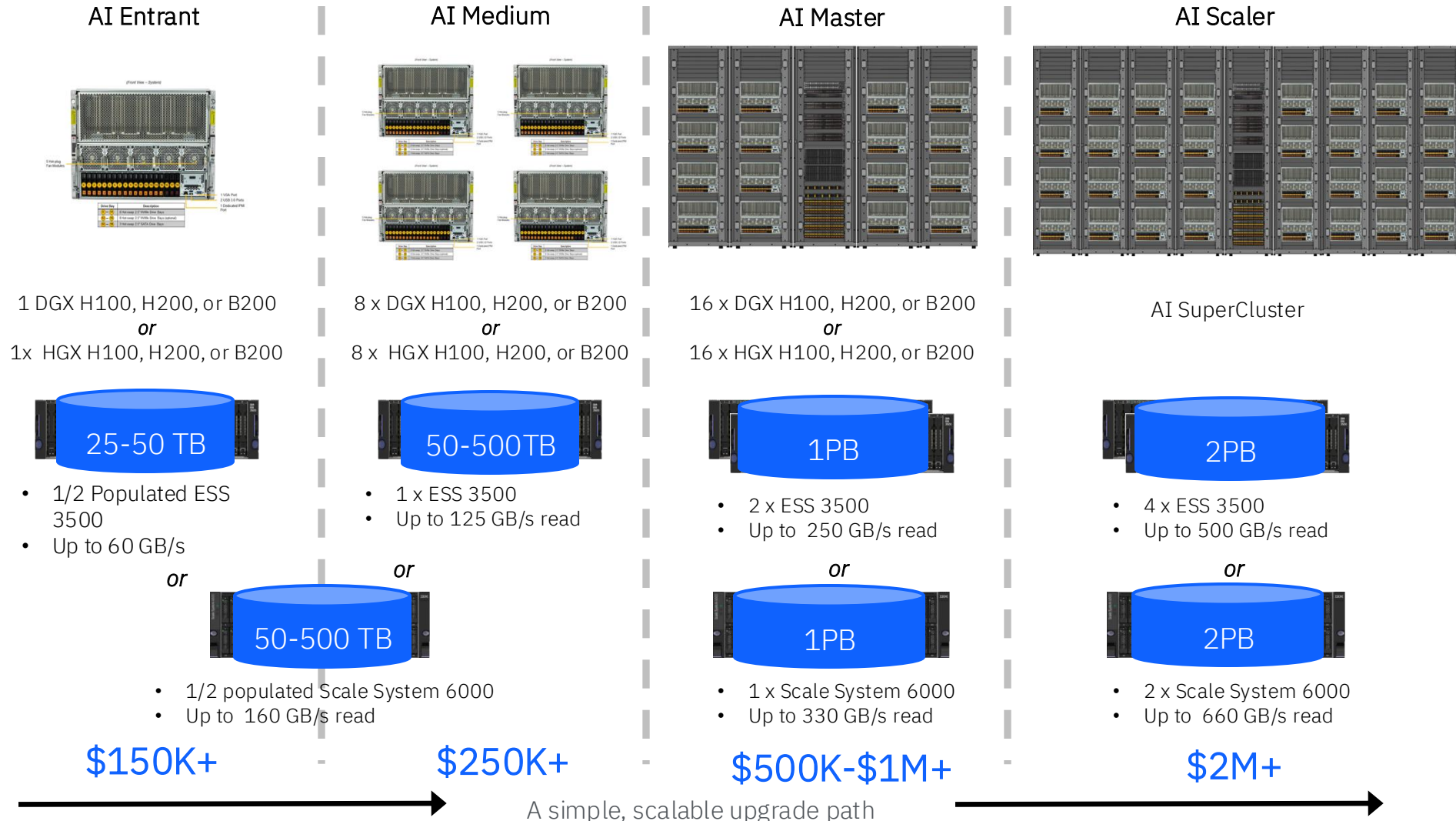
Front View



Rear View
and internal Components

IBM Storage for Data and AI & NVIDIA GPU Solutions

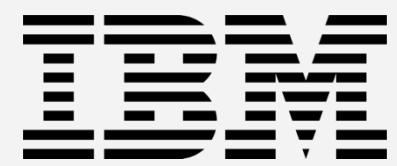
A full spectrum of scalable AI solutions



Promise of IBM Storage

- Simple building blocks** – Scalable seamless storage upgrade path as needs grow from 1st HGX to AI CoE HGX SuperCluster
- Global Data Platform** – Data fidelity capabilities to automate AI workflows
- Data Economics** – Eliminate copies and transparently tier
- Global Deployment** – Trusted and successful global enterprise level support and services

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



LLM Data Set Size

Checkpoint time reduced to 1% an hour => 36 seconds

Model Specific Example for Synchronous

Checkpoint

Tensor Model Parallel Size determines how many GPUs participate in the checkpoint.

For example, If Tensor Parallel size is set to 8, 1 out of 8 GPUs will participate in the checkpoint

~14 bytes per Parameter

Example

175B Parameter Model: ~2.4TB Data set size

512B Parameter Model: ~7.2TB Data set size

1T Parameter Model: ~14TB Data set Size

3 x IBM Storage Scale 6000 is 19.4 TB in 36 seconds

Total Number of GPUs is 4000 GPUs

Only 512 GPUs will participate in the checkpoint
(4000_GPUs / 8_Tensor Parallel_Size)

175B: ~4.8GB data set / GPU

512B: ~15GB data set / GPU

1T: ~28GB data set / GPU



Model Load

Using the same Tensor Parallel Size of 8 from the checkpoint

8x the data set size will need to be loaded across all GPUs

Example

175B Parameter Model: ~19TB Data set size

512B Parameter Model: ~58TB Data set size

1T Parameter Model: ~110TB Data set Size

3 x IBM Storage Scale 6000 load in < 2 min

Total Number of GPUs is 4000 GPUs

Data set per GPU is the same, but now all GPUs participate in the Model Load

175B: ~4.8GB data set / GPU

512B: ~15GB data set / GPU

1T: ~28GB data set / GPU

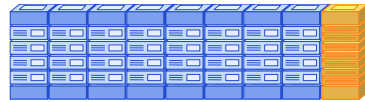
- ✓ READ: 320 GB/s
- ✓ WRITE: 155 GB/s

Blue Vela Compute Pods



IBM Blue Vela- HGX “SuperPOD” Storage Fabric (IBM Cloud/ IBM Research/NVIDIA) working together to delivery an AI solution

Storage Scale 6000



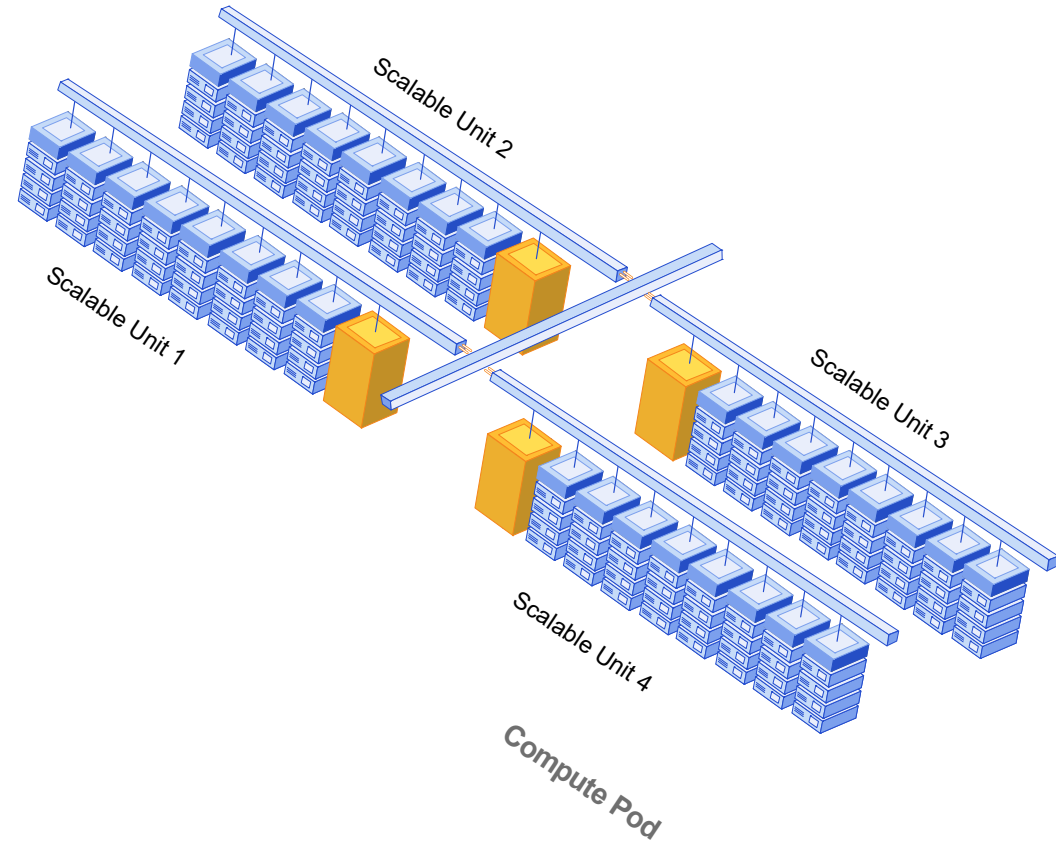
Scalable Unit

- 32 Compute Nodes
- 256 H100 GPUs



Compute Pod

- 4 Scalable Units
- 128 Compute Nodes
- 1024 H100 GPUs
- 82 TB of GPU Ram
- 12,288 Physical Cores
- 256 TB of RAM
- 3481 TB NVME Local Storage

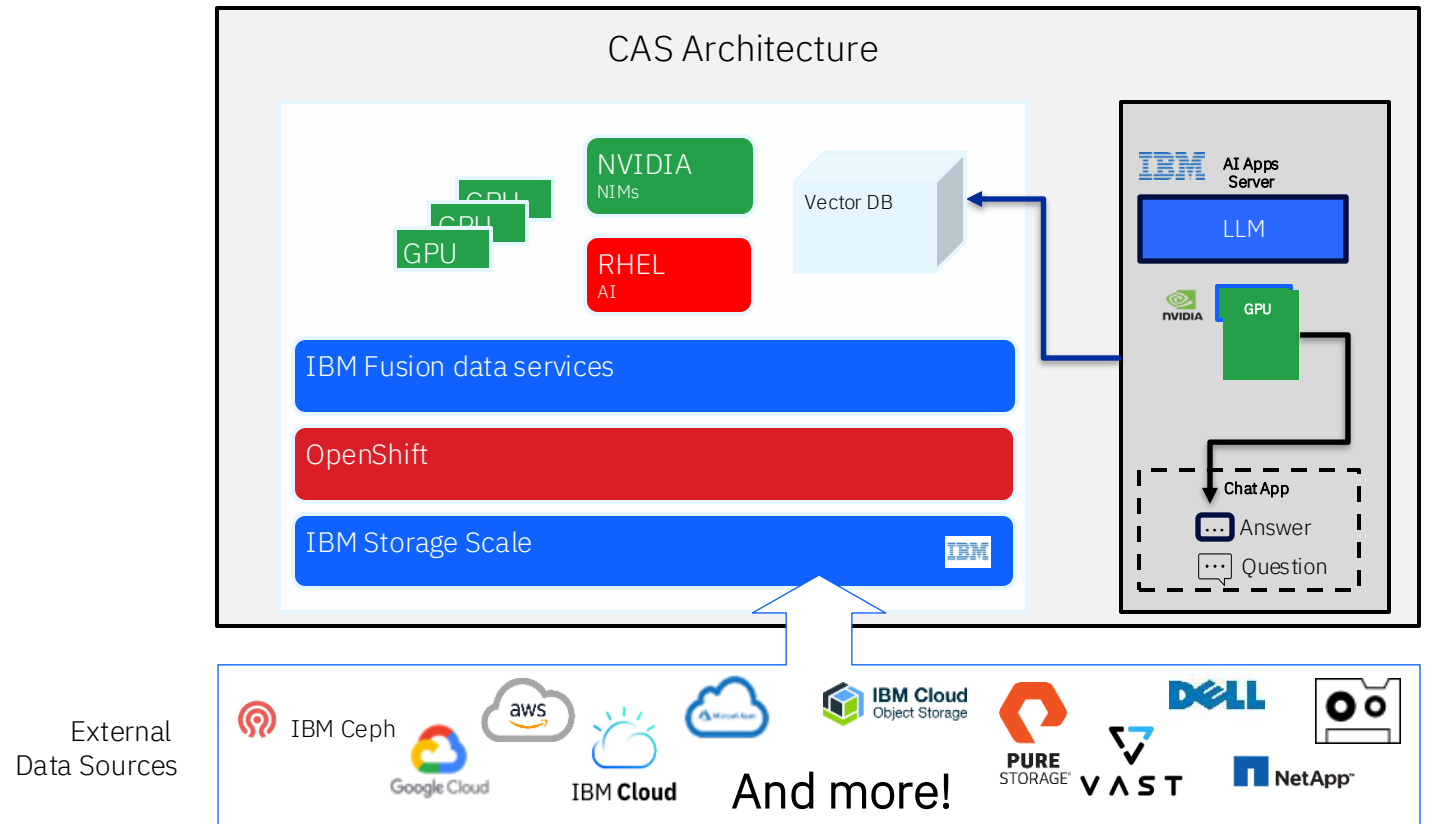


https://www.linkedin.com/posts/dannybarnett_todays-an-incredibly-proud-day-for-my-team-activity-7180654456361910274-4fvU?amp;utm_medium=member_ios

<https://arxiv.org/pdf/2407.05467>

IBM Content Aware Storage Software Architecture

- Provides flexibility to integrate data pipelines with IBM Fusion data services
- Integrates Storage Scale for data access and storage optimization
- Integrates advanced vector database and enables hardware acceleration





SCAN ME

IBM Storage Scale



SCAN ME

IBM Storage Scale System

Get started today!

Learn more about IBM Storage for AI:

- [IBM Storage Data and AI web page](#)

Learn more about IBM Storage for NVIDIA:

- [IBM Storage and NVIDIA web page](#)

Learn more about IBM Spectrum Scale and Elastic Storage System (ESS):

- [IBM ESS web pages](#)
- [IBM Spectrum Scale web page](#)

Learn more about IBM Cloud Object Storage (IBM COS):

- [Learn about the IBM COS story \(interactive experience\)](#)
- [IBM COS web page](#)

Learn more about IBM Spectrum Discover:

- [IBM Spectrum Discover web page](#)

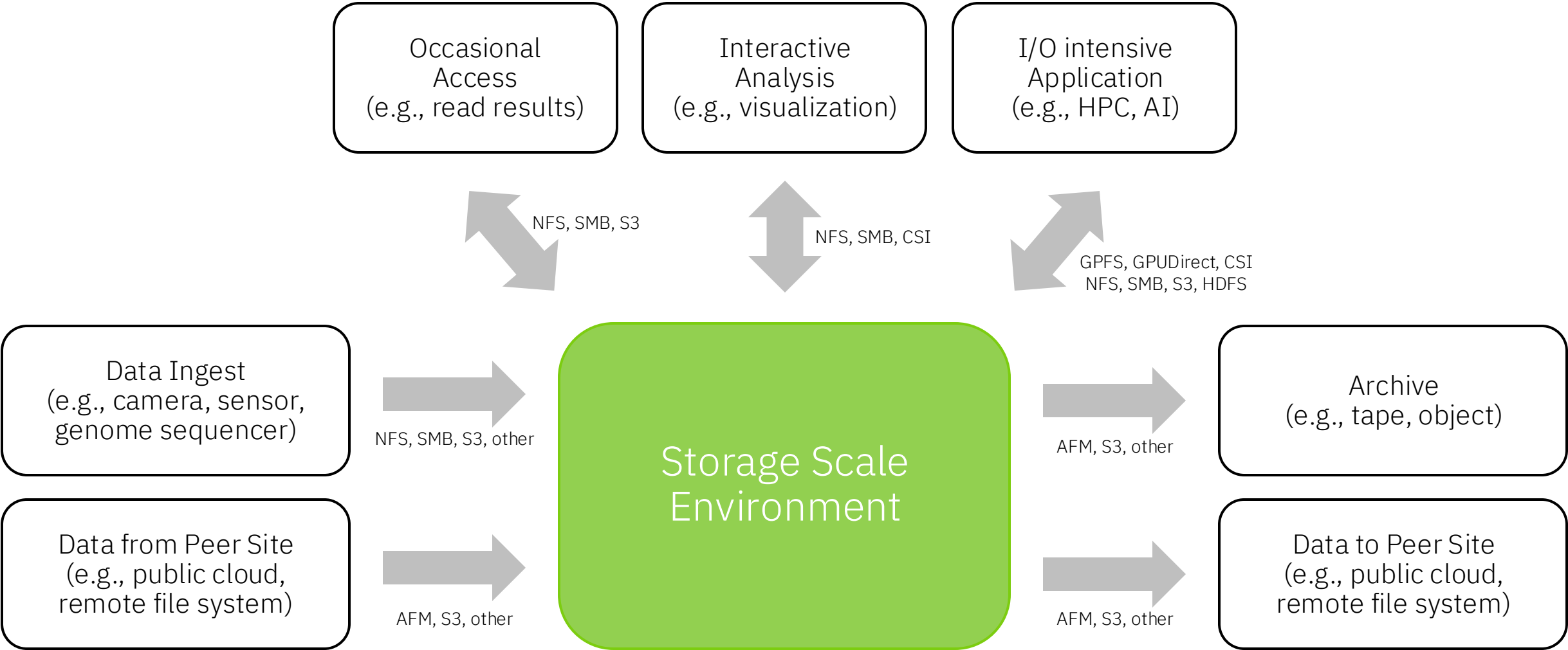
Learn more about IBM Spectrum Fusion:

- [IBM Spectrum Fusion web page](#)



Access Services

Data-Intensive Workflows for Unstructured Data



- Many unstructured data is generated and processed outside the data center.
- File and object protocols allows devices and applications to access and process data on remote servers and systems.
- Many of the remote applications and devices stick to one of the many file and object protocols (e.g., genome sequencers).

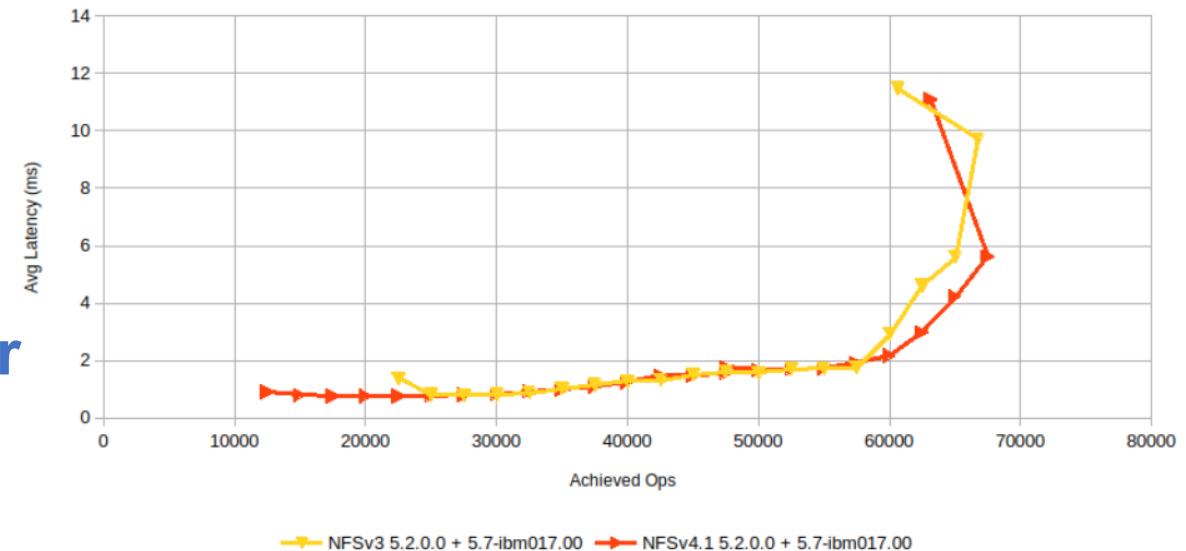
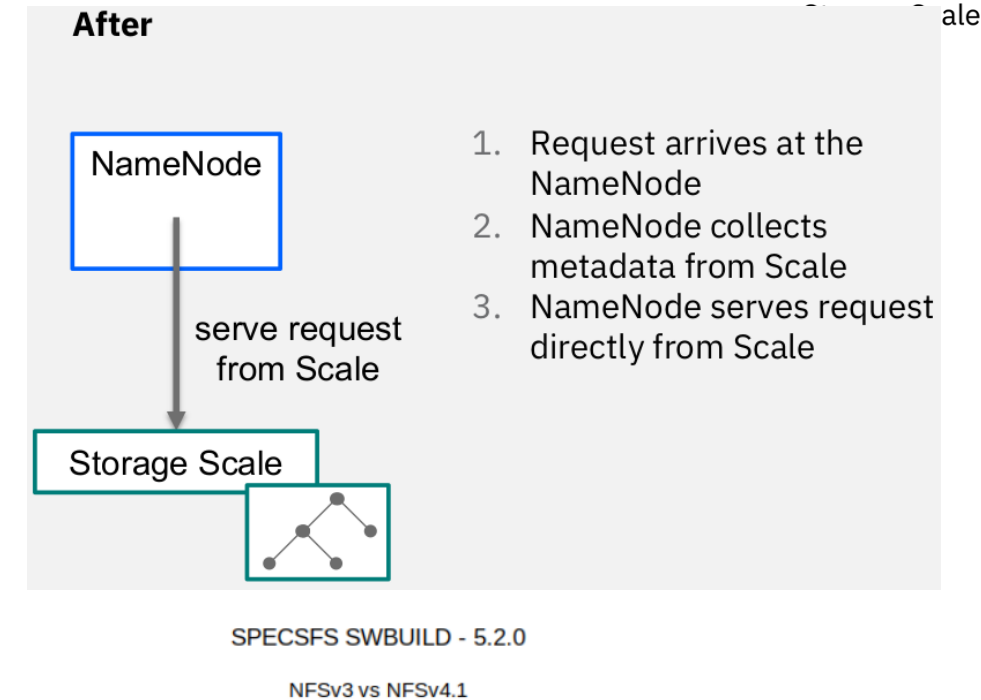
Access Services –NFS, SMB, HDFS

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:

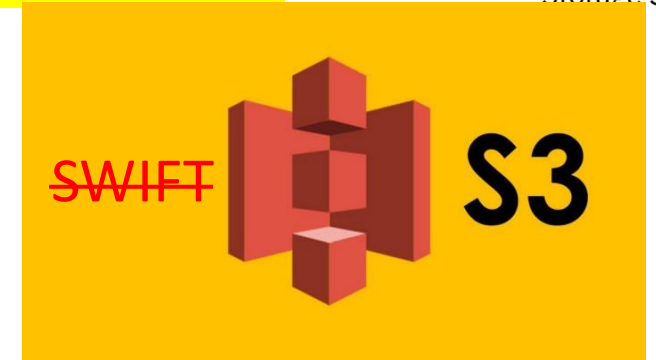
- NFS “meta data cache” component was revised resulting in significant performance improvements
- HDFS transparency metadata redesign
 - Full parallelism for RPC calls (GPFSNamesystem)
 - No more lock contention in NameNode
- **Continued partnership with Tuxera for high-performance SMB**
- **Evaluating MoSMB as well**



Access Services – High Performance Object 2.0!

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

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/rogerio-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 !

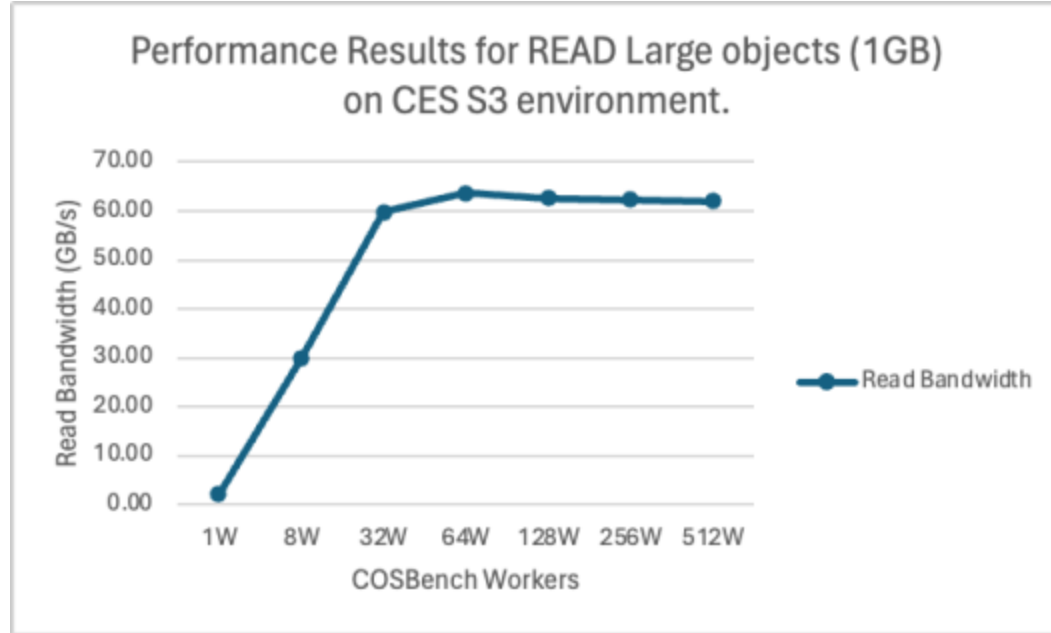
Deployment Requirements:

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

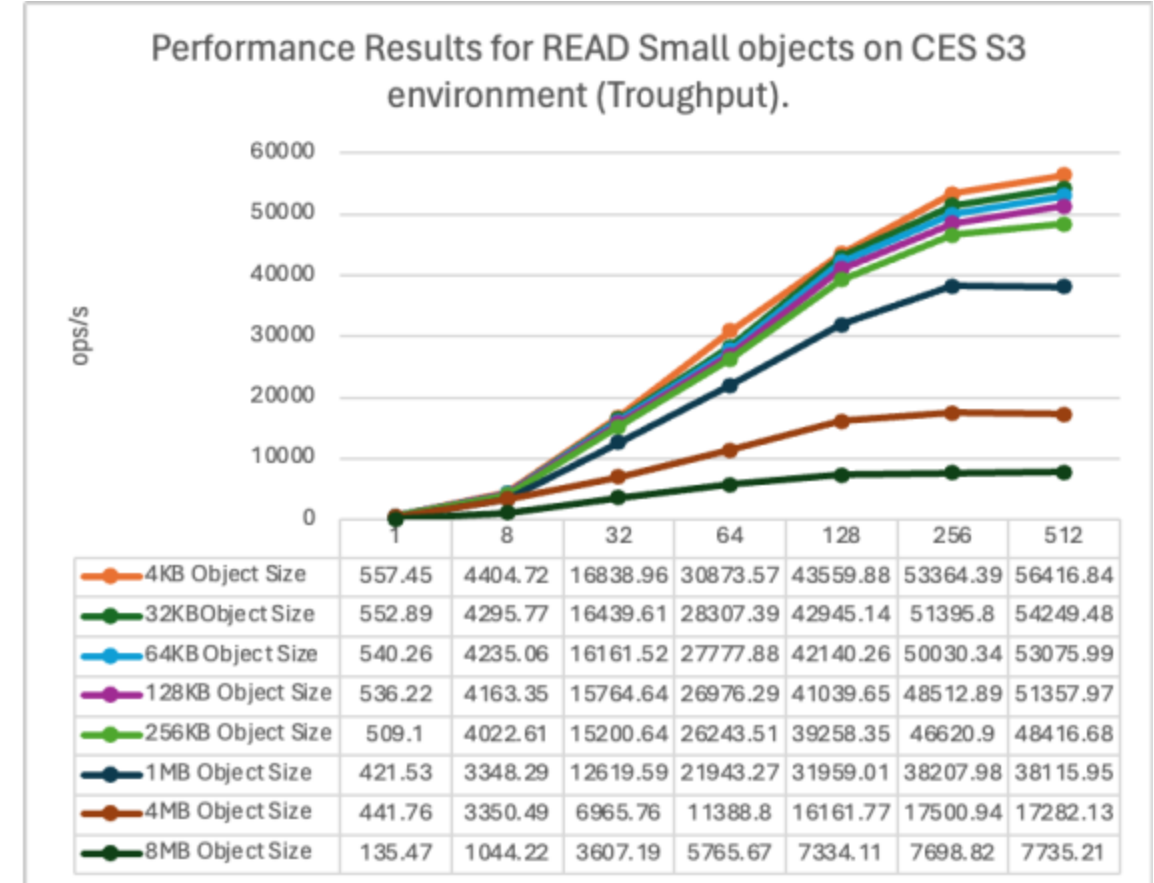
NO
Openshift
cluster
required !

Access Services – Object Performance



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.



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

Access Services – Container Native Storage Access (CNSA)



le

Improvements introduced in CNSA 5.2.1.0

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!

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

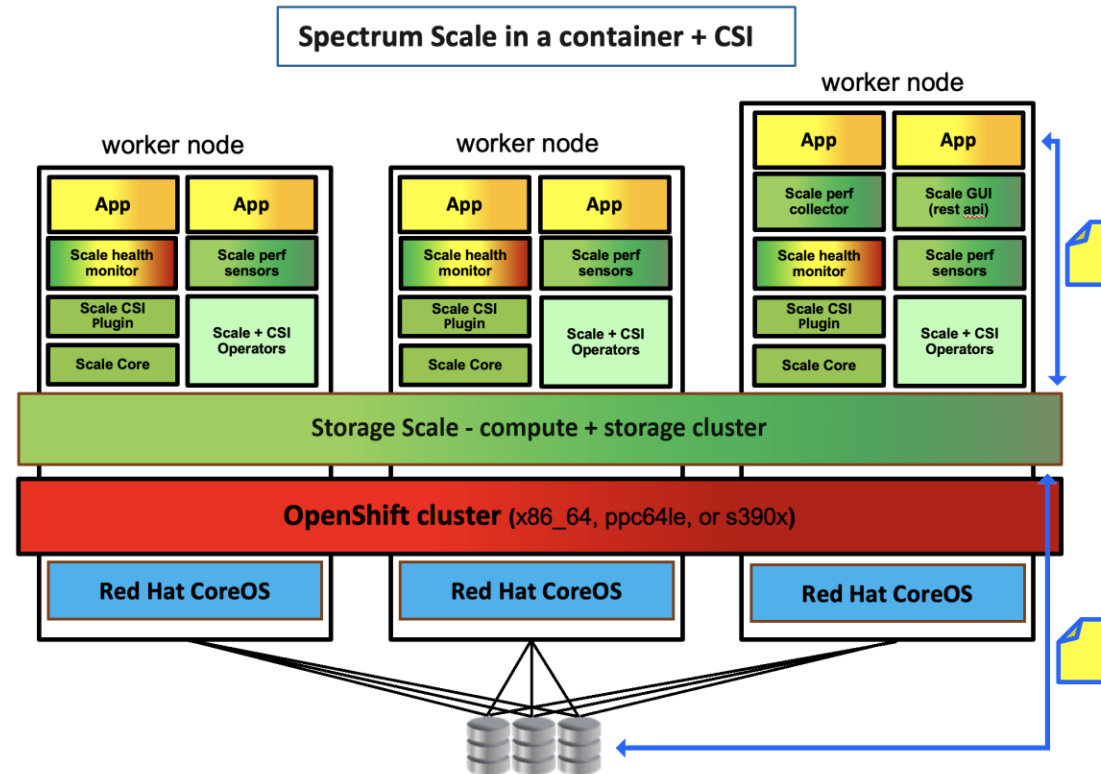
Support for RedHat OpenShift Container Platform 4.[13-16]

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.

Continued Tech preview of Infiniband RDMA - *note: currently for a single customer, as we are working with Nvidia and RH for more usable RDMA functionality within OCP*



Access Services – Container Storage Interface

Improvements introduced in CSI 2.11.0

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

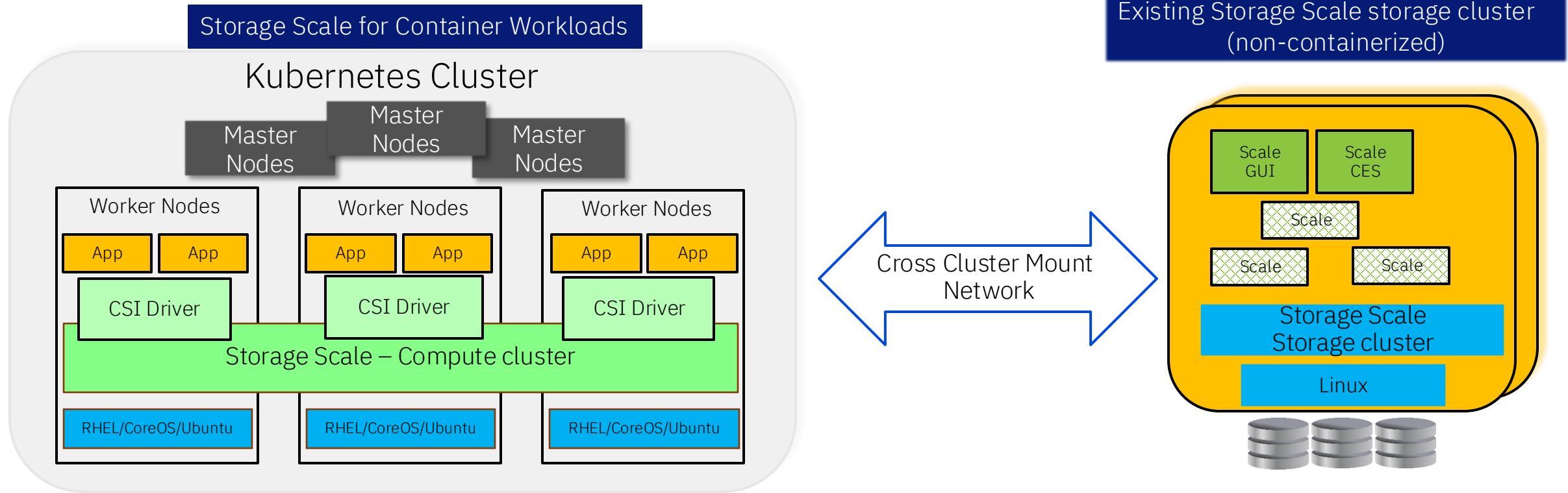
Improvements in the script for debug data collection

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



IBM Storage Scale

Container Storage Interface



- Scalability: Containerized compute cluster can scale with the Kubernetes cluster
- Speed: R/W benchmarks of Storage Scale CSI have shown same performance as non-containerized Storage Scale
- Compatibility: Integration with Run.AI and Nvidia Base Command Manager (BCM)
- Automation: Storage Scale and CSI operators allow automated cluster and storage provisioning
- Flexibility: Existing Storage Scale, ESS, ECE, clusters are used as storage via a remote mount, independent of OpenShift
- Open standards: CSI provides an open standard for direct access to Storage Scale storage

Abstraction and Acceleration

IBM Storage Scale

Data Acceleration Tier

Accelerates AI and Analytics by allowing data to be stored in two pools at once, enabling current use of:

(1) a GNR reliable copy

AND a performance copy, which could be:

(2a) a copy of data that is as close as possible to the computing client (where it's operated on) OR

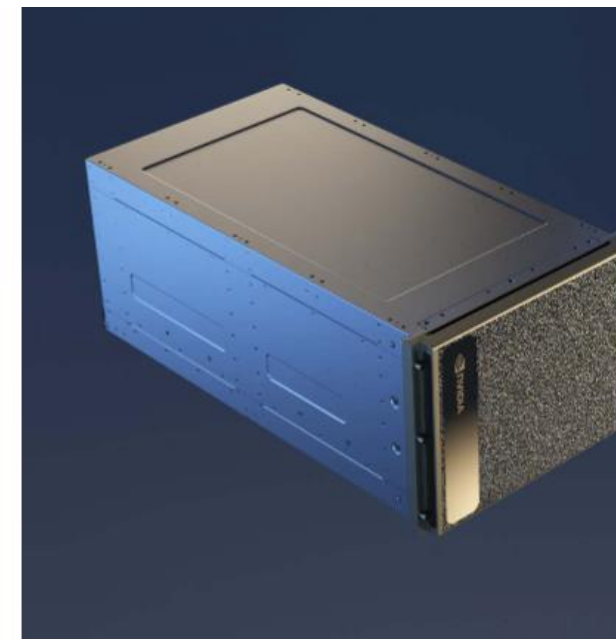
(2b) a copy of the data that leverages the small I/O performance benefits of NVMeoF.

The first release will support just (case 2b) SSS 6000 shared storage accessed via NVMeoF, and later will exploit (case 2a) local storage (e.g. local NVMe, Persistent Memory, or DRAM).

This first use of Asymmetric Replication on Storage Scale System is called the Data Acceleration Tier NVMeoF Performance Pool, providing one erasure-encoded copy of the data and one performance copy that can dramatically accelerate the read performance of smaller I/Os.

Can create a Shared (in model 2a) Co-operative Cache across all compute nodes. Any node can access all cached data, regardless of physical location.

NVIDIA DGX SuperPOD with IBM Scale System



Specifications	
GPU	8x NVIDIA H100 Tensor Core GPUs
GPU memory	640GB total
Performance	32 petaFLOPS FP8
NVIDIA® NVSwitch™	4x

Coming Soon

Introducing Data Acceleration Tier! (uStore) Caching and Acceleration to the compute

Coming Soon



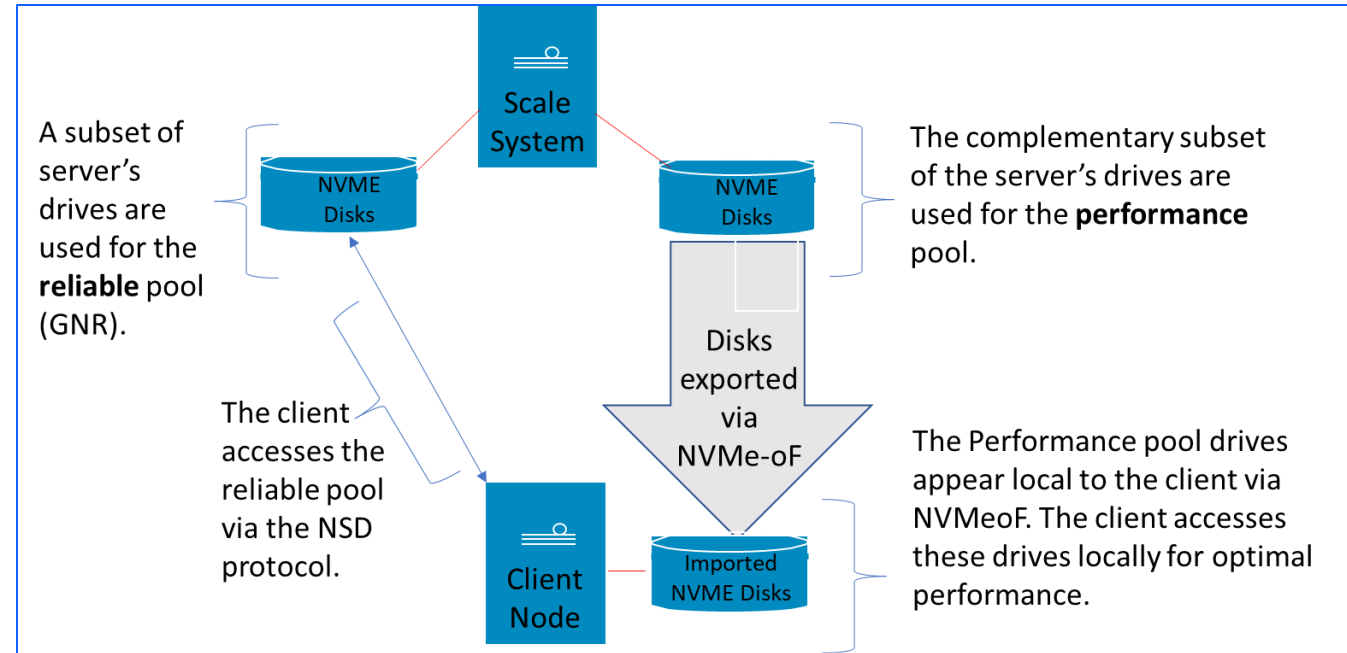
Accelerate AI and Analytics by storing the data as close to the compute as possible

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

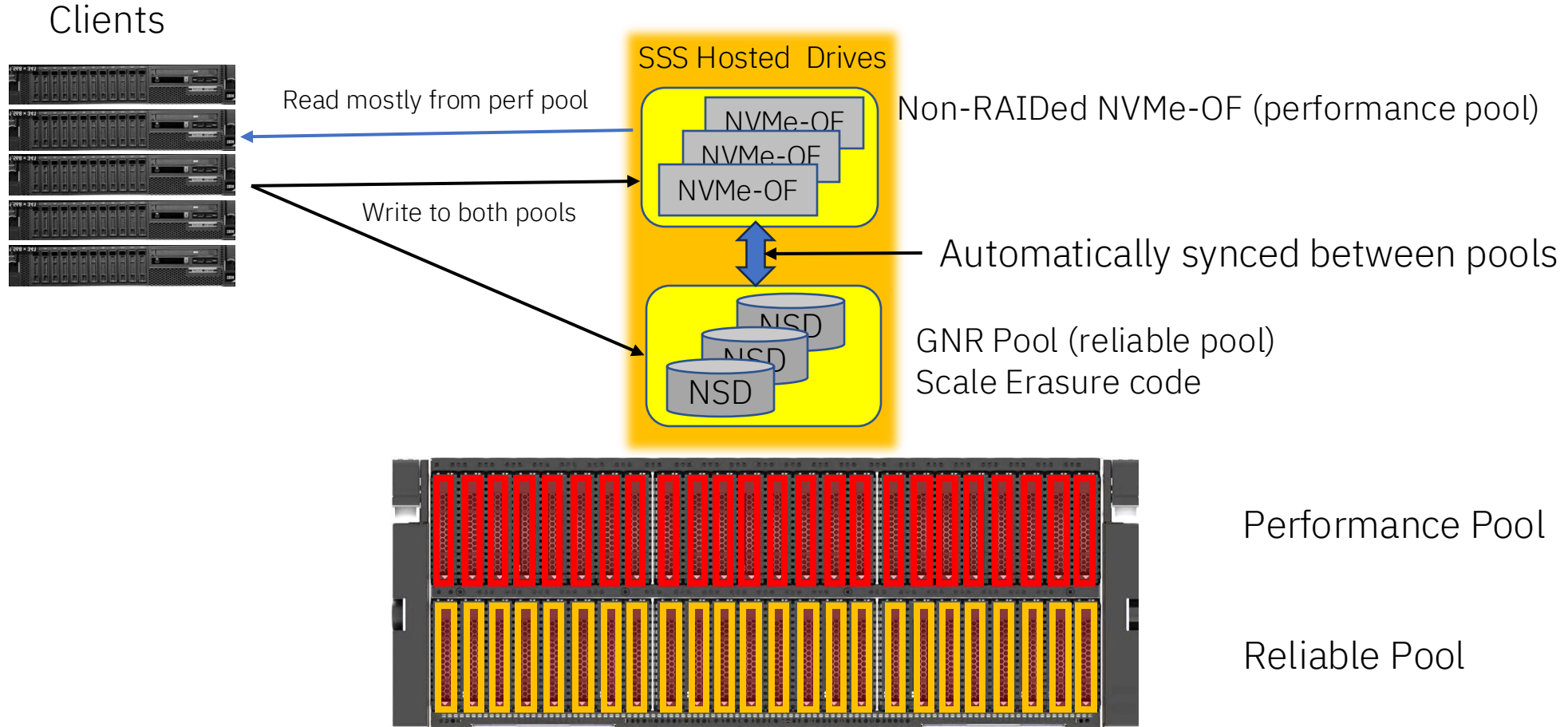
Creates a **Shared Co-operative Cache** across all compute nodes. Any node can access all cached data, regardless of physical location.

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**)



IBM Storage Scale

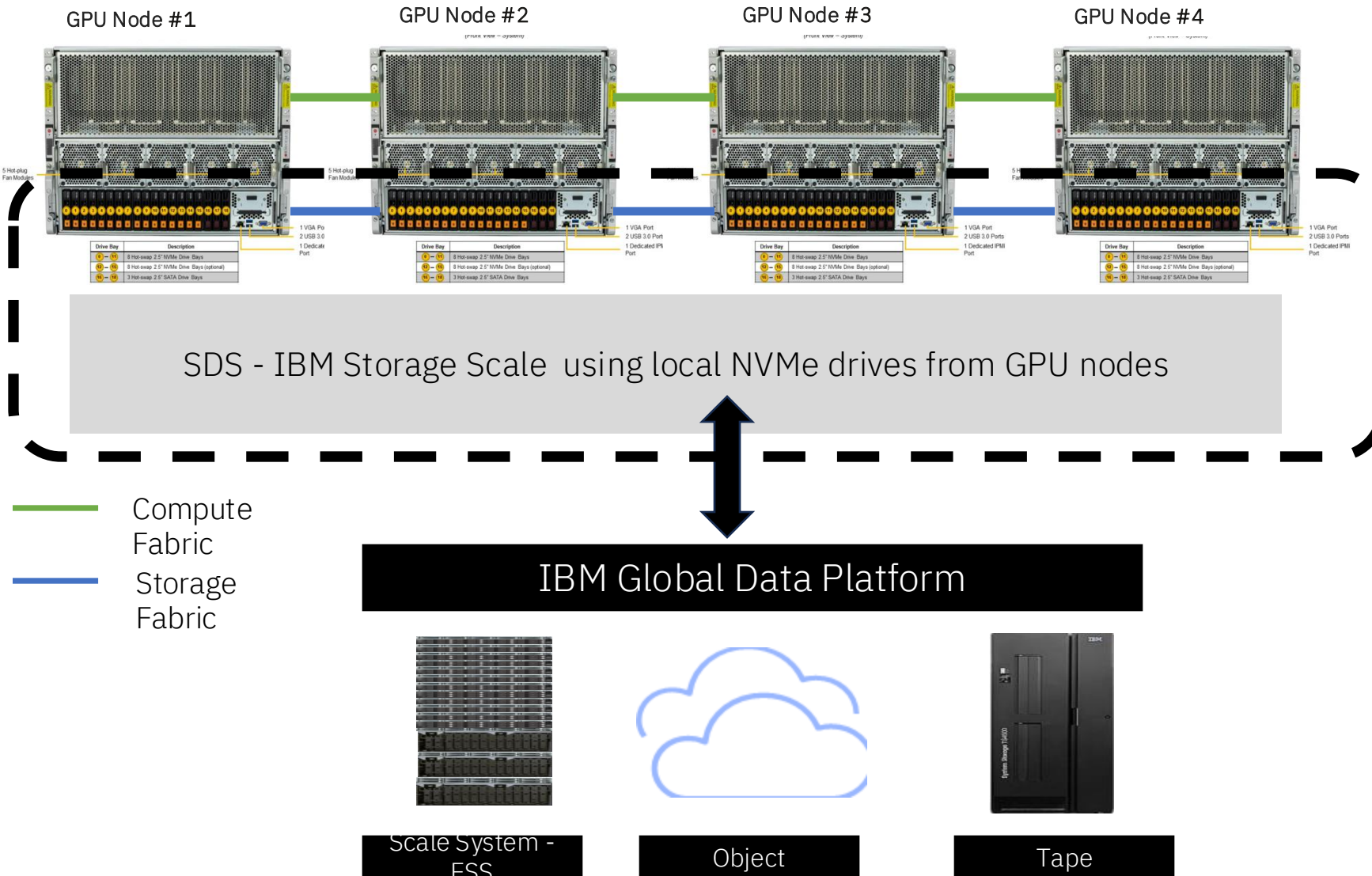
Data Acceleration



- Data is written both in performance pool and reliable at the same time – not need for Scale replication and data placement policies
- Provides very high IOPS with data resiliency of Storage Scale RAID (4KB read @ 13M IOPS)

A **NEW** approach to Data at the compute – History - Converged Solution Architecture

Coming Soon



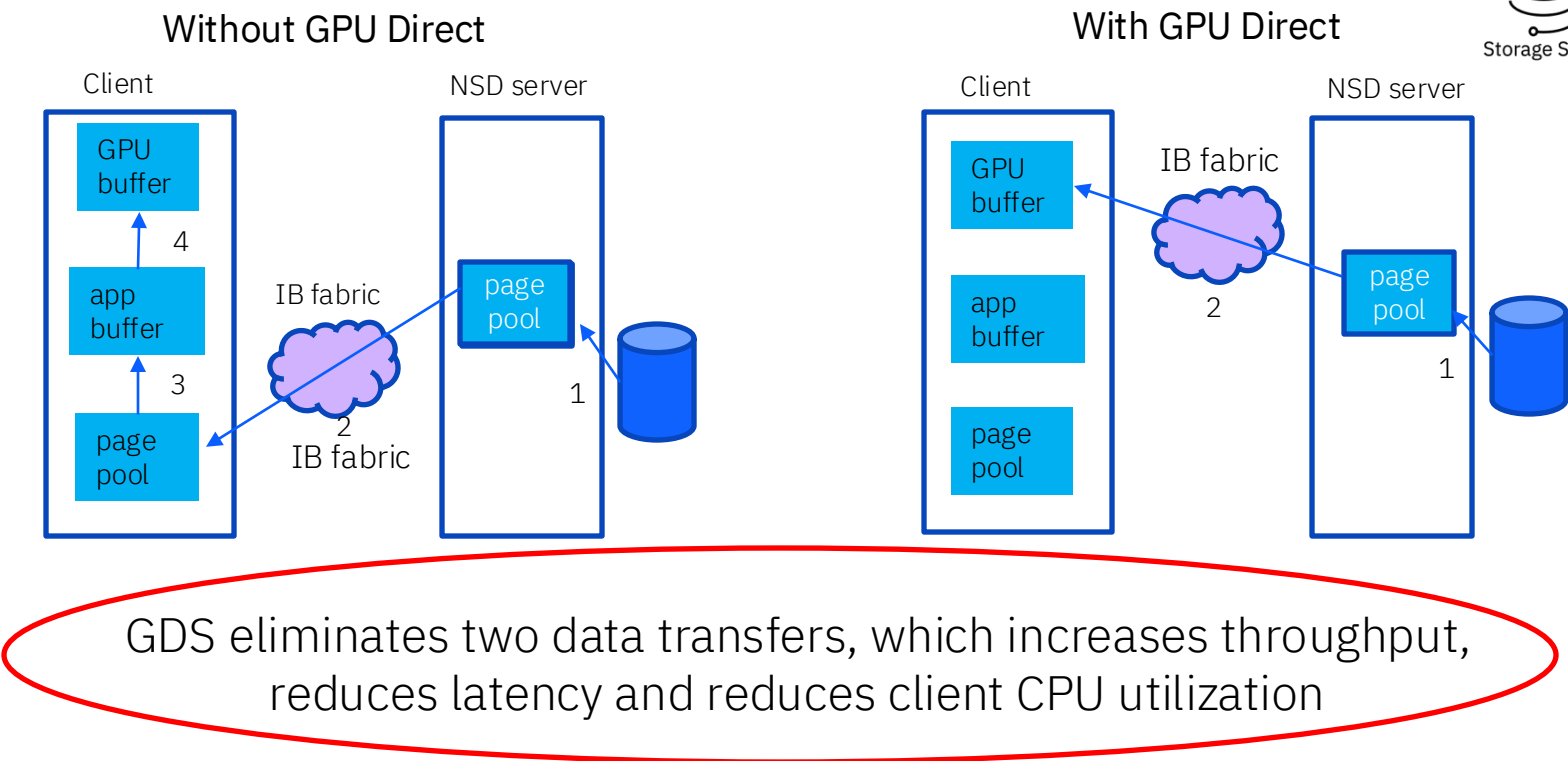
IBM Storage Scale configuration

- 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 drives across 4 node cluster
- 2 x 200 Gbps storage network per node

IBM Storage Scale

GPU Direct Storage

GPUDirect Storage (GDS) for Storage Scale enables an RDMA (remote direct memory access) path between GPU memory and storage.



NVIDIA Magnum IO

- Family of I/O Optimizations for GPU accelerated data centers.
- GPUDirect RDMA: Access peer node's memory without copying through host memory
- GPUDirect Storage: Transfer data between GPU and storage without involving CPU and CPU memory

NVIDIA CUDA Toolkit

- GDS API is in the CUDA toolkit
- A 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
- GDS APIs must be explicitly called by the applications
- Storage must be GDS enabled. If not, GDS APIs use regular data movement.

Why it matters

- AI, HPC, ML and analytics are data hungry and require a very high data throughput.
- GPUs can be starved by slow I/O due to multiple data transfers on the client.

IBM Storage Scale

High Performance Parallel Storage

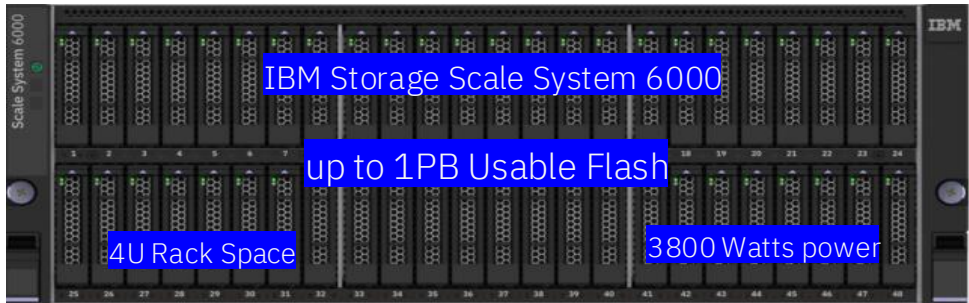
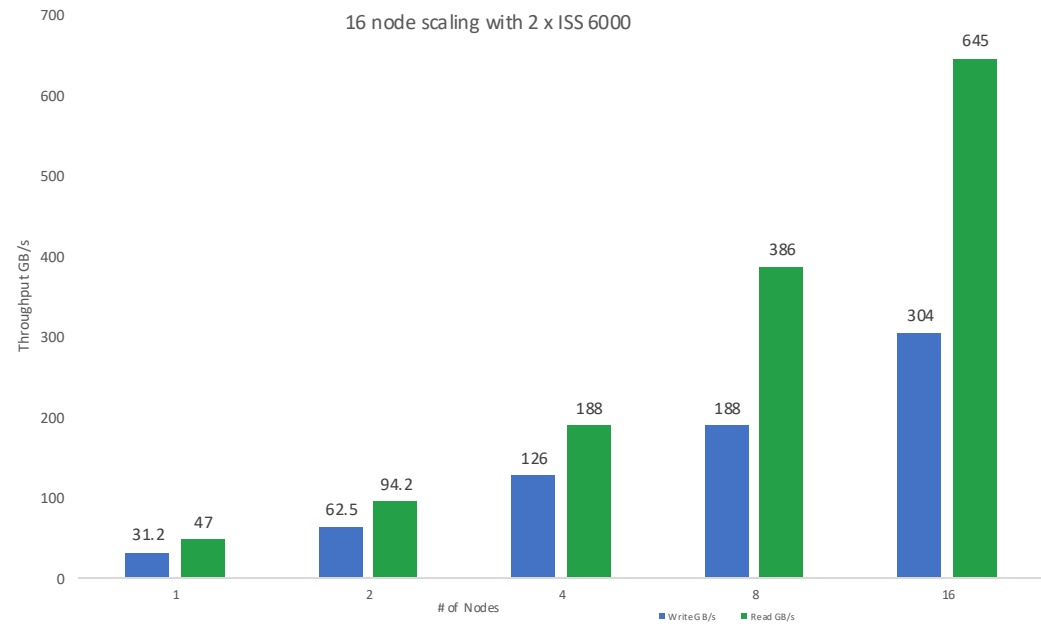
NVIDIA SuperPOD Storage Guidelines

<https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastrucre-h100/latest/storage-architecture.html>

Performance Characteristic ¹	Good (GBps)	Better (GBps)	Best (GBps)
Single node read	4	8	40
Single node write	2	4	20
Single SU aggregate system read	15	40	125
Single SU aggregate system write	7	20	62
4 SU aggregate system read	60	160	500
4 SU aggregate system write	30	80	250

- Peak performance for writes and read are needed for creating and reading checkpoint files
- Checkpoint is a synchronous process, and training stops during this phase
- High performance, resilient, parallel filesystem storage recommended for multi-threaded read and write operations across multiple nodes
- Checkpoints restart drastically reduces LLM training time with Storage scale

IBM Storage Scale system 6000 delivers “Best” in class performance



Abstraction and Acceleration Services – Dynamic Page Pool



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 every grace period.

Default configuration changes

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

IBM Storage Scale

Data Abstraction

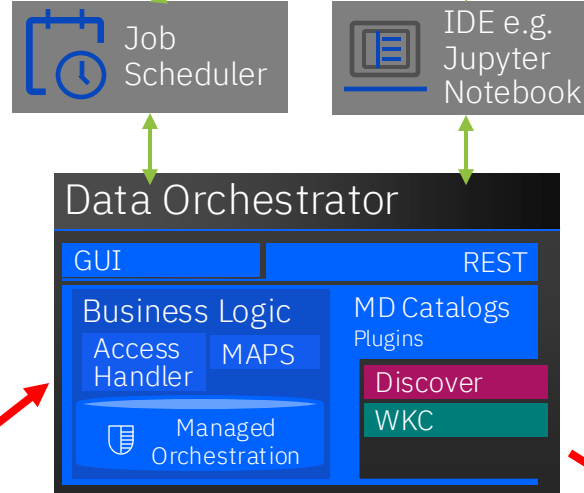
COMING SOON

Data Scientist

“I need pictures with curbstones on a rainy day”

Data Orchestrator:

- Query metadata catalog/vector database to determine data set
- Establish AFM connection
- Prefetch dataset
- Create PV and PVC for external data
- Create access control credentials
- Schedule work
- Monitor progress
- Post process clean up



```
26  "#ffffc8": "Utility vehicle 2",
27  "#e96400": "Sidebars",
28  "#6e6e00": "Speed bumper",
29  "#808000": "Curbstone",
30  "#ffc125": "Solid line",
31  "#400040": "Irrelevant signs",
32  "#b97a57": "Road blocks",
33  "#000064": "Tractor",
34  "#8b636c": "Non-drivable street",
35  "#400000": "Takes exception"
```

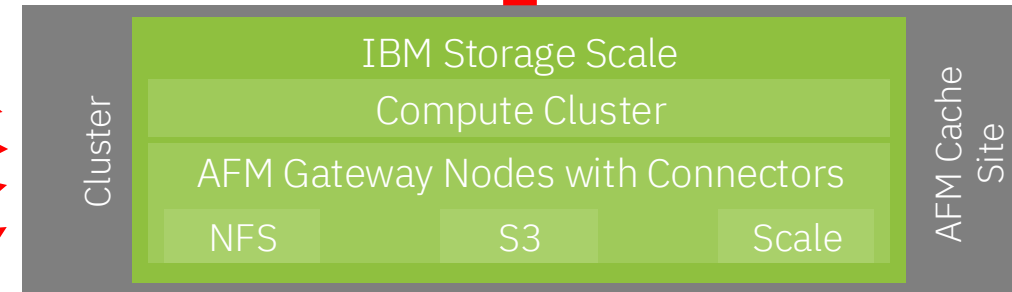
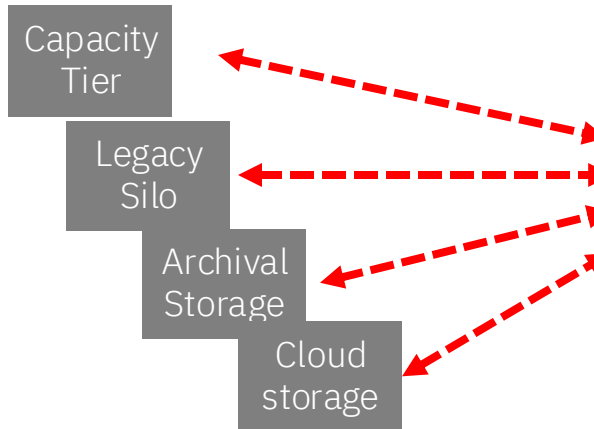


	A	P	Q	R	S
1	Filename	Sidebars	Speed bumper	Curbstone	Solid l
19995	95954.png	2841	0	0	17
19996	24650.png	1534	0	0	23
19997	75032.png	1241	0	0	22
19998	08079.png	1257	0	12203	20
19999	10288.png	3177	0	0	6
20000	65653.png	2121	0	0	15
20001	52979.png	0	3432	12333	18
20002	40246.png	0	0	9255	41

Dynamic Fileset

IBM Fusion Data Catalog
(aka. Discover)

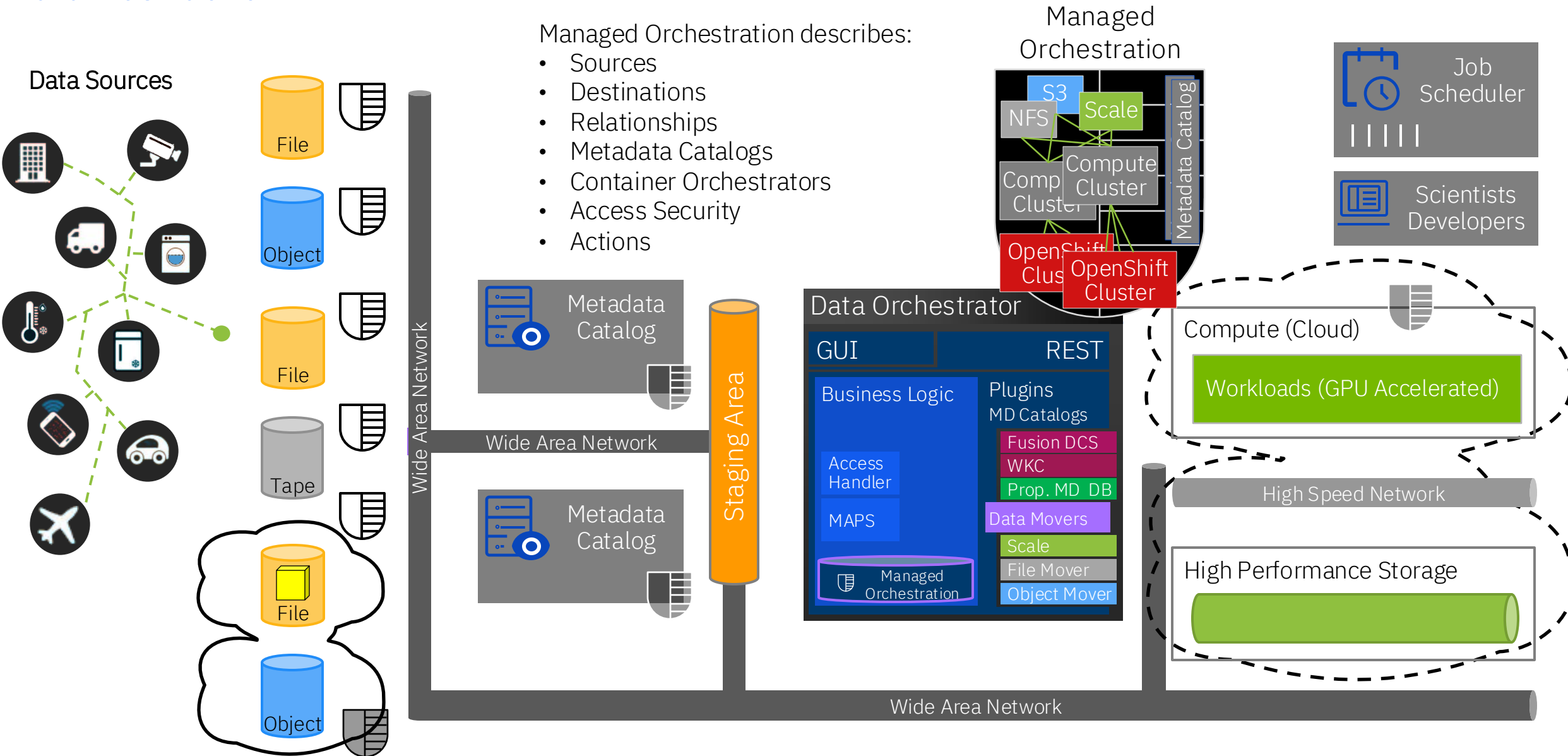
Data Orchestrator
eliminates storage
workflow complexity



IBM Storage Scale

Data Abstraction

Data Orchestrator Sample Use Case



IBM Storage Scale

Automated Deployments and Upgrades



Cloudkit: Resource provisioning and deployment of Storage Scale on public Clouds

- Command Line Interface to create Storage Scale clusters on public clouds (AWS, GCP, Azure (tech preview))
- Provides end to end automation to create and bring up an IBM Storage Scale cluster on public clouds (in minutes)
 - Automates infrastructure provisioning on the cloud
 - Automates the deployment of IBM Storage Scale on the cloud
 - Applies IBM Storage Scale best practises for deploying on the cloud
- Easy to use, guided interface

Install Toolkit: Installation, Deployment and Initial Configuration

- Fully automated CLI to install, Deploy and Configure Storage Scale on Bare Metal servers or Virtual Machines
- Provides ability to install, create and bring up a Storage Scale cluster
- Supports the automated creation of filesystems
- Supports the installation and initial configuration of advanced functionality such as Scale data access services (NFS, SMB, HDFS and Object protocols), AFM, File Audit Logging etc.

Upgrade

- “One button” rolling upgrade: Support for rolling upgrade of the Storage Scale cluster
- Offline parallel upgrade: Upgrade entire cluster parallely when the cluster is shutdown

Ansible based Install Toolkit Overview

Install toolkit Workflow

Define Cluster Topology

Use Storage Scale CLI commands to Cluster definition

- Add nodes to cluster
- Assign roles to nodes
- Define NSD
- Define File system

Install

Use Storage Scale install to perform:

- Installation required RPMs on all nodes
- Creates a Storage Scale cluster
- Creates NSD
- Sets up Management GUI
- Creates file system

Deploy

Use Storage Scale deploy to perform:

- Install, Configure and enable protocols

Upgrade

Use Storage Scale upgrade to perform:

- Online sequential upgrade of cluster
- Offline cluster upgrade

Toolkit can be used to automate deploy alone when install has happened manually

Toolkit can be used to upgrade an existing cluster that has been created manually

IBM Confidential



Cloudkit - Hashicorp Terraform!

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

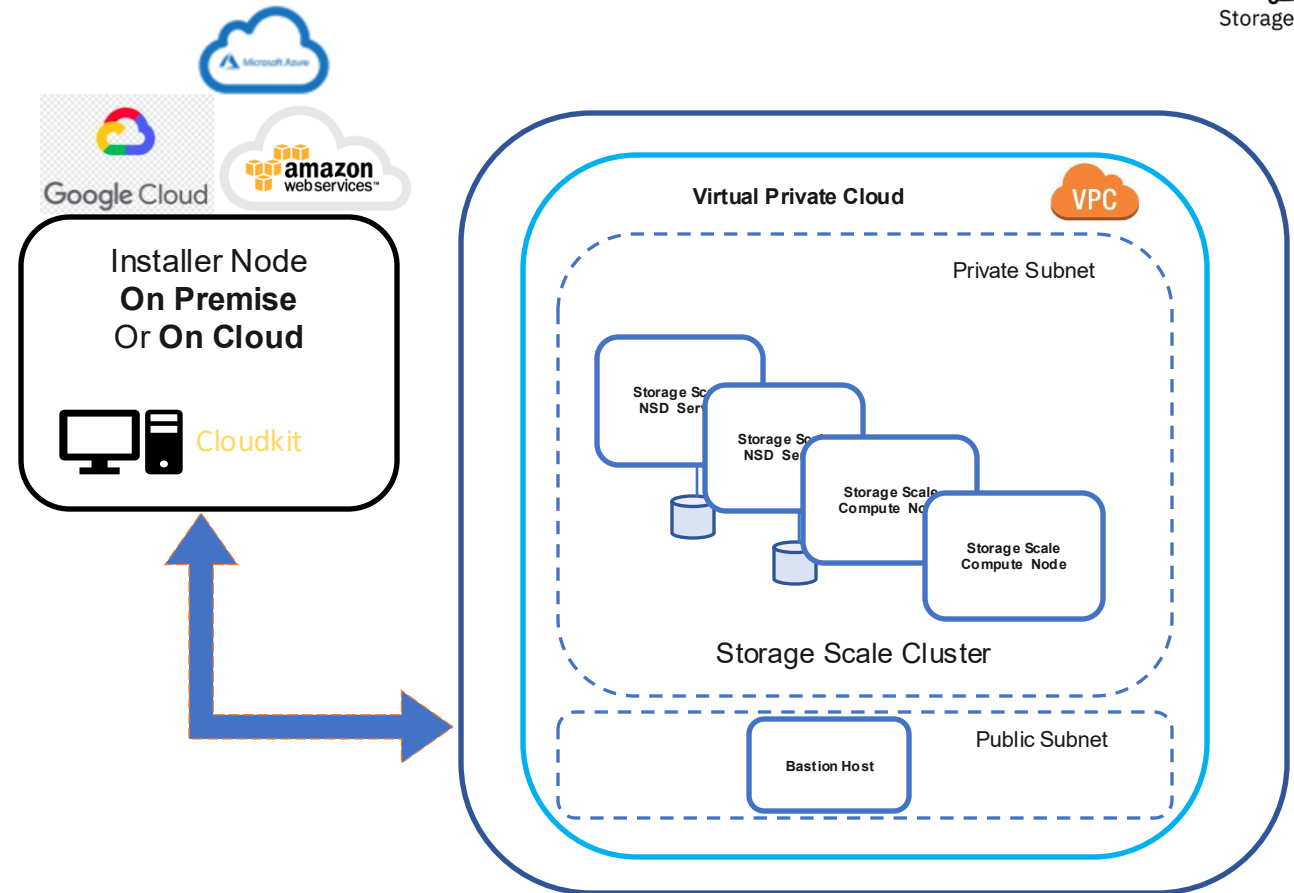
Advantages

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



- 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 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

Data Assurance

IBM Storage Scale

Assurance for Cyber Resiliency

IDENTIFY

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

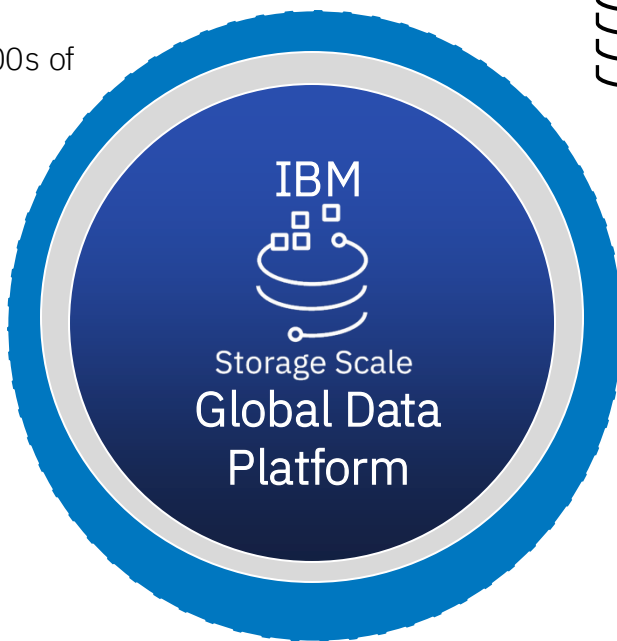
Governance

- 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

IBM Storage Scale

Native Filesystem Encryption

Data is encrypted while “at rest” on disk and decrypted on the way to reader – data not metadata

Encryption takes place on the node(s) from which the user drives the I/O

File content travels encrypted to the NSD server

MEKs can be accessed by nodes that have appropriate RKM credentials

Nodes that cannot access keys cannot access files, irrespective of file permissions

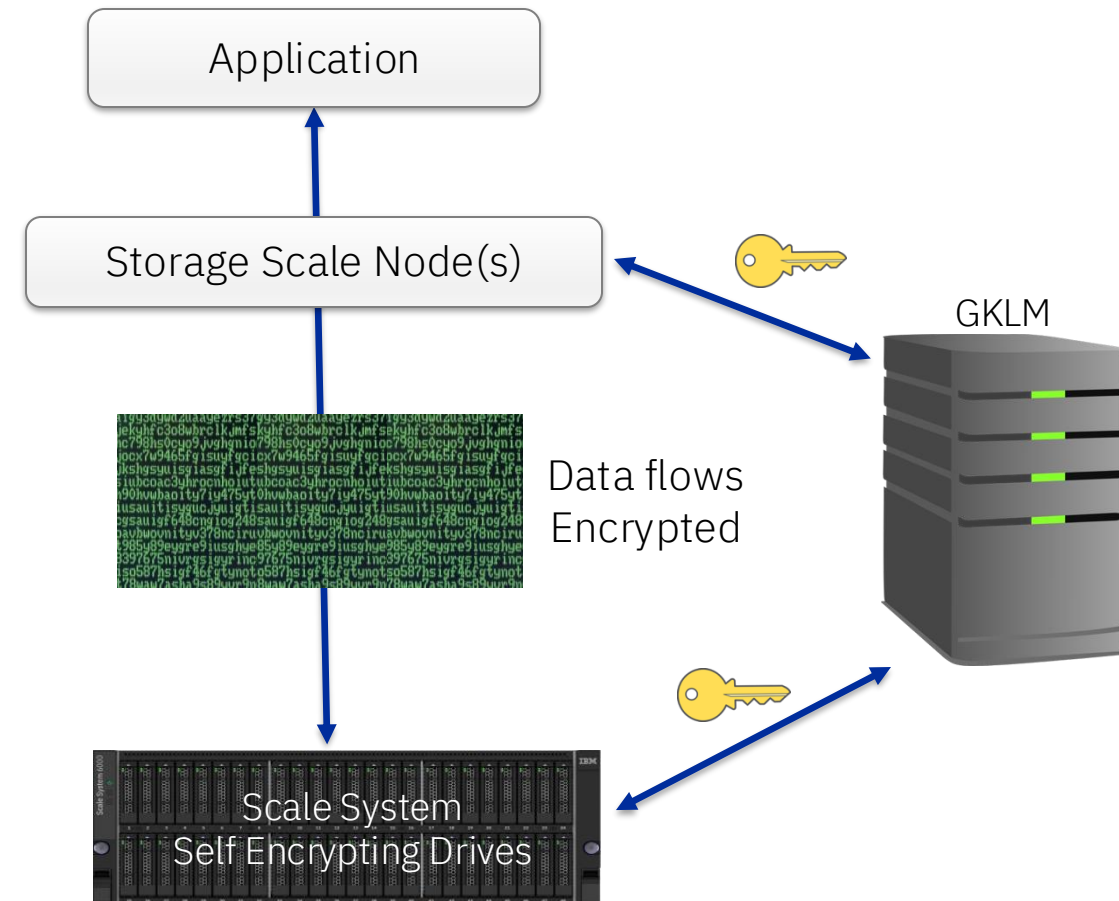
Granularity is per file or per fileset, as determined by encryption policies

Scale encryption can protect against attacks targeting disks
(theft/acquisition of improperly discarded disks)

Secure data deletion using cryptographic erasure

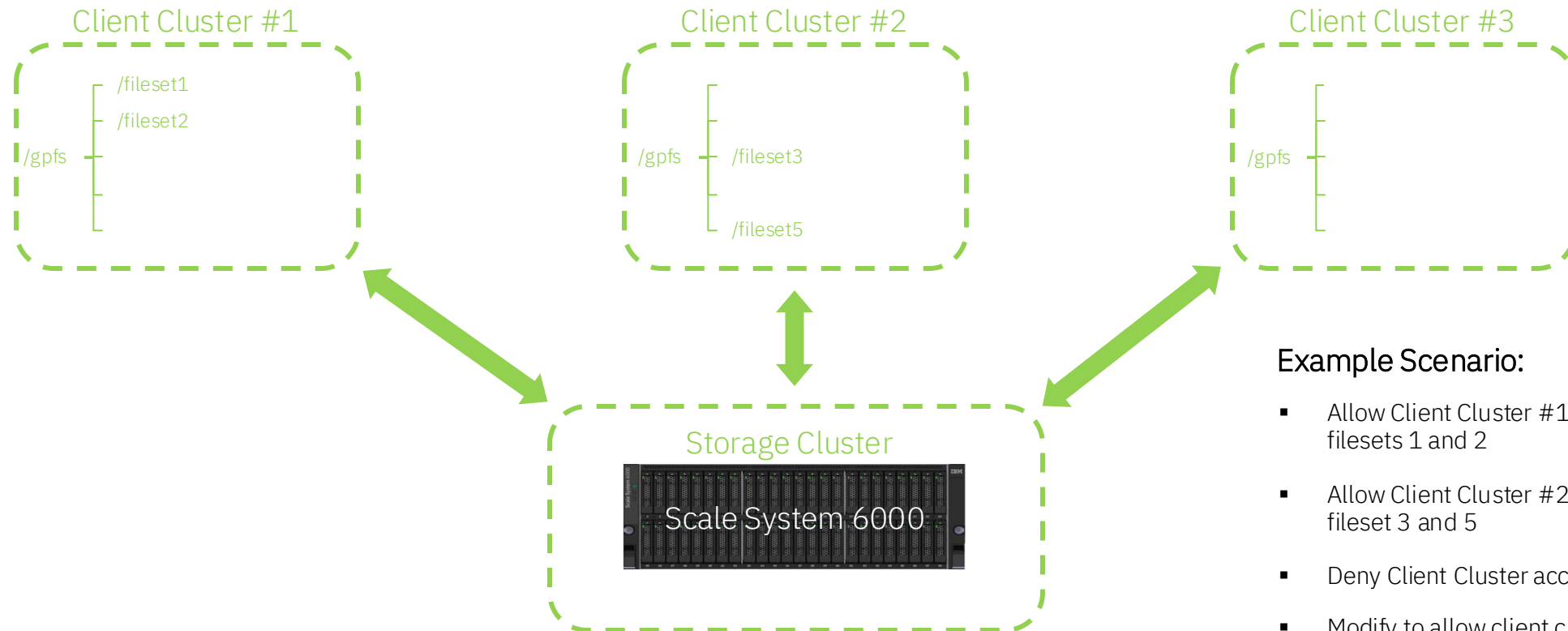
IBM Storage Scale System supports Self Encrypting Drives (SED) for data at rest protection

Investment in Quantum Safe algorithm support



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:

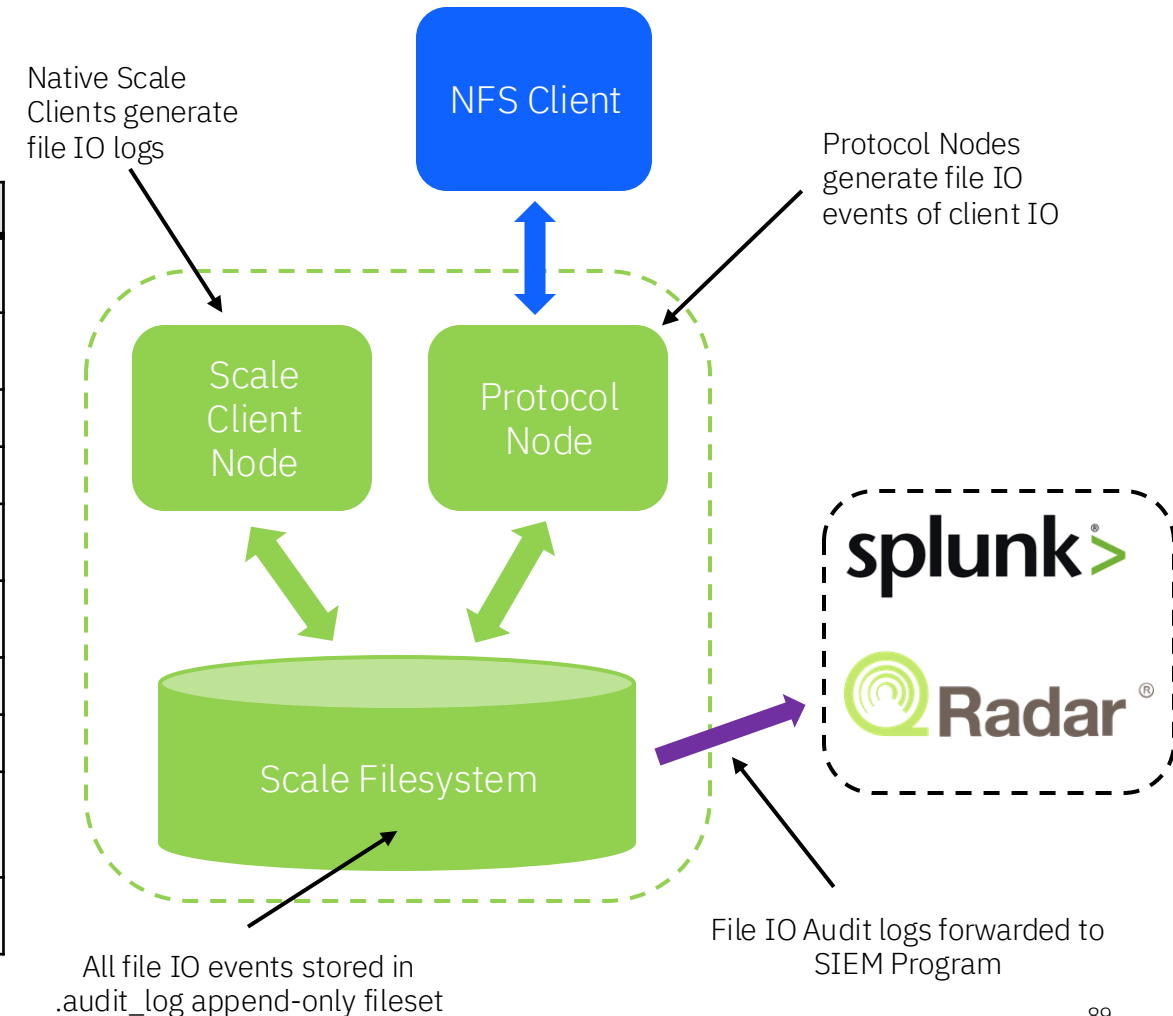
- Allow Client Cluster #1 to only see filesets 1 and 2
- Allow Client Cluster #2 to only see fileset 3 and 5
- Deny Client Cluster access to any filesets
- Modify to allow client cluster #3 to see fileset 2 and 4

File Audit Logging

- Lightweight File IO Event logs stored in JSON format
- All filesystem IO events captured from root, users, Protocols, etc...
- Audit logs stored in an append-Only fileset
- Events forwarded to SIEM program such as IBM Qradar or Splunk for analysis of known access patterns
- Fully-configurable based on needed file events

Video on how to forward to Splunk:
<https://www.youtube.com/watch?v=FGVsYysk1Q>

Event Name	Description	Examples
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



Command Audit Logging



- Log all Storage Scale Administrative Commands from GUI, CLI, and REST API
- Easy audit trail for tracking Storage Scale cluster changes
- GUI users can be authenticated to Active Directory or LDAP groups with different security roles

Command Audit Log

Actions Refresh Export

Command	Arguments	System User	GUI User	Access	Executi...	Start Time	End Time
mmnfs	export add /scale/swatfs1/software/ISO -...	root	admin	GUI	✓ Success	2023-05-31 16:39:13	2023-05-31 16:39:14
mmuserauth	service create --data-access-method file -...	root	admin	GUI	✓ Success	2023-05-31 16:30:44	2023-05-31 16:30:45
mmperfmon	config add --sensors /var/lib/mnfs/gui/tm...	root	admin	GUI	✓ Success	2023-05-31 16:28:59	2023-05-31 16:29:00
mmperfmon	config add --sensors /var/lib/mnfs/gui/tm...	root	admin	GUI	✓ Success	2023-05-31 16:28:36	2023-05-31 16:28:37
mmperfmon	config add --sensors /var/lib/mnfs/gui/tm...	root	admin	GUI	✓ Success	2023-05-31 16:28:18	2023-05-31 16:28:19
mmlinkfileset	swatfs1 ISO -J /scale/swatfs1/software/ISO	root	admin	GUI	✓ Success	2023-05-31 16:01:48	2023-05-31 16:01:49
mmcrfileset	swatfs1 ISO -t Filset for ISOs --inode-spac...	root	admin	GUI	✓ Success	2023-05-31 16:01:47	2023-05-31 16:01:48
mmlinkfileset	swatfs1 software -J /scale/swatfs1/software	root	admin	GUI	✓ Success	2023-05-31 16:01:09	2023-05-31 16:01:10
mmcrfileset	swatfs1 software -t Fileset for storing soft...	root	admin	GUI	✓ Success	2023-05-31 16:01:08	2023-05-31 16:01:09
mmmount	swatfs1 -N all	root	admin	GUI	✓ Success	2023-05-31 15:45:50	2023-05-31 15:45:51
mmmount	cesroot -N all	root	admin	GUI	✓ Success	2023-05-31 15:07:27	2023-05-31 15:07:28
mmnfs	config change IDMAPD_DOMAIN=SWAT.P...	root		CLI	✓ Success	2024-02-28 13:37:22	2024-02-28 13:37:23
mmnfs	config change DOMAINNAME=SWAT.PBM....	root		CLI	✓ Success	2024-02-28 13:36:50	2024-02-28 13:36:51
mmnfs	config change DOMAIN=SWAT.PBM.IHOST...	root		CLI	✗ Failure	2024-02-28 13:36:36	2024-02-28 13:36:37
mmnfs	config change LOCAL_REALMS=SWAT.PB...	root		CLI	✓ Success	2024-02-28 13:36:17	2024-02-28 13:36:18
mmces	service start NFS -N swatnas.pbm.ihost.co...	root		CLI	✓ Success	2024-02-27 16:14:15	2024-02-27 16:14:16

GUI User Groups

+ Create Group Actions

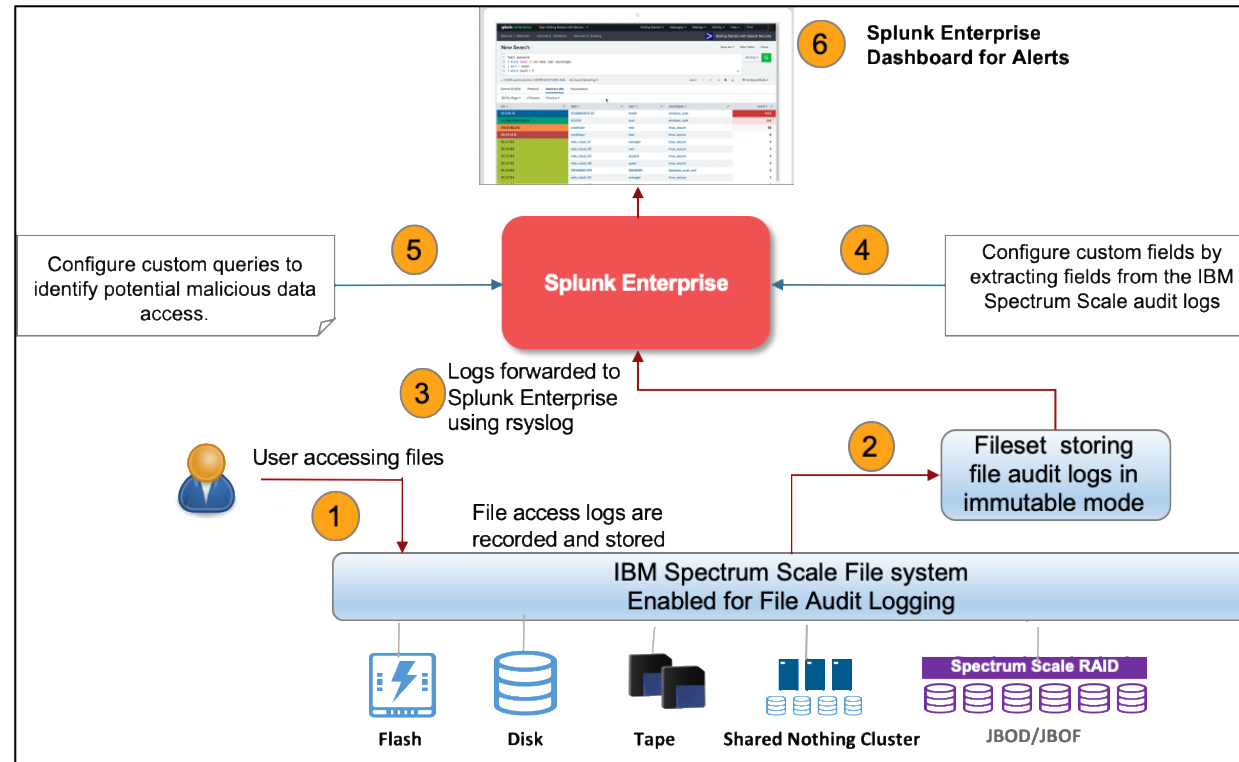
Name	Role	MFA	User Groups
+ Administrator (1)	Administrator		
+ SecurityAdmin (1)	Security Administrator		
+ StorageAdmin (1)	Storage Administrator		
+ SystemAdmin (1)	System Administrator		
+ Monitor (1)	Monitor		
+ SnapAdmin (1)	Snapshot Administrator		
+ DataAccess (1)	Data Access		
+ ProtocolAdmin (1)	Protocol Administrator		
+ UserAdmin (1)	User Administrator		
+ CsiAdmin (1)	CSI Administrator		
+ ContainerOperator (1)	Container Operator		
Scale GUI Admin	Administrator		
Scale GUI Monitor	Monitor		



These two groups linked to Active Directory

Splunk Enterprise (SIEM) Analysis

Enhanced Threat Detection



Solution Brief:

<https://www.ibm.com/downloads/cas/1OLV7L3Z>

Video on how to forward to Splunk:

<https://www.youtube.com/watch?v=FGVsYcck1Q>

Solution:

- Integration of IBM Storage Scale with Splunk Enterprise and Splunk Enterprise for Security
- Ability to send StorageScale File Audit Logs to Splunk Enterprise.
- Easy to configure Custom Queries using Storage Scale fields.
- Rules are configured and alerts are generated and displayed on Splunk dashboards

Modernization of Scale (MOS): Security

Security Improvements

Removal of SSH dependency



Removal of root requirement for control plane

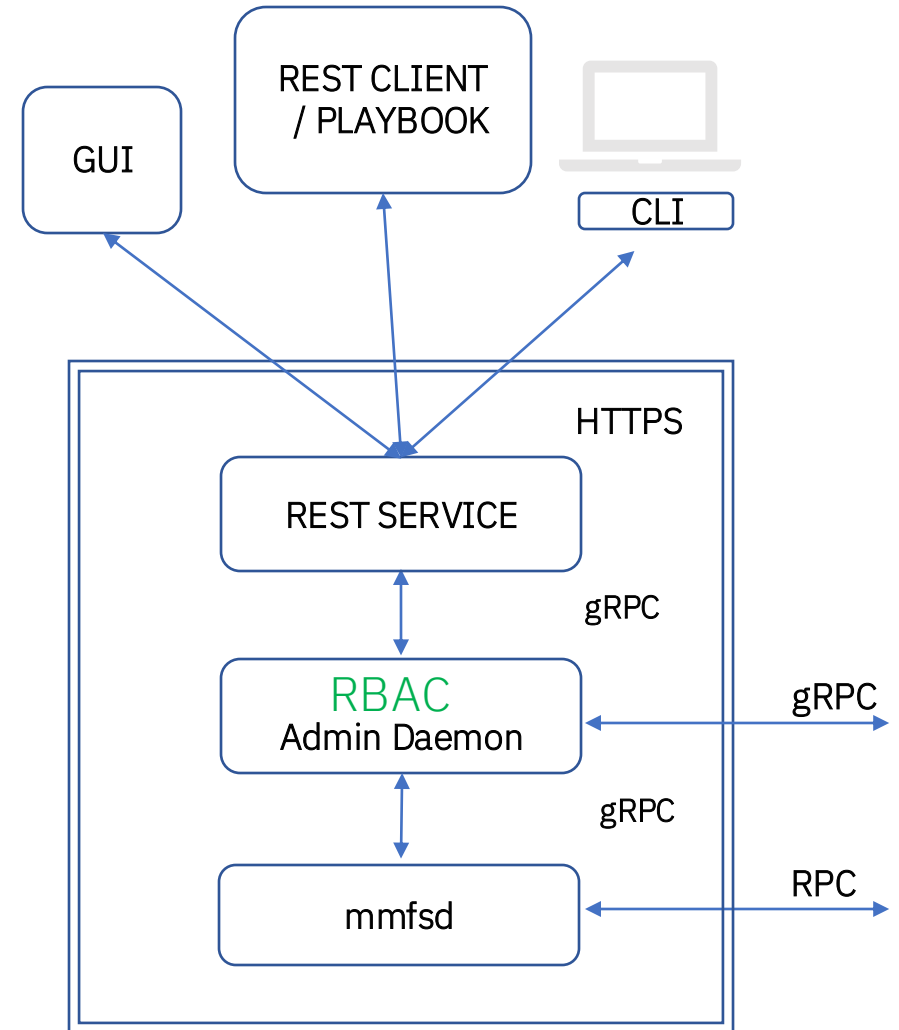
Remote Administration

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

Tech Preview



Scale for AI Workloads

NVIDIA and IBM are ready to go!

©CBS NEWS

60 MINUTES

https://ibm.biz/60_Minutes



Accelerate AI with NVIDIA and IBM Storage

The IBM Storage Scale System 6000 is the fastest integration point for NVIDIA DGX to address the challenges of AI data optimization.



Date: July 15, 2024

DGX SuperPOD and IBM Storage

IBM Storage Scale System 6000 All-flash systems with GPFS is an approved solution for DGX SuperPOD featuring DGX's with A100, H100, H200 and B200 GPU's.

Please let me know if you have any other questions.



Changing technology

GPUs used for AI are driving the need for larger data sets and faster data delivery



Data silos

Data is scattered and siloed throughout an organization making it difficult to gain access to relevant data for AI



Unknown threats

The veracity and accuracy of data are critical to AI and it must be protected from data breaches – accidental or otherwise



Costs

More data and faster delivery can mean new technologies and infrastructure that can strain budgets and sustainability strategies

IBM

Accelerate AI and data delivery

GPU Direct Storage with embedded AI accelerator delivering 310 GB/s and 13M IOPS

Eliminate data silos

Globally connect relevant data without data movement from across the organization (on-premises and off)

Data and cyber resilient storage

Six 9s of availability with globally dispersed erasure coding for always on and immutable data protection against accidents and threats

Meet sustainability goals and lower costs

Greater storage density on all-flash media with computational drives to offload CPU-intensive services across storage tiers

IBM Storage Scale System 6000

Storage for Data and AI

Customer Challenges



Escalating Storage and GPU costs

Storage Solution:

- ✓ Minimizes cost of training AI models
- ✓ Integrates existing dispersed storage silos
- ✓ Manages data in different cost optimized storage tiers



AI Data is Distributed

Storage Solution:

- ✓ Effectively integrates unstructured data
- ✓ Delivered when and where it is needed
- ✓ Transparent to AI workload



Data is a Critical Business Asset

Storage Solution:

- ✓ Resilient to cyber threats
- ✓ Quick recovery time
- ✓ Highly available

Optimizing the platform and infrastructure for AI/ML models

...across the whole AI workflow

Data preparation



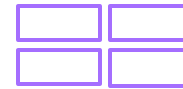
Workflow of steps
(e.g., remove hate and
profanity, deduplicate)

Distributed training



Long-running job on
massive infrastructure

Model adaptation



Model tuning with
custom data set for
downstream tasks

Inference



May have sensitivity to
latency, throughput,
power,

Hours to days

10-2000+ low to mid-end CPU cores
10+ low to mid-end GPUs per
10-100+ concurrent jobs



on-prem **Public clouds**

weeks to months

10-500+ high-end GPUs (per job)
10+ concurrent jobs



on-prem **Public clouds**

minutes to hours

1+ mid to high-end GPU (per job)
100+ concurrent jobs



on-prem **Public clouds**

sub-second API request

Single low-end GPU per fine tuning task
Fraction to multiple GPUs per inference,
or specialized accelerator
Thousands of API requests



on-prem **Public clouds** **Edge**

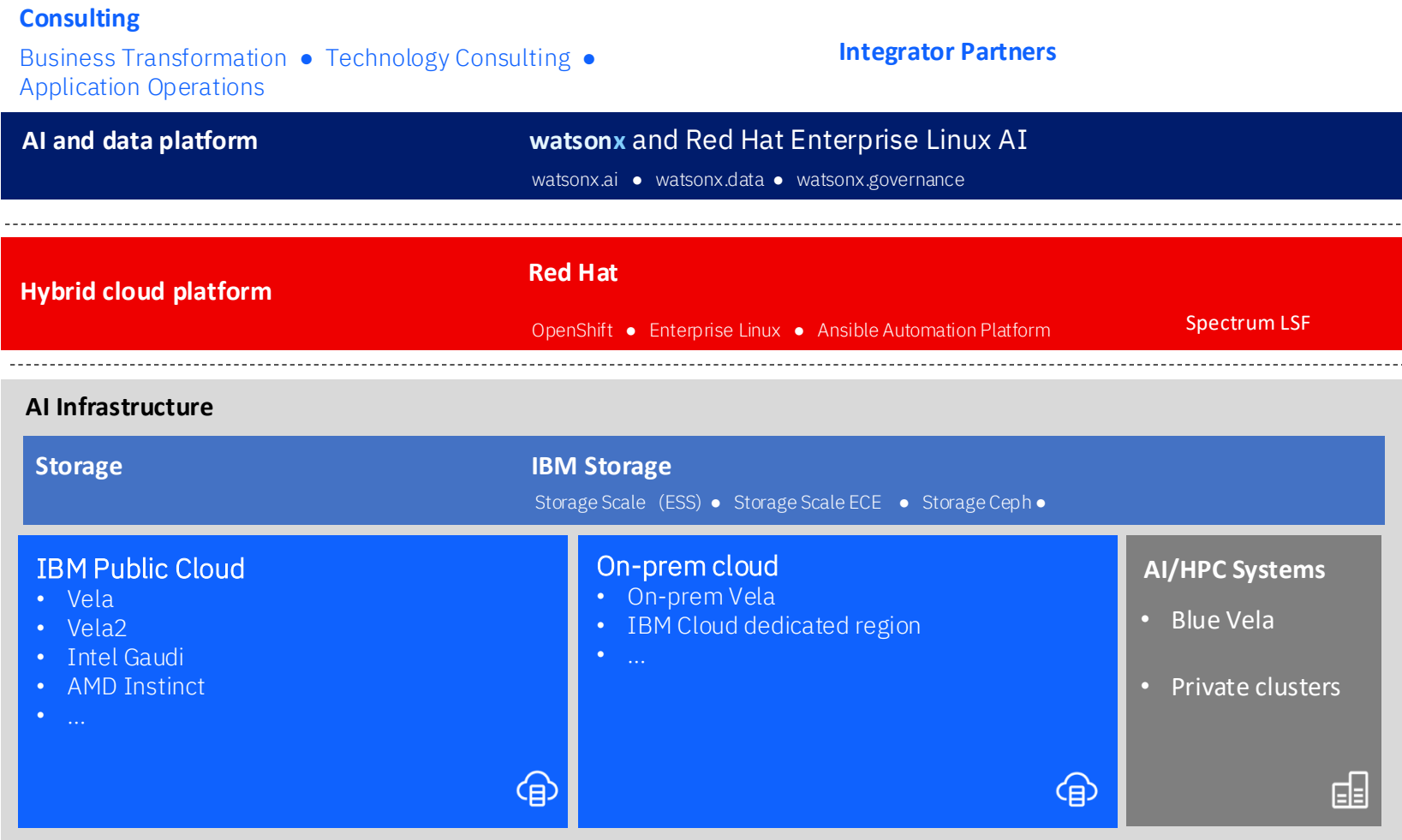
IBM's PoV on hybrid cloud and AI Infrastructure



Our AI infrastructure stack spans between public cloud, private cloud, and on-prem HPC systems to meet the distinct needs of current and future AI workloads.

Common storage and platform layers provide the foundation for AI and data platforms.

This hybrid AI infrastructure and platform enables business to scale, optimize, and deliver AI solutions.





<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.



“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.”

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



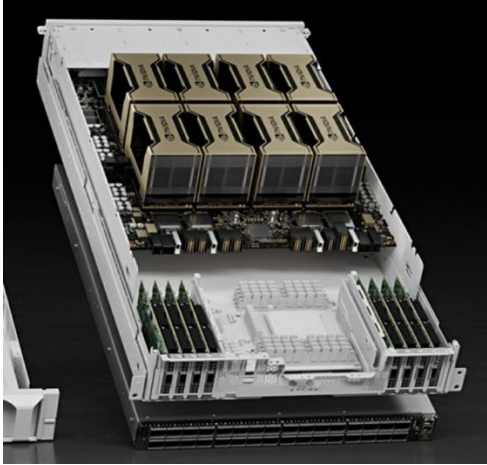
NEWS PROVIDED BY
CoreWeave →
Jan 15, 2025, 08:00 ET

<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



IBM Scale System is Nvidia Certified Storage!



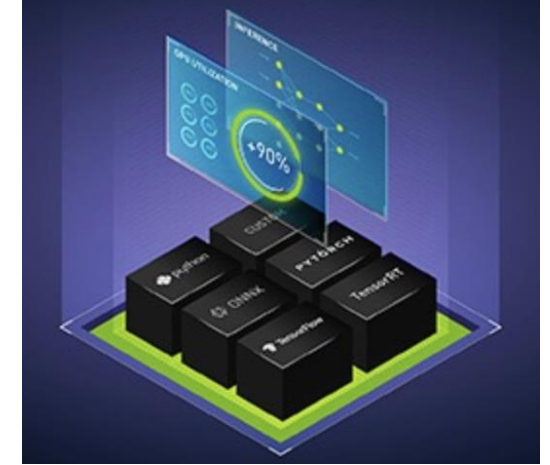
Nvidia
GPUDirect
Storage (GDS)
certified



Nvidia
BasePOD
certified



Nvidia
SuperPOD
certified



Nvidia
Cloud Partner
(NCP) certified

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 for Data & AI Solutions with NVIDIA

IBM Storage & NVIDIA Collaborations

- DGXH100 SuperPOD RA : 2023
- **1st SuperPOD BCM installation: 2022**
- DGX BasePOD validated storage partner : 2022
- DGX A100 SuperPOD installations: 2021
- DGX A100 SuperPOD RA : 2021
- GPUDirect 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)

- OEM HGX BasePOD
- IBM Storage Fusion System
- IBM Storage Scale S/W
- IBM ESS3500 and SSS6000

NVIDIA Compute

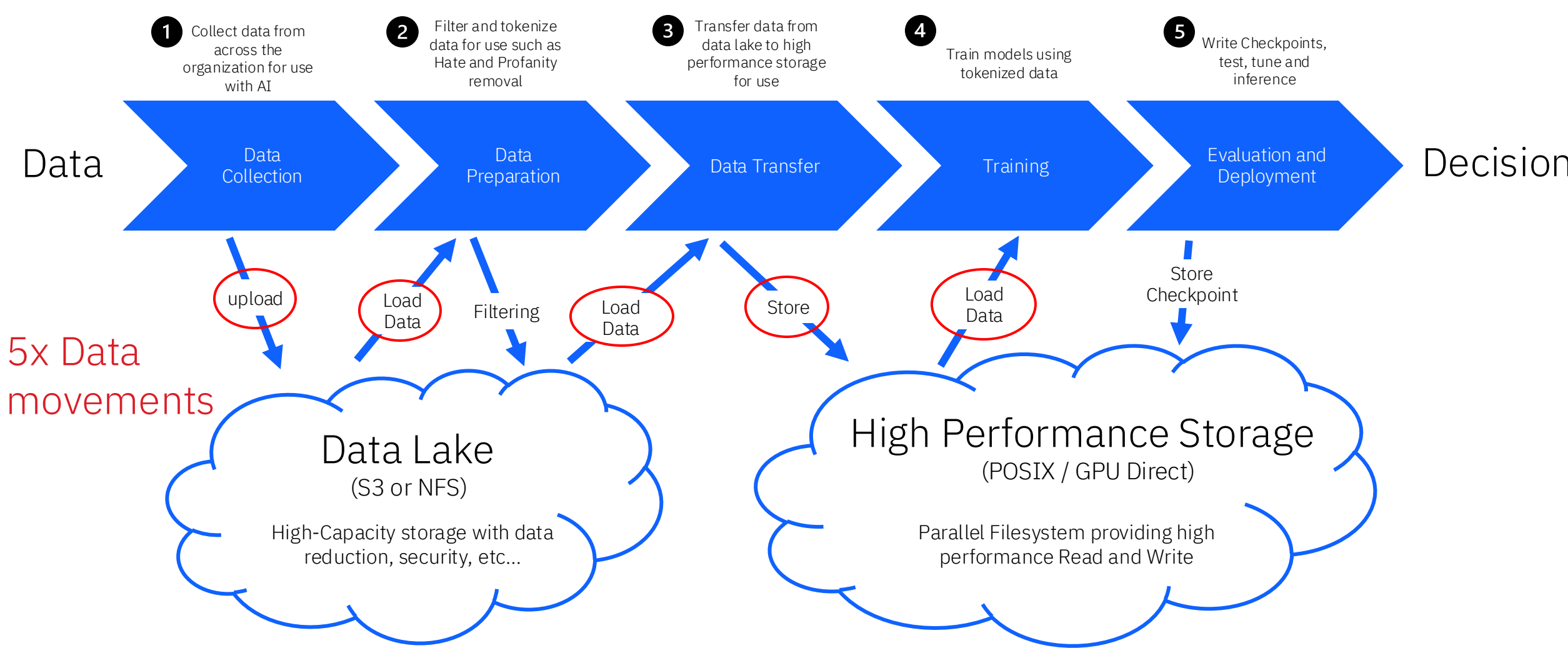
- Every NVIDIA DGX generation
- IBM Storage Sale systems (ESS)
- BasePOD, OVX, and SuperPOD Reference Architectures
- GraceHopper support

Recent Installations

- Enterprise AI Pipeline: Spark to HDFS to AI
- Burst to cloud for GPU sharing
- Julich: Exascale NVIDIA GH200
- IBM Model Factories: Vela / Blue Vela:

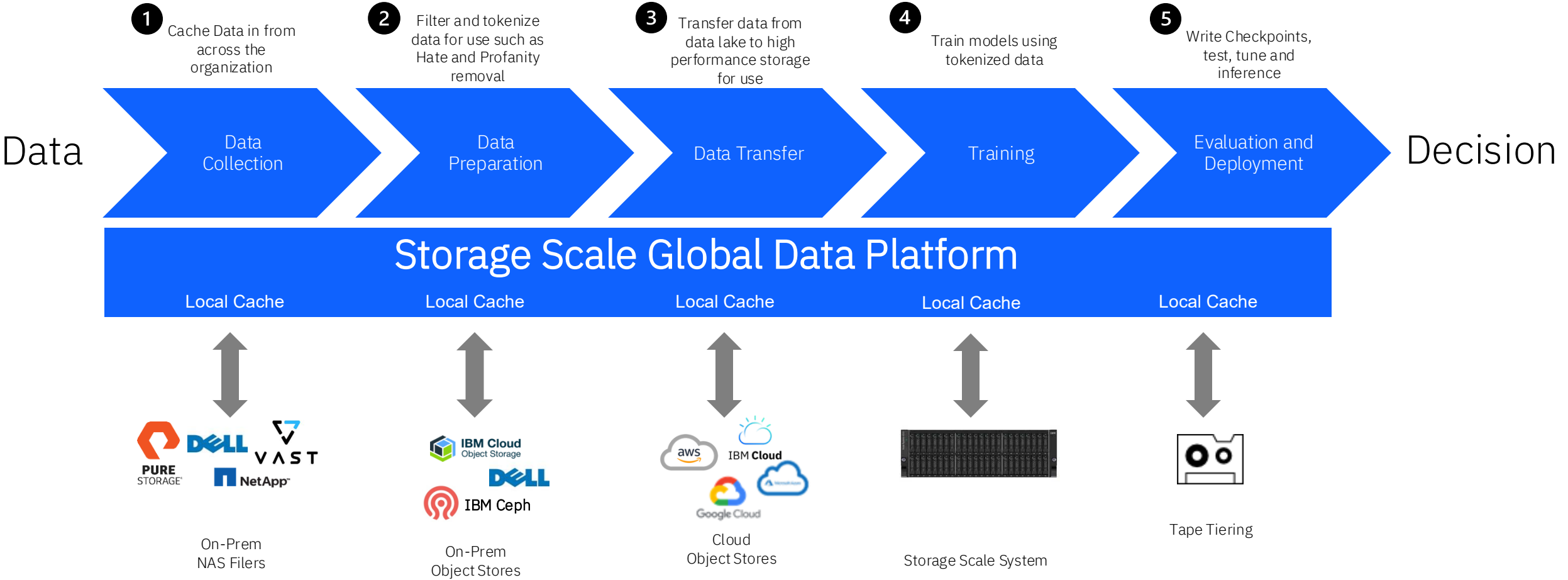
Storage for Data and AI

Phases of AI



Storage for Data and AI

Phases of AI – Scale Optimized



Cache Data as needed - No Copies!

IBM's Purpose-Built Storage Solution for AI

The world's fastest systems need the world's best storage.
IBM has the best storage for NVIDIA GPUs

Highest Performance Platform

- Fastest performance for reads, writes, and density
- Linearly scalability for future growth

A Robust Enterprise Platform

- Six 9's for all apps: AI, Analytics, HPC, Back-up, Archive, Cloud
- Cyber-resilient, encryption, WORM, and immutability

Collapse Layers & Simplify Workflows

- Eliminate extra copies and share data globally with all protocols
- Data cataloging and tiering for economics and data flexibility

High Efficiency and Low TCO

- Minimal power, cooling, and rack space without sacrificing performance
- Leading **GB/s-per-kW** and **GB/s-per-RackU**
- 5-Year TCO lower than leading competitors

IBM Storage Scale System 6000

A single 4U node with active-active controllers and redundant hardware to maximize always on data



Ultimate Performance and Scalability

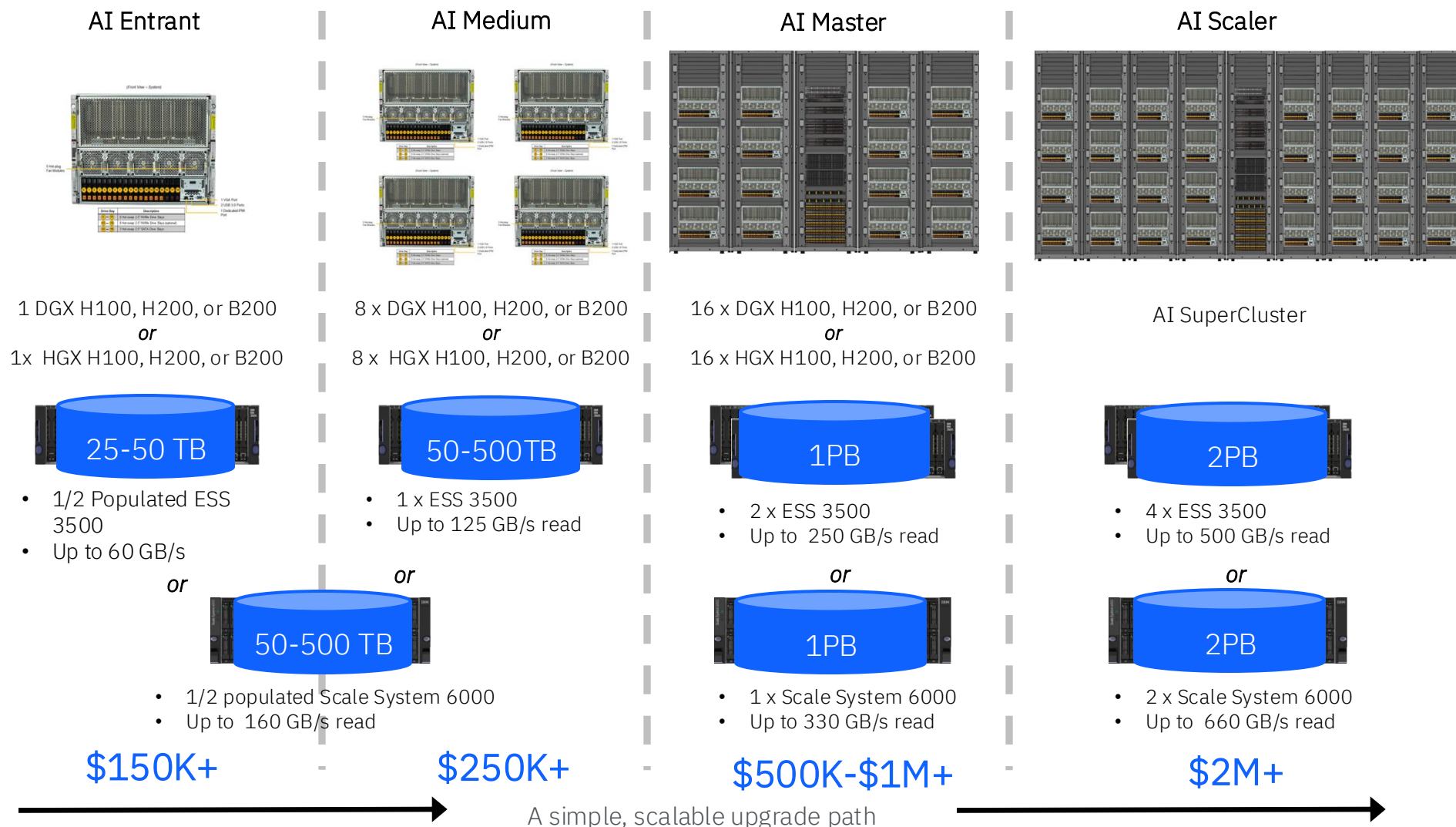
up to 330 GB/s read performance per node

up to 155 GB/s write performance per node



IBM Storage for Data and AI & NVIDIA GPU Solutions

A full spectrum of scalable AI solutions



A simple, scalable upgrade path

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

Promise of IBM Storage

- Simple building blocks** – Scalable seamless storage upgrade path as needs grow from 1st HGX to AI CoE HGX SuperCluster
- Global Data Platform** – Data fidelity capabilities to automate AI workflows
- Data Economics** – Eliminate copies and transparently tier
- Global Deployment** – Trusted and successful global enterprise level support and services

Storage Requirements for AI

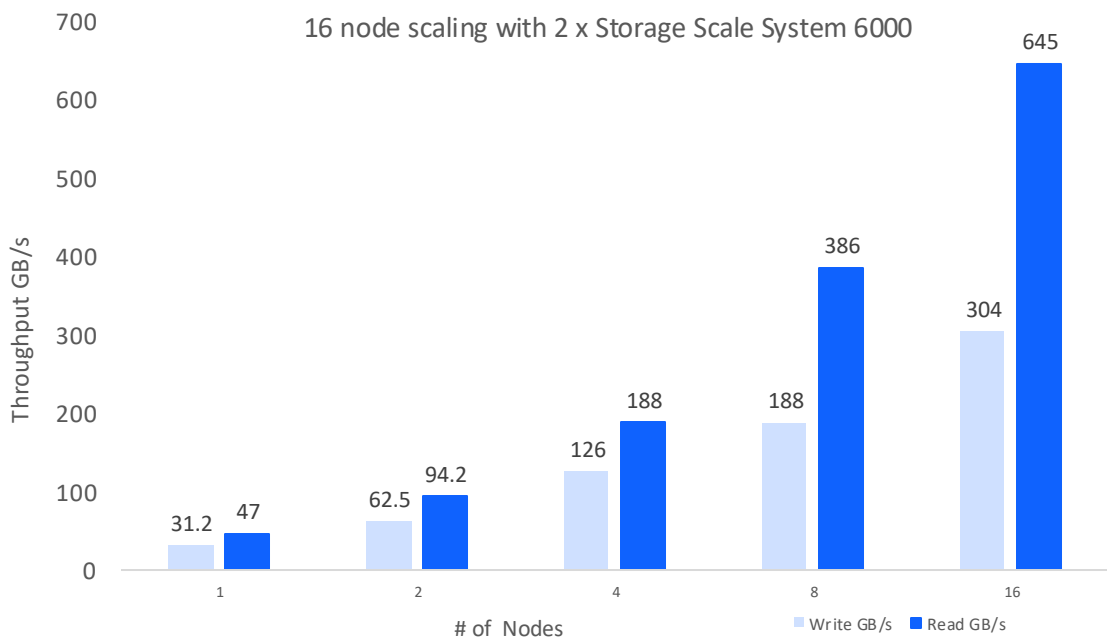
High Performance Parallel Storage

NVIDIA H100 SuperPOD Storage Guidelines

Performance Characteristic1	Good (GBps)	Better (GBps)	Best (GBps)
Single node read	4	8	40
Single node write	2	4	20
Single SU aggregate system read	15	40	125
Single SU aggregate system write	7	20	62
4 SU aggregate system read	60	160	500
4 SU aggregate system write	30	80	250

- Peak performance for writes and read are needed for creating and reading checkpoint files
- Checkpoint is a synchronous process, and training stops during this phase
- High performance, resilient, parallel filesystem storage recommended for multi-threaded read and write operations across multiple nodes
- Checkpoints restart drastically reduces LLM training time with Storage scale

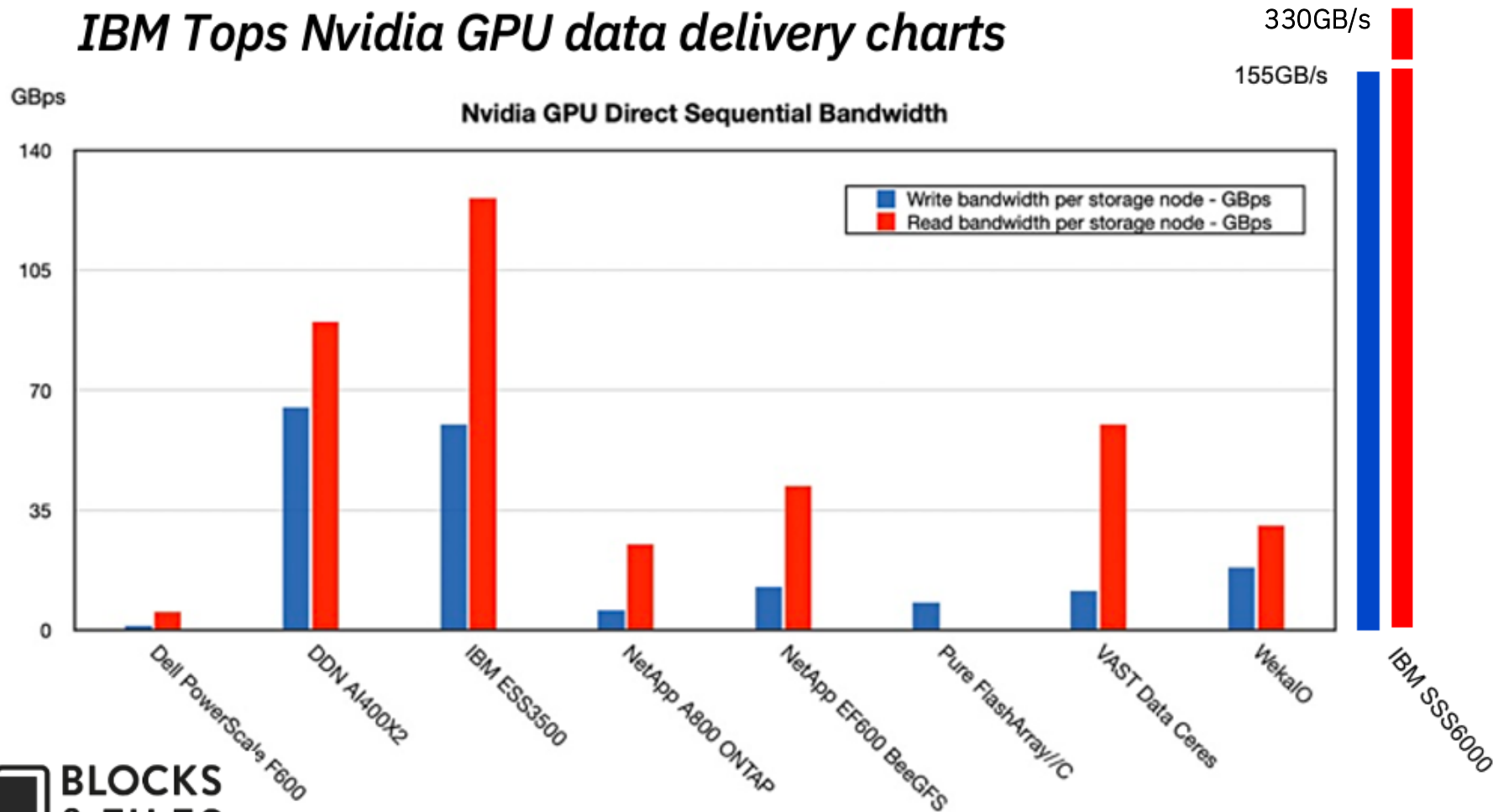
IBM Storage Scale System 6000 delivers “Best”



- ✓ Fastest performance for reads, writes, and density
- ✓ Leading GB/s-per-kW and GB/s-per-RackU
- ✓ 9x NDR network ports to support DGX SuperPod
- ✓ 5-Year TCO lower than leading competitors

IBM Storage Scale System 6000 sets new marks for performance

IBM Tops Nvidia GPU data delivery charts



- ✓ IBM 6000 is more than 2x more performant than the current 3500
- ✓ Read: 330 GB/s
- ✓ Write: 155 GB/s
- ✓ Latest in networking
- ✓ Ready to support GB200 platforms



BLOCKS
& FILES.

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

IBM Storage Scale

Scale System AI Training and Checkpoint Comparison

Nvidia DGX H100 SuperPod 4x Iteration Training Time



*Assumes Nvidia H100 SuperPod with 256x 80GB GPUs with 20TB total GPU memory + 30min epoch runtime per training cycle

**Performance based on [io500](https://io500.benchmark.com) benchmarks and <https://blocksandfiles.com/2023/08/15/ibm-nvidia-gpu-data-delivery/>

IBM Storage for Data and AI / © 2025 IBM Corporation

LLM Data Set Size

Checkpoint time reduced to 1% an hour => 36 seconds

Model Specific Example for Synchronous

Checkpoint

Tensor Model Parallel Size determines how many GPUs participate in the checkpoint.

For example, If Tensor Parallel size is set to 8, 1 out of 8 GPUs will participate in the checkpoint

~14 bytes per Parameter

Example

175B Parameter Model: ~2.4TB Data set size

512B Parameter Model: ~7.2TB Data set size

1T Parameter Model: ~14TB Data set Size

3 x IBM Storage Scale 6000 is 19.4 TB in 36 seconds

Total Number of GPUs is 4000 GPUs

Only 512 GPUs will participate in the checkpoint
(4000_GPUs / 8_Tensor Parallel_Size)

175B: ~4.8GB data set / GPU

512B: ~15GB data set / GPU

1T: ~28GB data set / GPU



Model Load

Using the same Tensor Parallel Size of 8 from the checkpoint

8x the data set size will need to be loaded across all GPUs

Example

175B Parameter Model: ~19TB Data set size

512B Parameter Model: ~58TB Data set size

1T Parameter Model: ~110TB Data set Size

3 x IBM Storage Scale 6000 load in < 2 min

Total Number of GPUs is 4000 GPUs

Data set per GPU is the same, but now all GPUs participate in the Model Load

175B: ~4.8GB data set / GPU

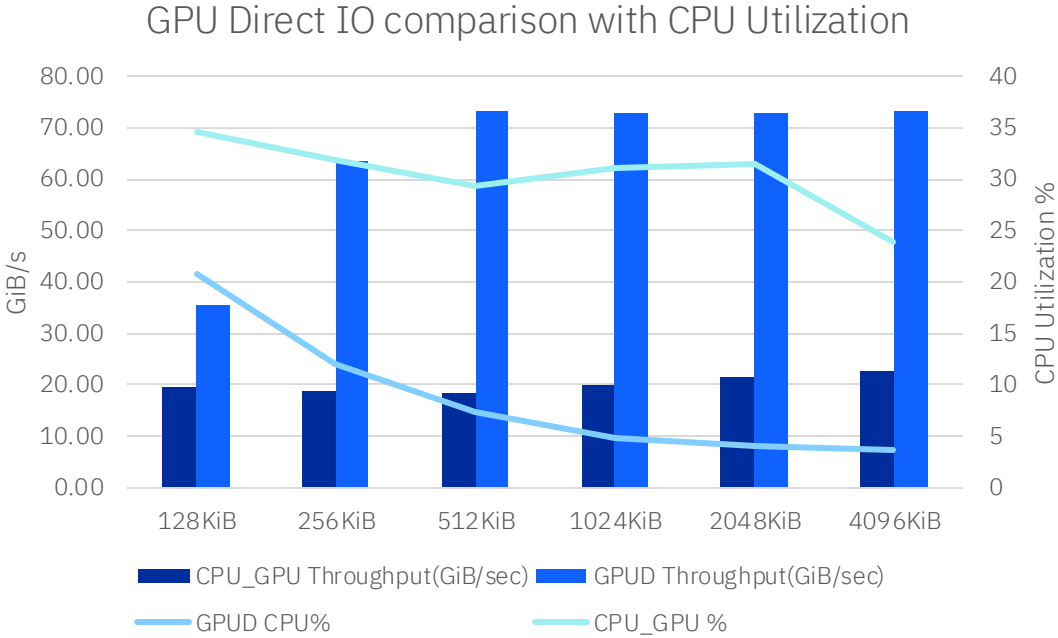
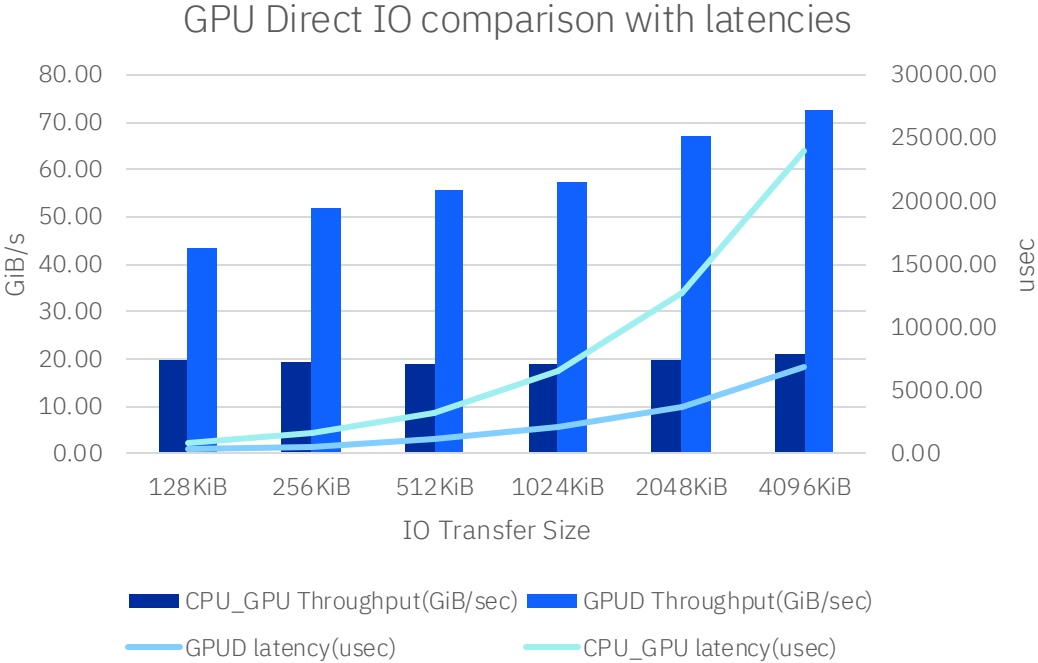
512B: ~15GB data set / GPU

1T: ~28GB data set / GPU

- ✓ READ: 330 GB/s
- ✓ WRITE: 155 GB/s

IBM Storage Scale

GPU Direct Storage Comparison



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

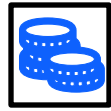
Up to 3.5x Higher Bandwidth; 50% reduction in latencies

Content Aware Storage (CAST)

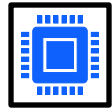
AI's Hunger for Data will Exacerbate the Current Challenges of Unstructured Data



Cost-
Capac
ity



Cost-
Perfor
manc
e



Chang
e
efficie
ncy



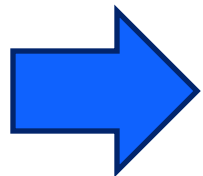
Time
to
acces
s



Data
gover
nance
(secur



Robus
tness
(availa
bility,



Current storage systems

because they have

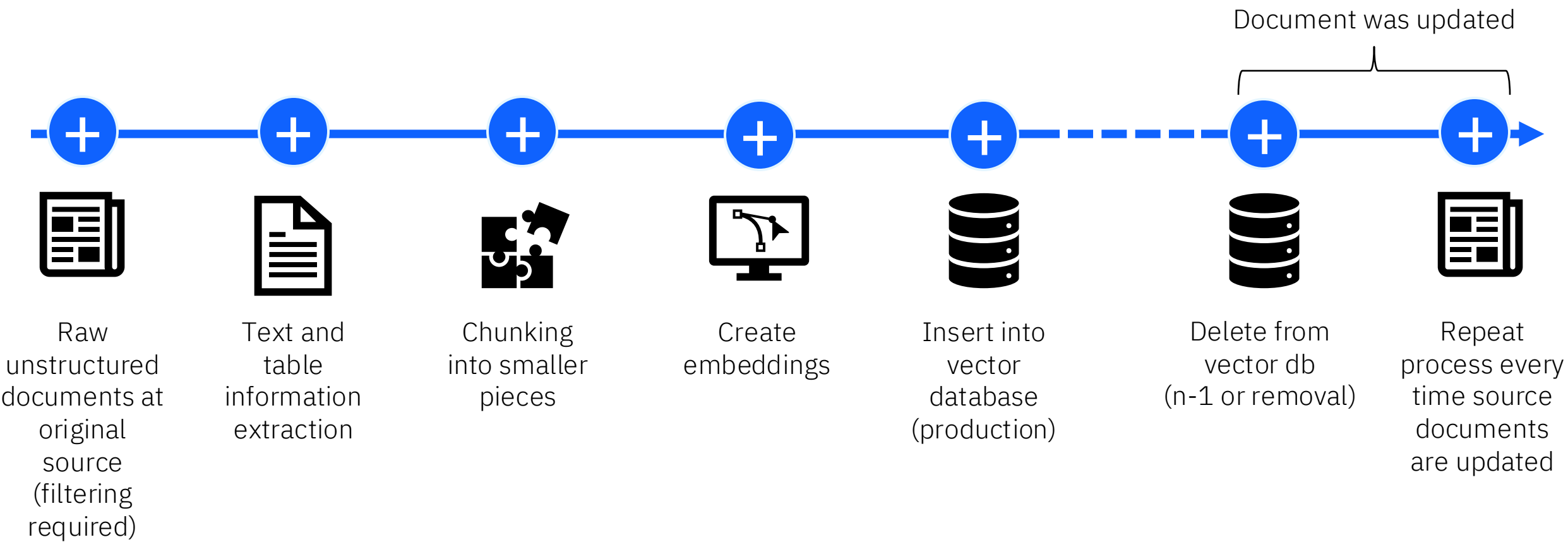
no understanding of content

can not address these challenges

, e, y, servic
rebalanc search compl eabilit

High-level RAG Pipeline

Customer steps for data preparation
for new or updating documents



Why is it hard?

Scanned documents

Invoice account: 501-C02567
Terms of payment: End of Month + 60 days
Due date: 31/01/2018
Method of payment: 501-PS047141 from 28/11/2017
Delivery: 501-PS047141 from 28/11/2017

Order account: 501-S00479
Order number: 501-S00479
Customer: ARROW FINANCE HQ
VAT Reg No: NA-GB-230

Invoice address: Gulf Business Machines Abu Dh
W.L.L. P.O. Box 37543 Abu Dhabi United Arab Emirates

Delivery address: Gulf Business Machines Abu W.L.L. P.O. Box 37543 Abu Dhabi United Arab Emirates

Invoice No: 501-SPI045340 Tax Point Date: 28/11/2017

Item number	Description	Quantity	List price	Discount %
ARW_INT_BM_SFS3	IBM FlashSystem V9000 Storage Implementation	1.00	2,200.00	

Sales tax code: Serv tax/Free item, 3rd city Amount origin: 2,200.00 VAT amount: 0.00

EXPORT SERVICES

Handwritten notes: GIBP/11/320.40, 20/800/800, S. Arin

Noise artifacts

Variety of tables

Fig.7: Sensitivity to price and volume movements

	Net Profit	
For each +/-1% chg in coal price	FY14F	FY15F
Indo Tembung	-0.40%	-1.89%
Adaro	-0.50%	-1.43%
Haum	-1.00%	-1.83%

	Net Profit	
For each +/-1% chg in volume	FY14F	FY15F
Indo Tembung	-1.00%	-1.10%
Adaro	-1.10%	-1.10%
Haum	-1.20%	-1.20%

	Previously
- FCCR (limitation on indebtedness)	3x
- Debt/EBITDA	4x
- Bekasi Power debt basket	5mn
- General debt basket	10mn
- Permitted investment basket	-
- Interest reserve account establishment	Yes

Table with visual clues only

Table with graphic lines

Multi-row, multi-column headers

	Three months ended September 30			N
(In thousands)	2015	2014	Change	
Salaries and benefits	171,506	153,123	12.1%	
Employee share purchase plan	24,101	19,879	21.0%	
Employee profit share	32,974	19,039	73.2%	
Share-based payment plans	4,123	4,348	(5.2%)	
Total	230,786	196,499	17.4%	

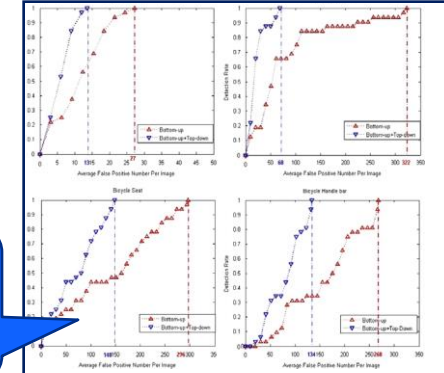
Presentation on the Condensed Consolidated Statement of Earnings:

	2015	2014	Change
Airport operations	30,916	28,145	9.8%
Tighten conditions and average pool changes	67,546	59,251	13.8%
Sales and distribution	10,496	17,312	11.0%
Marketing, general and administration	27,063	22,818	18.6%
Infra	36,116	33,828	6.7%
Maintenance	15,737	15,256	3.2%
Employee profit share	32,974	19,039	73.2%
Total	230,786	196,499	17.4%

Nested row headers

Tables with Textual content

Various elements



Line Charts
Histogram
Pie Charts

Fuzzy Text & Skew

ORACLE

SERVICES AGREEMENT

The Services Agreement (the "Agreement") is between Oracle Corporation Redwood City, California 94065 ("Oracle") and IBM Corporation, Kingston, NY 12401 ("IBM").

I. Services

Oracle will provide to Client, in the United States, the Services specified on a Work Order, under the terms of this Agreement.

II. Definitions

2.1. "Work Order" shall mean Oracle's standard form for ordering Services (entitled "Work Order" or "Order Form") and shall specify the Services and applicable fees. Each Work Order shall be governed by the terms of this Agreement and shall reference the Effective Date specified below.

2.2. "Services" shall mean work performed by Oracle for Client pursuant to a Work Order, agreed to by the parties, under this Agreement. The schedule for Services will be agreed upon by the parties, subject to availability of Oracle personnel.

III. Charges, Payment, and Taxes

3.1. Fees for Services

Unless otherwise expressly specified in the applicable Work Order, Services shall be provided on a time and material ("T&M") basis at Oracle's T&M rates current when the Services are performed. If a dollar limit is stated in the applicable Work Order for T&M Services, the limit shall be deemed an estimate for Client's budgeting and Oracle's resource scheduling purposes; after the limit is expended, Oracle will continue to provide the Services on a T&M basis, if a Work Order for continuation of the Services is signed by the parties.

3.2. Incidental Expenses

Client shall reimburse Oracle for reasonable travel, communications, and out-of-pocket expenses incurred in conjunction with the Services.

Per capita poultry consumption

Country	Chicken consumption (kg/capita/year)	GDP
Malaysia	37.3	
Singapore	36.2	
Thailand	12.5	
China	9.2	
Philippines	8.4	
Vietnam	7.2	
Indonesia	6.1	
India	2.3	

Colored background

Signatures

Logos

Signature of [Name] on [Date]

Signature of [Name] on [Date]

Logos of [Company 1] and [Company 2]

Parsing example

S3 Deep Archive Storage on Diamondback



Reduce the Cost of Archiving Data, Without Impacting Data Access

IBM S3 Deep Archive on IBM Diamondback provides S3 Glacier storage class access at up to 85% savings compared to AWS S3 Glacier hosted storage.

■ Highlights

Interoperable with all S3 Glacier Flexible Retrieval storage class supporting applications.

No tape skills required.

Flexible Retrieval performance at a fraction the cost of S3 Deep Archive storage classes.

The rising costs of storing data continues to be a challenge. "many data center managers will be forced to use tape..." according to Furthur Market Research, "as ultra-low-cost, sustainable storage alternatives."¹ Some data center managers resist tape as difficult to deploy and operate due to the need for specific tape software and skills. IBM S3 Deep Archive flips the rhetoric with simple deployment and a standardized interoperable interface, without sacrificing the ultra-low-cost storage position.

IBM S3 Deep Archive brings S3 Glacier Deep Archive Flexible retrieval classes on-premises at a fraction of the cost of even AWS S3 Glacier Deep Archive. Approximately 80% of organizations have S3 skills.² IBM S3 Deep Archive is delivered configured and ready for deployment as an S3 target for all deep archive data.

- S3 low-cost, secure, durable availability zone
- No supporting application modifications
- Subscription like capacity, without the monthly charges
- Zero egress fees

Table 1. Total Acquisition Cost Compare

	IBM S3 Deep Archive	
	9PB	27PB
Total Acquisition Cost per TB	\$22.56	\$11.89
5-year \$/TB year	\$4.51	\$2.37
AWS S3 Glacier Flexible Retrieval \$/TB year	\$43.20	
Savings compared to AWS	89%	94%

Unstructured text

S3 Deep Archive Storage on Diamondback

Reduce the Cost of

Archiving Data, Without

Impacting Data Access

T56000

IBM S3 Deep Archive on IBM Diamondback

provides S3 Glacier storage class access at up

to 85% savings compared to AWS S3 Glacier

hosted storage.

Highlights

Interoperable with all S3 Glacier Flexible Retrieval

storage class supporting applications.

No tape skills required.

Flexible Retrieval performance at a fraction the cost of S3 Deep Archive storage classes.

The rising costs of storing data continues to be a challenge. "many data center managers will be forced to use tape..." according to Furthur Market Research, "as ultra-low-cost, sustainable storage alternatives."¹ Some

data center managers resist tape as difficult to deploy and operate due to the need for specific tape software and skills. IBM S3 Deep Archive flips the rhetoric with simple deployment and a standardized interoperable interface, without sacrificing the ultra-low-cost storage position.

IBM S3 Deep Archive brings S3 Glacier Deep Archive Flexible retrieval classes on-premises at a fraction of the cost of even AWS S3 Glacier Deep Archive. Approximately 80% of organizations have S3 skills.² IBM S3 Deep Archive is delivered configured and ready for deployment as an S3 target for all deep archive data.

- S3 low-cost, secure, durable availability zone
- No supporting application modifications
- Subscription like capacity, without the monthly charges
- Zero egress fees

Table 1. Total Acquisition Cost Compare

Table with the following rows:
Total Acquisition Cost per TB, IBM S3 Deep Archive: 9PB = \$22.56, IBM S3 Deep Archive: 27PB = \$11.89
5-year \$/TB year, IBM S3 Deep Archive: 9PB = \$4.51, IBM S3 Deep Archive: 27PB = \$2.37
AWS S3 Glacier Flexible Retrieval \$/TB year, (IBM S3 Deep Archive: 9PB, IBM S3 Deep Archive: 27PB) = \$43.20
Savings compared to AWS, IBM S3 Deep Archive: 9PB = 89%, IBM S3 Deep Archive: 27PB = 94%

Table

IBM AI Content Aware STorage (CAST)

Imagine the scenario:

“Clients store documents in a bucket....

Minutes later, their chatbots can answer questions with this new information just being imported...”

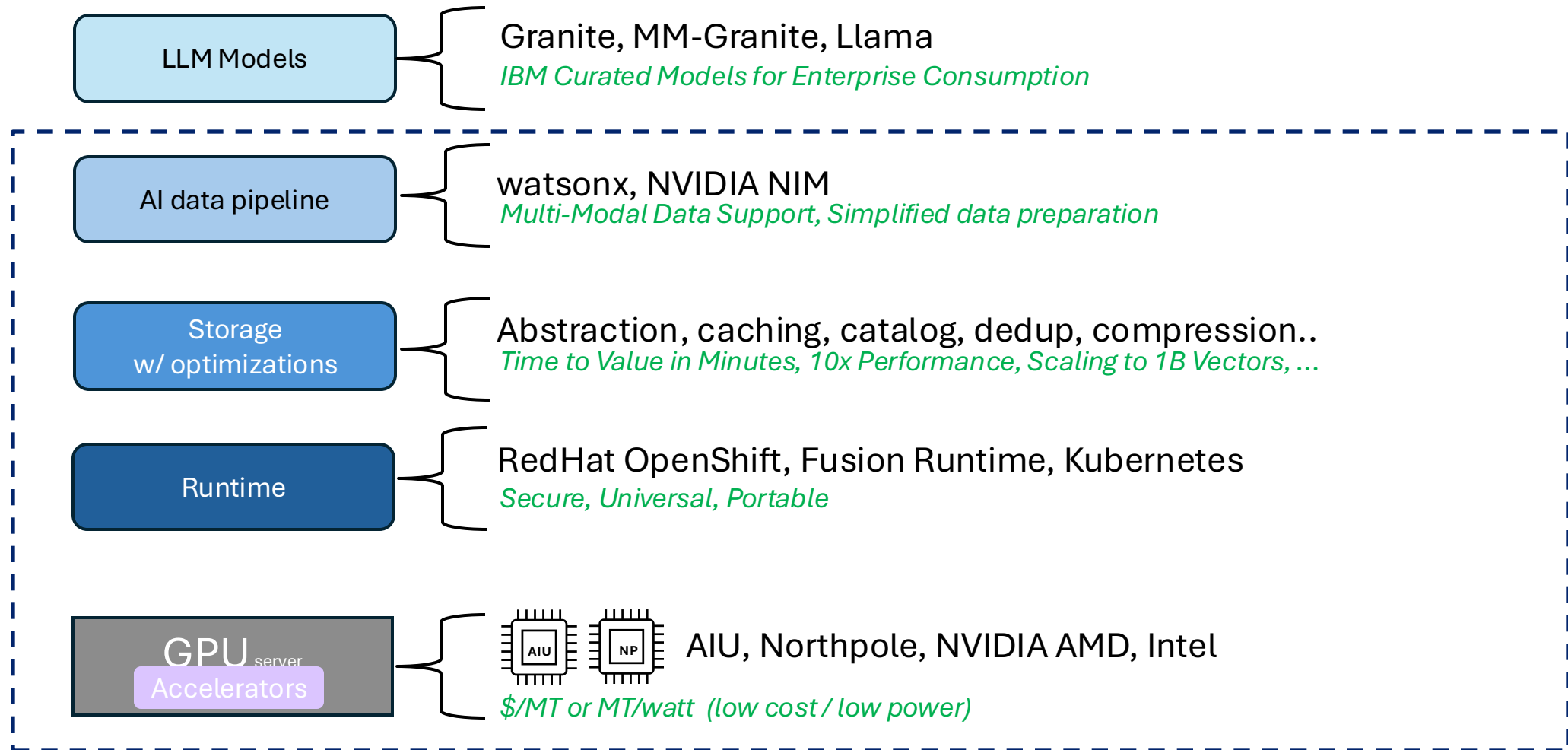
Value props of CAST v01:

- a) Enable all legacy storage into RAG storage. No data copy needed.
- b) 10X efficiency (cost, performance, energy) improvement for RAG patterns

CAST v01 approaches:

- a) Only process (chunking and vectorizing) changed data and update vectorDB accordingly
- b) Leverage storage monitoring capability to detect data changes in object, file and HDFS (storage virtualization for unstructured data)

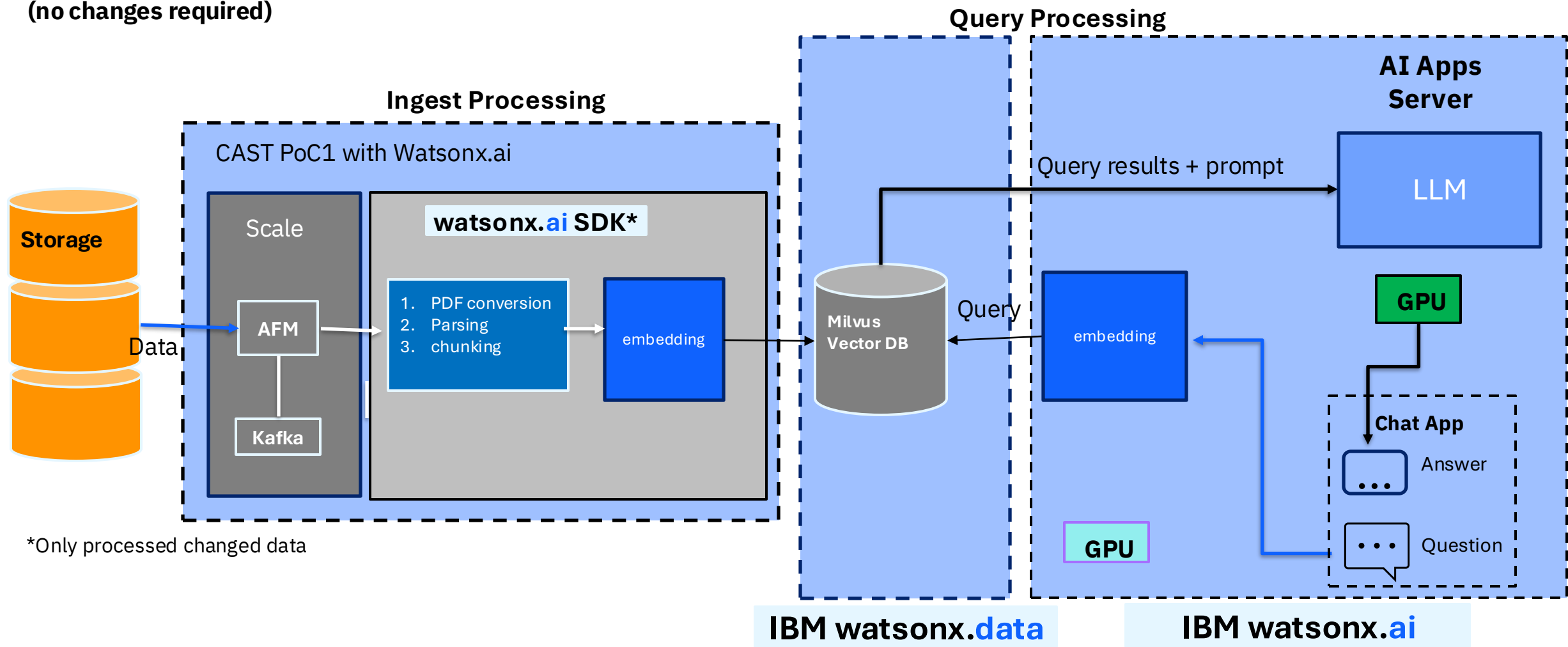
CAST architecture



CAST PoC1 CAST with watsonx

Customer imports documents into their secure S3 bucket (no changes required)

Minutes later, ai apps can answer questions on the newly imported documents



IBM watsonx.ai grounding feature with IBM AI CAST

IBM watsonx

2581212 - Storage Tech S...
Dallas
KN

Projects / springside-cloud / khanhatest_mmrag

Edit vector index

Details

Sample questions

Test

Vector index details

Vector store	watsonx.data
Vector store connection	khanhatest-milvus-wxd-ibm-cloud
Embeddings model	ibm/slate-30m-english-rtrvr
Milvus database	default
Milvus collection	wx_khanhatest_mmrag 13 files
Milvus collection schema	Document name field: document_name Text field: page_content Page number field: page

About this asset

Name

khanhatest_mmrag
Vector index

Description

What's the purpose of this asset?

Asset details

Vector store
watsonx.data

Connection
[khanhatest-milvus-wxd-ibm-cloud](#)

Milvus database
default

Milvus collection
wx_khanhatest_mmrag (13 files)

Last modified

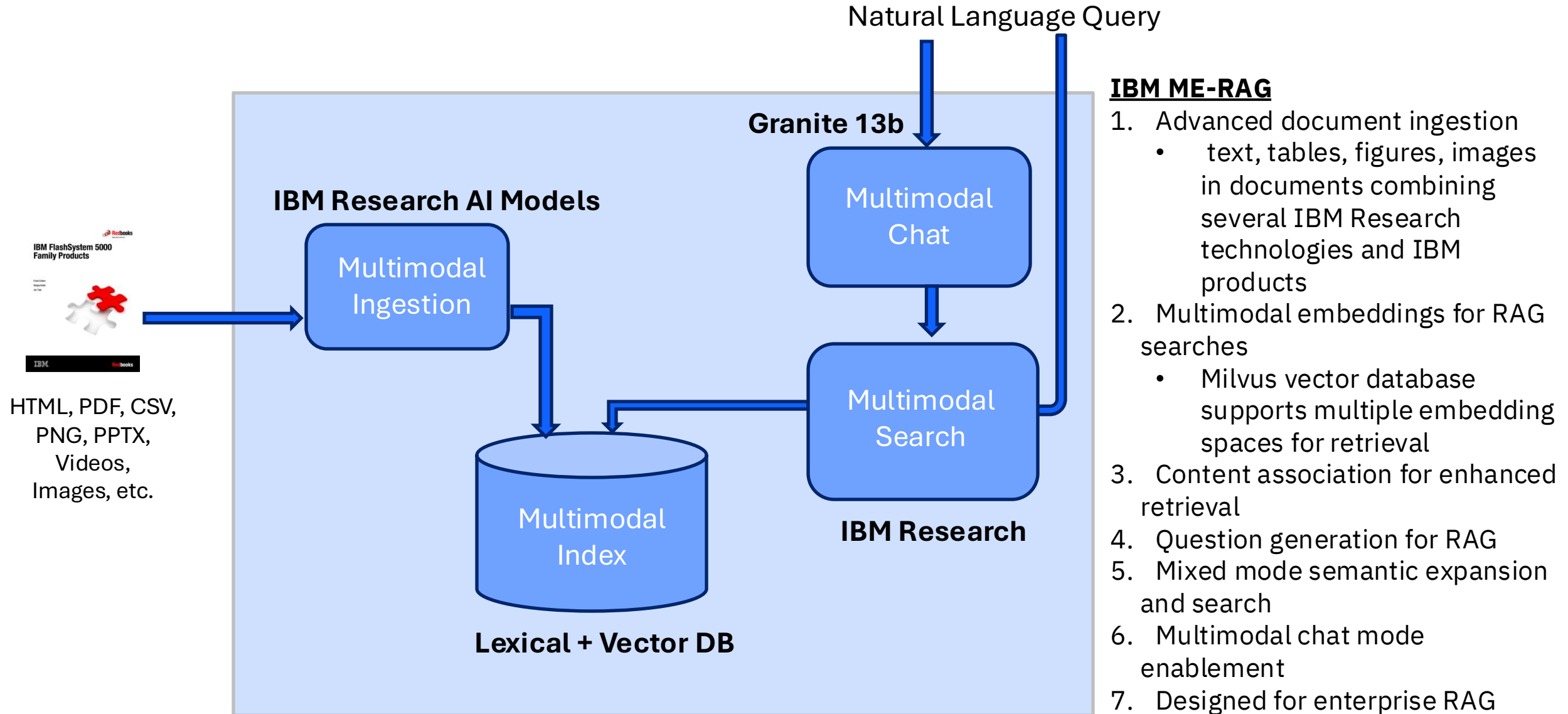
3 wk ago by Khanh Ngo

Created on

3 wk ago by Khanh Ngo

Open in Prompt Lab

PoC 2: ME-RAG as a Data Service in CAST

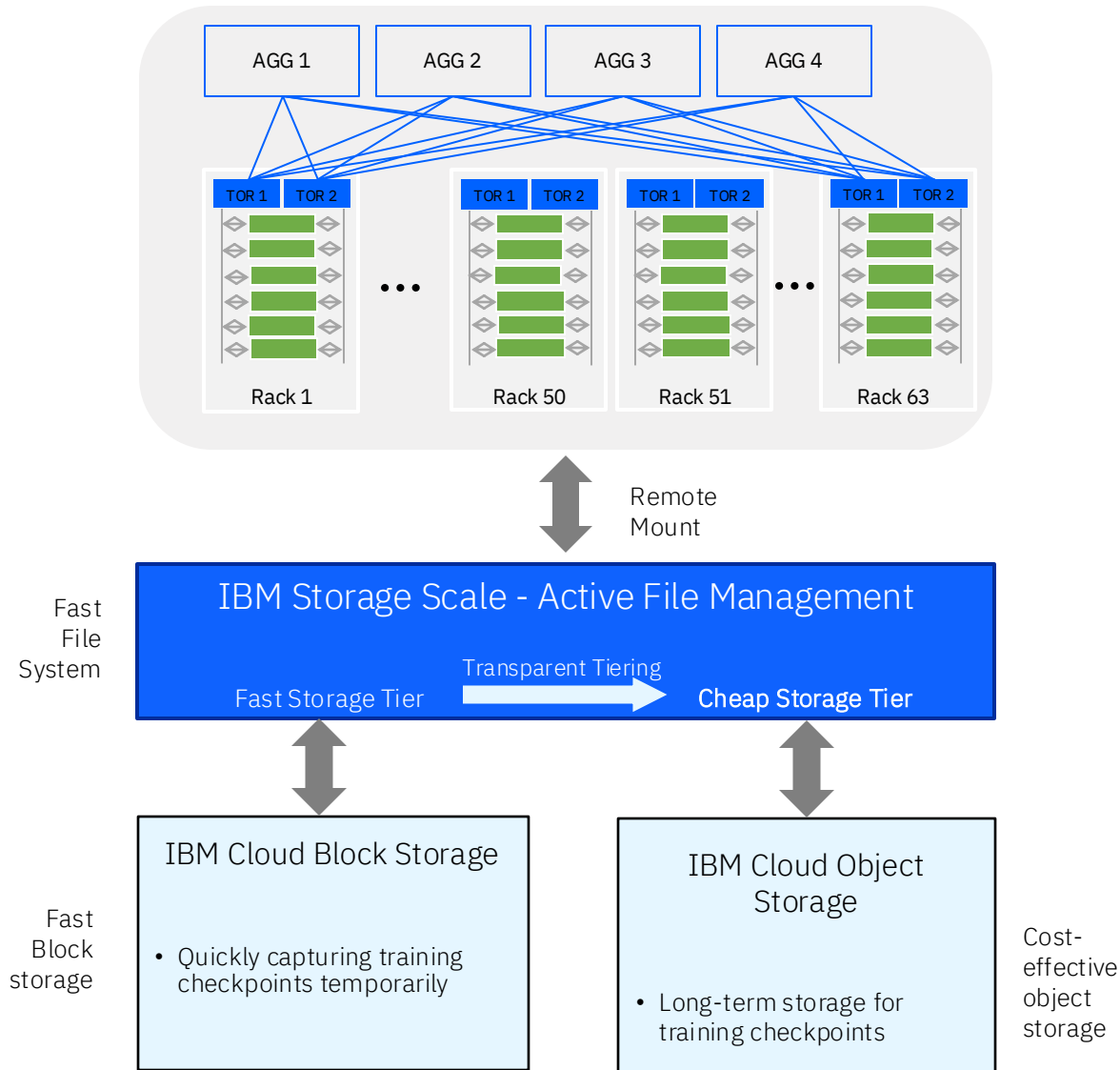


MDU-RAG Demo

IBM Blue Vela

IBM Storage Scale

An integral part of the Vela architecture



- Built completely on **IBM Cloud** infrastructure
- Dedicated **IBM Storage Scale cluster** on IBM Cloud instances
 - Container Native Storage Access (CNSA) on GPU compute cluster
 - 200 nodes, 1600 GPUs
 - Shared POSIX file system semantics
 - One volume for training data
 - Fit complete training dataset
 - One volume for checkpointing
 - Can accumulate ~10 days of checkpointing
- Fast storage tier: **block storage** in IBM Cloud
 - 140 x 1TB x attached volumes
- Large cost-effective data repository using **IBM Cloud Object Storage**
 - Two-tier architecture where AFM transparently moves data between the object storage and file system

Raw performance improvements:

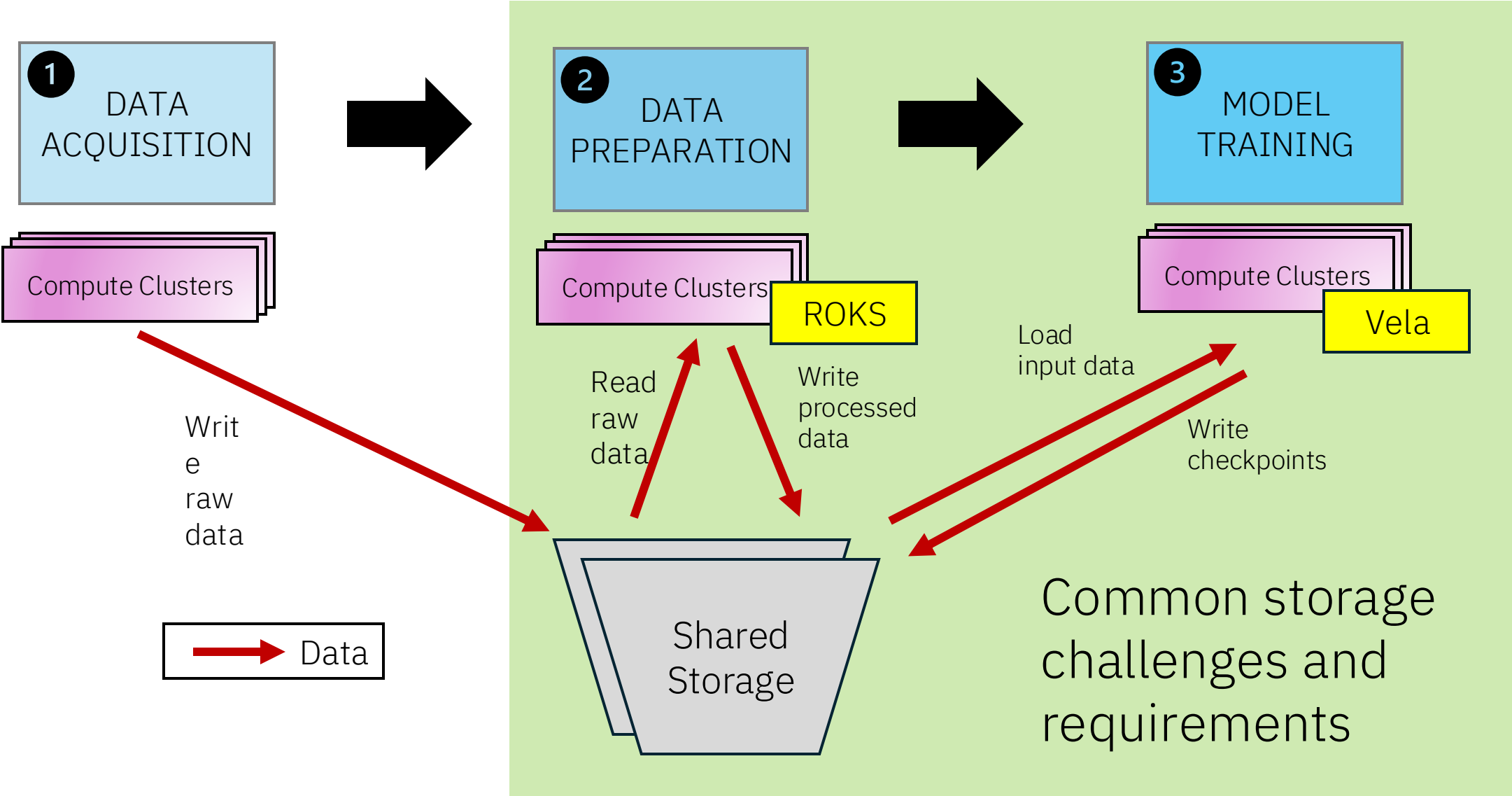
- 3x write bandwidth compared to COS-only (15GB/s vs 5GB/s)
- 40x read bandwidth over NFS (40GB/s vs 1GB/s)

Training performance improvements:

- Storage Scale improved training step time variation by 5X

<https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster>
<https://research.ibm.com/blog/vela-ai-supercomputer-updates>

Data Access in Data and AI workflows

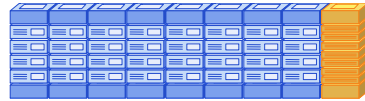


Blue Vela Compute Pods



IBM Blue Vela- HGX “SuperPOD” Storage Fabric (IBM Cloud/ IBM Research/NVIDIA) working together to delivery an AI solution

Storage Scale 6000



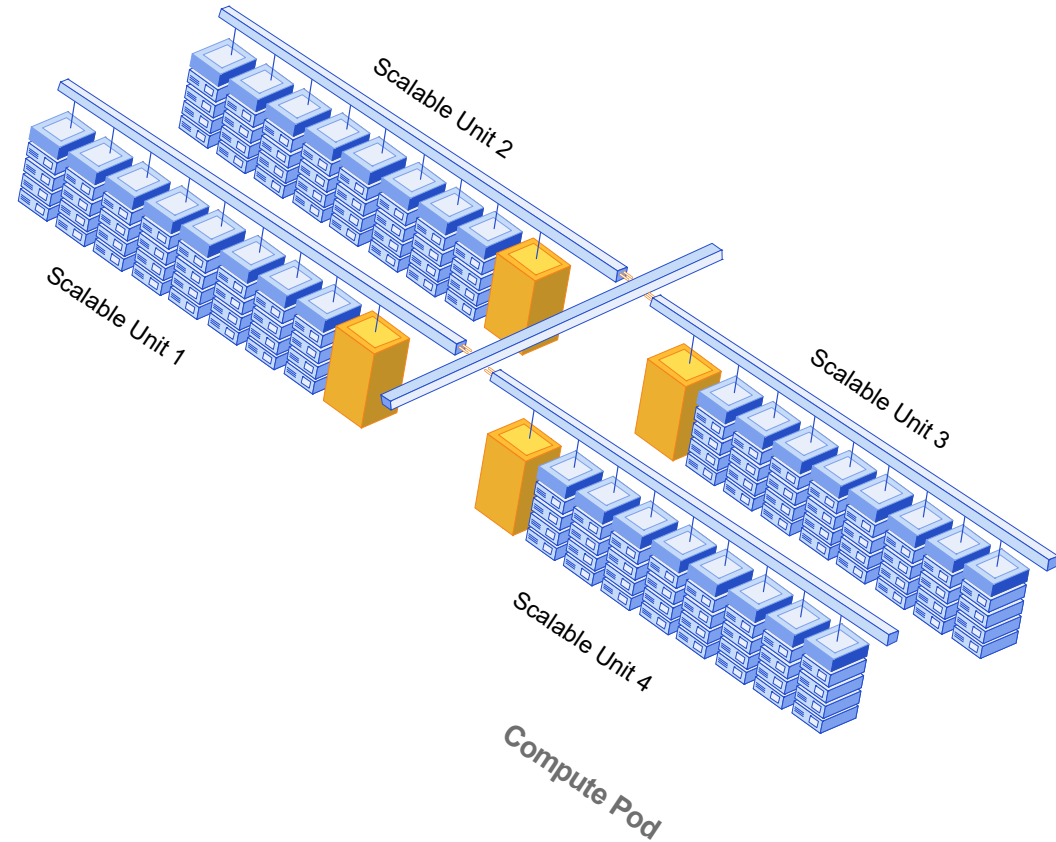
Scalable Unit

- 32 Compute Nodes
- 256 H100 GPUs



Compute Pod

- 4 Scalable Units
- 128 Compute Nodes
- 1024 H100 GPUs
- 82 TB of GPU Ram
- 12,288 Physical Cores
- 256 TB of RAM
- 3481 TB NVME Local Storage



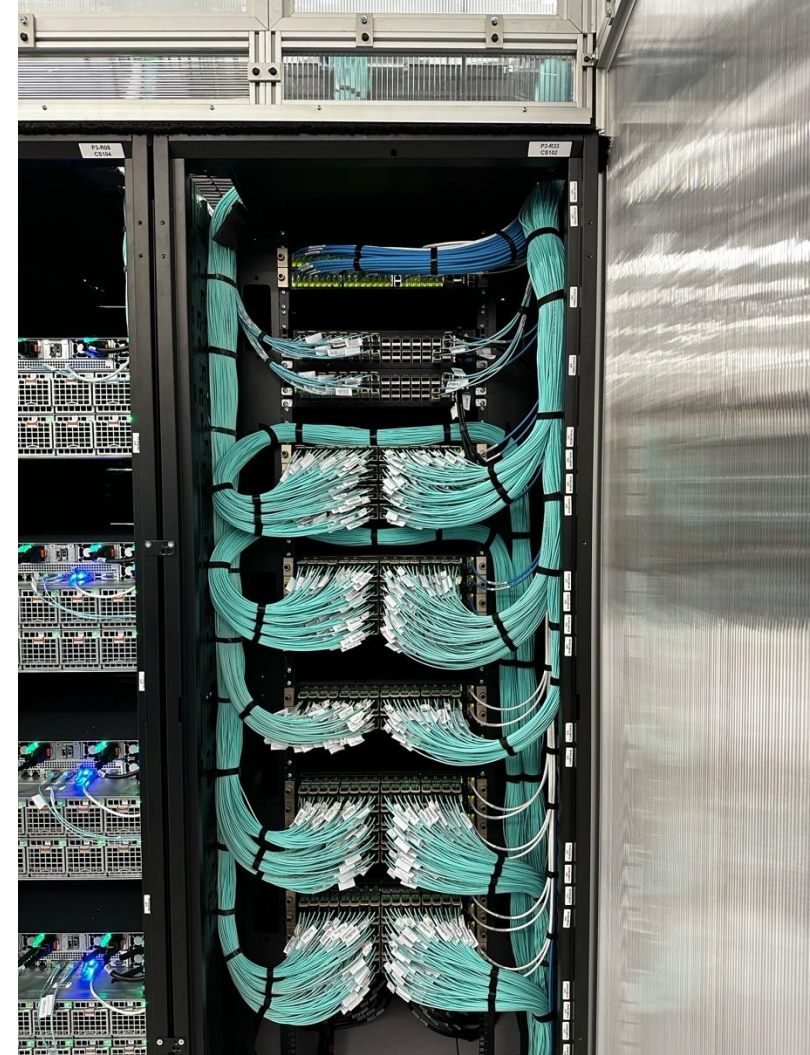
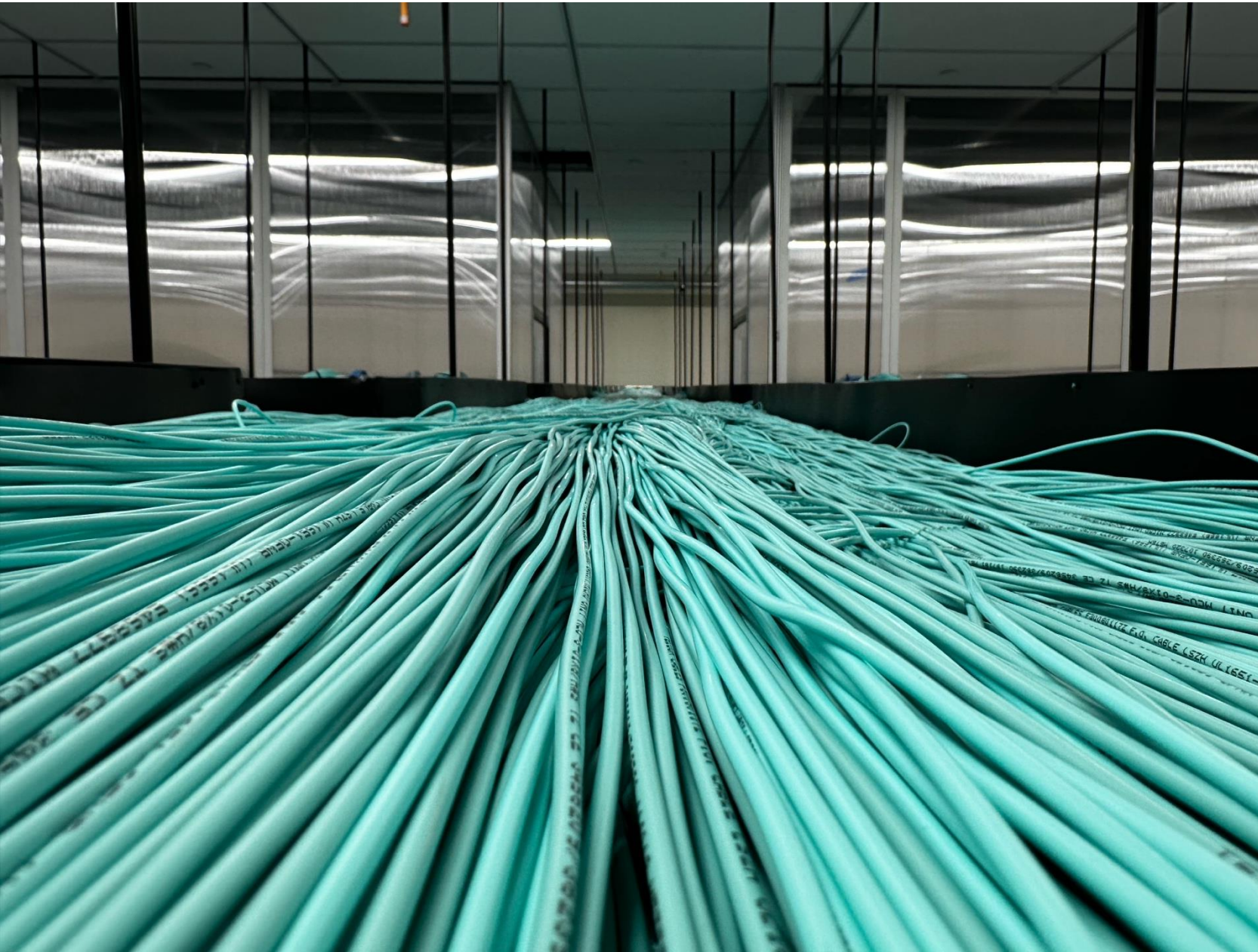
https://www.linkedin.com/posts/dannybarnett_todays-an-incredibly-proud-day-for-my-team-activity-7180654456361910274-4fvU?utm_medium=member_ios

<https://arxiv.org/pdf/2407.05467>

Blue Vela Compute Pods



Blue Vela

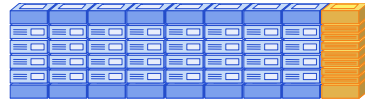


Blue Vela Compute Pods



IBM Blue Vela- HGX “SuperPOD” Storage Fabric (IBM Cloud/ IBM Research/NVIDIA) working together to delivery an AI solution

Storage Scale 6000



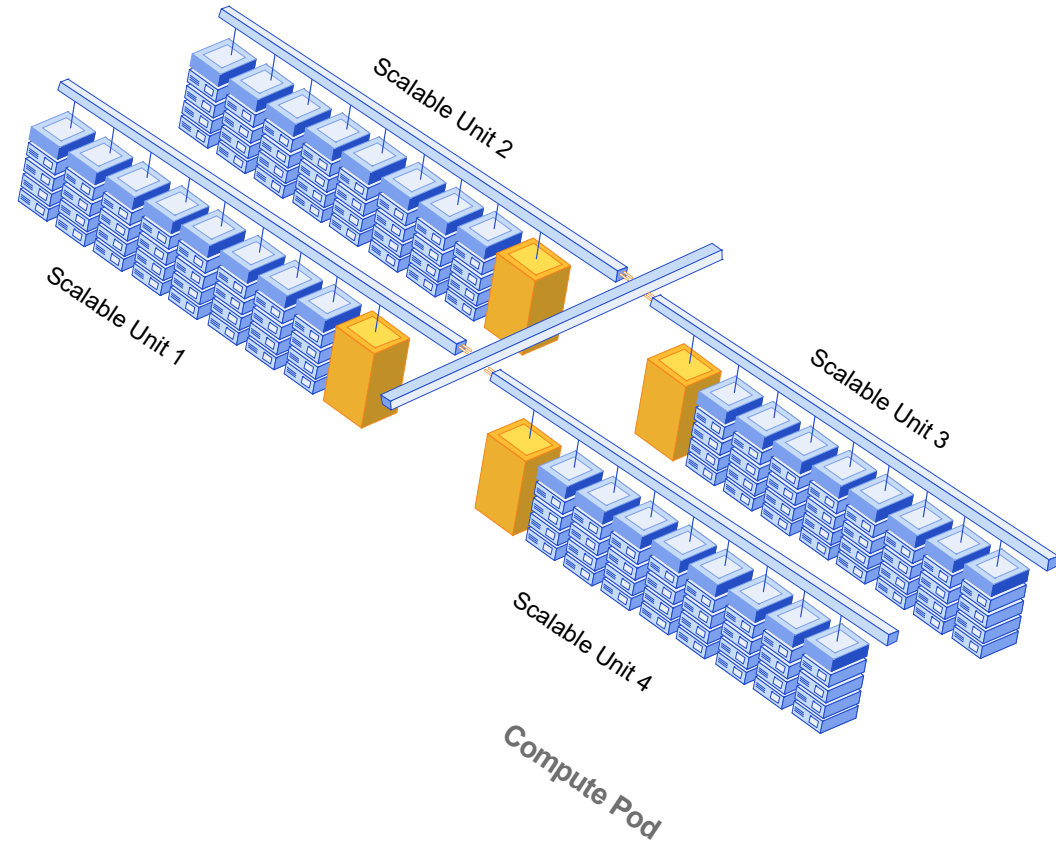
Scalable Unit

- 32 Compute Nodes
- 256 H100 GPUs



Compute Pod

- 4 Scalable Units
- 128 Compute Nodes
- 1024 H100 GPUs
- 82 TB of GPU Ram
- 12,288 Physical Cores
- 256 TB of RAM
- 3481 TB NVME Local Storage



https://www.linkedin.com/posts/dannybarnett_todays-an-incredibly-proud-day-for-my-team-activity-7180654456361910274-4fvU?utm_medium=member_ios

<https://arxiv.org/pdf/2407.05467>

IBM Storage Scale

IBM Blue Vela

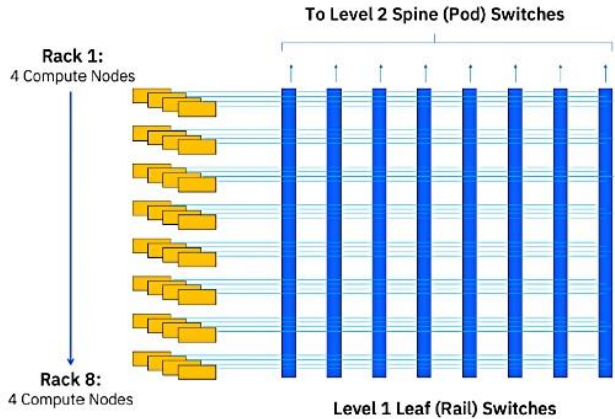


<https://blocksandfiles.com/2024/08/02/big-blues-storage-scale-using-blue-vela-ai-supercomputer/>
<https://arxiv.org/pdf/2407.05467>

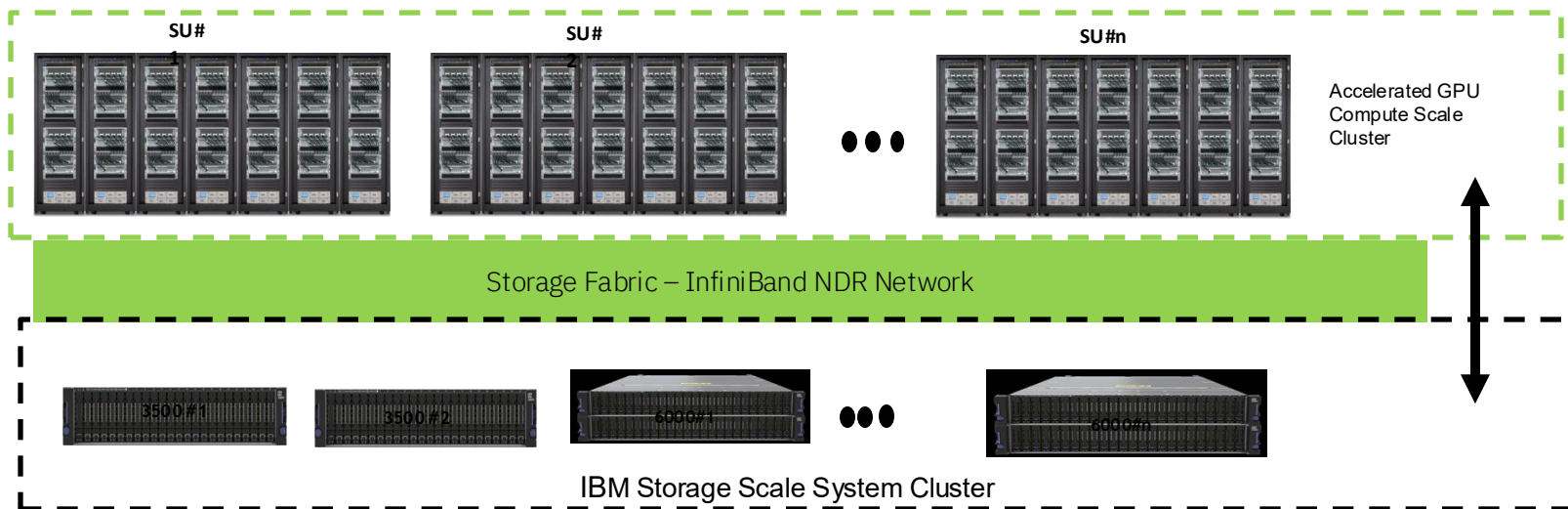


Scalable Unit InfiniBand Fabric

- 8 Compute Racks
 - Each with 4 compute nodes
- 32 Compute Nodes
 - Each with 8 Data IB connections
- 256 Total Endpoints
- 8 Level 1 Leaf (Rail) switches
 - Each with 32 endpoint connections
 - Each with 32 uplinks
- 256 downlinks, 256 uplinks



IBM Blue Vela- HGX “SuperPOD” Storage Fabric (IBM Cloud/ IBM Research)



- IBM Cloud and Infrastructure
- AI Supercomputer Scalable up to 1024 H100 HGX Systems
- 1st Phase 1000s of HGX GPUs

- AI and Data platform to deliver enterprise AI service
- Training LLM models with 100B+ parameters
- Faster results – quality & speed of the training models.

<https://arxiv.org/pdf/2407.05467>



Danny Barnett

VP of Emerging Technology Engineering, IBM Research
1w

Today's an incredibly proud day for my team and me. We just handed over the first tranche of H100 GPUs in our AI training supercomputer ("Blue Vela") to our model research team. Many thanks to our partners [Dell Technologies](#), [QTS Data Centers](#) and [NVIDIA](#) for helping us get to this point.

Thank you to our executive sponsors [Dario Gil](#), [Rick Lewis](#) and [Rohit Badlaney](#) for funding us (a lot of thanks due there) and for clearing the way for us to execute quickly. How quickly? Well, we took delivery of our first server at the beginning of December 2023 and we're running our first productive workload 1st April. So pretty quickly given the size and complexity of these things.

This was a huge team effort but a special shout-out to two colleagues without whom this definitely wouldn't have happened or happened so quickly: [Felix Eickhoff](#) and [Brian Belgodere](#)



https://www.linkedin.com/posts/dannybarnett_todays-an-incredibly-proud-day-for-my-team-activity-7180654456361910274-4fvU?utm_medium=member_ios