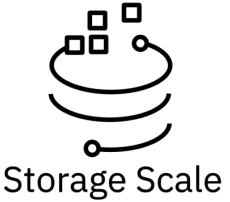
IBM Storage & Storage Scale Strategy

IBM Storage Scale Days 2024

March 5-7, 2024 | Stuttgart Marriott Hotel Sindelfingen

Ted Hoover
Product Manager, Storage for Data and AI

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.

Traditional storage infrastructure cannot keep up with digital business

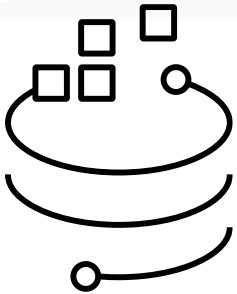
90% of organizations are engaged in an enterprise-wide digital transformation

Agile infrastructure for app modernization

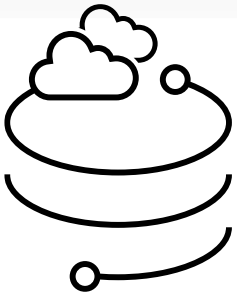
Unstructured data management for AI and analytics

Security, resiliency, and governance of data

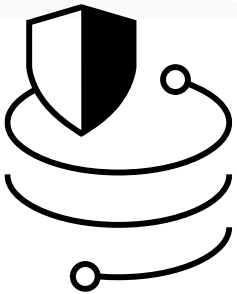
IBM Storage



Storage for
Data and AI



Storage for
Hybrid Cloud



Storage for
Data Resilience

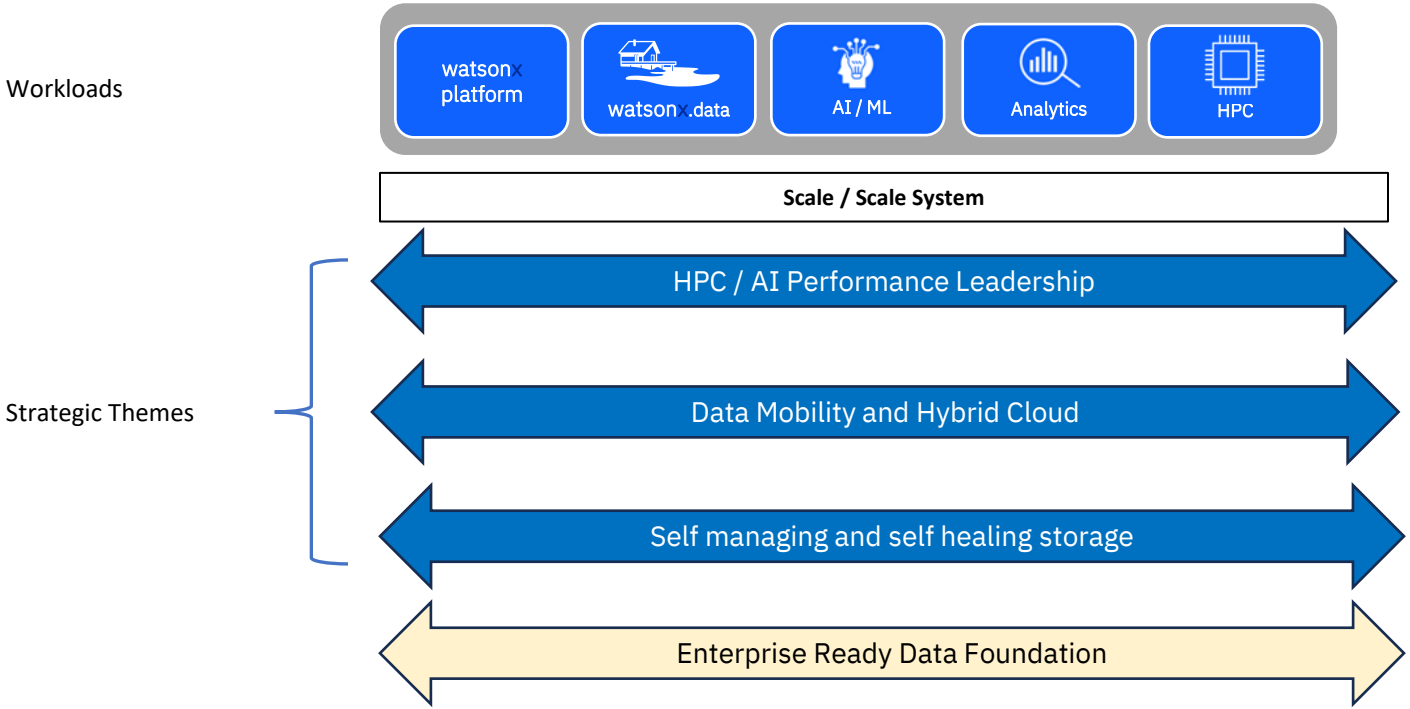
High Performance Storage for Analytics, AI, & HPC

- Workload Requirements for Data Intensive Computing
  - Traditional Modelling / Simulation
  - Growth of AI (ML, DL) as a new paradigm in high performant workloads
    - LLM & HPC - most demanding, modern AI space
  - High Performance Data Analytics
  - Hybrid (Data Lakehouse, AI Augmented Modelling/Simulation, etc)

- Expansion Beyond the Traditional On-prem Datacenter, to the Edge, and to the Cloud
  - Data Driven Workflows
  - Data Architecture that Scales
  - Emergence of Usage Based Consumption Models

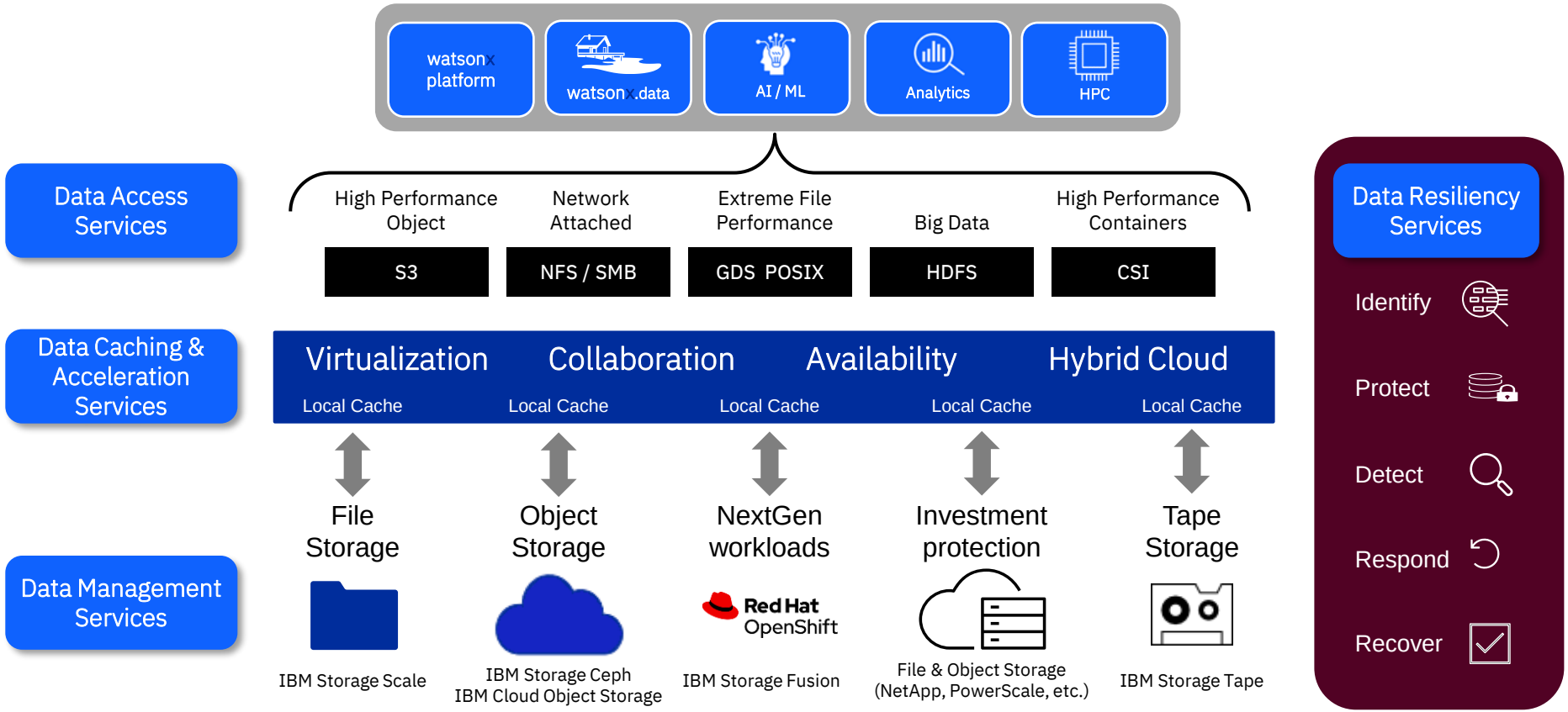
- New Data Challenges
  - Data Governance, Management, and Orchestration
  - Continued Data Growth
  - Increased Performance (Throughput, Bandwidth, IOPS)

Storage Scale Strategic Themes



IBM Storage Scale

A Global Data Platform for Unstructured Data for AI Foundation Models & Data Lakes



Storage for Data and AI Customer Challenges

$$$

GPU resources are expensive

I need a storage solution that:

- Minimizes the cost of training AI models by delivering a faster time to solution

AI Data is distributed

I need a storage solution that:

- Can effectively integrate unstructured data
- Delivered when and where it is needed
- Transparent to the AI workload

$$$

Escalating Storage Costs

I need a storage solution that:

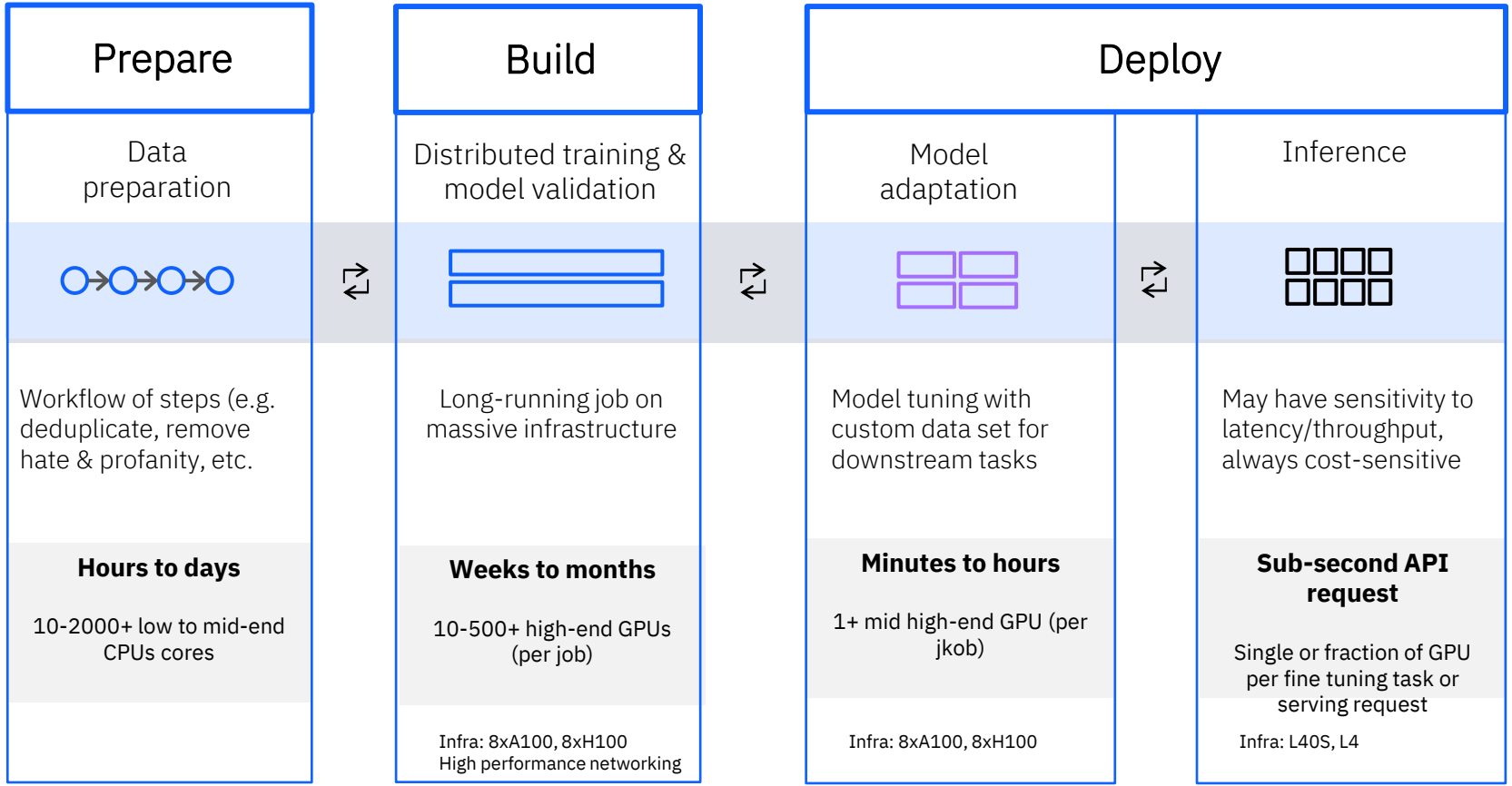
- Allows me to integrate existing dispersed storage silos
- Allows me to seamlessly manage data in different cost optimized storage tiers
- Effectively manage the data lifecycle

Data is a critical business asset

I need a storage solution that:

- Protects my data from security threats, unplanned disasters and always available
- Resilient to cyber threats, can be brought back online quickly and highly available to keep my business running

Optimizing infrastructure
Across the whole AI workflow



Watsonx Data and AI platform

watsonx.ai

An enterprise studio to build AI applications quickly and with fraction of data

Train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models.

watsonx.data

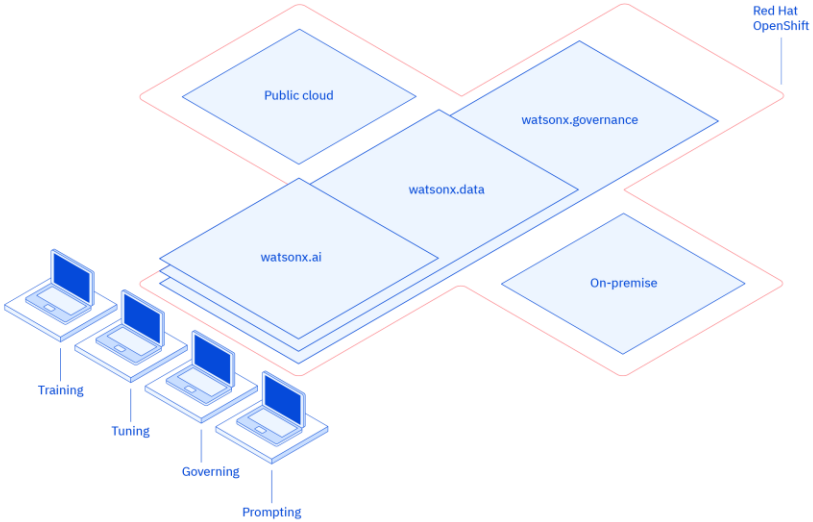
Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture to access and share data across hybrid cloud through a single point for entry.

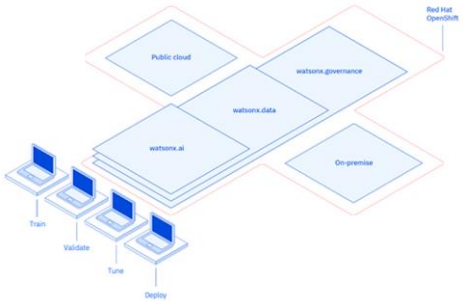
watsonx.governance

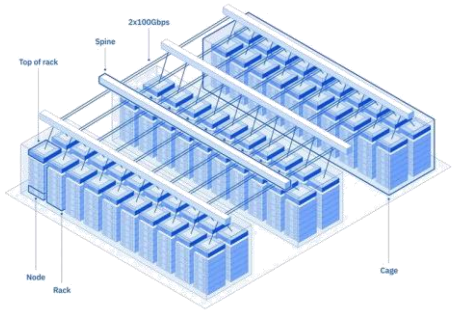
Trusted – Responsible, transparent, and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to enable responsible, transparent, and explainable AI workflows.



Storage requirements for AI





AI Tuning/Inferencing

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

AI Training

Maximum Performance	Efficient GPU support High bandwidth Low latency
Scalability	Linear capacity scaling High density

IBM Storage for Data and AI & NVIDIA GPU Solutions

A full spectrum of scalable AI solutions

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

AI Entrant

1 x DGX A100/H100
or
1x NVIDIA Certified Server

- 12 NVMe Half Populated 3500
- Up to 60 GB/s Read
- **1 x 6000 w/ 12 NVMe**
- **Up to 80+ GB/s read**

AI Medium

4 x DGX A100/H100
or
4 x NVIDIA Certified Servers

- 1 x 3500
- Up to 125 GB/s read
- **1 x 6000**
- **Up to 310 GB/s read**
- **Up to 155 GB/s write**

AI Master

8 x DGX A100/H100
or
8x NVIDIA Certified Servers

- 2 x 3500
- Up to 250 GB/s read
- **1 x 6000**
- **Up to 310 GB/s read**
- **Up to 155 GB/s write**

AI Scaler

NVIDIA SuperPOD
32 x DGX H100
or
32 x NVIDIA Certified Servers

- 2 x 3500
- Up to 250 GB/s read
- **2 x 6000**
- **Up to 620 GB/s read**
- **Up to 310 GB/s write**

IBM Storage:

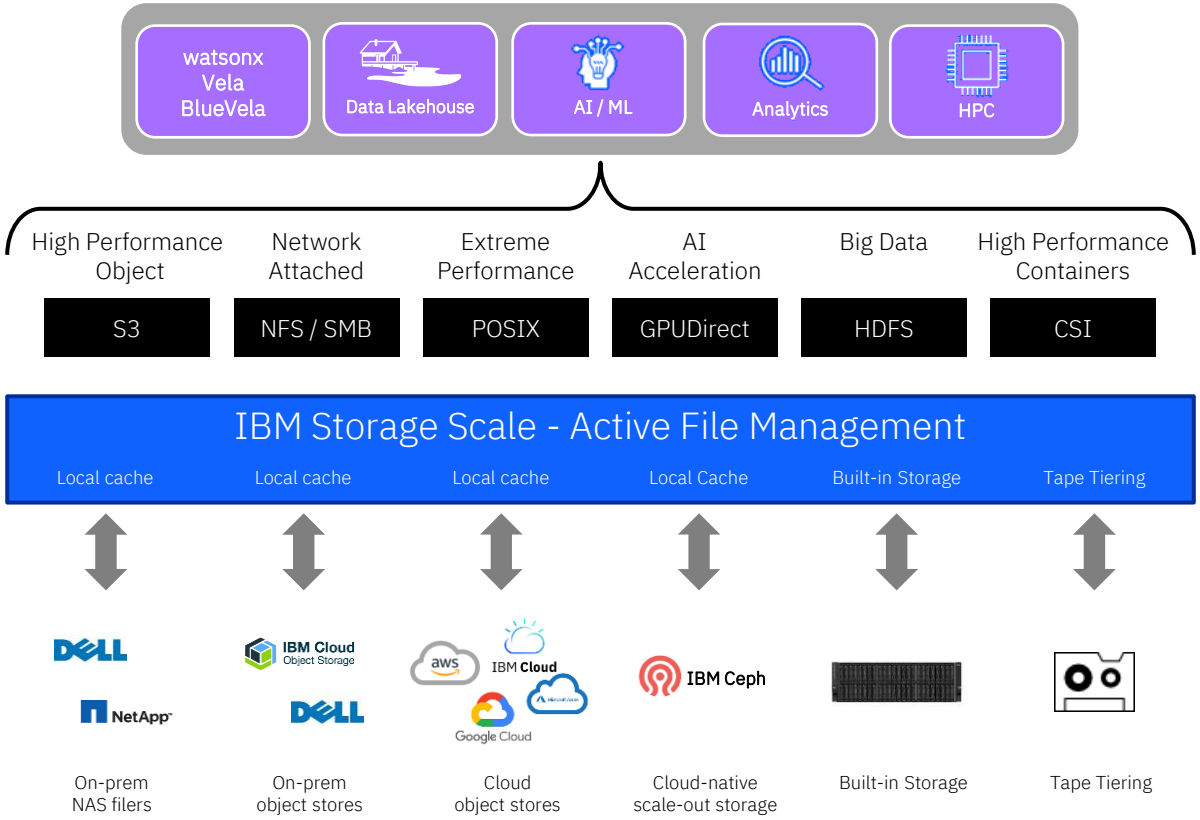
- *Simple building block – simple, scalable seamless upgrade path*
- *Enterprise features– performance, scalability, data protection and security*
- *Global Data Platform Services – Integrate with current storage, multi-site active-active, edge to cloud to core, single namespace across multiple installations*
- *IBM expertise and services*
- *Successful deployments across the globe –*

  Telco, Automobile, Banking and Finance, Healthcare, Retail, Academic/ Research and Public Sector

A simple, scalable upgrade path

IBM Storage Scale a Global Data Platform for Unstructured Data

Storage Abstraction and Acceleration



Multi-Protocol Support

Simultaneous multi-protocol access including GPUDirect support

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

Storage Acceleration

Automatic, transparent caching of back-end storage systems

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

Storage Abstraction

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

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

IBM Fusion

The “Easy Button” for watsonx on-premises

Quickest way to deploy watsonx

- Enterprise ready with HA/DR and full-stack backup/restore
- Simplify operations with automated day-2 monitoring, and maintenance services
- **Integrated system, engineered for watsonx**

AI and Storage acceleration

- Abstracts storage from multiple existing storage systems into a single global namespace
- Accelerate data queries by 7x to 9x with intelligent caching
- **GPU nodes with NVIDIA GPUs**

Flexibility to meet client needs

- Optional data lake for storing newly-created data or migrating from existing storage systems
- Easily add additional storage capacity with scale-out architecture
- **Easily add cores for additional performance**

Compute

Storage

Network

GPU nodes
AI acceleration

Compute/storage nodes
Runtime platform for
- OpenShift
- Fusion Data Services
- watsonx

Built-in IBM Storage Scale
Storage abstraction and acceleration

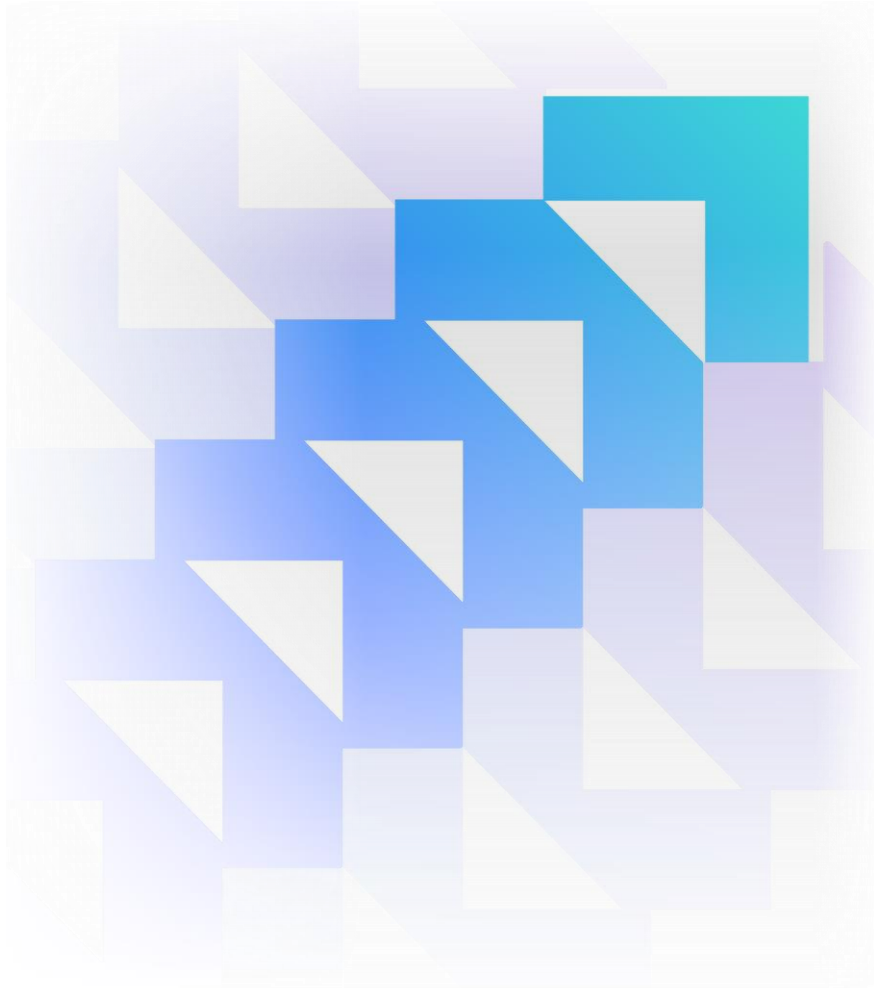
Storage nodes
Expandable data lake

High speed switches
Connectivity to the data center and existing storage

Management switches
Appliance management /monitoring

IBM Fusion HCI

IBM Storage Scale Scale System Roadmap



Usability and Security Enhancements

Tech Preview
GA 2Q24

Security Improvements

Removal of SSH dependency

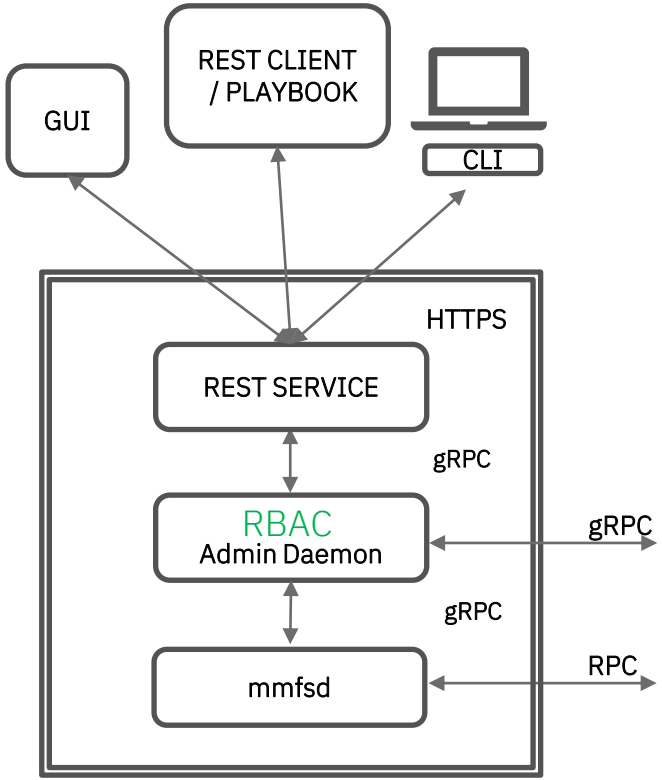
Removal of root requirement for control plane

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



Usability Enhancements v2

Multi-Tenancy Improvements

Resiliency

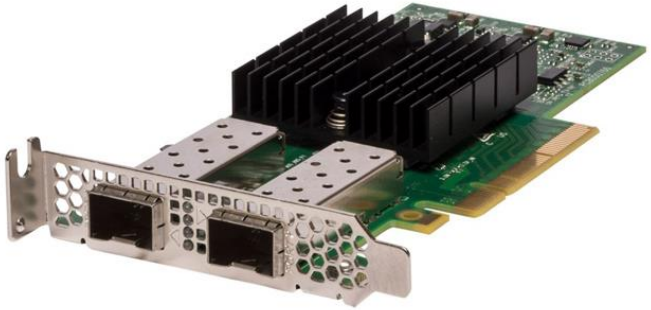
- Failure Isolation / Blast Radius
- Network Fault handling
- Bad / Slow Nodes
- Rapidly changing configuration

Performance Isolation

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

Manageability

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



Network Resiliency

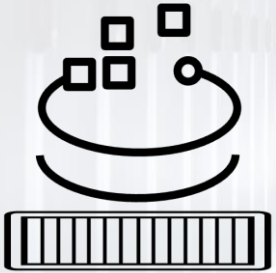
Scale is a clustered file system and depends on timely and reliable TCP/IP communication between all nodes in the cluster to ensure data integrity and good performance

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

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

Thank you for using

Storage Scale



Storage Scale System

Ted Hoover

Mail: hoov@us.ibm.com

Telefon: 845-433-7693

LinkedIn:
https://www.linkedin.com/in/ted-hoover-75234310/