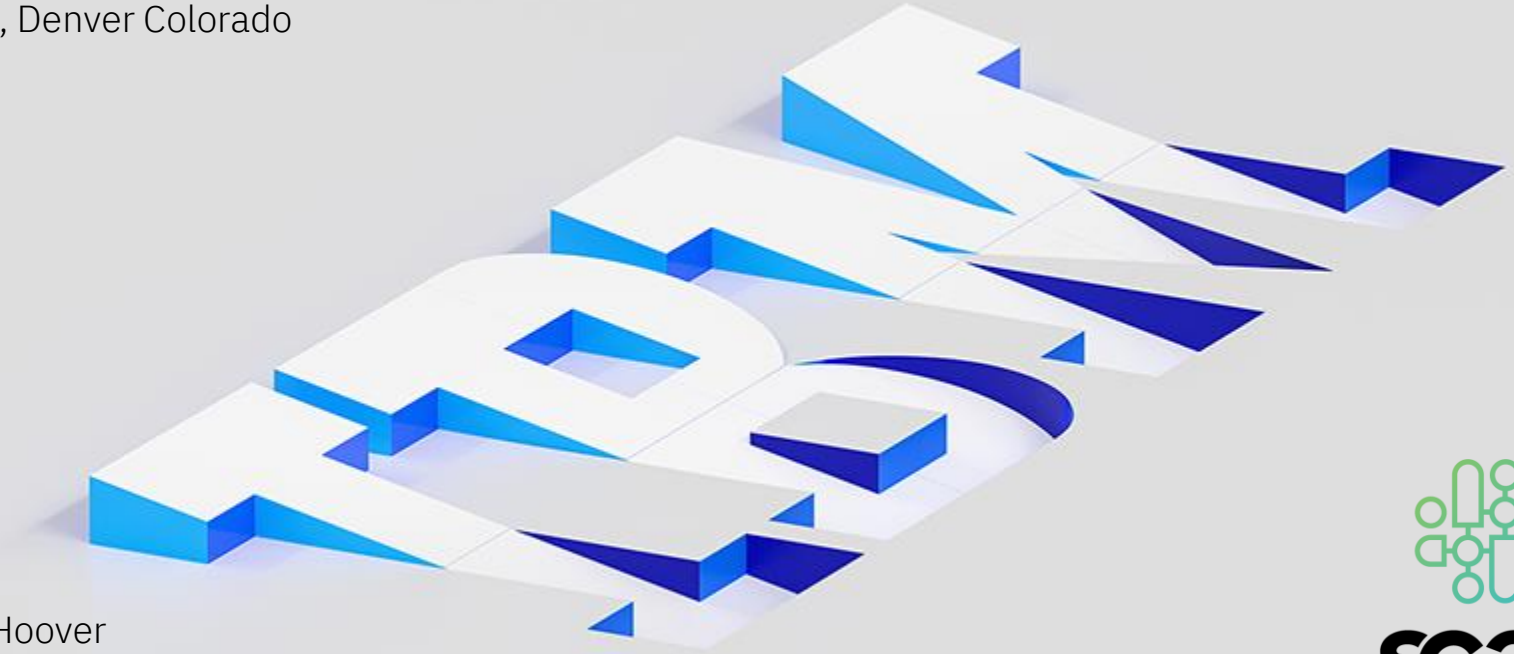
IBM Storage Scale Strategy

IBM Storage Scale User Group 2023
SC23, Denver Colorado



Ted Hoover
Product Manager, Storage for Data and AI

Wayne Sawdon
CTO, IBM Storage Scale & Scale System

SC23
Denver, CO | i am hpc.

IBM

Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

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)

Requires a Global Data Platform for Scale-Out File & Object Data

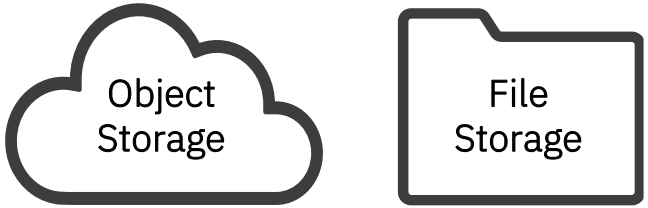
What's driving the need for a global data platform for unstructured data?

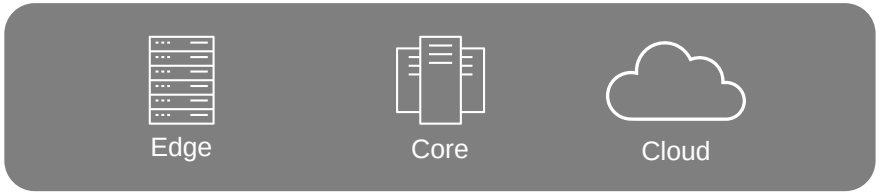
Data centric application development

- Agile application development cycle times
  - Days/weeks vs. months/year
- Example: The emergence of AI/ML use cases
  - Data hungry apps and GPUs, need access to more data, faster
  - As new applications and use cases roll out, data silos occur
  - Need unified and consistent approach to accessing data throughout AI/ML Pipeline – in both native object and file storage repositories
- Data fabric initiatives with requirements to provide consistent services across diverse infrastructure

The diverse IT infrastructure options available

- Many choices, from edge to core data center to public cloud
- Containers to simplify hybrid cloud infrastructure choices
- Drives the need for a single source of truth across diverse infrastructure that facilitates secure access while eliminating data redundancy and inconsistencies.

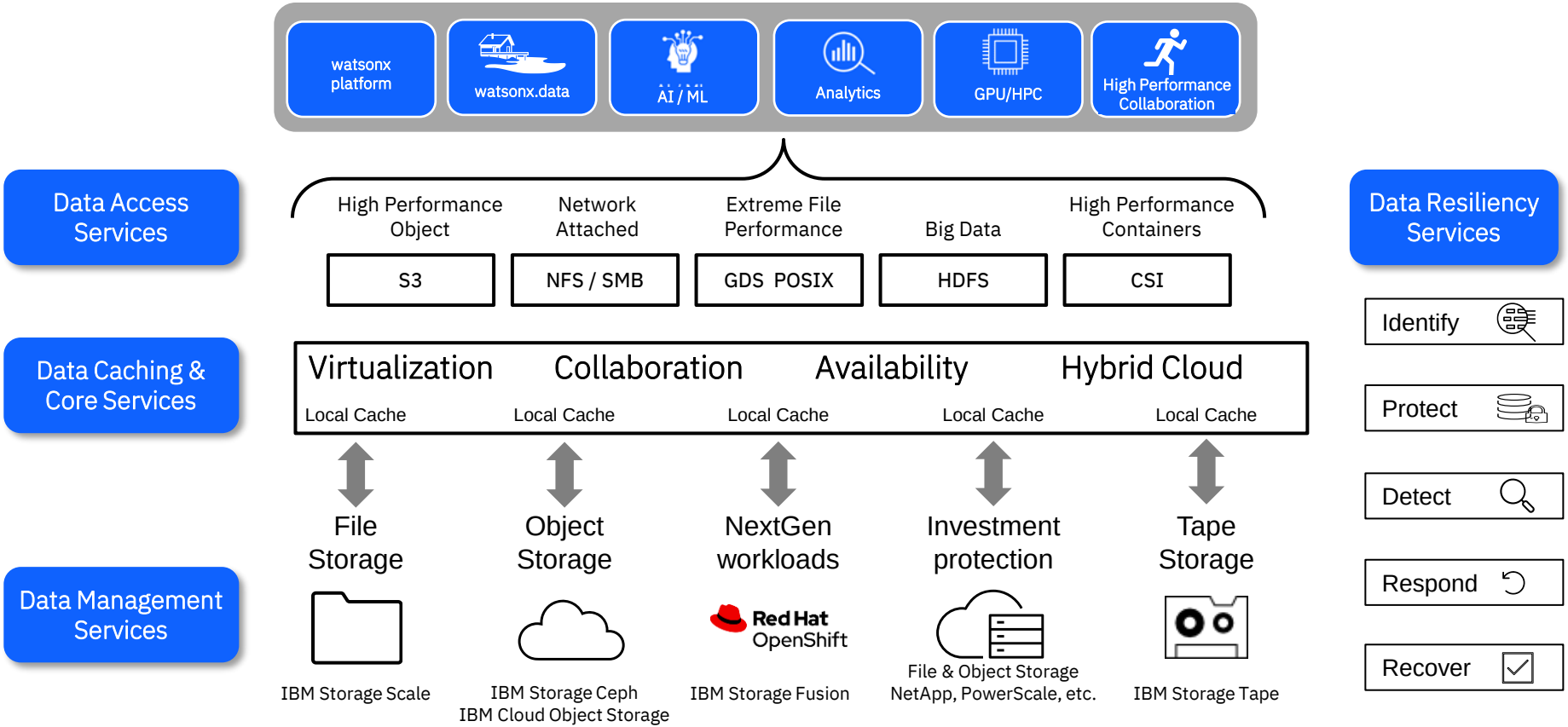




"No AI without IA"
No Artificial Intelligence without an Information Architecture

A Global Data Platform for Unstructured Data

Unifying File and Object Storage to Provide Common Data Services



IBM Storage Scale – Storage Accelerator for AI

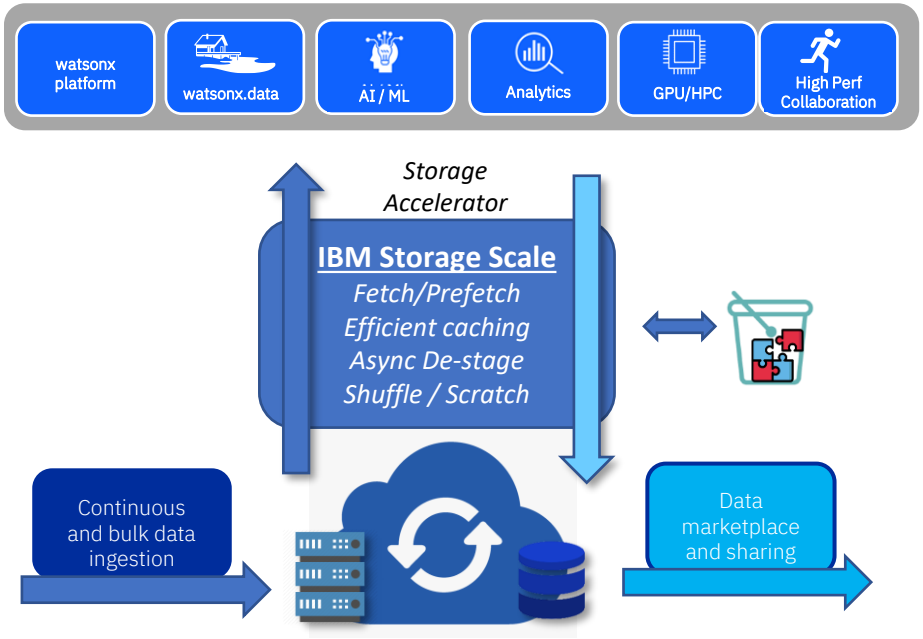
Client challenges:

- *AI training on GPU based compute clusters bottlenecked by storage infrastructure*
- *Large existing data lakes of unstructured data are distributed across on-prem, edge, & cloud*
- *Difficult to integrate distributed data lakes with data warehouses*
- *AI data pipeline is multi protocol and distributed*
- *Data often lives where it was generated*
- *Increasing data volume & data architectures that does not scale*

Solution:

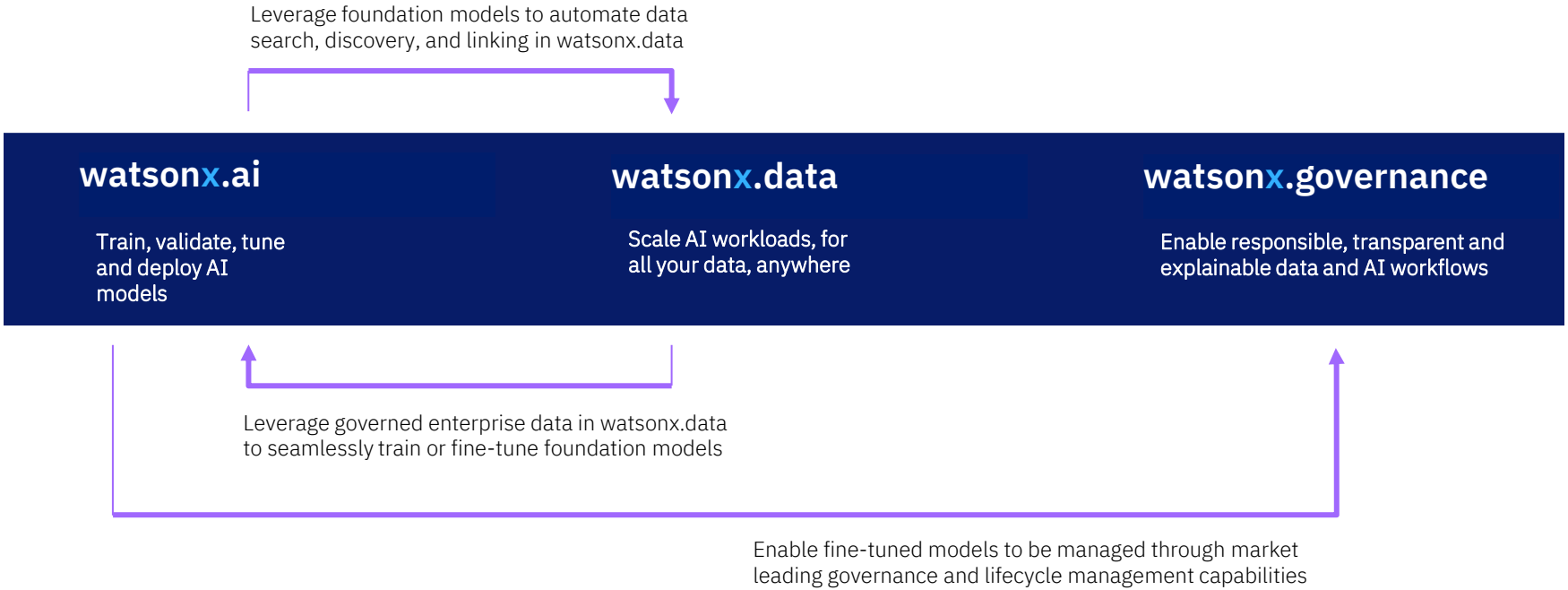
Storage Scale and Storage Scale System

- *Delivers high performant data access for large training models*
- *Data Caching Services enables data access from High Performance Tier*
- *Data Caching Services enables the integration of one or more data sources under a single name space*
- *Supports multiple stages of the AI data pipeline*
- *On-Prem, Edge, Cloud*



Put AI to work with **watsonx**

Scale and accelerate the impact of AI with trusted data.



IBM's Global Data Platform for AI with NVIDIA®

Engineered and optimized for data science productivity

CLOUD

NVIDIA CERTIFIED

NVIDIA-Certified Systems

EGX/HGX Servers

NVIDIA DGX BasePOD

NVIDIA DGX SuperPOD

Easy to Start

Easy to Expand

Parallel Performance Access Services	Multi-site/Multi-vendor Caching Services	Increased Efficiency Management Services	Cyber Resilient Security Services
Speeds AI Results	*Connects AI Data*	*Optimizes AI Data*	Protects AI Data
Up to 1.8TB/s and 30M IOPS per rack	Breaks down silos with a Global Data Platform	Policy based data placement and reduction	IBM Safeguarded Copy and Cyber Vault

IBM

Storage Scale on the Cloud

Access Data from Multiple Interfaces

Access Data from Many Sources

Deliver on the Value of Spectrum Scale

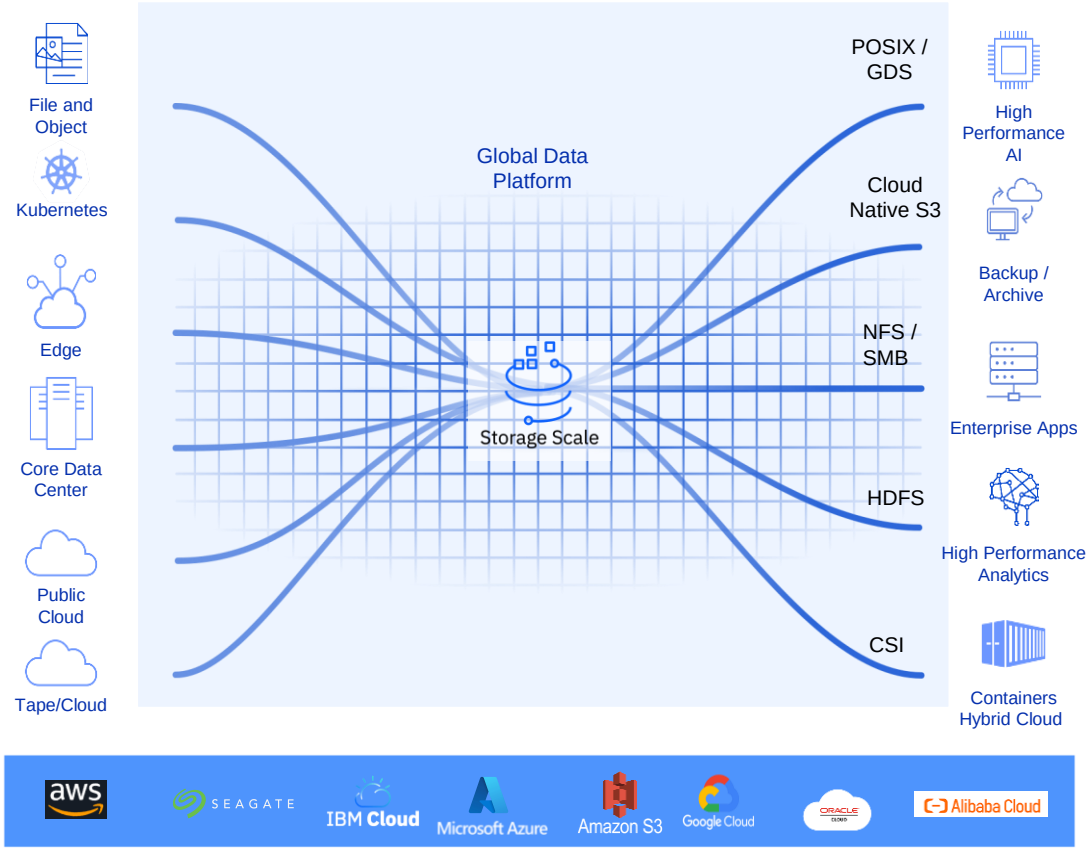
Hybrid Cloud Use Cases
- Backup / Archive
- Tiering
- Bursting
- Data Sharing

Deployment Models
- Lift and shift
- Container Native
- Managed Service
- Hybrid

Workload Enablement
- Analytics, AI, Containers

Ecosystem Integration

Spectrum Scale CloudKit



Modernization of Scale: Security

Tech Preview 4Q23

Security Improvements

Removal of SSH dependency

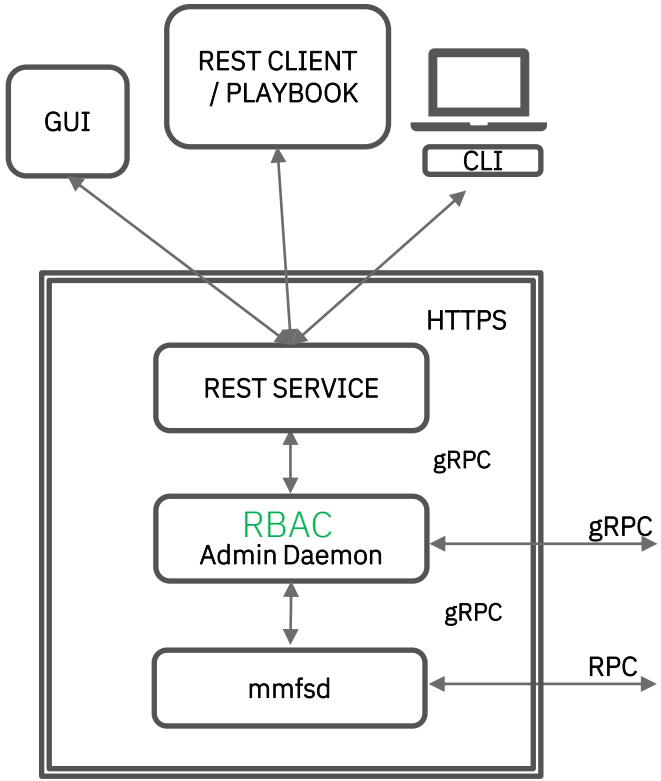
Removal of root requirement for control plane

Remote Administration

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

Control Plane Designed For Applications / Operators

Retain CLI for human management



Modernization of Scale v2

Multi-Tenancy Improvements

Resiliency

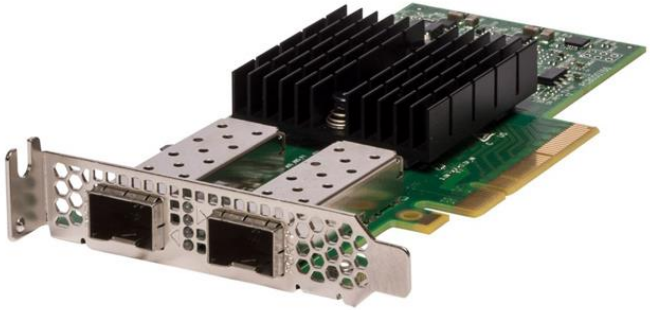
- Failure Isolation / Blast Radius
- Network Fault handling
- Bad / Slow Nodes
- Rapidly changing configuration

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

IBM Storage Scale System 6000

Gen 5 Dual Canister 4U-48 NVMe

- AMD Genoa, dual socket 48 cores / canister
- New x86 utility node (EMS and protocols)
- NDR/CX7 support
- 48x U.2 NVMe / FCM (PCIe Gen 4 drives)
- Up to 1.5PB of NVMe flash and 1.8PB FCM flash
- Up to 5.5PB of compressed FCM flash
- HDD JBOD expansion option (up to 18PB)

NVMe Supported Drives

- 3.84 TB
- 7.68 TB
- 15.36 TB
- 30.74 TB
- ***19.2 / 38 TB FCM 4.0**

HDD SED Supported Drives

- 12 TB SAS HDD
- 16 TB SAS HDD
- 20 TB SAS HDD
- 22 TB SAS HDD

Performance and Sustainability

- **2x throughput improvement**
- **NVMeoF support**
- **Hybrid performance and capacity support**
- **Containerized protocol support on IO nodes**

***1H24**

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!

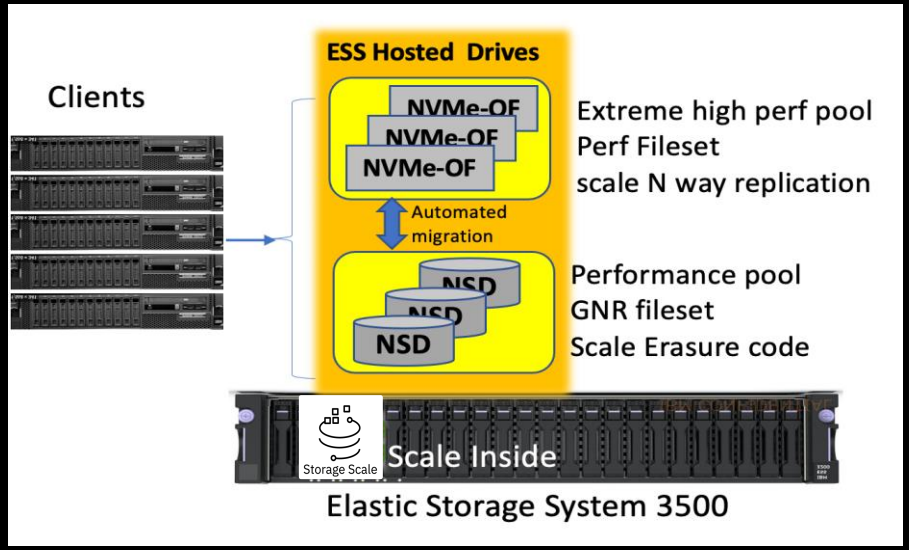
Internally tiered storage
-> MRAM -> SLC -> 3D QLC
Performance comparable to TLC

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%

Integrated NVMe-OF Extreme performance Tier



Measured over 16 M IOPs and 110 GB/s

Use Case

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

High performance Scratch / Shuffle space

System Config

3.84 TB, 7.68 TB, 15.36 TB or 30.74 TB

4x CX6-VPI Adapters / canister

Performance and Features

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


Thank You