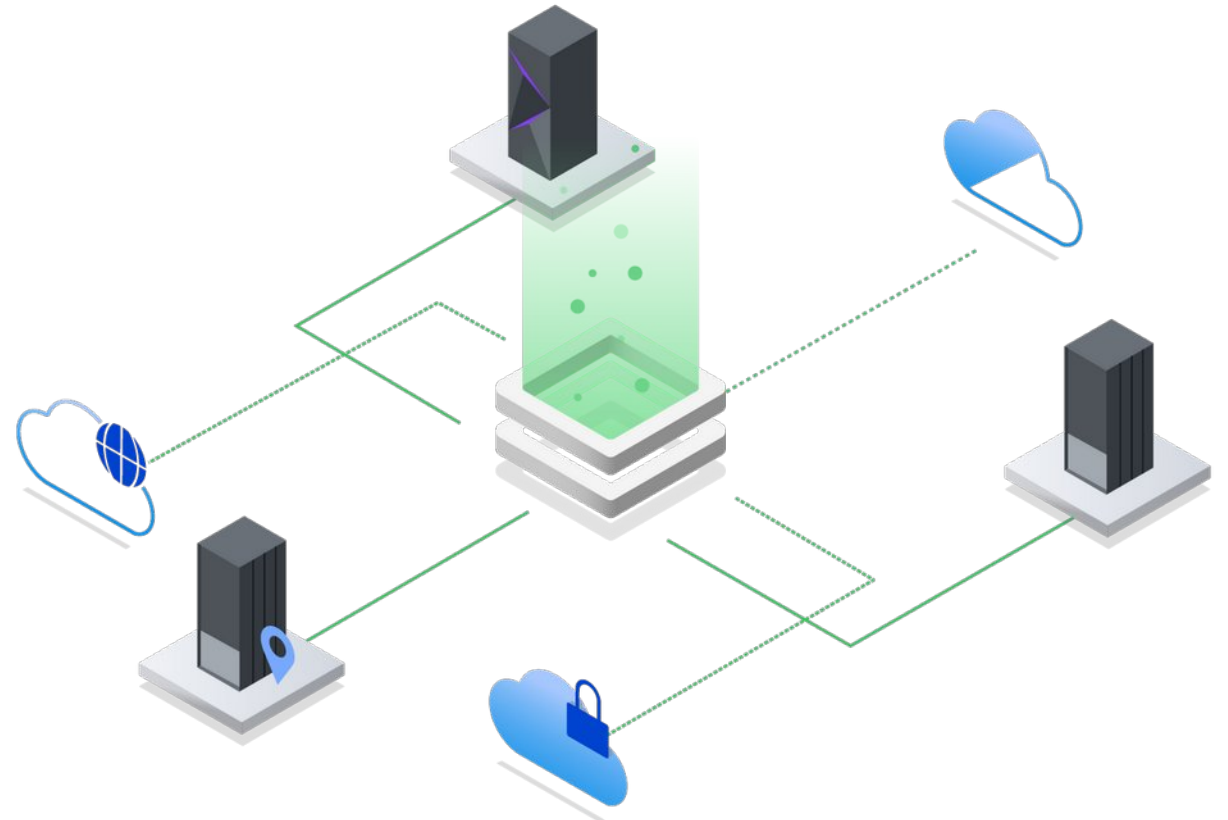


IBM Storage Scale General Overview



Matthew Klos
Client Solutions Engineering Leader
Storage Solutions Team



IBM Storage Scale

Scale 101

- Storage Scale was formerly known as Spectrum Scale, General Parallel Filesystem (GPFS), or Tigershark
- Scale has evolved and grown over its 27+ year history
- Scale has 1000s of customers worldwide and has been used in some of the largest organizations
- Storage Scale isn't just for large customers; many Scale deployments are small to medium sized ranging in the 10TB -1PB capacity
- Storage Scale is not just a filesystem, Scale incorporates many data management capabilities, multi-protocol support, data virtualization, and much more!



IBM Storage Scale

Workloads



High Performance

- HPC filesystem
- AI training
- Big-Data and Analytics
- Research Computing



Enterprise

- Network Attached Storage (NAS)
- AI Inference
- GRID computing
- Project storage



Capacity

- Backup target
- Archival Storage
- HPSS

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

General Overview

Beyond the usual file system capabilities

Under the file system

- Use any block storage usable by the OS.
- Information Lifecycle Management features enable tiering – putting data on the most appropriate storage, as quickly as possible.
- Built-in encryption and compression
- Tape and cloud-based storage usable as tiers.
- Storage Scale RAID gives unrivaled data integrity.

Above the file system – since not everything will run Scale directly...

- Export the data as SMB shares, through NFS, or as Object storage.
- Transparently serve as HDFS NameNode and DataNodes for Hadoop or Spark applications.
- Container Storage Interface (CSI) extends storage access to containers.
- GPUDirect access to accelerate AI and analytics workloads.

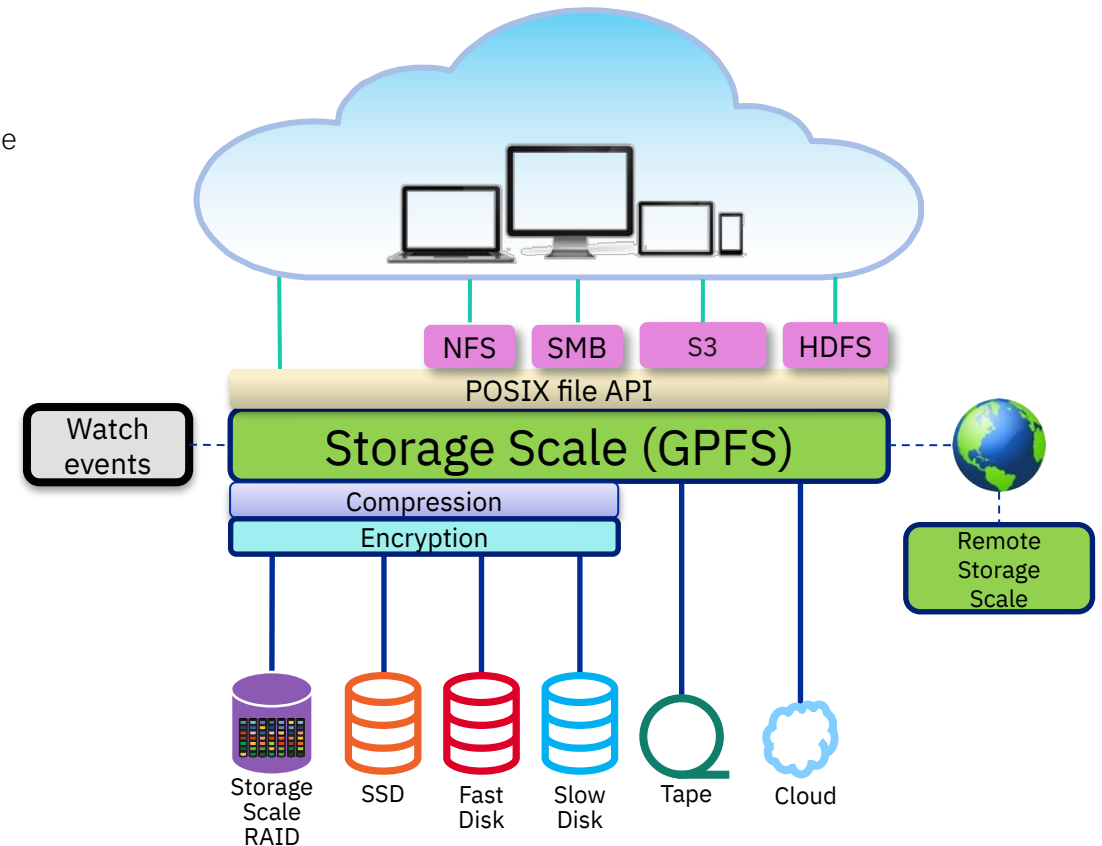
Beside the file system

- Peer with remote Storage Scale file systems for efficient remote access.
- Replication capabilities for robust Disaster-Recovery capabilities.
- Watch events in the file system with Watch Folders and File Audit Logging.

In a single name space

- Vastly simplify administration and reduce costs by eliminating storage islands.

Constantly getting better and faster!



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

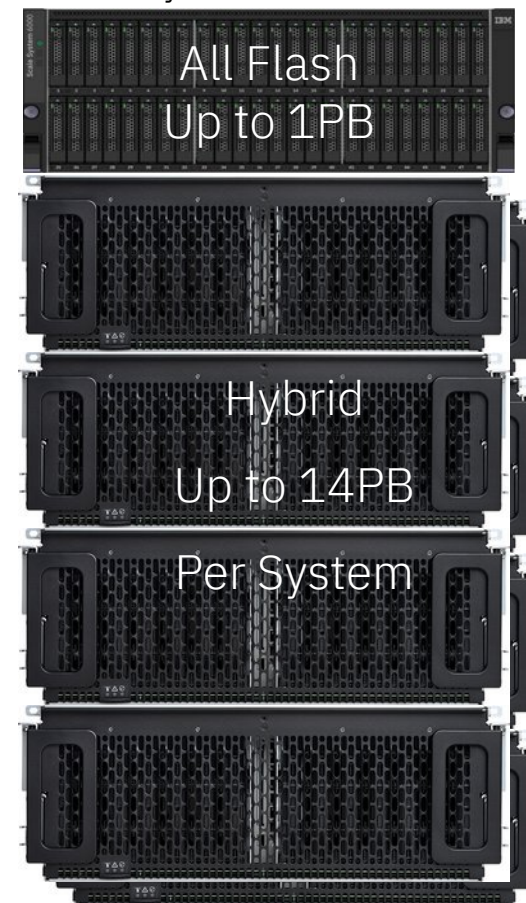
2 Storage Scale System

- Building block storage solution based on GNR
- NVMe, HDD, or Hybrid building blocks
- Highly Scalable – add more building blocks for more Capacity and Performance
- Up-to 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

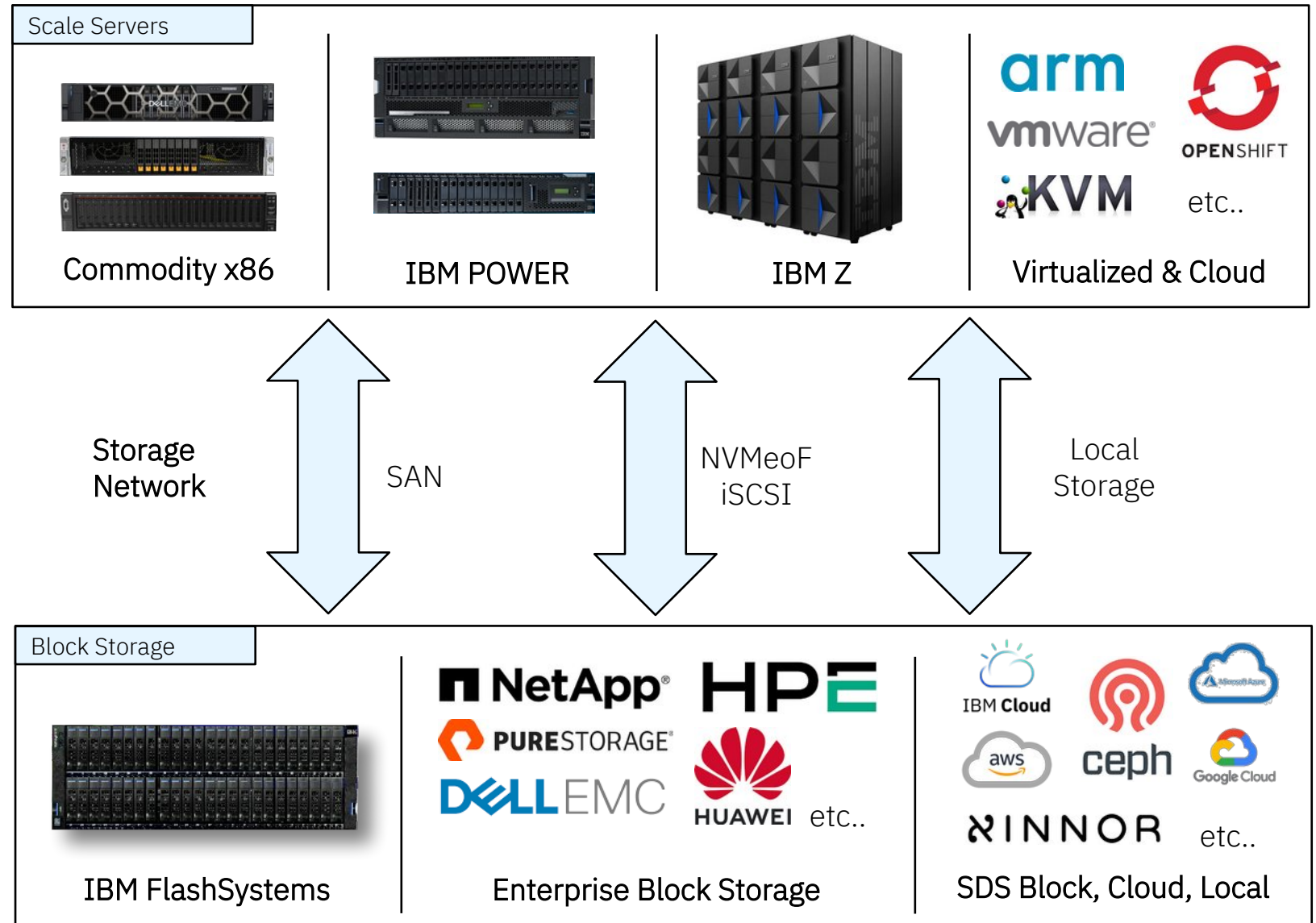


IBM Storage Scale

Scale is software, Deploy anywhere!

1. Software Defined Storage

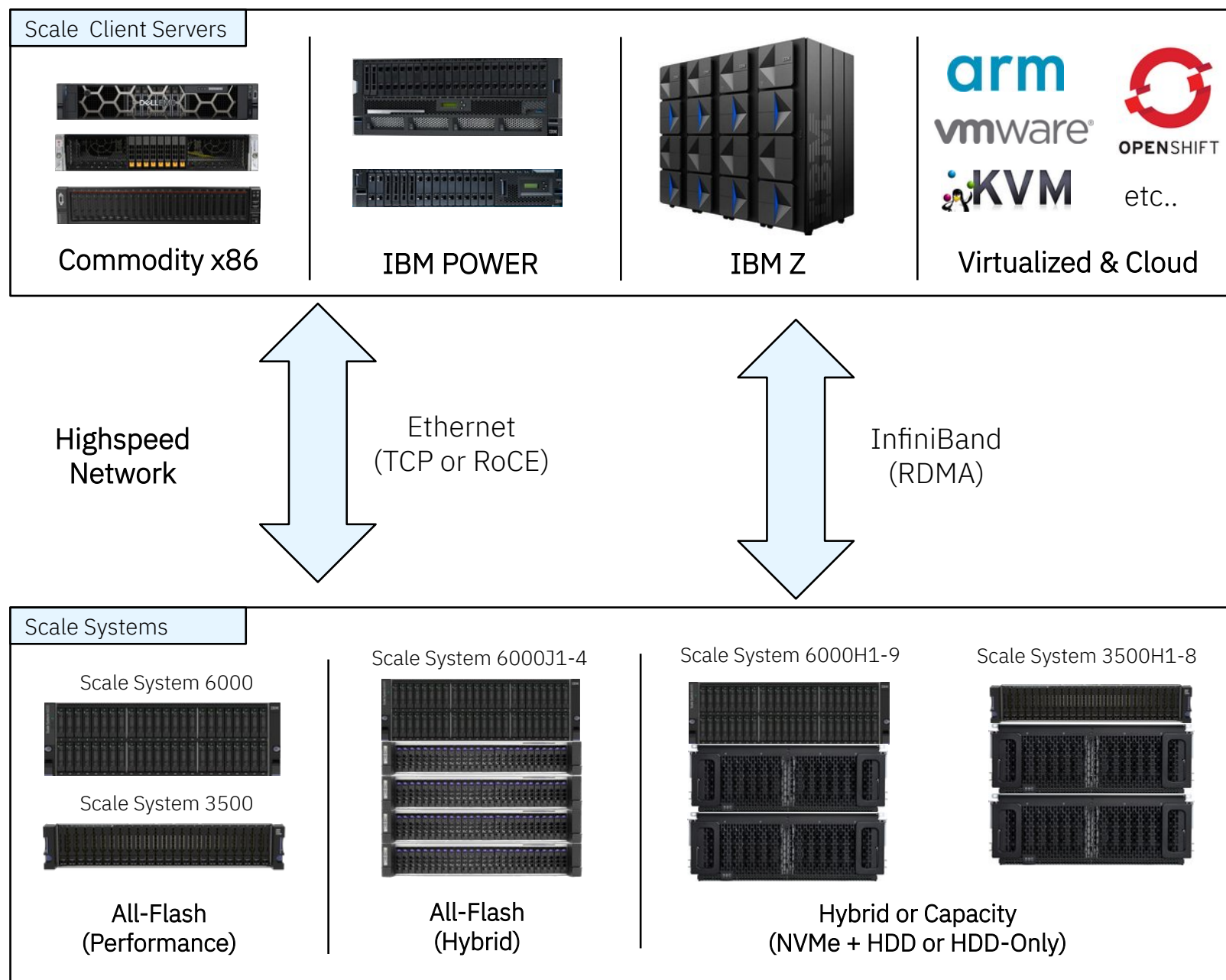
- Bring your own servers and storage
- Leverage existing storage and compute infrastructure
- Deploy On-Prem, Cloud, or Hybrid cloud
- Supports variety of system types x86, Power, s390x, arm
- Baremetal, VMs, Containers
- Highly Scalable and flexible!
- Add more storage and NSD servers, gain more performance and capacity
- Licensed per-TB with Data Access Edition (DAE) or Data Management Edition (DME)
- MB/s to TB/s in performance
- GBs to EBs in Capacity



IBM Storage Scale

2. Storage Scale System

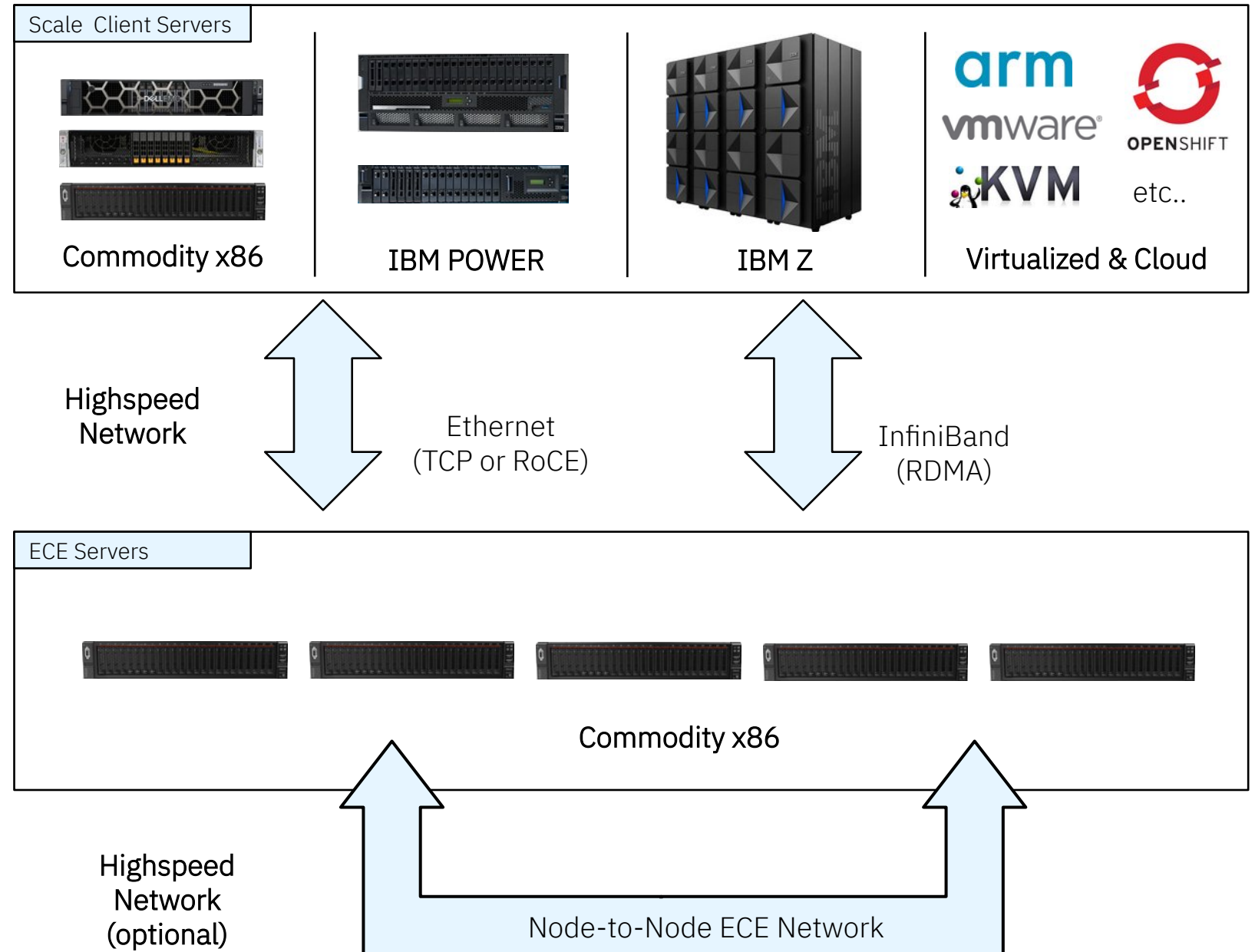
- IBM engineered “appliance” with tested and validated software stack
- Customizable building blocks with ability to expand over time
- Multiple media types supported including TLC, QLC and FCM4 flash and NL-SAS HDDs
- Erasure coded storage (GNR) supporting wide range of data protection algorithms
- Add more Scale Systems to gain capacity and performance
- Optional Premium Expert Care with Remote Code load
- Licensed per-TB with Data Access Edition (DAE) or Data Management Edition (DME)
- High performance 330GB/s in 4U



IBM Storage Scale

3. Erasure Code Edition (ECE)

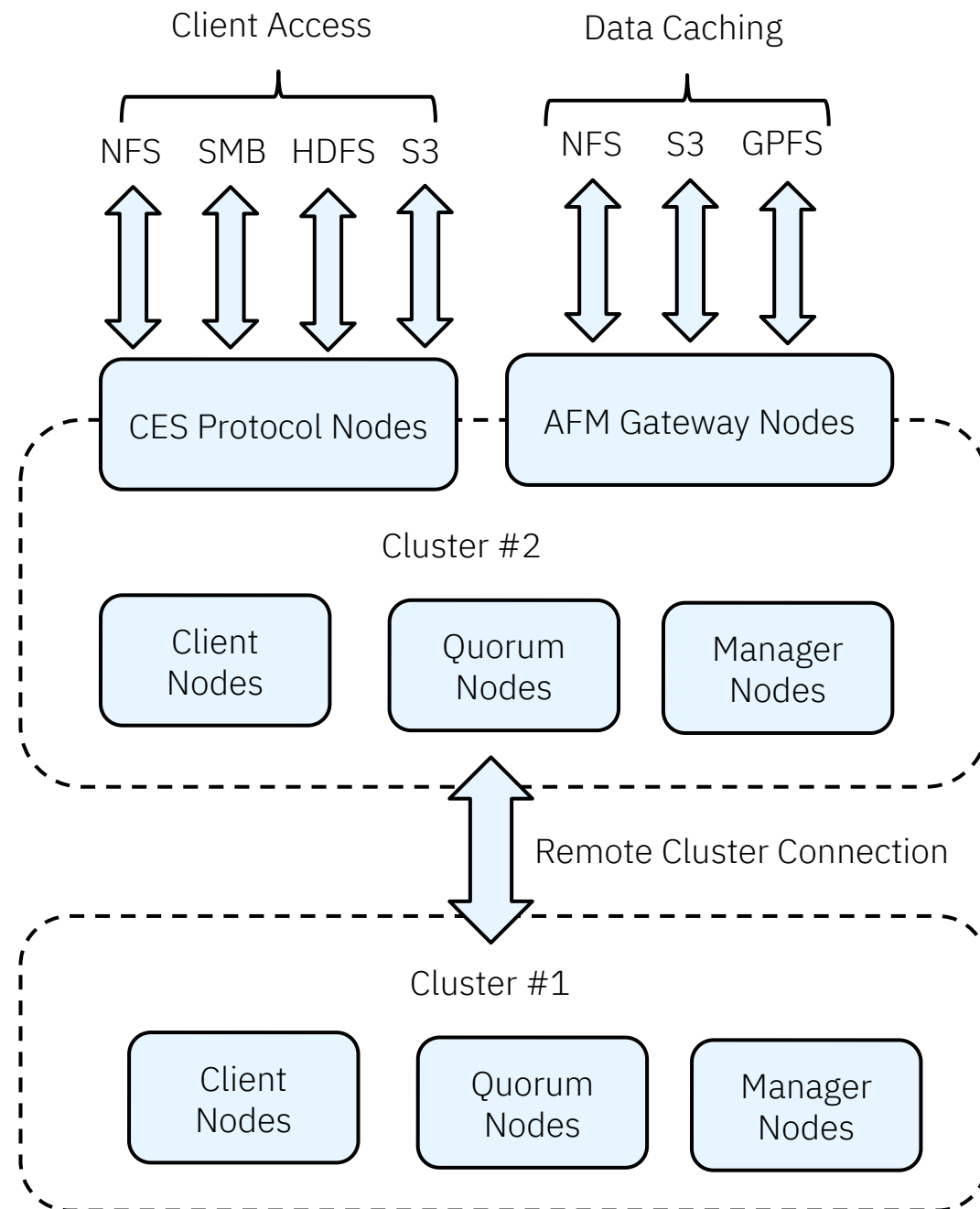
- Bring your own storage-rich servers
- Potentially leverage existing storage and compute infrastructure
- Deploy On-Prem, Cloud, or Hybrid cloud
- Baremetal or VMs (x86 only)
- Multiple media types supported including Flash and Spinning media
- Erasure coded storage (GMR) supporting wide range of data protection algorithms
- Add more storage-rich servers gain more capacity and performance
- Licensed per-TB with only one edition
- Highly flexible but some restrictions to consider:
<https://www.ibm.com/docs/en/storage-scale-ece/6.0.0?topic=edition-storage-scale-erasure-code-hardware-requirements>



IBM Storage Scale

Cluster Overview

- Fundamentally, a Storage Scale cluster is simply a group of nodes logically configured together
 - 1 to 16,000+ nodes
- Nodes in the Scale cluster can serve many roles
 - Manager Nodes
 - Quorum Nodes
 - Protocol Nodes
 - Gateway Nodes
 - Client Nodes
- All nodes in a cluster can access the same storage and are tightly coupled together
- There can be many filesystems in a single scale cluster
- A Scale cluster can share a filesystem with another cluster called a “Remote Cluster”
 - Common to have a “Storage Cluster” and 1 or many “Client clusters”

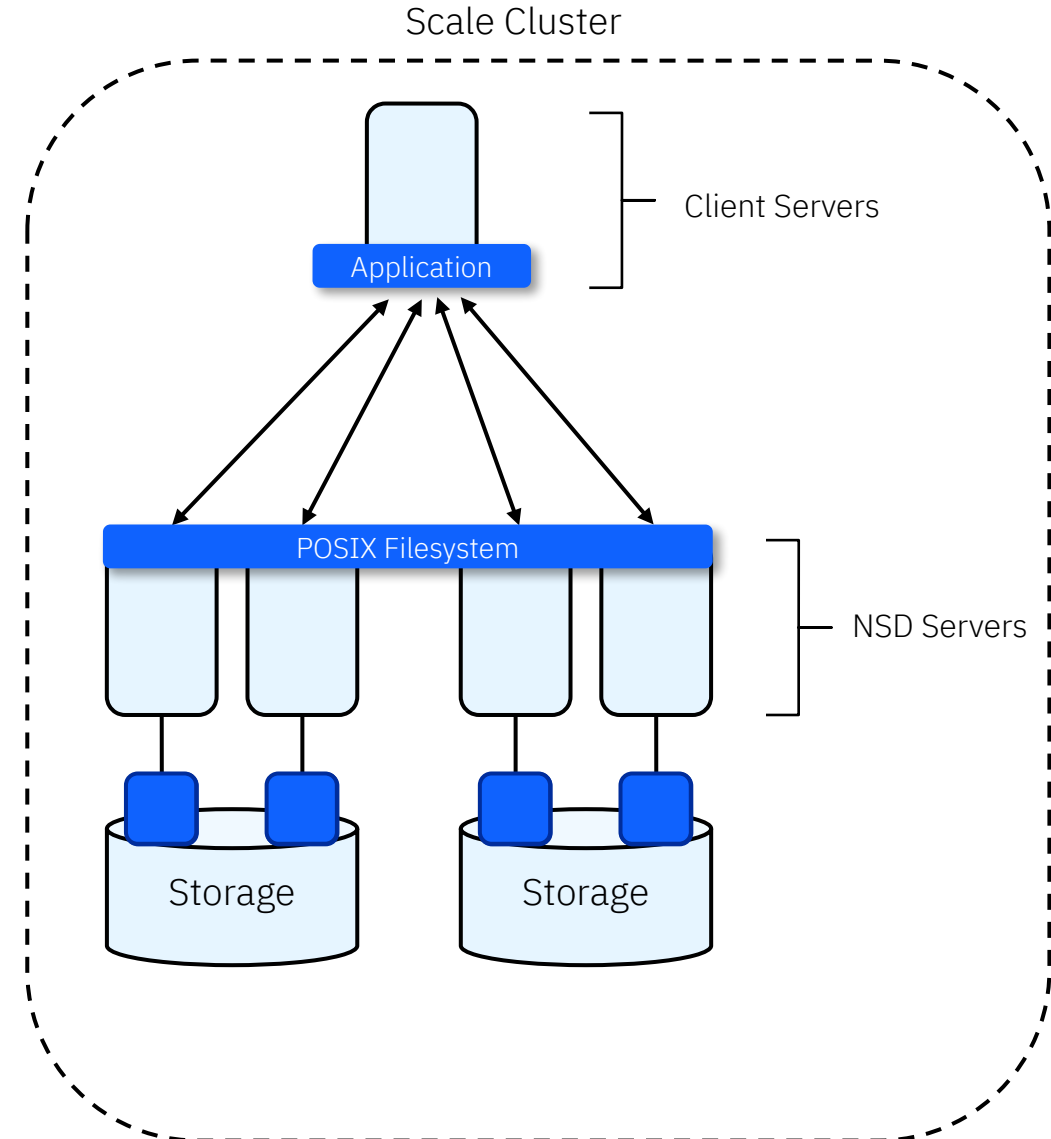


IBM Storage Scale

Parallel Architecture

Network Shared Disks (NSDs)

- An NSD is a virtualized piece of storage for use in Scale filesystem
- A Scale filesystem is made up many NSDs
- NSDs can be designated as dataOnly, metadataOnly, dataAndMetadata, or descOnly
- All NSD servers export to all clients in active-active manner
- IBM Storage Scale stripes files across NSD servers and NSDs in units of file-system block-size
- Easy to scale file-system capacity and performance while keeping the architecture balanced



IBM Storage Scale

Parallel vs. NAS

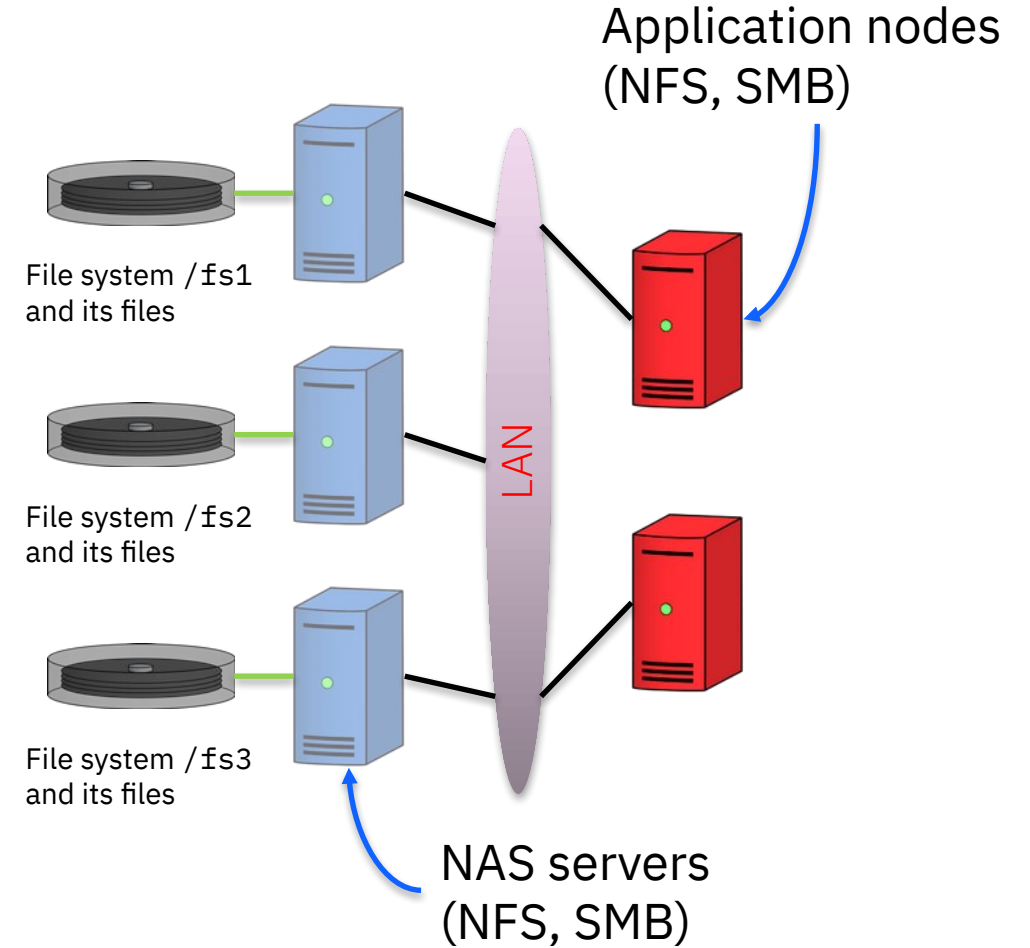
A traditional NAS file system is associated with a fixed set of disks.

When additional capacity is needed, new storage needs to go into *yet another new file system*.

This complicates storage management and limits potential gains in performance.

- Adding storage will not remove a “hot spot” file system unless workloads are actively moved to use the new storage – this is not automatic.

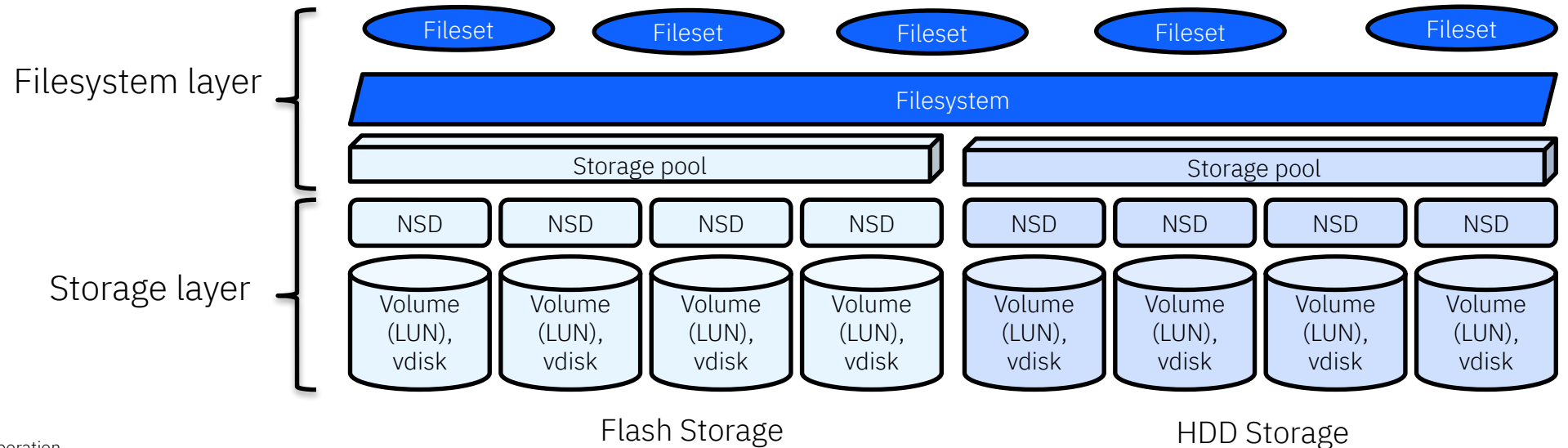
This is not a weakness with Scale...



IBM Storage Scale

Filesystem Fundamentals

- There can be multiple filesystems in a single Storage Scale cluster
- A filesystem is made up of 1 or more NSDs
- Filesystems can have multiple pools with NSDs of different types (NVMe, FCM, HDD, etc..)
- Filesystems have a blocksize and subblock size
- Filesets can be used to create “sub-filesystems” within a single filesystem for multitenancy
- Information Lifecycle Management (ILM) policies can move data between pools and control placement



IBM Storage Scale

Stretch Cluster Overview

This is a **single Storage Scale cluster** is configured using nodes and storage from two data centers.

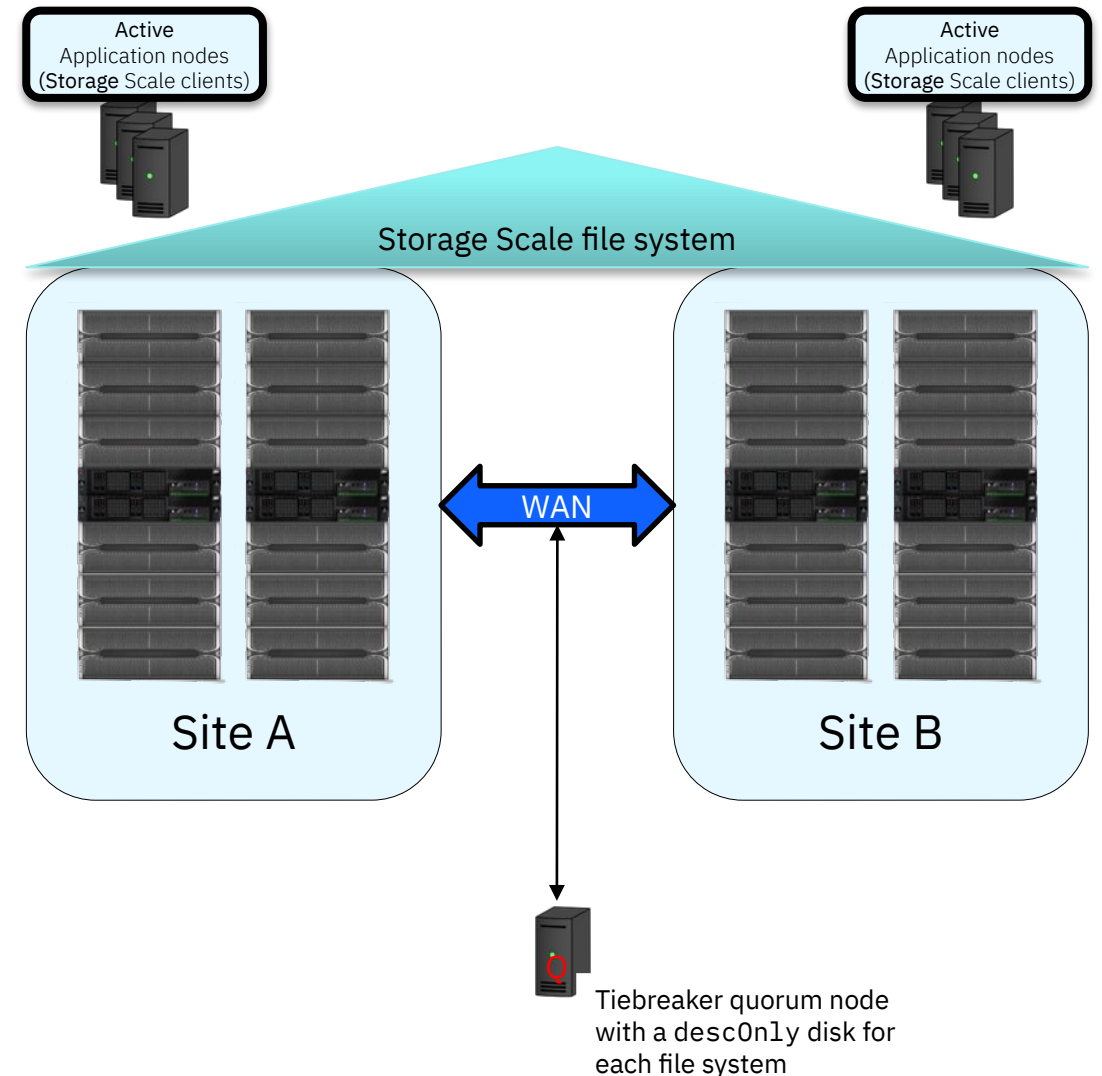
In other words, it is “stretched” between two sites, connected by a WAN.

File systems in such a cluster are available to systems at both sites and may be actively used concurrently by both sites.

File systems may judiciously use “failure group” replication to ensure both sites have a current instance of all the data.

Careful design can ensure one site remains active, even if the other site (or link) fails.

The result is an **active/active highly available** synchronously replicated Storage Scale file system.



IBM Storage Scale

Stretch Cluster Features

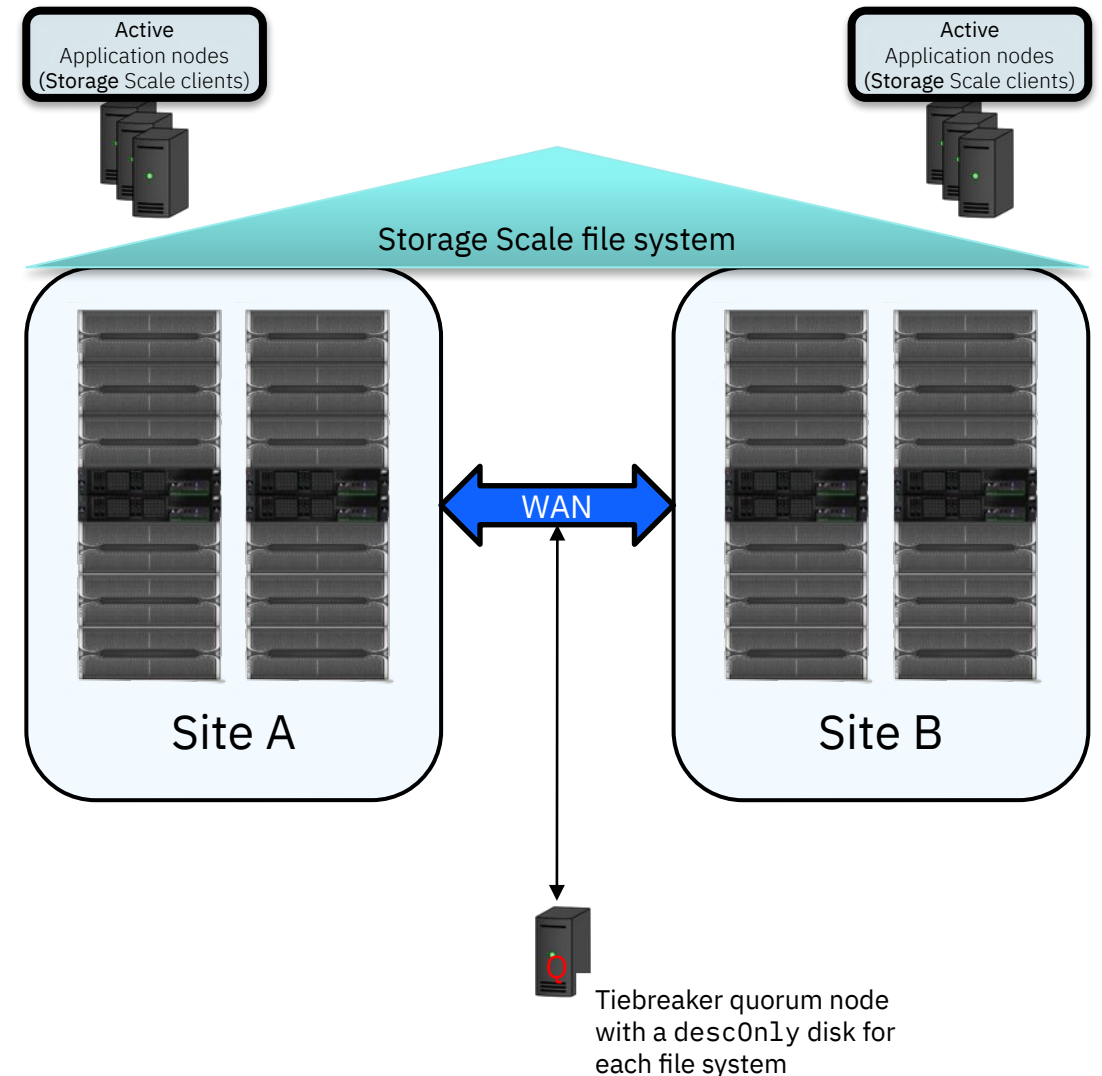
Active/active – Applications may work concurrently at both sites, on the same data and files, with complete “single system” consistency, with the same performance at each site.

High availability – after a site failure, applications at the other site may continue with an RPO of 0.

Reads are kept local – this reduces the latency otherwise imposed by the WAN.

Writes are synchronous – there is no possibility of updates getting out of order or the sites becoming inconsistent.

Recovery after failure is fast – Applications can begin working at both sites even while recovery is taking place.



IBM Storage Scale

Administrative Commands

Commands and binaries are stored in
/usr/lpp/mmfs/bin

All commands intended for customer use have the prefix **mm-**.
Generally, the form is:

mm{verb}{object}

{verb} is an action like [create](#), [add](#), [delete](#), [change](#), [list](#), ...
{object} is class of object of interest, like [cluster](#), [node](#), [nsd](#),
[disk](#), [file system](#), ...

Examples:

- **mm**[lscluster](#)
- **mm**[lsconfig](#)

New is the Native REST API and its `scalectl` command, an emerging role-based replacement for many **mm-** commands.

```
# export PATH=/usr/lpp/mmfs/bin:$PATH
```

```
# mmlscluster
```

```
GPFS cluster information
```

```
=====
```

```
GPFS cluster name:      slsr.local
GPFS cluster id:        14559237835526968155
GPFS UID domain:        slsr.local
Remote shell command:   /usr/bin/ssh
Remote file copy command: /usr/bin/scp
Repository type:        CCR
```

Node	Daemon node name	IP address	Admin node name	Designation
1	a-scale01	10.0.200.11	a-scale01	quorum-manager
2	b-scale01	10.0.200.21	b-scale01	quorum-manager
3	tiebreak	10.0.200.31	tiebreak	quorum
4	a-scale02	10.0.200.12	a-scale02	
5	a-scale03	10.0.200.13	a-scale03	
6	b-scale02	10.0.200.22	b-scale02	
7	b-scale03	10.0.200.23	b-scale03	

Did you know?

- The **mm-** prefix hearkens to Scale's origin as the "Multi-Media File System".
- Many internal commands start with the prefix **ts-**, for an earlier Tigershark File System (based on even earlier Shark). Don't use the **ts-** commands unless directed by support!

IBM Storage Scale

Multitenancy

Roles of Providers

Storage administrators provision resources, configure encryption, set quotas, and manage access controls to ensure security and efficiency

Tenant Responsibilities

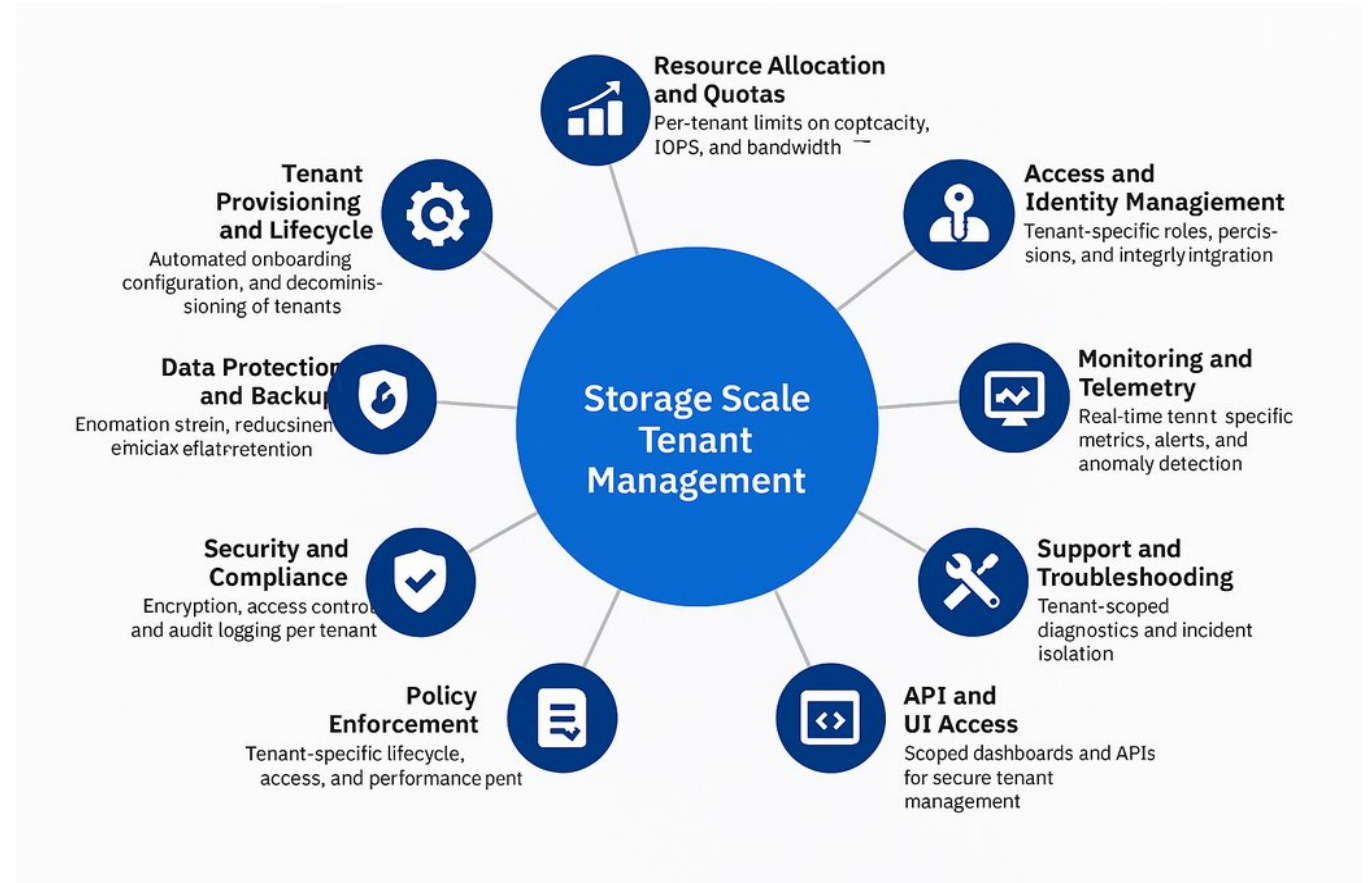
Tenant administrators handle data management, including ACLs, backups, and file-level restores, ensuring data integrity and availability.

Delegated Operations via REST API

The REST API roadmap expands delegated operations, enabling tenants to perform more independent tasks securely and efficiently.

Separation of Duties Benefits

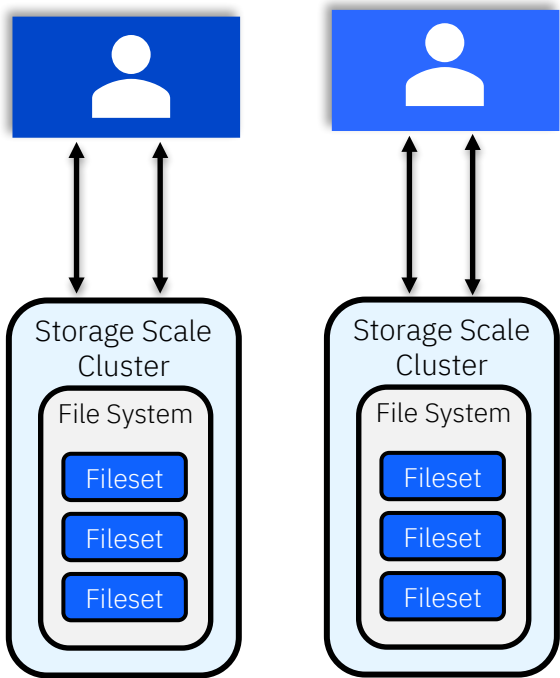
Clear separation between providers and tenants improves security and operational efficiency by assigning appropriate responsibilities.



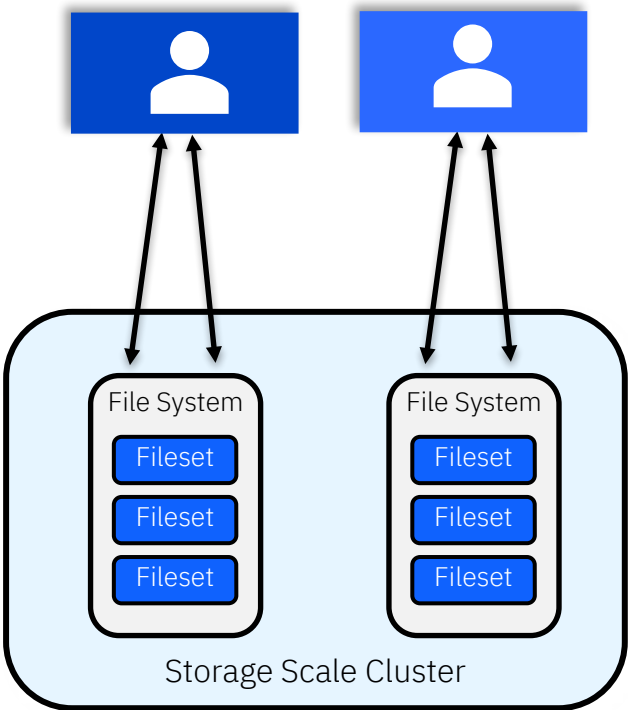
IBM Storage Scale

Multitenancy Architectures

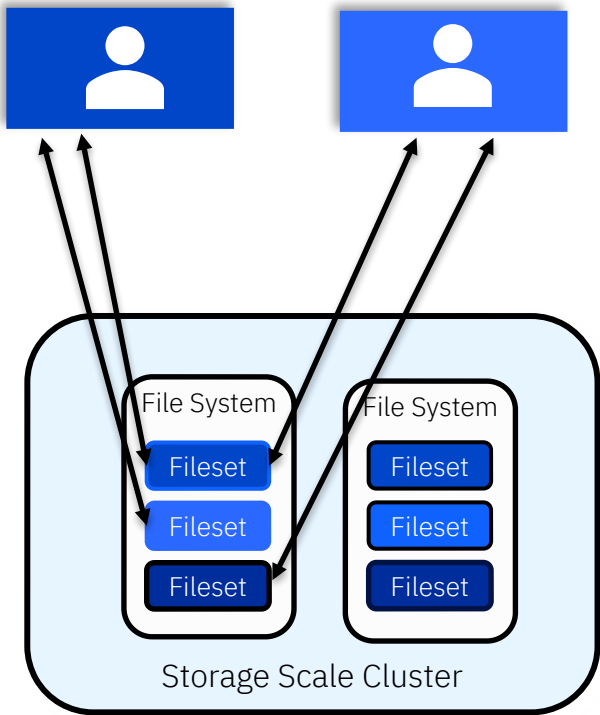
One Storage Cluster
per Tenant



One Scale Filesystem
per Tenant



One Scale Fileset
per Tenant



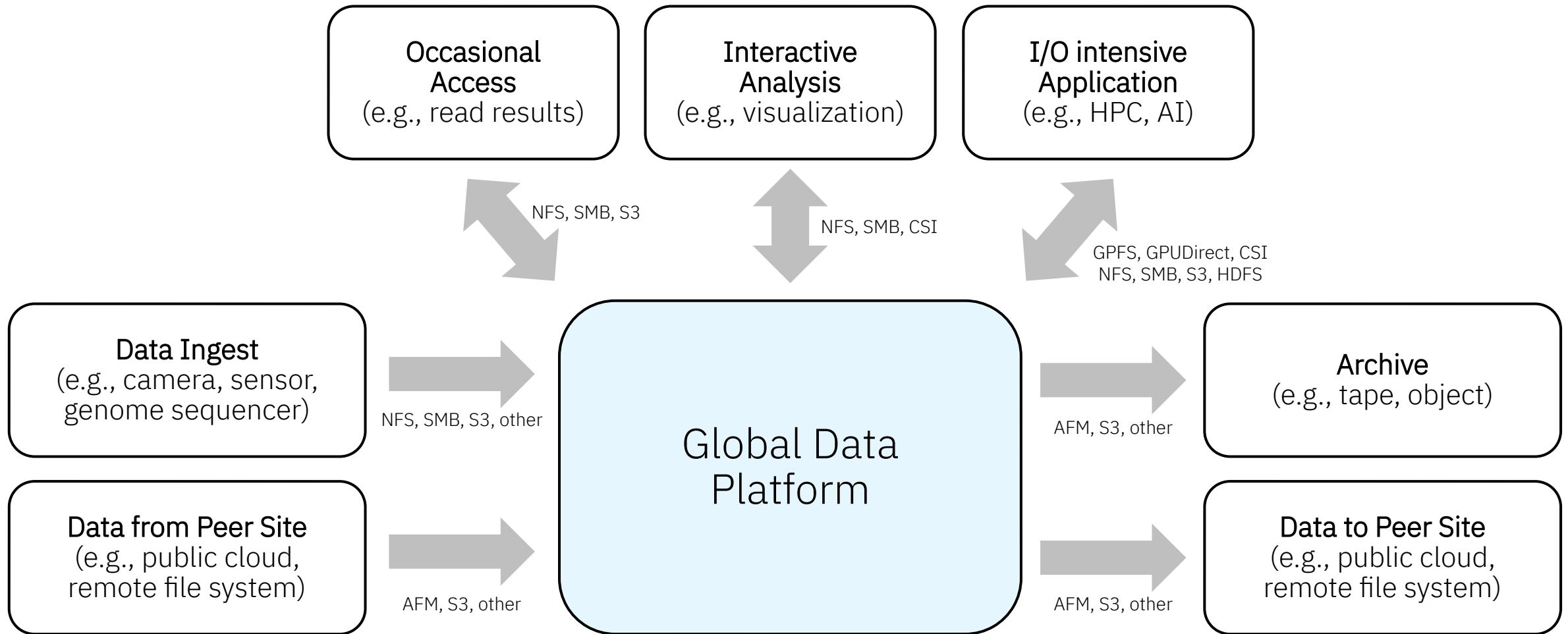
IBM Storage Scale

General Summary

- Storage Scale is software and can be deployed in many ways
- Storage Scale supports many node types, operating systems, networks, etc...
- Storage Scale is a high-performance filesystem that can grow to massive capacities however; not all deployments are massive or high performance – Scale can support many workload types
- Storage Scale is deployed in a cluster of 1 or more nodes
- NSDs are used to create filesystems and there can be mixed media types (NVMe, FCM, HDD, etc....) in a single filesystem leveraging ILM for placement and migration between pools
- Additional NSDs and NSD servers can be added to expand capacity and performance
- Filesystems can be partitioned into filesets for easier management and multi-tenancy



Storage Scale Multi-Protocol Access

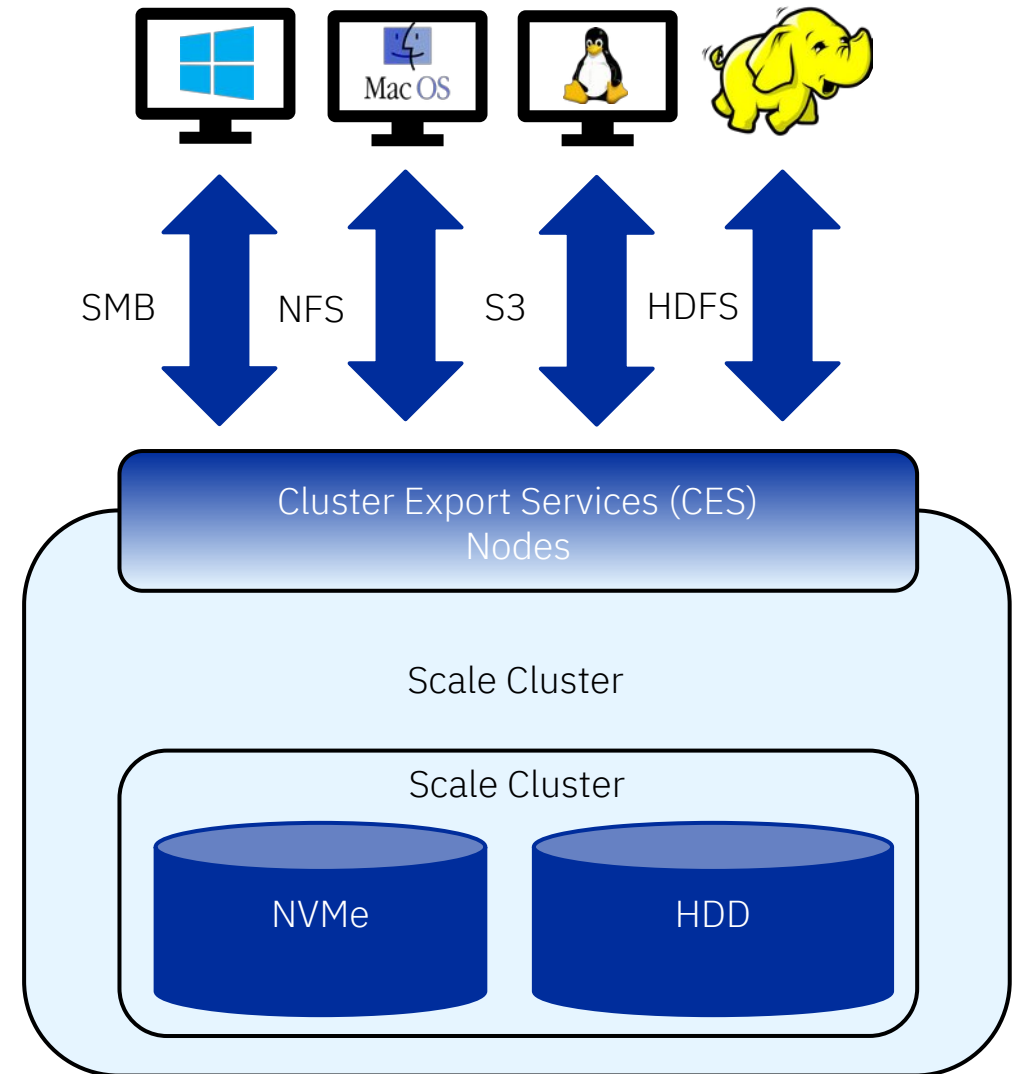


- Many unstructured data is generated and processed outside the data center.
- File and object protocols allows devices and applications to access and process data on remote servers and systems.
- Many of the remote applications and devices stick to one of the many file and object protocols (e.g., genome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

Why export Storage Scale through other protocols?

Storage Scale provides a unified POSIX name space with advanced capabilities and performance. However:

- There are application nodes that need data access running on an OS that doesn't support Storage Scale, e.g., MacOS, FreeBSD, Solaris.
- There are application nodes running software that require NFS, SMB, HDFS, or S3 protocols, e.g., web services.
- In many cases, we would like to be able to access the same data both directly (through Storage Scale code) and through one of the other protocols – we do not want to create separate storage silos for the different protocols.



IBM Storage Scale

Protocols Overview

- Storage Scale allows for high performance multi-protocol access in addition to the native POSIX Scale Client
- Not every application is suited for the the Scale client such as Windows environments, edge systems, IOT, peer-site access, etc...
 - In fact, some of Scale's largest customers access Scale solely through multi-protocol access leveraging the best of Scale's filesystem capabilities on many client-less nodes such as large-scale enterprise NAS
- Storage Scale allows for the core filesystem to be access through NFS, SMB, S3, HDFS along with the POSIX client accessible through GPU Direct storage and within container environments such as OpenShift and Kubernetes
- Storage Scale allows for simultaneous cross protocol access
 - A Windows client can mount a Scale filesystem via SMB
 - A Linux client can mount via NFS
 - An AI or HPC cluster can mount with the POSIX Scale Client
- **All clients see the same data simply through a different file protocols**

IBM Storage Scale

How to Export Protocols

NFS

- Cluster Export Services (CES) – Ganesha (user space)
- NFS, configured with /etc/exports
- Cluster NFS (cNFS) – “stabilized” so avoid

SMB

- Cluster Export Services (CES) – Samba
- 3rd party, e.g., Tuxera

S3

- Cluster Export Services (CES) – S3
- ~~Move away from older CES Object (Swift) and DAS~~

HDFS

- Cluster Export Services
- Scale’s HDFS transparency outside CES control

Containers

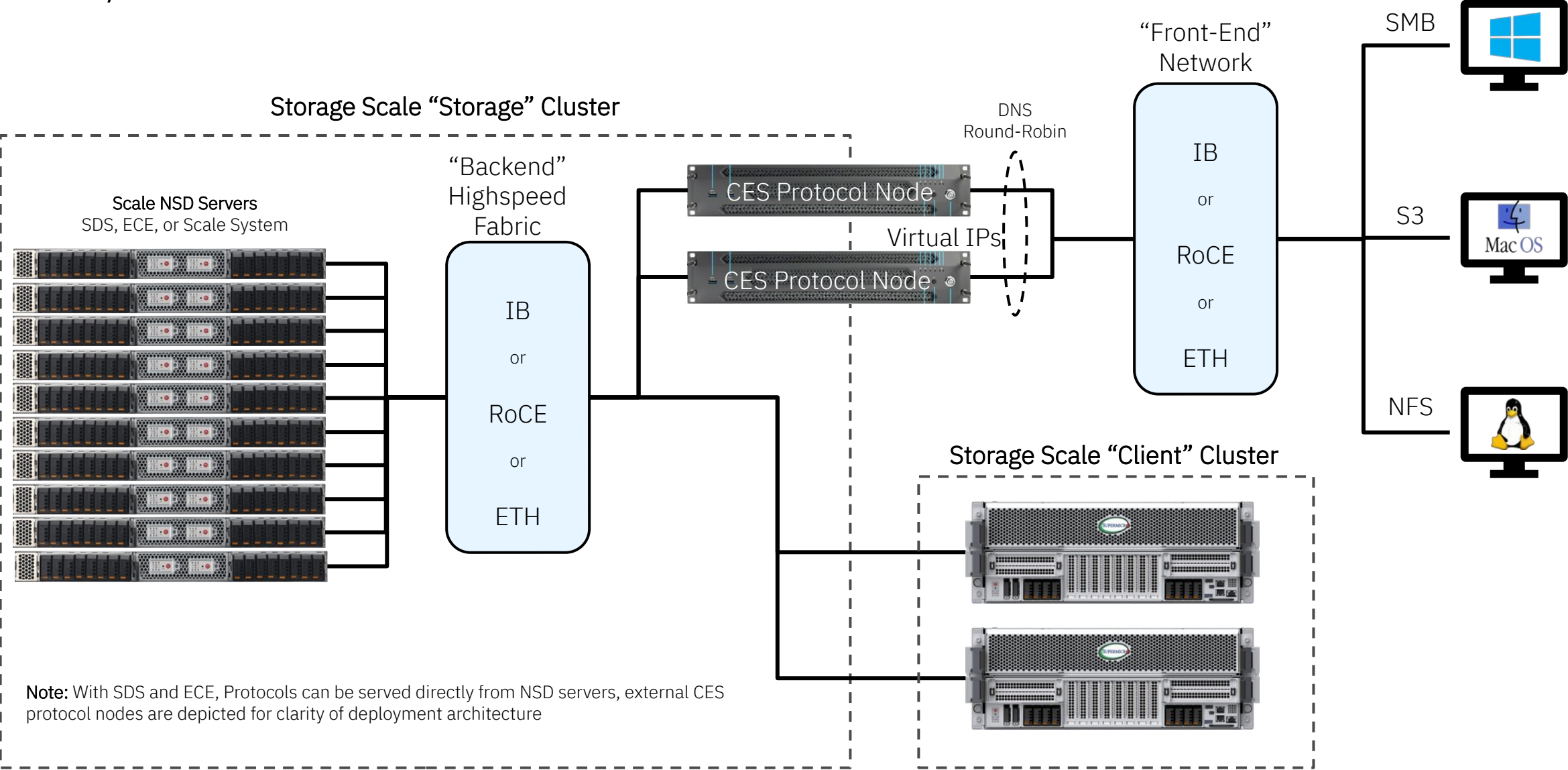
- Container Storage Interface (CSI)
- Container Native Storage Access (CNSA)

GPU Direct

- Use NVIDIA CUDA and RDMA networking

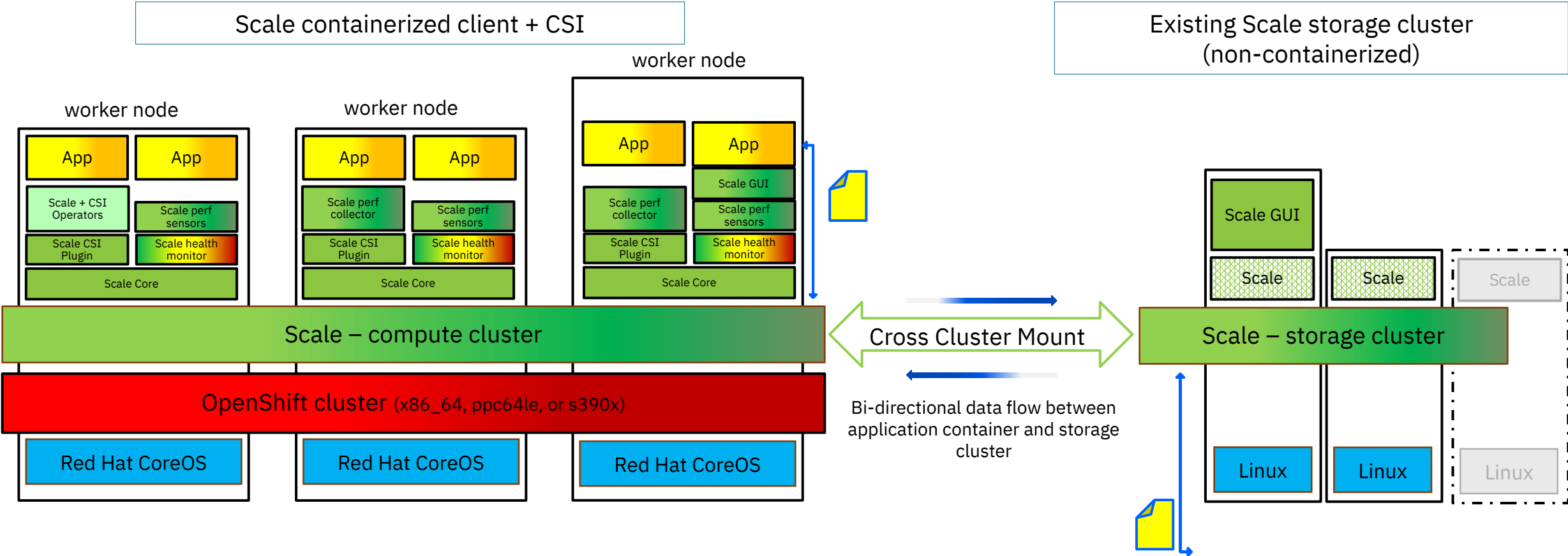
IBM Storage Scale

Example Protocols Architecture



IBM Storage Scale

CNSA Architecture

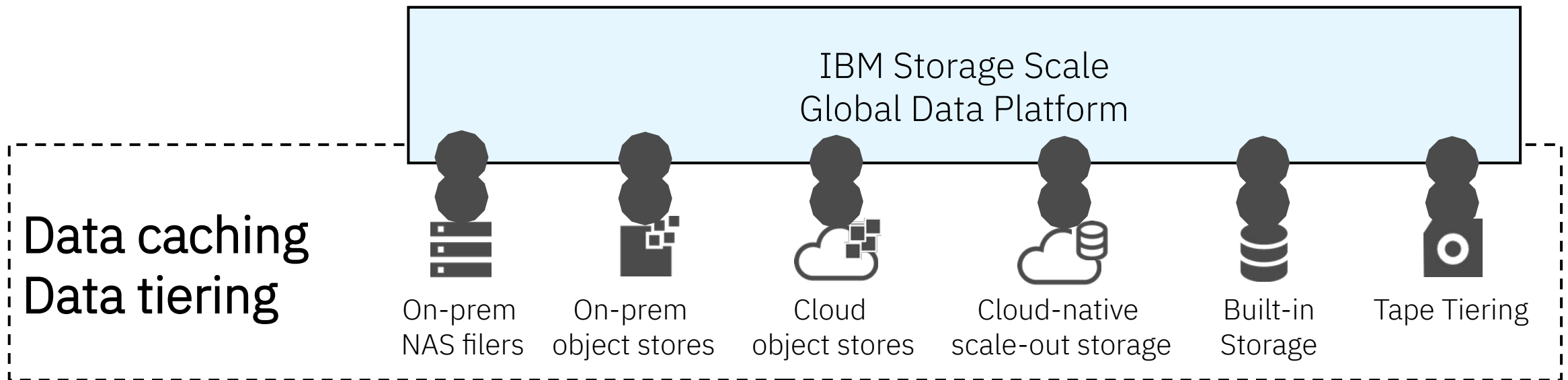


IBM Storage Scale

Data Abstraction and Tiering

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

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

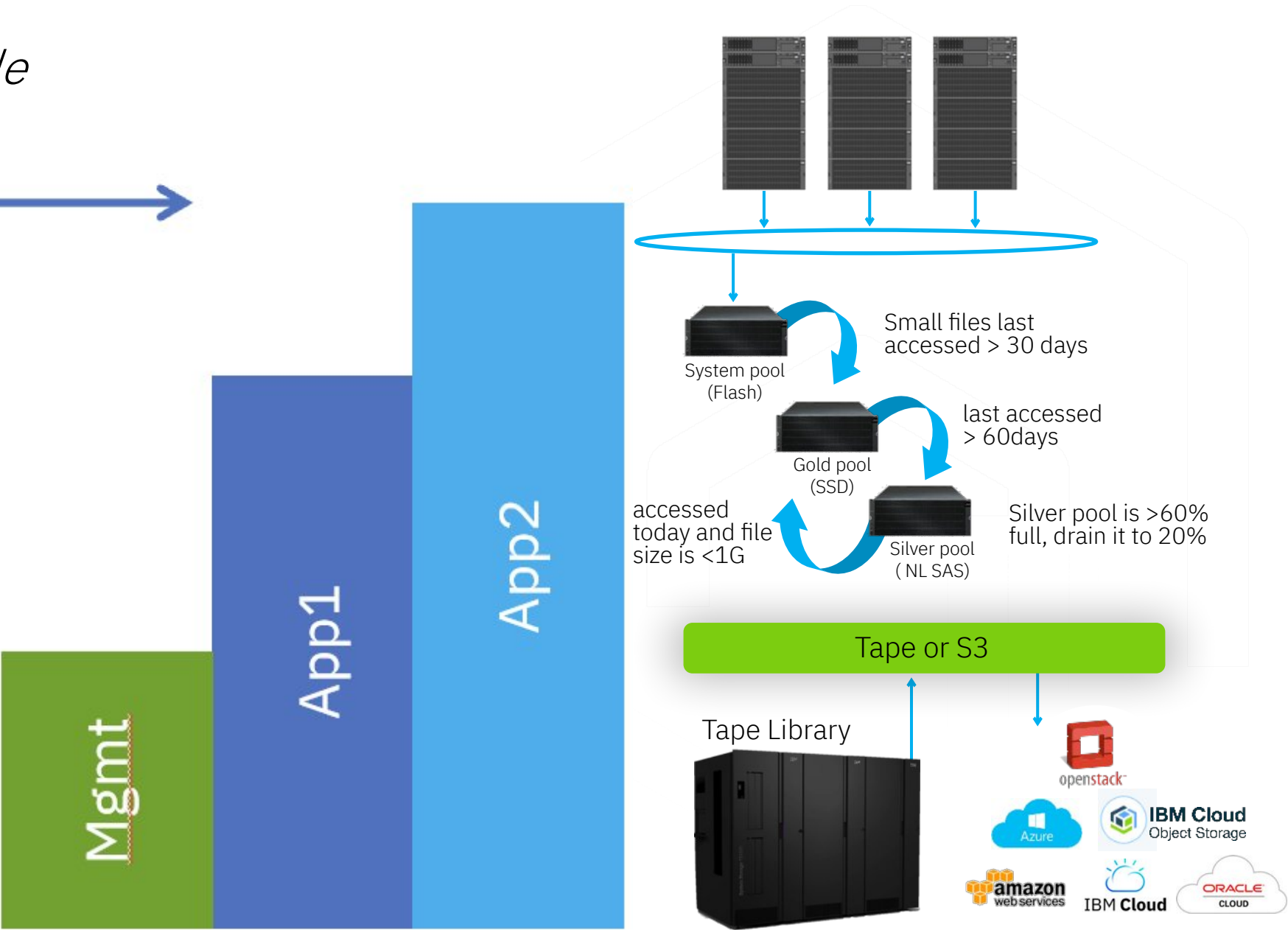


IBM Storage Scale

Data Tiering and Lifecycle

Total IO Capability
B/W or IOPS

Management
Capped at 20%
(example)



IBM Storage Scale

AFM Overview

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

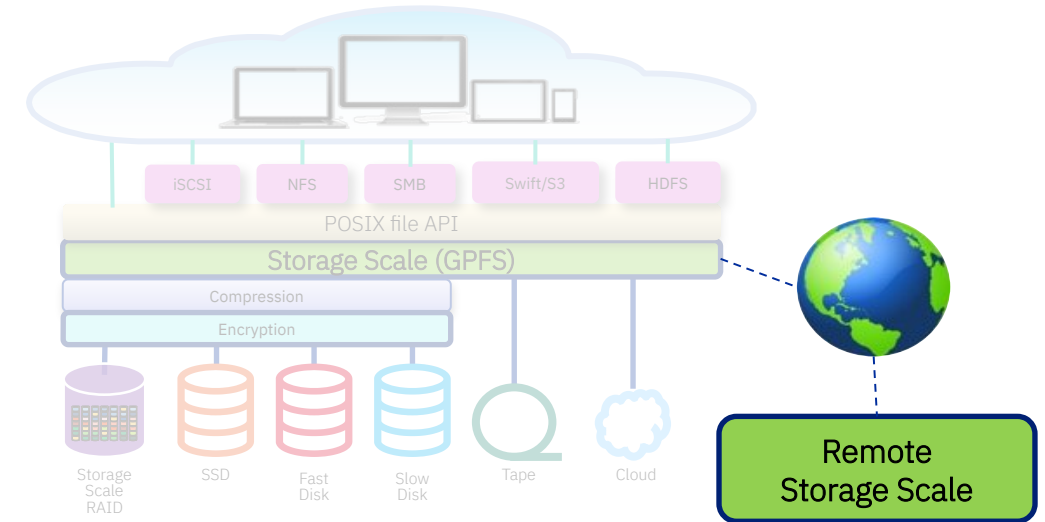
- Tolerates unreliable, high-latency networks (like a WAN).
- *Caches* copies of data from a remote file system into the local Storage Scale cluster.
- Cached files have the same read and write performance as other local files.
- However, there is no locking between different caches.

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

- Unifies heterogeneous remote storage

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

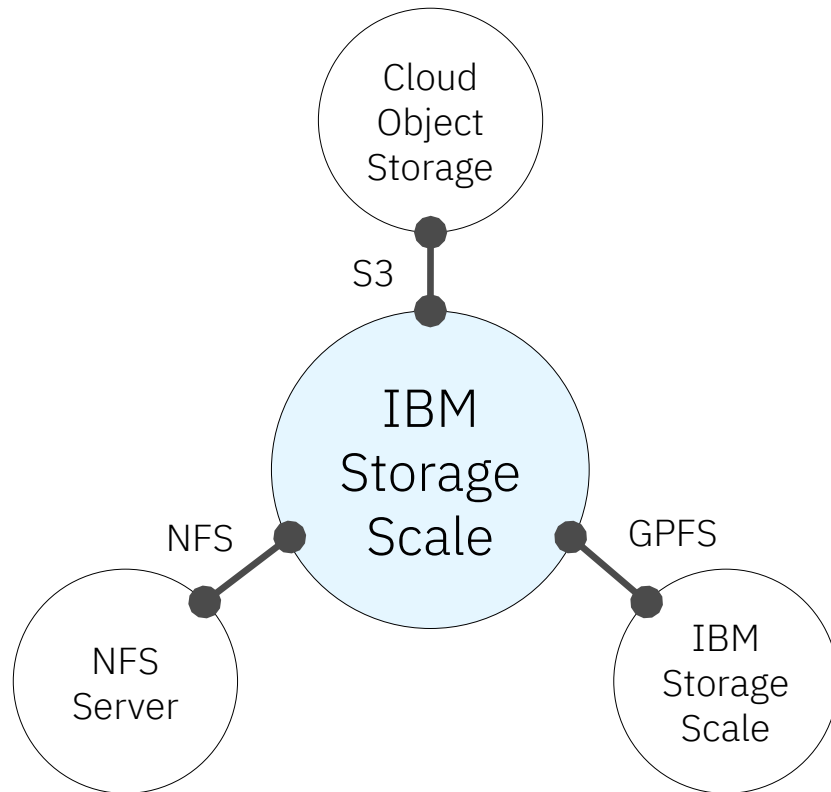
- Bidirectional awareness for failover and failback with data integrity
- Recovery Point Objectives (RPO) for application consistency points



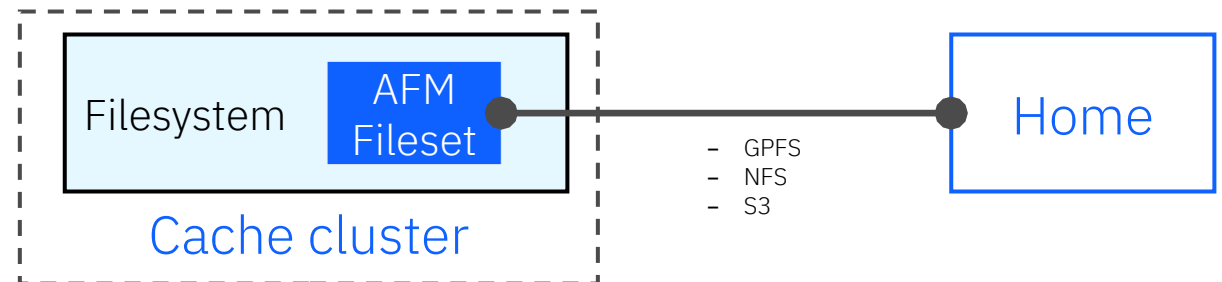
IBM Storage Scale

Active File Management (AFM)

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.



IBM Storage Scale

Monitoring

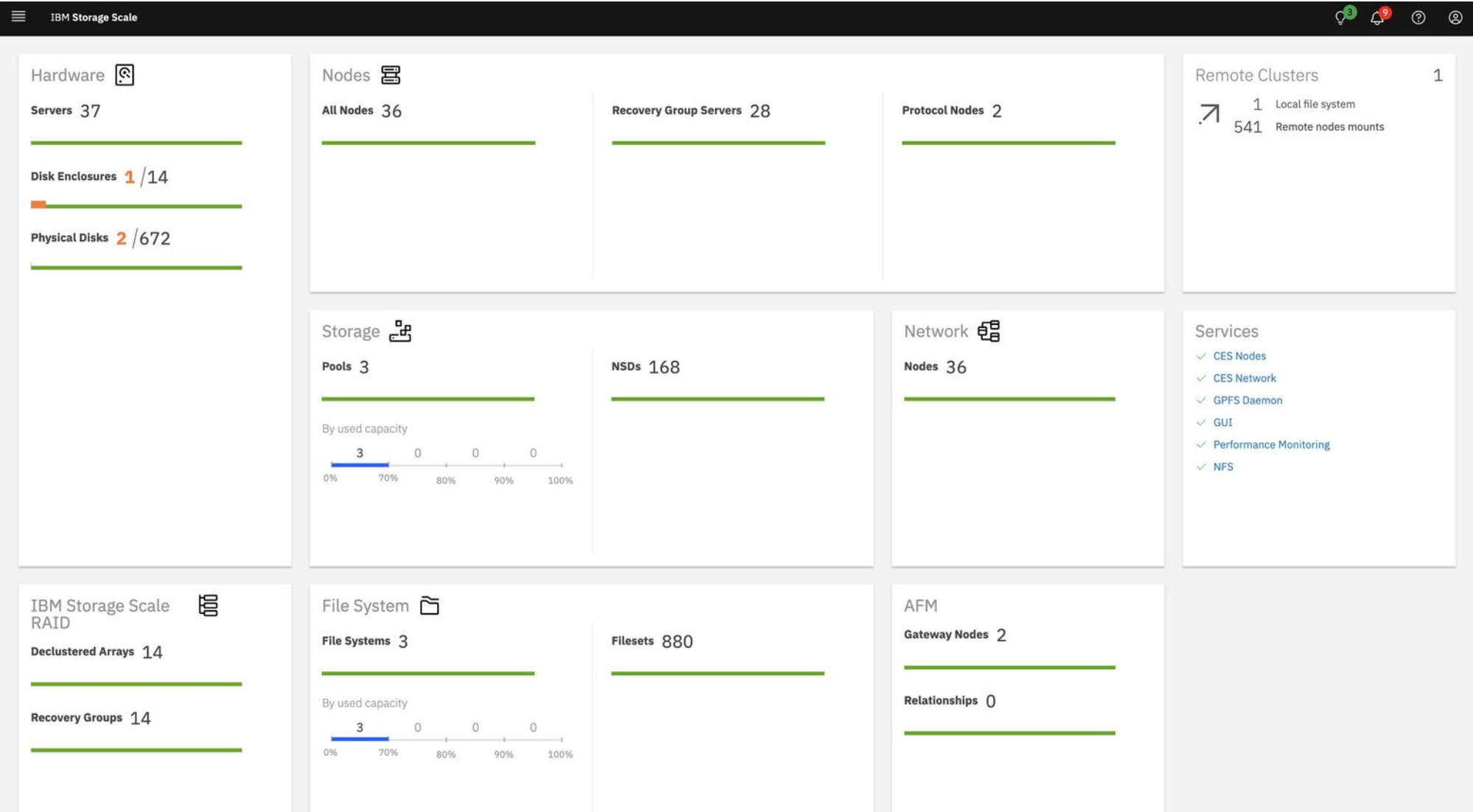
- The **mmhealth** command allows health to be monitored from the command line.
- Every node is running the mmsysmon process to monitor health and log events.
- The GUI can automatically send reports to administrators about ESS and Storage Scale health.
- Storage Scale has a collection of callbacks that can be configured to take action (like email) upon “interesting” events.
- Storage Scale bridge to Grafana

```
[root@swat19 ~]# mmhealth cluster show
```

Component	Total	Failed	Degraded	Healthy	Other
NODE	11	0	1	10	0
GPFS	11	0	1	10	0
NETWORK	11	0	0	11	0
FILESYSTEM	4	0	0	4	0
DISK	16	0	0	16	0
AFM	14	0	0	14	0
CALLHOME	1	0	0	1	0
CES	2	0	0	2	0
CESCLUSTER	1	0	0	1	0
FILEAUDITLOG	11	0	0	10	1
FILESYSMGR	4	0	0	4	0
GUI	1	0	0	1	0
HEALTHCHECK	1	0	0	1	0
NATIVE_RAID	3	0	1	2	0
PERFMON	11	0	0	11	0
THRESHOLD	11	1	0	10	0

IBM Storage Scale

GUI



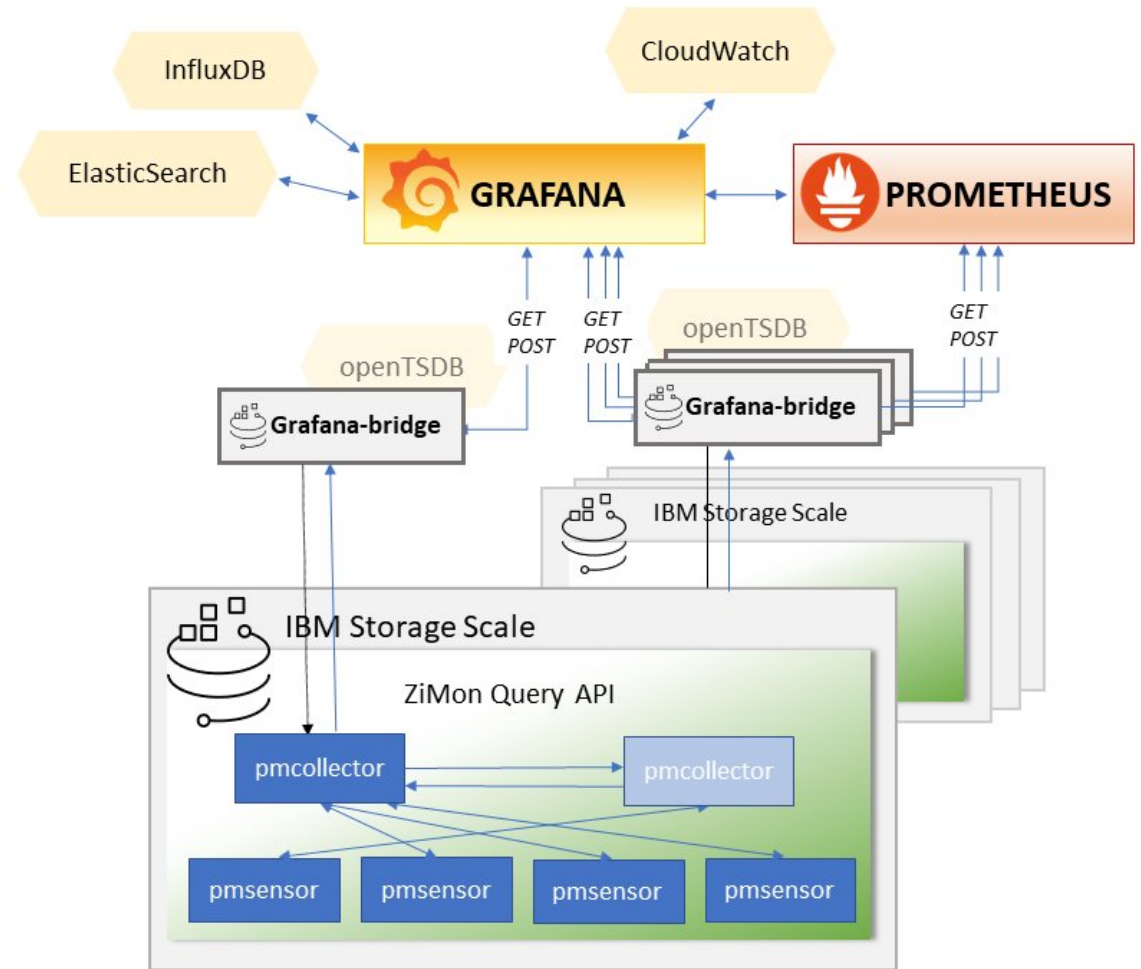
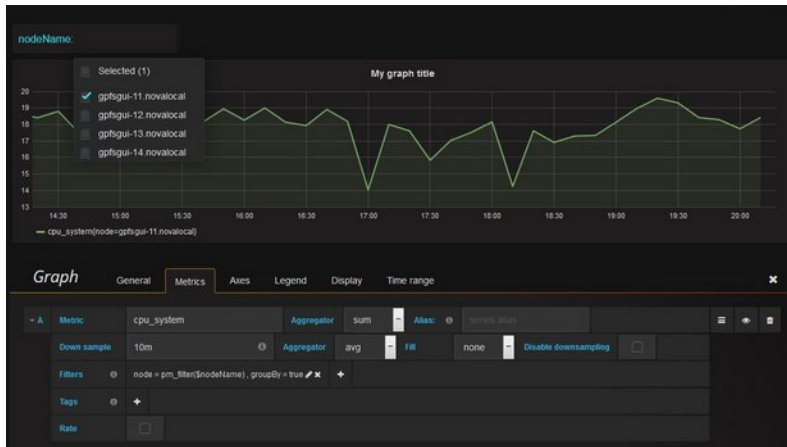
IBM Storage Scale

Grafana Bridge

Many organizations already use Open Source tools such as [Grafana](#) to visualize current metrics from many sources.

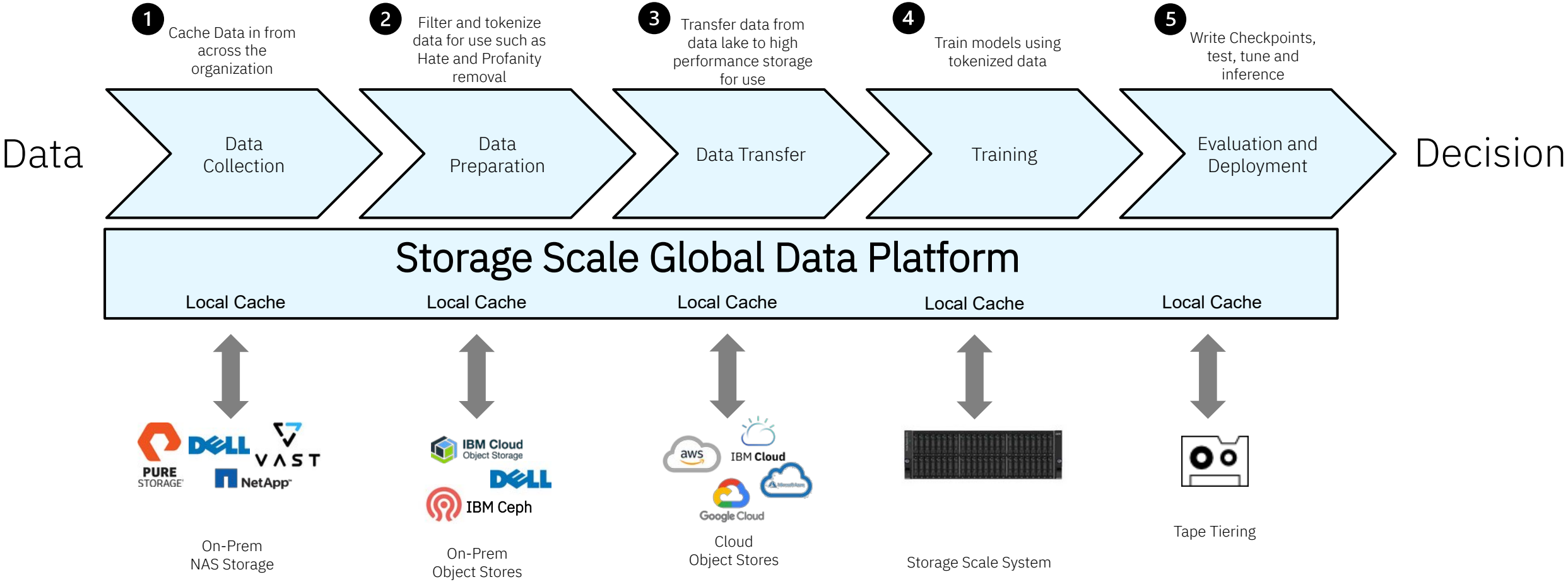
The Grafana bridge is provided by IBM as Open Source software at: github.com/IBM/ibm-spectrum-scale-bridge-for-Grafana

This exports the performance monitor data as an [openTSDB](#) provider.



Storage for Data and AI

Unified Data Platform for AI



Cache Data as needed from IBM and non-IBM Storage

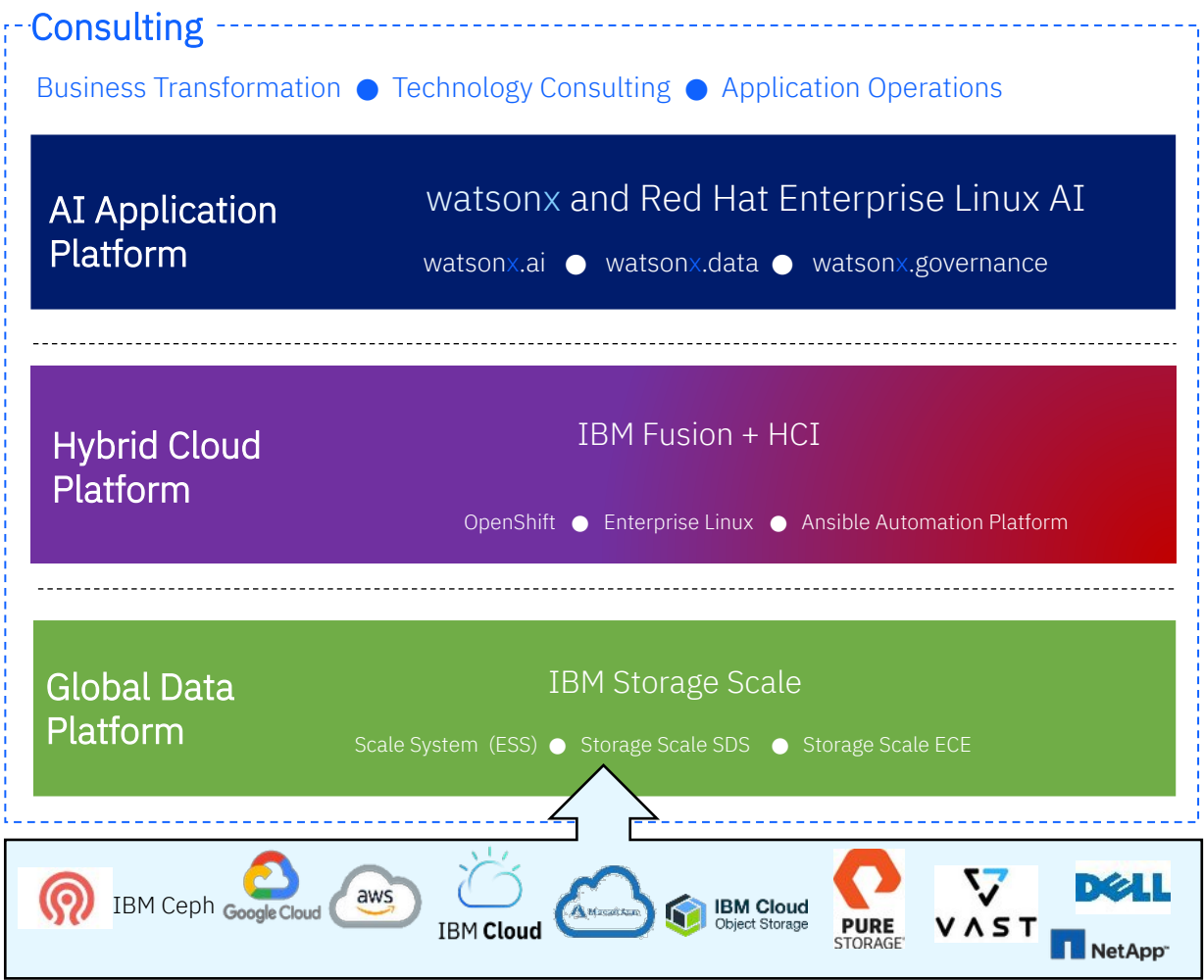
IBM's Enterprise Data and AI Solution

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 Data & AI Solution



IBM Fusion HCI



watsonx



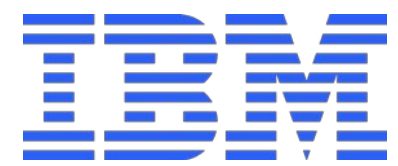
OPENSIFT



NVIDIA

IBM Scale System 6000





IBM Storage Scale

Creating a Global Data Platform

- **Access** data regardless of the protocol and client architecture or location
- **Accelerate** IO with advanced parallel data access
- **Abstract** customers existing storage infrastructure on-pre, cloud, and hybrid
- **Awareness** to datatypes, locations, and file contents leveraging AI
- **Assure** data is secure as it flows through the entire platform following NIST

Global Data Platform

