ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: Storage Scale 101 – The Basics and Architecture

Speaker:
Lindsay Todd, PhD
Principal Storage Technical Specialist – ATG Storage Team
lindsay@us.ibm.com

Accelerate with ATG Technical Webinar Series

Advanced Technology Group experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with ATG webinars - Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

2025 Upcoming Webinars – Register Here!

Tape Considerations with IBM Z Cyber Vault - June 17th, 2025

Quantum-Safe SAN Security: Future-Proofing your IBM Storage Networks with Brocade Gen 7 Security – June 26th, 2025

Important Links to Bookmark:

Accelerate with ATG - Click here to access the Accelerate with ATG webinar schedule for 2025, view presentation materials, and watch past replays dating back two years. **https://ibm.biz/BdSUFN**

ATG MediaCenter Channel - This channel offers a wealth of additional videos covering a wide range of storage topics, including IBM Flash, DS8, Tape, Ceph, Fusion, Cyber Resiliency, Cloud Object Storage, and more. **https://ibm.biz/BdfEgQ**

Offerings

Client Technical Workshops

- **Cyber Resilience with IBM Storage Defender: July 16 (Virtual)**
- IBM Fusion & Ceph
- IBM Storage Scale & Storage Scale Functions
- IBM DS8000 G10 Advanced Functions
- IBM FlashSystem Deep Dive & Advanced Functions

TechZone Test Drive / Demo's

- IBM Storage Scale and Storage Scale System GUI
- IBM Storage Virtualize Test Drive
- IBM DS8900F Storage Management Test Drive
- Managing Copy Services on the DS8000 Using IBM Copy Services Manager Test Drive
- IBM DS8900F Safeguarded Copy (SGC) Test Drive
- IBM Cloud Object Storage Test Drive - (Appliance based)
- IBM Cloud Object Storage Test Drive - (VMware based)
- IBM Storage Protect Live Test Drive
- IBM Storage Ceph Test Drive - (VMware based)

Please reach out to your IBM Representative or Business Partner for more information.

***IMPORTANT* The ATG team serves clients and Business Partners in the Americas, concentrating on North America.**

Announcing the 2025 IBM TechXchange Conference

Our theme this year is simple but powerful: **GO / Explore. Build. Launch.**

IBM TechXchange 2025
October 6-9 | Orlando, FL
The learning event for developers & technologists on a mission
EXPLORE.
BUILD.
LAUNCH.
GO/

For more information, please visit - https://www.ibm.com/community/ibm-techxchange-conference/

Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code **5151 0447**
or

Direct link https://www.menti.com/alhsf3bgvxu6
or

QR Code

ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: Storage Scale 101 – The Basics and Architecture

Speaker:
Lindsay Todd, PhD
Principal Storage Technical Specialist – ATG Storage Team
lindsay@us.ibm.com

Meet the Speaker

-

Lindsay Todd – As a Principal Storage Technical Specialist with IBM's Advanced Technology Group (ATG) Storage team, Lindsay provides deep technical expertise with Storage Scale (GPFS). Lindsay earned his PhD in Computer Science from Rensselaer Polytechnic Institute, where, as a systems programmer he supported research computing and high-performance computing, heavily used GPFS (now Storage Scale), and was an adjunct Computer Science faculty. His curiosity continues as he explores and uses Storage Scale for ATG infrastructure needs, creates and tests innovative architectures for emerging problems, and helps clients understand it and use it to build solutions to their unique business problems.

Email: lindsay@us.ibm.com
Linked-in: https://linkedin.com/in/rltodd

Agenda

- Storage Scale: the Global Data Platform
- Storage Scale: a General Parallel File System
  - Core capabilities of the Scale file system
  - Below: Storage pools and policies
  - Above: Accessing Scale through non-POSIX protocols
  - Beside: Remote access and caching
  - Introspection: Management, monitoring, and auditing
- Storage Scale System
- Conclusion

Storage Scale

Storage Scale

Storage Scale: the Global Data Platform

… a Global Platform For Storage

Storage Scale – IBM's flagship (file) storage for unstructured data

Pushing the limits (over 7 years ago...)

2.5 TB/sec single stream IOR
as requested from Oak Ridge National Lab (ORNL)

- 1 TB/sec 1MB sequential read/write
  as stated in CORAL RFP
- Single Node 16 GB/sec sequential read/write
  as requested from ORNL
- 50K creates/sec per shared directory
  as stated in CORAL RFP
- 2.6 Million 32K file creates/sec
  as requested from ORNL

CORAL
COLLABORATION
OAK RIDGE • ARGONNE • LIVERMORE

500 PB capacity accessed by > 4600 nodes

- 200 petaflops peak for modeling and simulation
- 3.3 ExaOps peak for AI and data analytics

https://www.olcf.ornl.gov/summit/

Still pushing limits – Blue Vela supercomputer – IBM as “client zero”

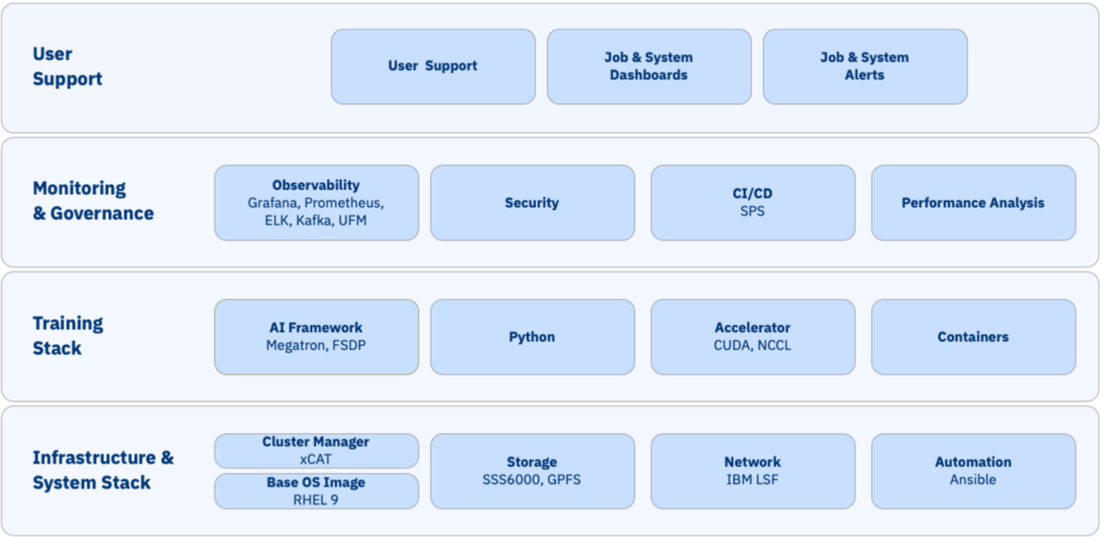
The **Vela** supercomputer is built in IBM Cloud for AI model training.

- Storage is IBM Cloud Object Storage
- IBM Storage Scale builds a caching layer with AFM

The next generation, **Blue Vela**, is a purpose-built on-prem supercomputer that exceeds what was possible with Vela.

- Blue Vela is used for continuous model training of the the IBM Granite LLM, with regular releases to Hugging Face.
- Storage is IBM Scale System 6000

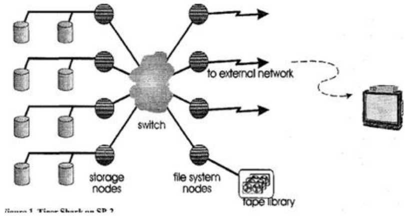
https://arxiv.org/abs/2407.05467



User Support		User Support	Job & System Dashboards	Job & System Alerts
Monitoring & Governance	Observability Grafana, Prometheus, ELK, Kafka, UFM	Security	CI/CD SPS	Performance Analysis
Training Stack	AI Framework Megatron, FSDP	Python	Accelerator CUDA, NCCL	Containers
Infrastructure & System Stack	Cluster Manager xCAT Base OS Image RHEL 9	Storage SSS6000, GPFS	Network IBM LSF	Automation Ansible

The Many Lives of IBM Storage Scale…

Began as the **Tiger Shark File System** in 1993. Designed to support multi-media applications (where the CLI prefix “mm…” comes from)



Leveraged **Vesta** filesystem’s ability to partition files and enable parallel access to create **GPFS** (General Parallel File System) in 1998

IBM®
GPFS

Leveraged some aspect of GPFS in an appliance-based implementation of the **Scale out File Services** (SoFS) for NAS (SONAS) in 2007

IBM **Spectrum Scale**

Rebranded as **“Spectrum Scale”** in 2015

Rebranded as **“IBM Storage Scale”** in 2022

Storage Scale

Leading use cases for IBM Storage Scale and Scale System

A Strategic

Storage for AI, Big Data, and Analytics
High performance and scalability

1. Artificial Intelligence (AI)
2. Data-intensive technical computing
3. Big Data and Analytics, Hadoop

B Strategic

Data lake(house), industry applications
Enterprise data architecture

4. Unified storage for “Data lake(house)”
5. Select ISV and industry solutions

C Tactical

Data optimization and resiliency
Enterprise data optimization and data management.
Infrastructure pre-requisites for segments A and B

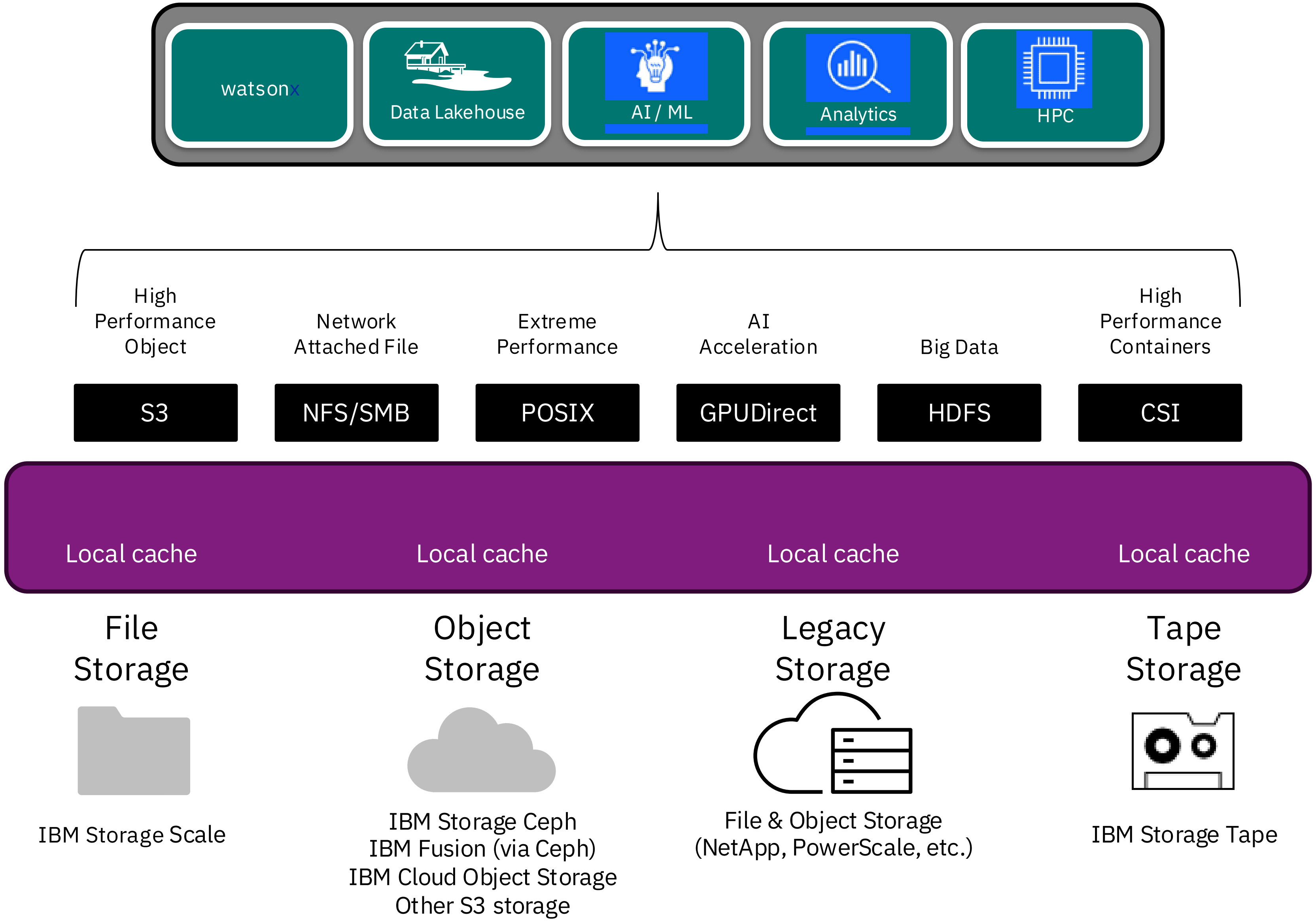
6. Archive
7. Information Lifecycle Management
8. Back-up / restore

IBM Storage Scale as the Global Data Platform: services for unstructured data

Data Access Services

Data Acceleration

Data Abstraction



Data Assurance Services

Data Administration Services

Data Aware Services

Storage Scale

Storage Scale: a General Parallel File System

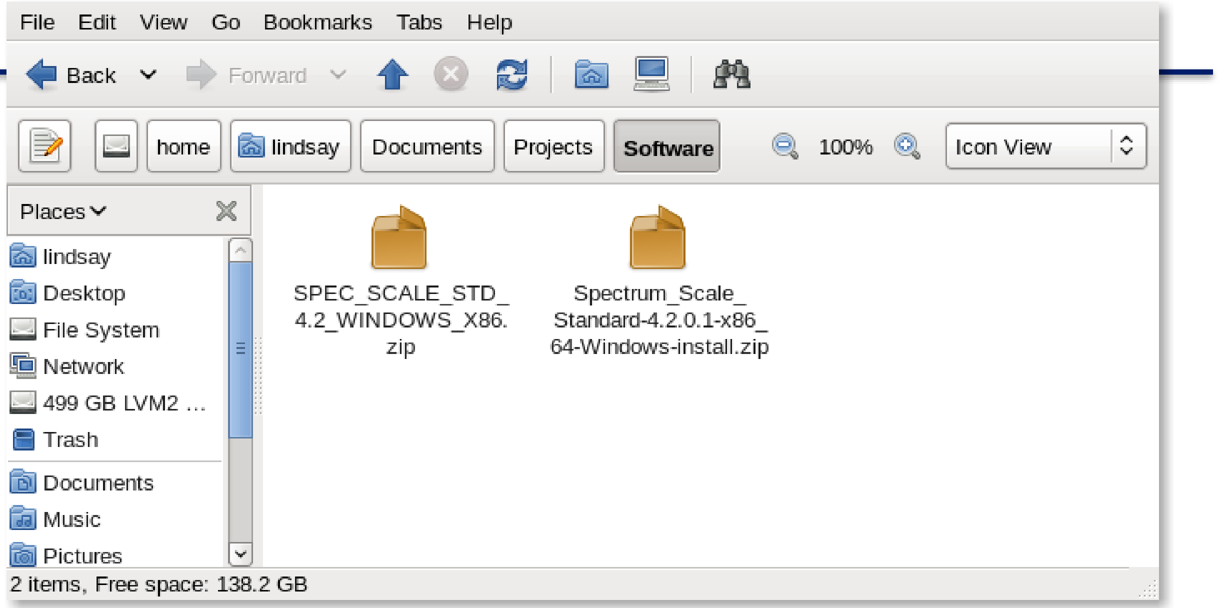
… and the unifying power of a common global file system

Scale is a global file system

Storage Scale is ***software***:

- implementing a **highly scalable distributed *parallel* POSIX file system**,
- able to run on systems under **Linux, AIX, and Windows** on **x86**, **ARM**, **IBM Power**, and **IBM Z**
- and **building upon any block storage** (from local disk to SAN-attached, not necessarily from IBM),
- with **advanced data management features** built into it that go well beyond the capabilities of traditional file systems,
- able to **tier** and **cache** from many storage types,
- and with **features layered on top** of it to make that storage accessible through NFS, SMB, and S3 protocols.

Never underestimate the power of unification that comes with a global file system.



SPEC_SCALE_STD_4.2_WINDOWS_X86.zip

Spectrum_Scale_Standard-4.2.0.1-x86_64-Windows-install.zip

2 items, Free space: 138.2 GB

```
$ cd /home/lindsay/Documents/Projects/Software
$ ls
SPEC_SCALE_STD_4.2_WINDOWS_X86.zip
Spectrum_Scale_Standard-4.2.0.1-x86_64-Windows-
  install.zip
$
```

Unifying around a global file system

Beyond the core capabilities of a file system...

Under the file system

- Use any block storage usable by the OS, grouped into storage pools.
- 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

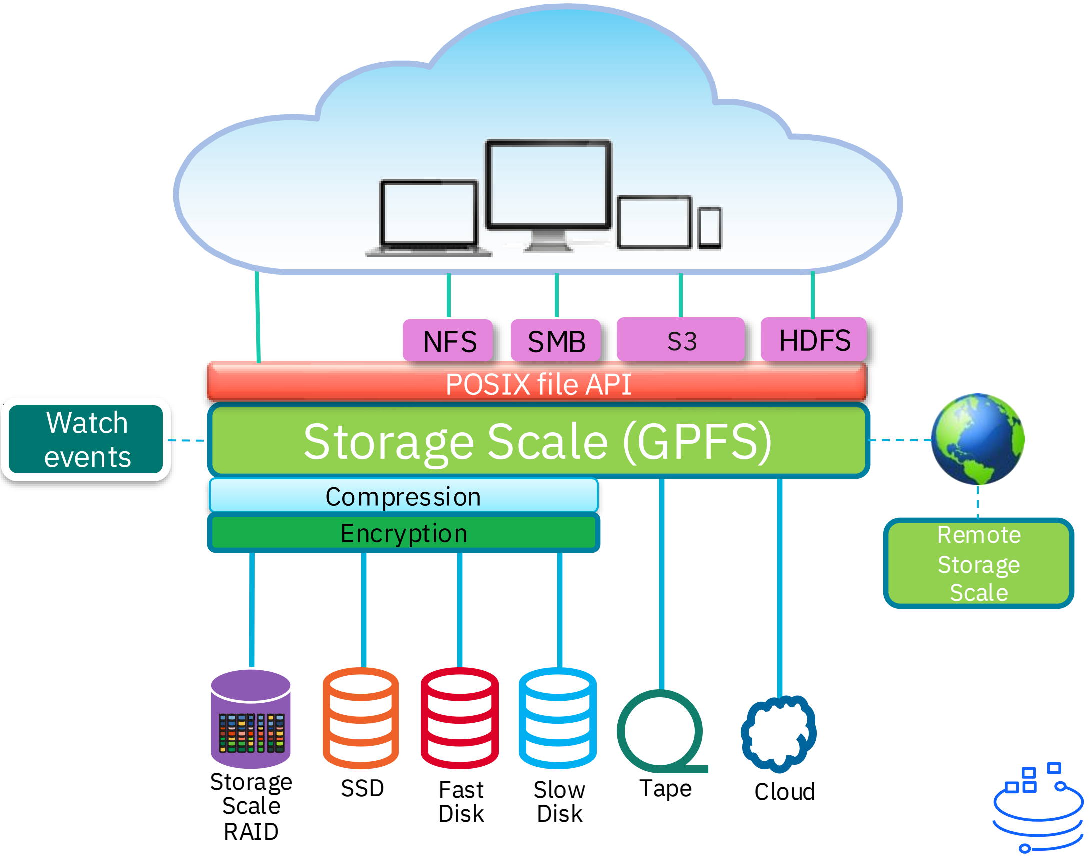
- Peer with remote Storage Scale file systems for efficient remote access.
- Replication capabilities for robust Disaster-Recovery capabilities.

Introspection

- Watch events in the file system with Watch Folders and File Audit Logging.
- Monitor and manage the file system

In a single name space

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



Storage Scale

Scale: a parallel *clustered* file system

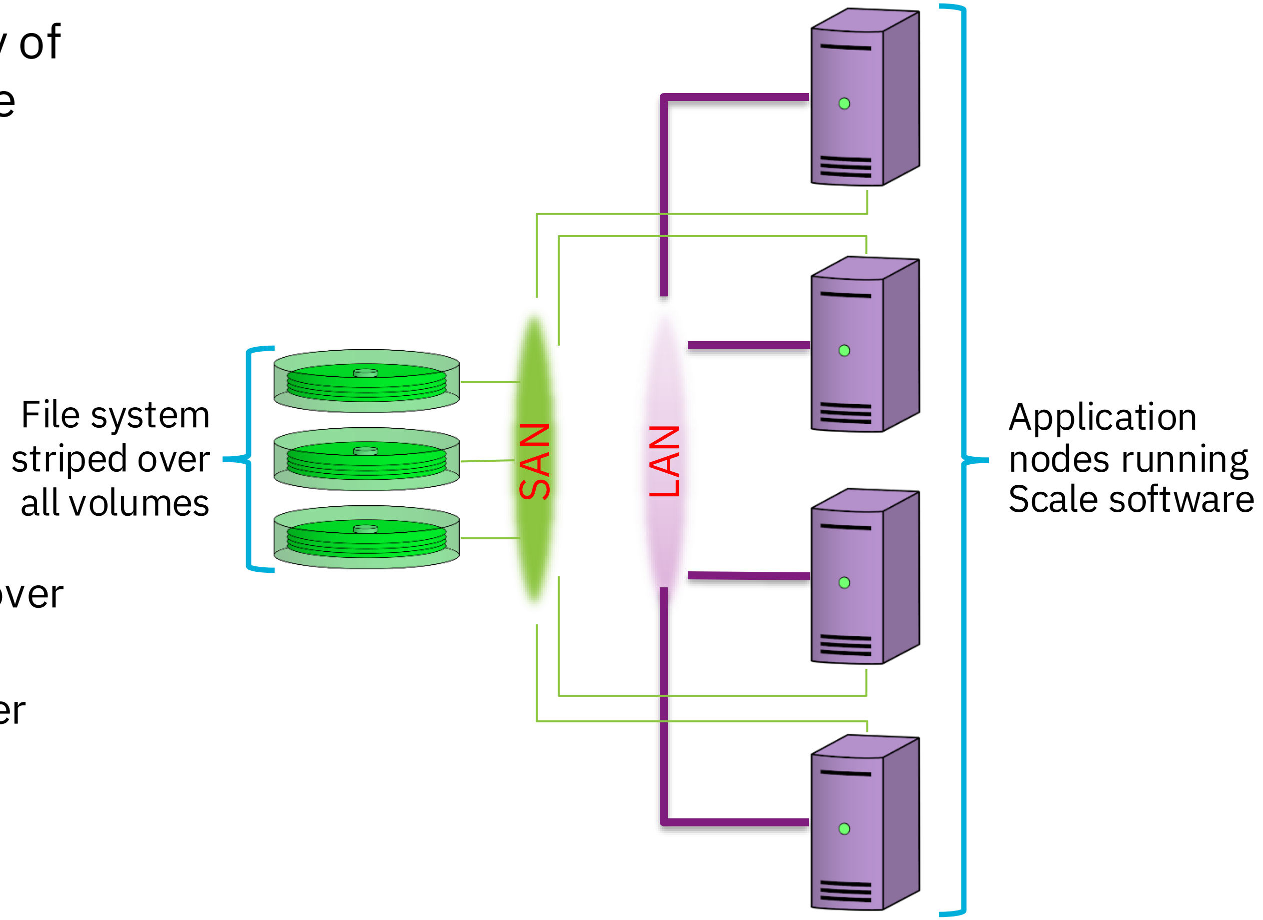
Scale is fundamentally a clustered file system:

- All nodes in the cluster share in the responsibility of accessing the disk volumes to manage the the file system and the data stored in it.
  - There are no “client” and “server” roles, as with distributed file systems like NFS and SMB.

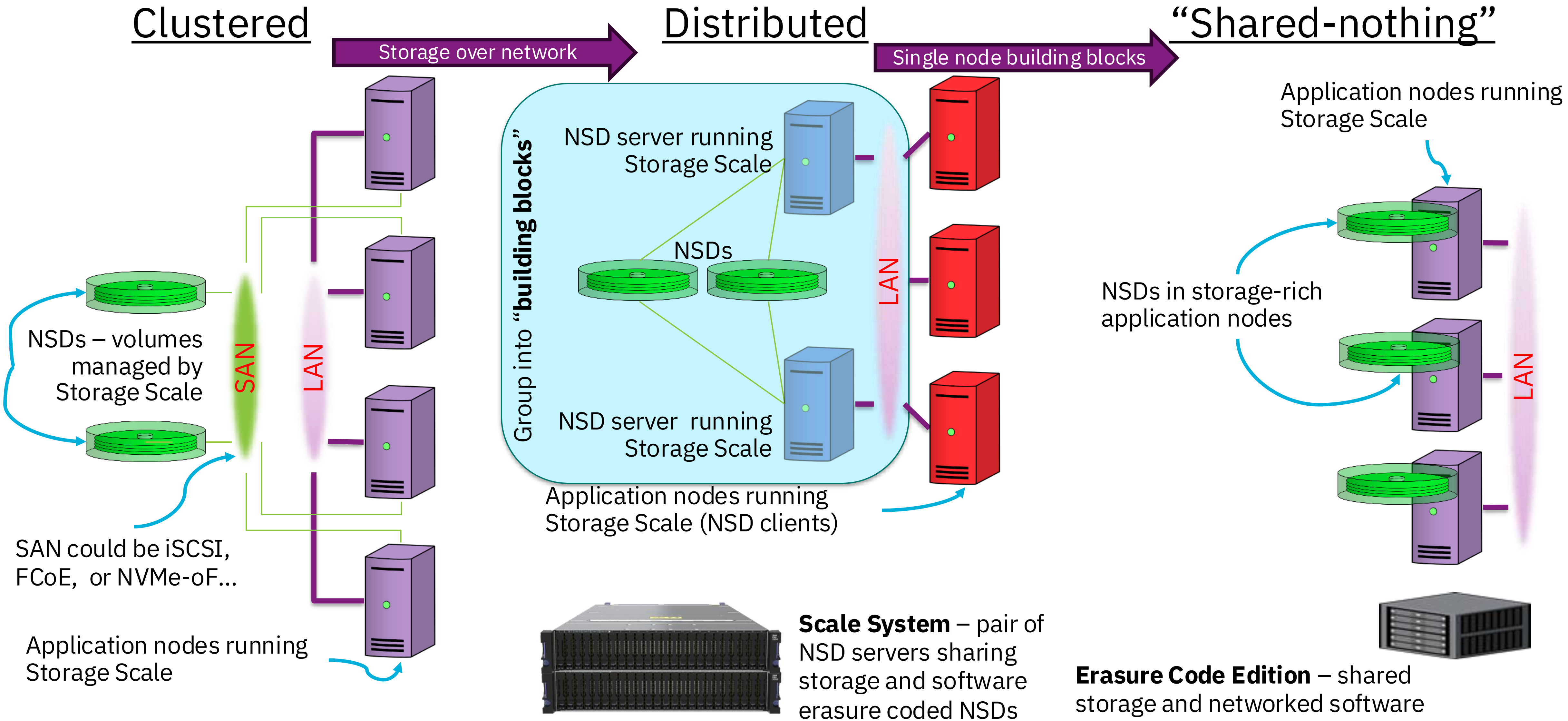
Scale is a **parallel** file system:

- The file system is striped over potentially many disk volumes.
- Even individual files (once large enough) are striped over multiple volumes.
- Uniquely, Scale even stripes file system metadata over many volumes.

All disk volumes controlled by a Scale cluster are called **NSD**s.


File system striped over all volumes
SAN
LAN
Application nodes running Scale software

High-level deployment strategies



NSDs – volumes managed by Storage Scale

SAN could be iSCSI, FCoE, or NVMe-oF…

Application nodes running Storage Scale

NSD server running Storage Scale

NSD server running Storage Scale

Application nodes running Storage Scale (NSD clients)

Application nodes running Storage Scale

NSDs in storage-rich application nodes

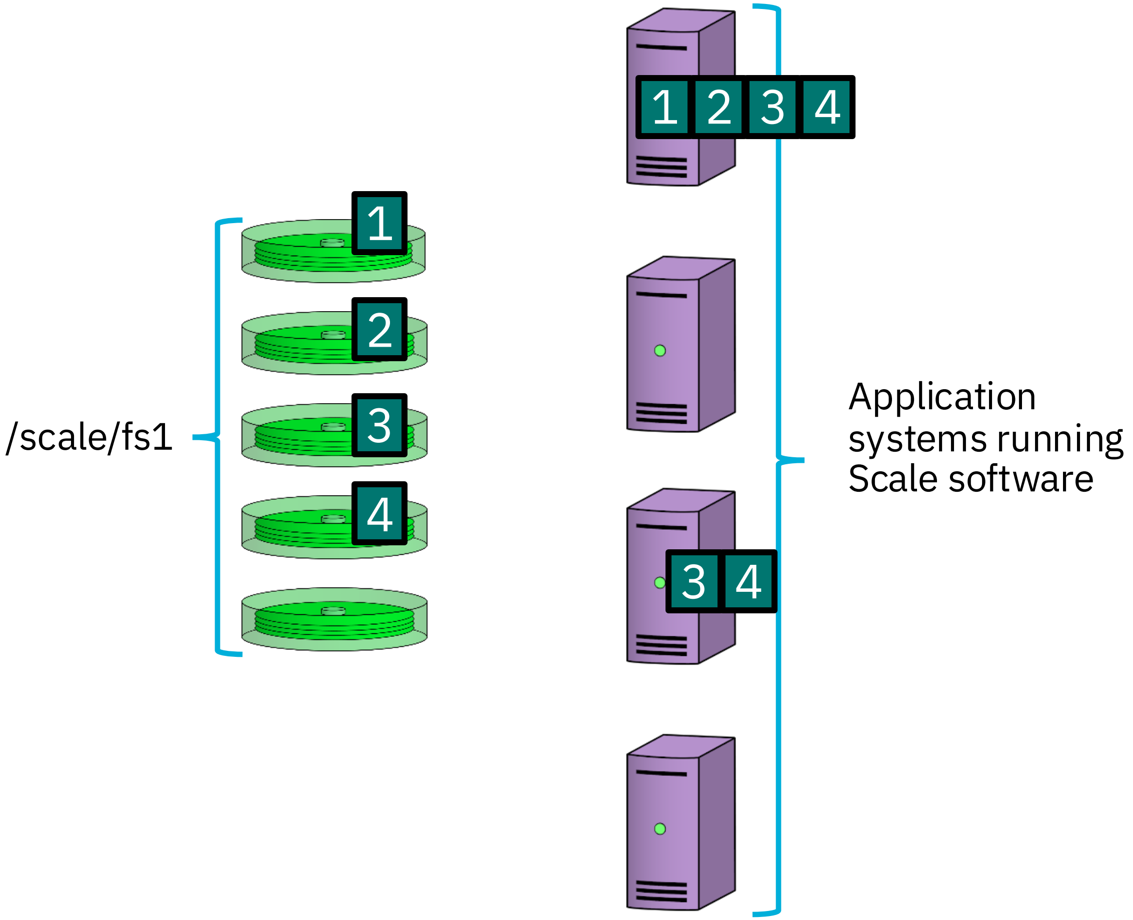
Scale System – pair of NSD servers sharing storage and software erasure coded NSDs

Erasure Code Edition – shared storage and networked software erasure coded NSDs

Parallel I/O – the secret to performance with Scale

Striping a file system over multiple disks is the key to Scale’s performance.

- Large files are written over multiple disks – and those writes are sent in ***parallel***.
- The **cluster manager** ensures that application nodes are coordinating their access to the disks.
- Reads can also happen in parallel –Scale’s **prefetch** mechanism leverages this.
- Different nodes may access the same file – the **Distributed Lock Manager (DLM)** will make ensure that the nodes have a consistent view of the file system and the data in the files.
- This scales: Adding disks (which can be done online) adds both *capacity and performance* to the original file system, allowing multitudes of systems to safely access huge amounts of data (and metadata) quickly.



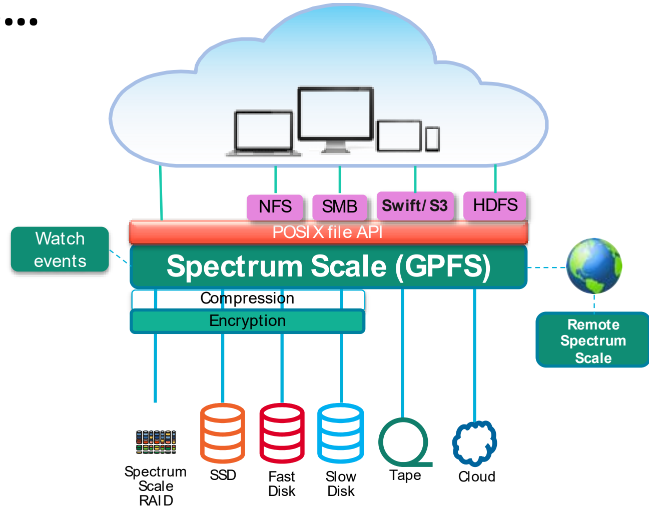
Storage Scale

Core capabilities of the Scale file system

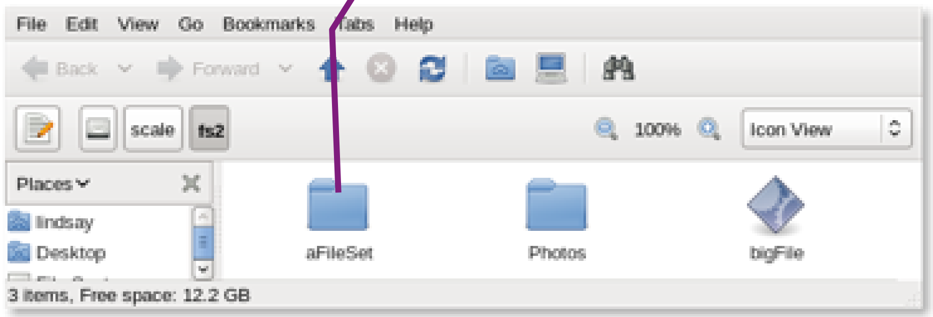
Storage Scale – a POSIX file system, and much more

Scale is POSIX file system features, and more:

- **Access Control Lists**
- **Filesets** – partitioning a file system into pieces, without losing performance but gaining manageability.
  - Filesets can have quotas.
  - Filesets can be made immutable or append-only.
- **Snapshots** – efficient read-only copy of file systems or filesets.
- **Quality of Service**
- **Encryption**
- **Performance and scalability features**
- **More advanced features…**



Fileset looks just like a directory



3 items, Free space: 12.2 GB

```
$ pwd
/scale/fs2
$ ls -l
total 20608
drwxr-xr-x. 4 root root     4096 Jan 17 10:07 aFileSet
-rw-r--r--. 1 root root 21098464 Jan 13 15:57 bigFile
drwxr-xr-x. 2 root root     4096 Jan 13 15:56 Photos
$ cd aFileSet
$ aFileSet]# ls -l
total 61824
drwxr-xr-x. 2 root root     4096 Jan 17 10:06 aDir
drwxr-xr-x. 2 root root     4096 Jan 17 10:07 anotherFileSet
-rw-r--r--. 1 root root 63295392 Jan 17 10:07 biggerFile
[root@gpfs01 aFileSet]#
```

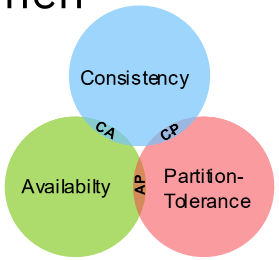
Scale clusters

A Scale **cluster** is a group of systems, the **nodes**, with Scale software installed and configured into a single administrative grouping:

- All nodes have a common view of the data.
- The nodes are tightly coupled, trusting each others' authentication of users.
- A cluster can have several Scale file systems.
- A cluster may share a file system with an authenticated remote cluster.

Clustering ensures all nodes in the cluster have a *consistent* view of the Scale file systems, even when more than one node is actively accessing a file system (or even the same file).

Careful! "Cluster" is a word used in many different ways in different contexts. Here we mean not merely a group of systems, but rather systems configured to work together as a single unit by running Scale.



CAP theorem – Brewer 1998,
Proven Gilbert, Lynch 2002
See https://en.wikipedia.org/wiki/CAP_theorem

Terms to know

Cluster – tightly coupled group of nodes.
- Configured cluster

Quorum nodes
- Cluster manager
- Active cluster

Storage – cluster resources
- NSDs

File systems
- File system manager
- Token management

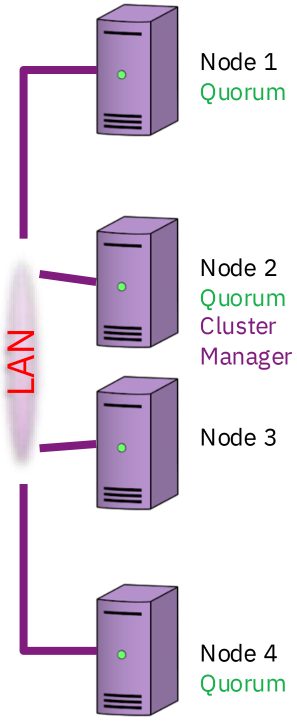


Table stakes: Rigorous POSIX consistency

For performance, modern operating systems buffer file I/O in memory caches.

POSIX operating systems, like Linux, mandate consistency between different threads (on the same system) reading and writing the same file.

- If one thread writes to a file, another thread can read that part of the file and see what was written.
- This must work even without using `lockf` or `fsync`, even if the *operating system* internally buffers file I/O in memory caches.

Scale maintains this same POSIX-guaranteed consistency *even when the threads are running on different nodes of the cluster*.

- Scale also caches file data, for performance.
- *Cache consistency* is maintained via the **distributed lock management** system, using **locks** and **tokens**.

<table>
<tr><th>Thread 1</th><th>Thread 2</th></tr>
<tr><td>x = os.open(‘/scale/aFile’,
os.O_RDWR|os.O_CREAT)
os.write(x, 100*b’A’)</td><td></td></tr>
<tr><td></td><td>y = os.open(‘/scale/aFile’,
os.O_RDWR)
os.seek(y, 100, os.SEEK_SET)
os.write(y, 100*b’B’)</td></tr>
<tr><td>os.lseek(x, 0, os.SEEK_SET)
os.write(x, 100*b'C')
os.lseek(x, 80*(2**20),
os.SEEK_SET)
os.write(x, 100*b'D')</td><td></td></tr>
<tr><td></td><td>os.lseek(y, 95, os.SEEK_SET)
print(os.read(y, 10))
→ b'CCCCCBBBBB'</td></tr>
</table>

With Scale, it doesn’t matter if threads 1 and 2 are on the same node, or on different nodes of the active cluster.

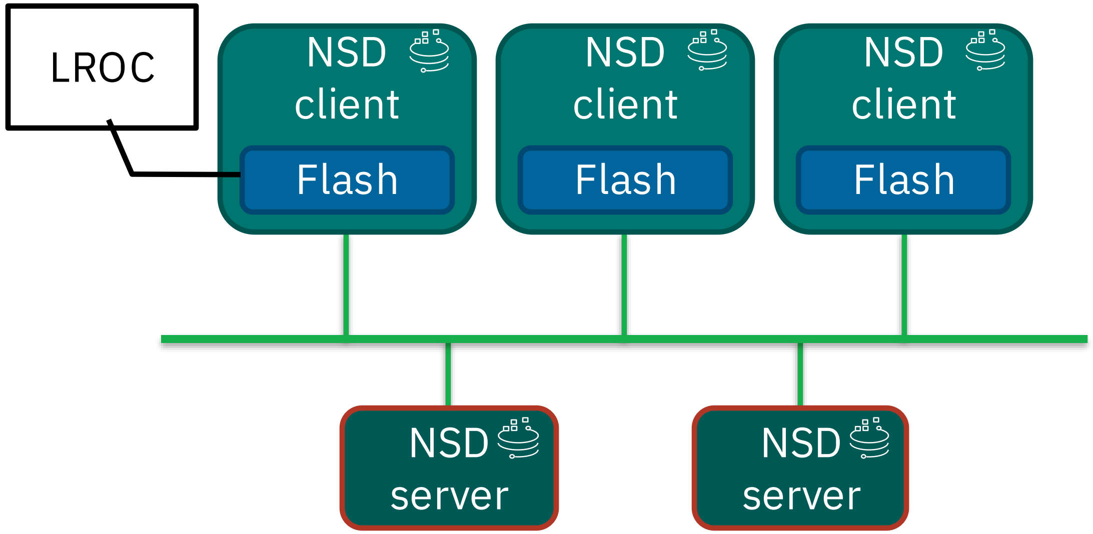
NFS does not guarantee this consistency.

Features for high-performance and scalability

Read-ahead deduced from application I/O patterns

Client-side caching:

- Pagepool memory for read-ahead buffers, write-behind buffers, inode cache
- Optional **Local Read Only Cache** (LROC) on local SSD/Flash/NVMe – and control if used with encryption.



Highly Available Write Cache (HAWC) leverages recovery log to accelerate small synchronous writes.

Subblock sizes kept small even with large block sizes, achieving performance with space efficiency

Multiple TCP and RDMA connections between cluster nodes

- Uses all lanes in bonded connections
- Full performance of high-speed network links

Scalable metadata and distributed lock management

Scalable metadata operations

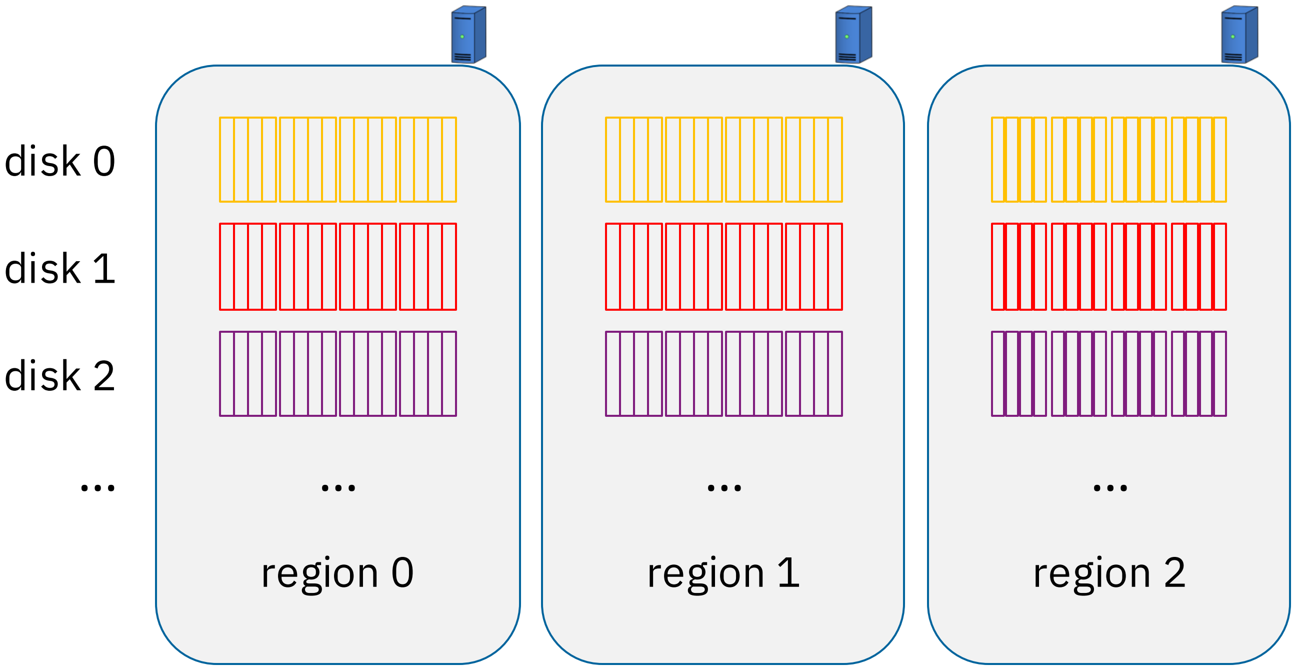
Scaling file updates

- File system **metadata** resides on NSDs, just like data (and can even be coresident with data).
- For each file system, the cluster manager chooses a node to be the **file system manager**.
- For each open file, the file system manager tracks a node to be the **metanode** for that file (usually the node that has had the file open the longest).
  - The metanode manages all metadata for the file (inode updates, indirect block updates, etc.)
- File system manager (and token managers) track which nodes have tokens on files.
  - A node needing a conflicting token will be directed to negotiate with nodes holding conflicting tokens.

Careful! "Metadata" is a word meaning different things in different contexts. Here we mean the specific internal data needed to define the file system.

Scaling space management

- File system manager allocates segments of the block allocation map to nodes.
- Nodes can manage space allocation using their segment, in parallel.



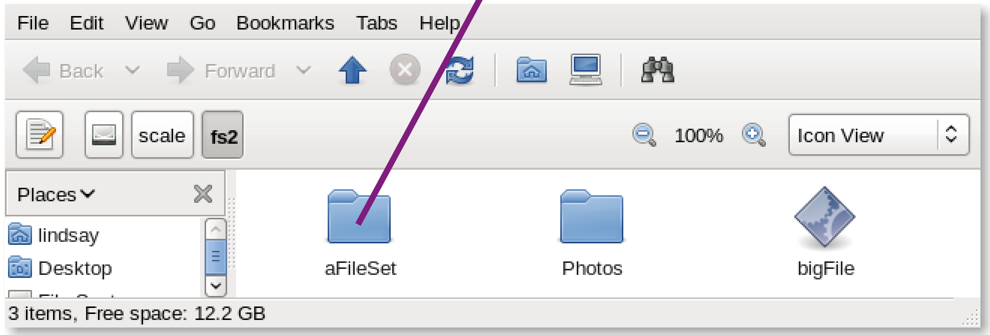
Filesets

A **fileset** might be considered to be either a “sub-file system” or a “glorified directory”:

- It **makes use of all the disks of the main file system**, so there is no loss of performance (unlike smaller file systems, forcing us to allocate fewer disks to each).
- It **may be unlinked and relinked to different locations** in the file system (like a file system `umount` and `mount`).
- A fileset can have a ***quota*** for users, groups, or on the aggregate contents of the entire fileset itself.
- ***Backups*** and ***snapshots*** may be taken of individual independent filesets.
- May be associated with ***Quality of Service*** (QoS) classes.
- A fileset may be given ***immutability*** settings to control how strictly retention and expiration times of its files are enforced.
- Policies can consider fileset membership of files.
- Exported file systems can control visibility by fileset.

You still have regular POSIX directories available in the file system – not everything must go into filesets.

Looks just like a directory



3 items, Free space: 12.2 GB

```
$ pwd
/scale/fs2
$ ls -l
total 20608
drwxr-xr-x. 4 root or     4096 Jan 17 10:07 aFileSet
-rw-r--r--. 1 root row 21098464 Jan 13 15:57 bigFile
drwxr-xr-x. 2 root big File     4096 Jan 13 15:56 Photos
$ cd aFileSet
$ aFileSet]# ls -l
total 61824
drwxr-xr-x. 2 root big File     4096 Jan 17 10:06 aDir
drwxr-xr-x. 2 root air     4096 Jan 17 10:07 anotherFileSet
-rw-r--r--. 1 root exclude list 63295392 Jan 17 10:07
  biggerFile
[root@gpfs01 aFileSet]#
```

Security: ACLs, immutability, encryption

Besides POSIX “mode bits”, Scale supports **Access Control Lists** (ACLs), both POSIX and NFS4 styles.

- Authentication and assigning process credentials (uid, gid, etc.) are handled by the operating system and assumed to be consistent across the cluster.

Immutability attributes on a file can prevent its modification, allow only appending, and/or set a retention time.

- User-defined **extended attributes** can be used for additional data-management processes, such as tracking file provenance.

Encryption can be policy-driven at the file system level or through leveraging the storage system’s encryption capabilities (such as SED drives in a Scale System).

Access Control Lists

Certified **Immutability**

Cohasset Associates

SEC 17a-4(f), FINRA 4511(c), CFTC 1.31(c)-(d)
Compliance Assessment

IBM Spectrum Scale

Abstract

Encryption

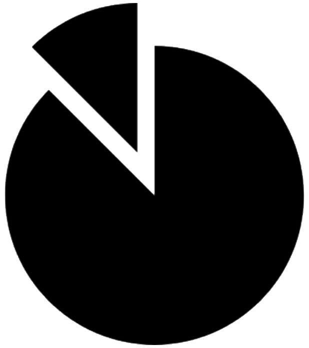
Usage management: Compression, Quotas, and Quality of Service

- Cold files can be **compressed** under the control of policies, allowing different files to be compressed with the most appropriate compression algorithm (perhaps none at all!).
- **Quotas** restrain how much storage an individual user or group can use of a fileset or a file system.
  - Or put a quota on the total size of a fileset.
- **Quality of Service** (**QoS**) restrains how much file system throughput may be used by applications using an individual fileset.
  - Or restrict how much throughput may be used for maintenance tasks, reducing their impact on applications and compute workloads.

File compression

Quotas

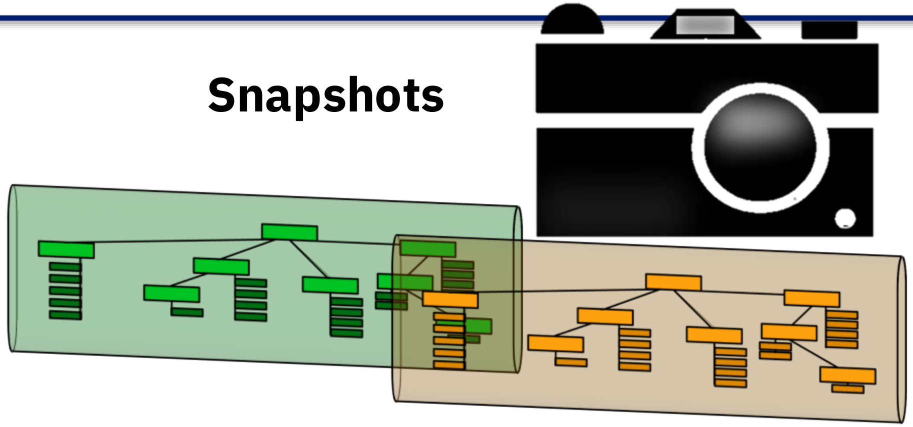
Quality of Service



Recovery: Snapshots and backups

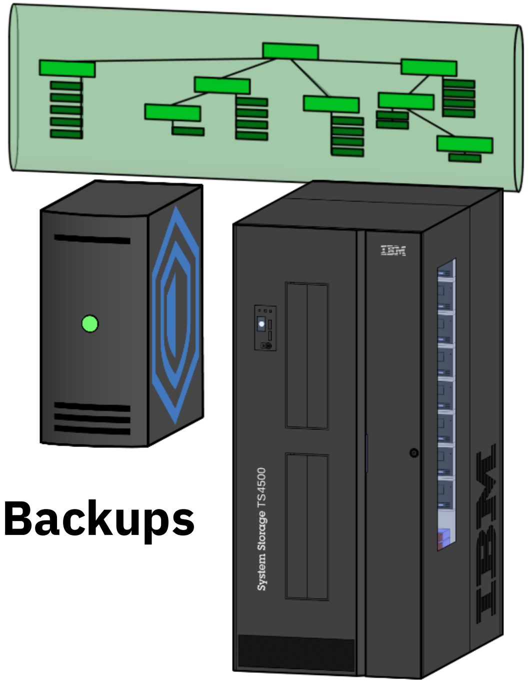
Snapshots are ***read-only*** copies of either entire *file systems* or individual *filesets*.

- File systems may have 256 global snapshots.
- Additionally, each fileset may have 256 snapshots.
- Copy-on-write semantics ensure snapshots are both time- and space-efficient.



The built-in mmbackup utility leverages Scale's fast, parallel inode scanning, Scale's innate parallelism, and a shadow database to drive IBM Storage Protect to efficiently **backup/restore** files and metadata:

- Create, access, and modify times
- Size of the file
- Mode bits and owner, owning group
- Immutability attributes
- Access Control Lists
- Extended attributes



Storage Scale "stretched cluster" using synchronous replication between failure groups

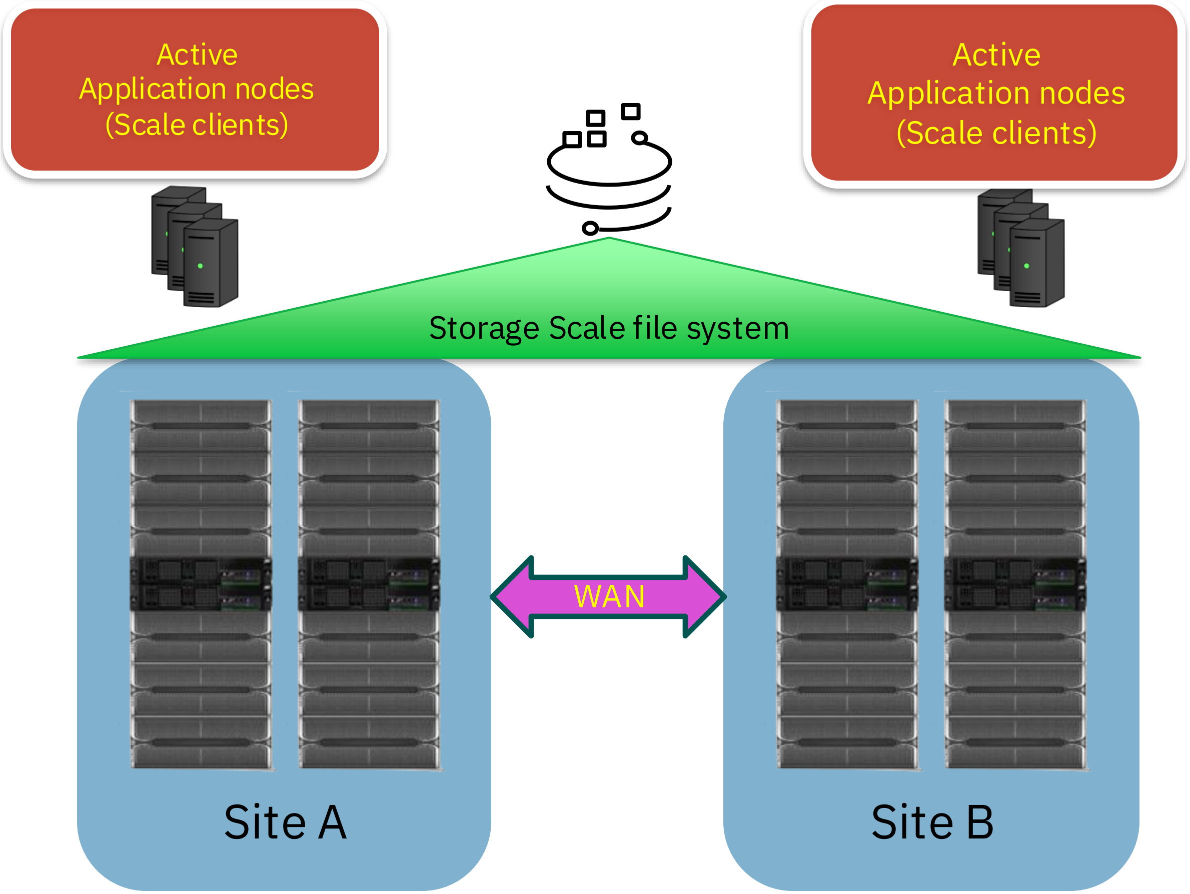
A *single* Scale cluster may be configured using nodes and storage from two data centers.

- In other words, it is "stretched" between two sites, connected by a WAN.

The file systems may judiciously use **"failure group" replication** to ensure both sites have a current replica of all the data.

- Careful design can ensure one site remains active, even if the other site (or the link between sites) fails.
- Both sites may **concurrently** access the file system for both reading and writing.

A stretched cluster provides an **active/active highly available** synchronously-replicated Scale file system across two data centers.

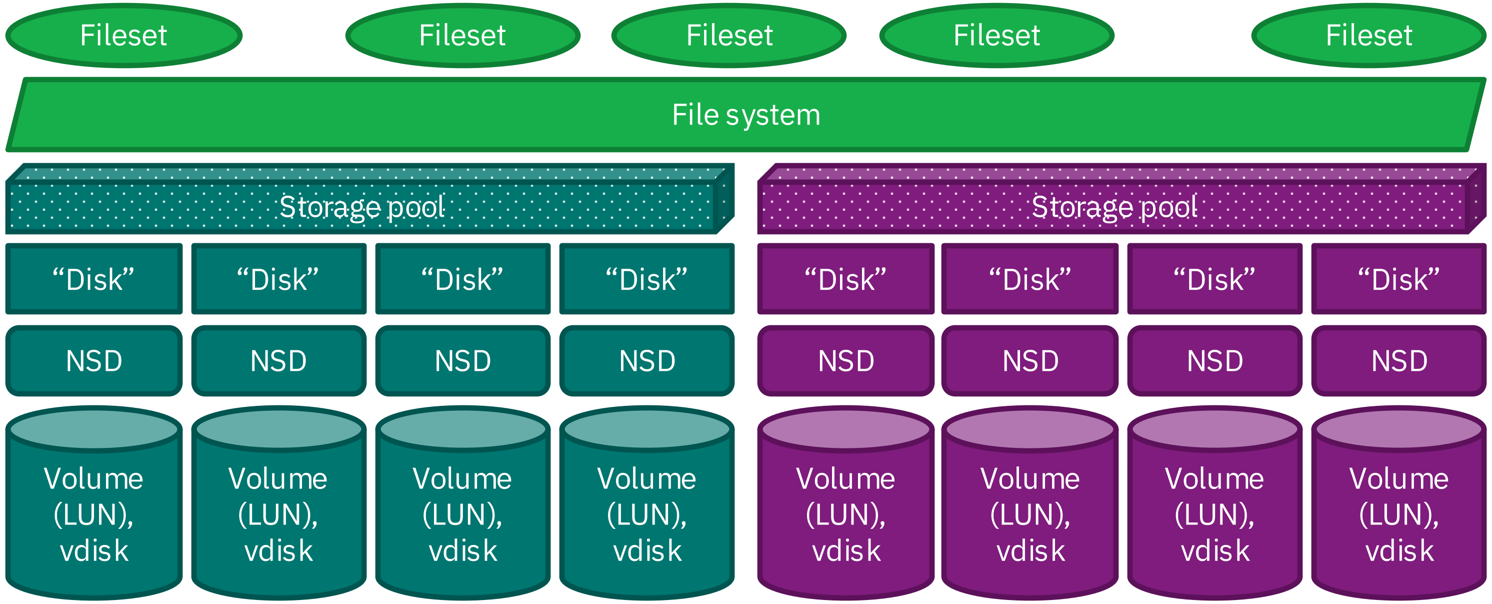


Storage Scale

Below the file system:
Storage pools and policies

Positioning the data to where it is most appropriate

A file system needs storage – organizing disks into storage pools

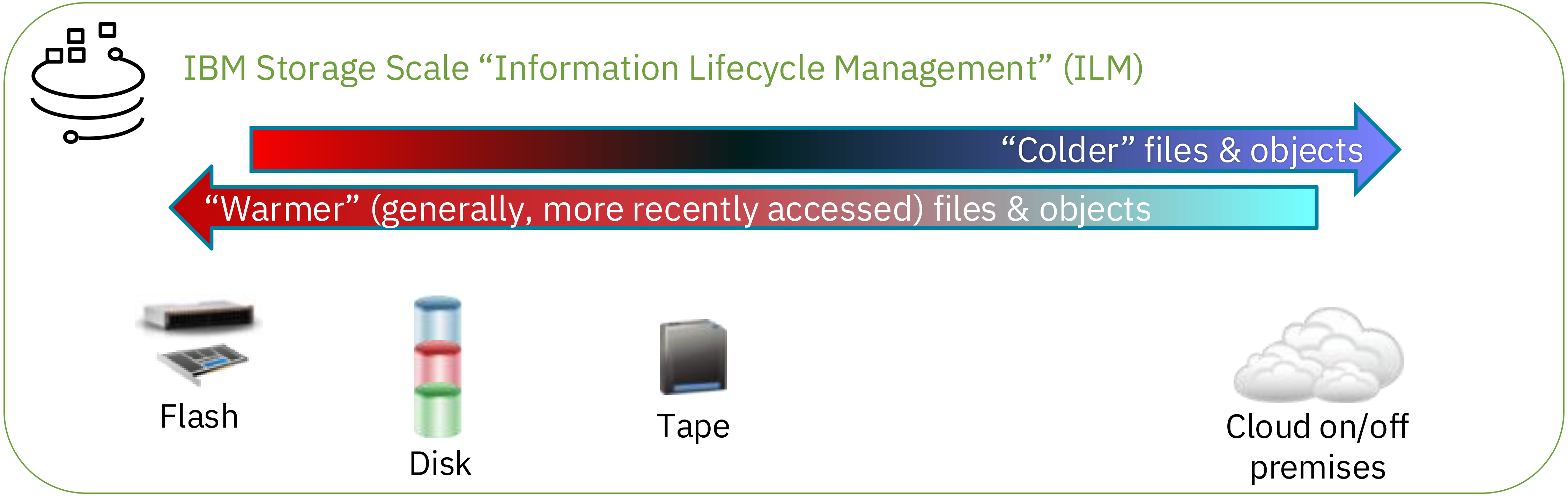


Storage pools enable organizing policy-driven “tiered storage”

*Automate data **placement**, **movement**, and **removal***

Powerful policy-driven automated tiered-storage management

- Use different storage pools of storage devices along with policies that control placement, movement, and deletion of data across those pools



IBM Storage Scale “Information Lifecycle Management” (ILM)

“Colder” files & objects
“Warmer” (generally, more recently accessed) files & objects
Flash
Disk
Tape
Cloud on/off premises

Policies and storage pools

Policies for different purposes:

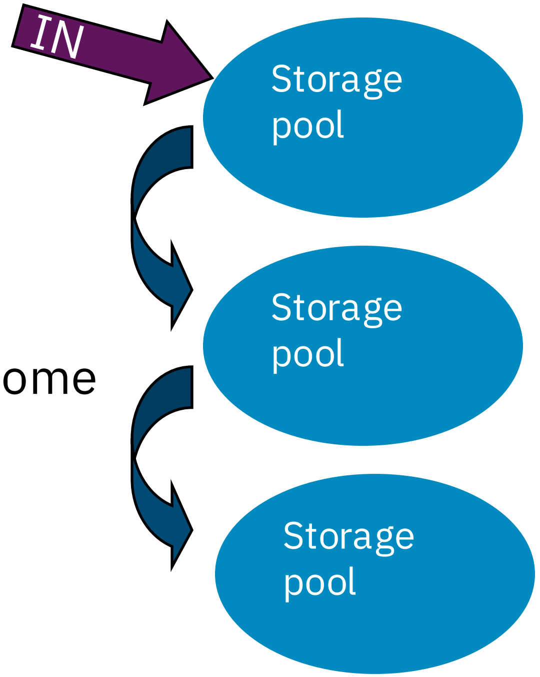
- **File placement policies** – Rules determining where and how file data is placed when a file is created
- **File management policies** – Rules for migration, deletion, encryption, etc. of files based on chosen criteria

Policies select files on which to operate from a diverse set of criteria, such as:

- File name, file type, path name, fileset, pool, name of the file, or snapshot
- Owner or group of the file
- Age, size, "file heat" of the file
- Extended attributes on the file

The attribute must be known when the policy is run (so, e.g., a placement policy can't know the ultimate length of a file, but it does know the owner)

Placement policy – select which pool the new file will "land"

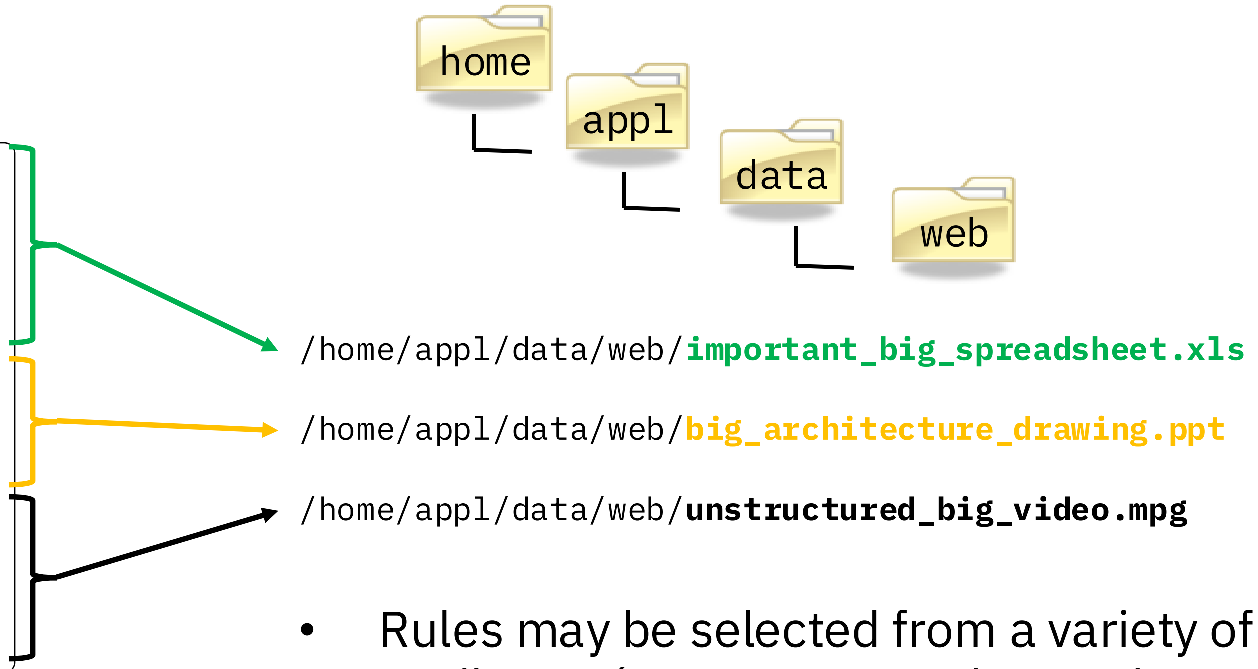


Management policies – process files selected by some criteria

Examples of ILM placement policy rules

Placement policy

```
RULE boss
    SET POOL system
    WHERE UPPER(NAME) LIKE '%.XLS' OR
          USER_NAME = 'boss'
RULE tech
    SET POOL ess
    WHERE UPPER(NAME) LIKE '%.PPT'
RULE fun
    SET POOL nlsas
    WHERE UPPER(NAME) LIKE '%.MPG'
```



/home/appl/data/web/**important_big_spreadsheet.xls**

/home/appl/data/web/**big_architecture_drawing.ppt**

/home/appl/data/web/**unstructured_big_video.mpg**

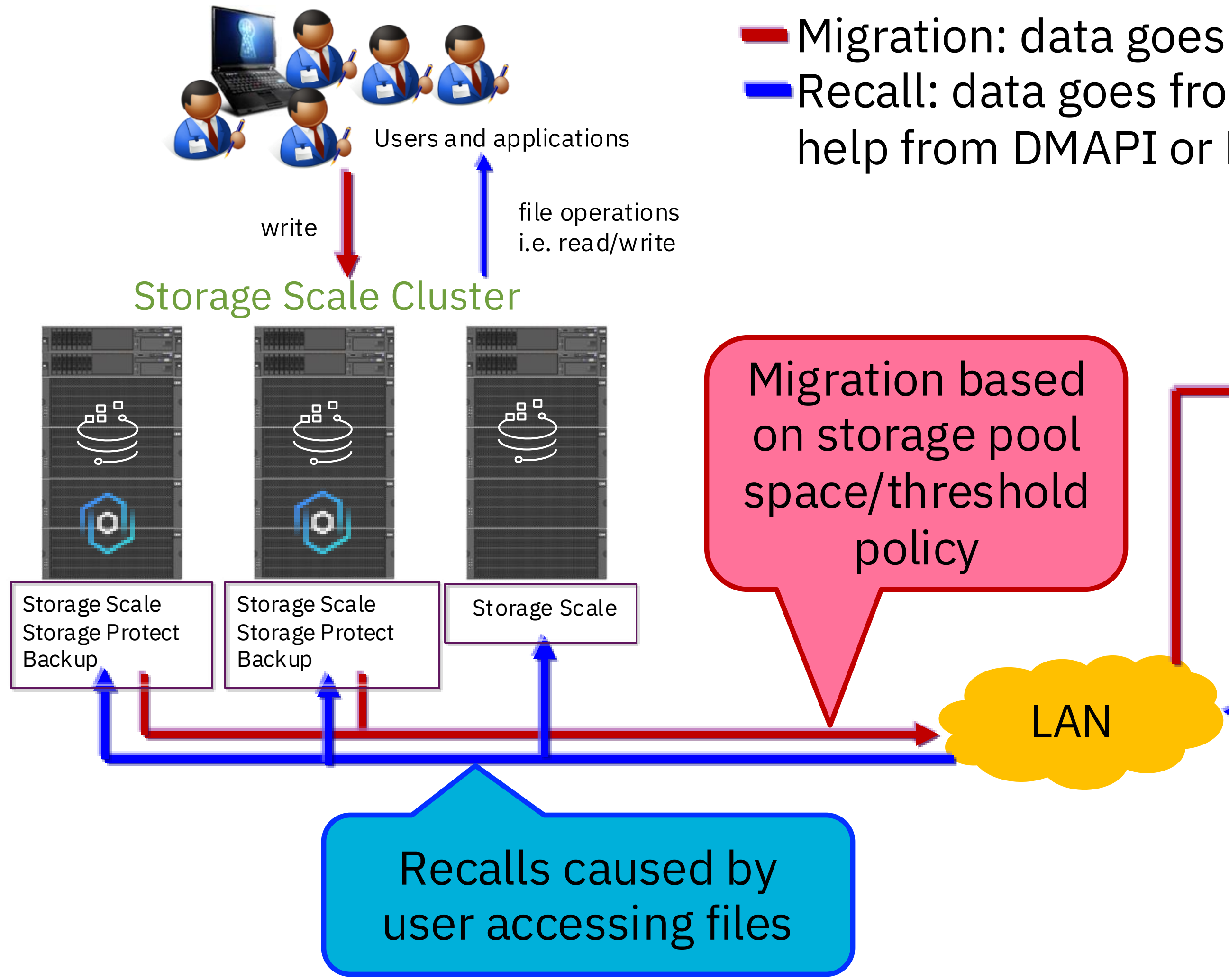
Policy syntax is somewhat like SQL, with an action performed when selection criteria matches.

- Rules may be selected from a variety of file attributes (owner, group, size, pathname, fileset, last access/modified time, “file heat”, extended attributes, etc.)
- Besides determining pools, rules support file deletion, moving to “external pools”, triggering actions, encryption, compression.

Policy-driven Hierarchical Storage Management (HSM)

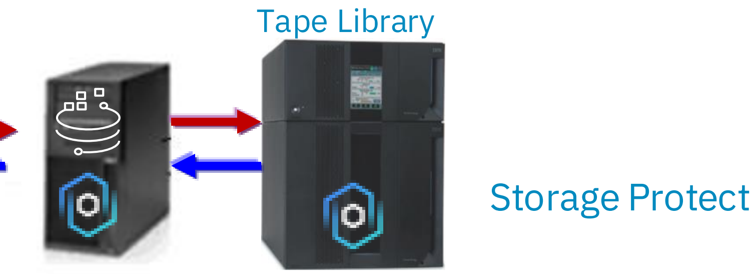
🟥 Migration: data goes from Scale to Protect server (policy driven)

🟦 Recall: data goes from Protect server to Scale (on demand, with help from DMAPI or Light-Weight Events)



Migration based on storage pool space/threshold policy

Recalls caused by user accessing files



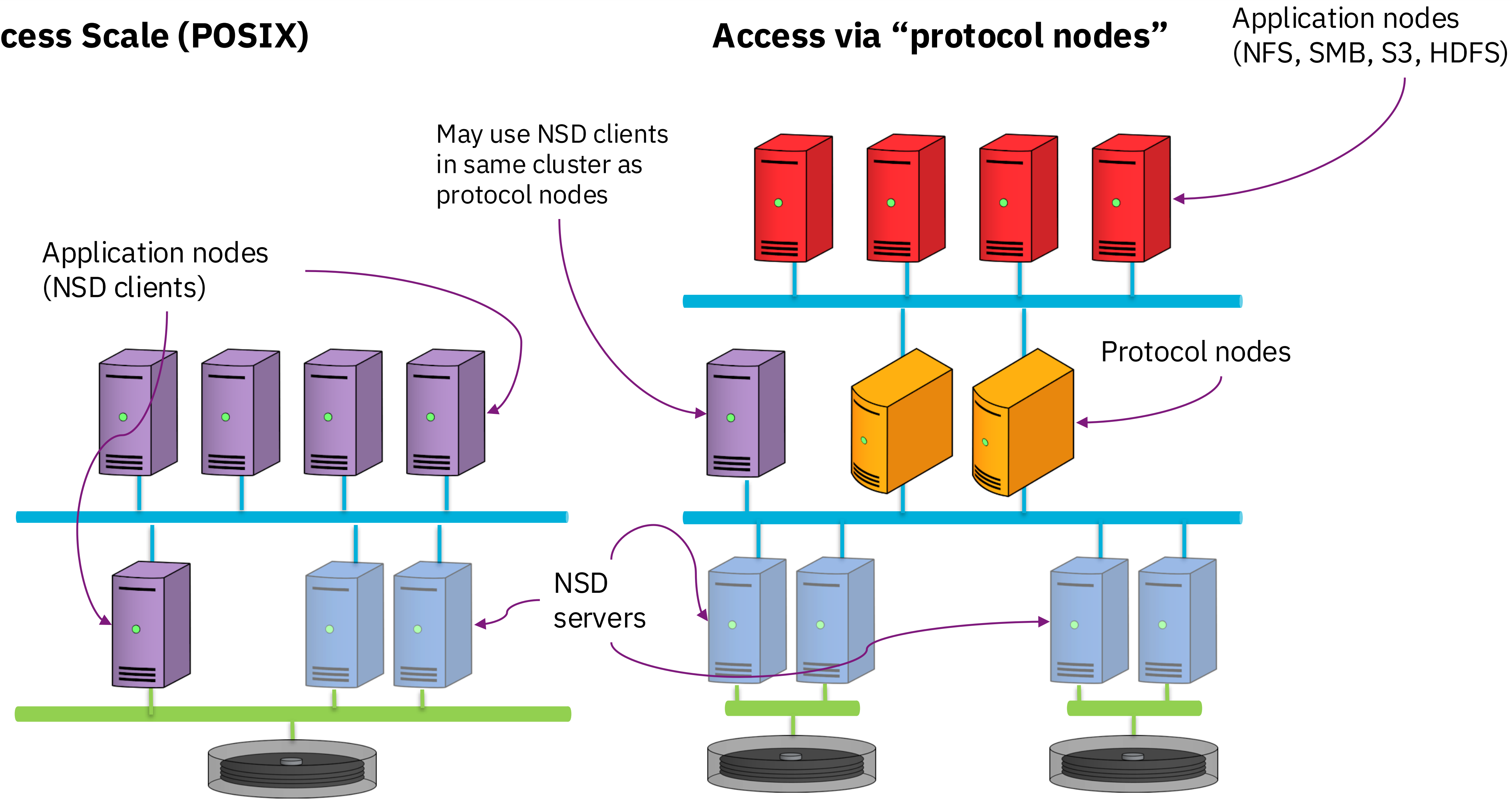
- Policy-driven migration can be done directly to tape.
- Multiple Protect and HSM clients can be moving data to Protect server.
- Recalls are triggered by file access.
- Migration and recalls are distributed and handled by Protect/HSM clients

Storage Scale

Above the file system: Accessing Scale via S3, NFS, SMB, HDFS, CSI, GPUdirect, etc.

… providing Scale to places it can’t run directly

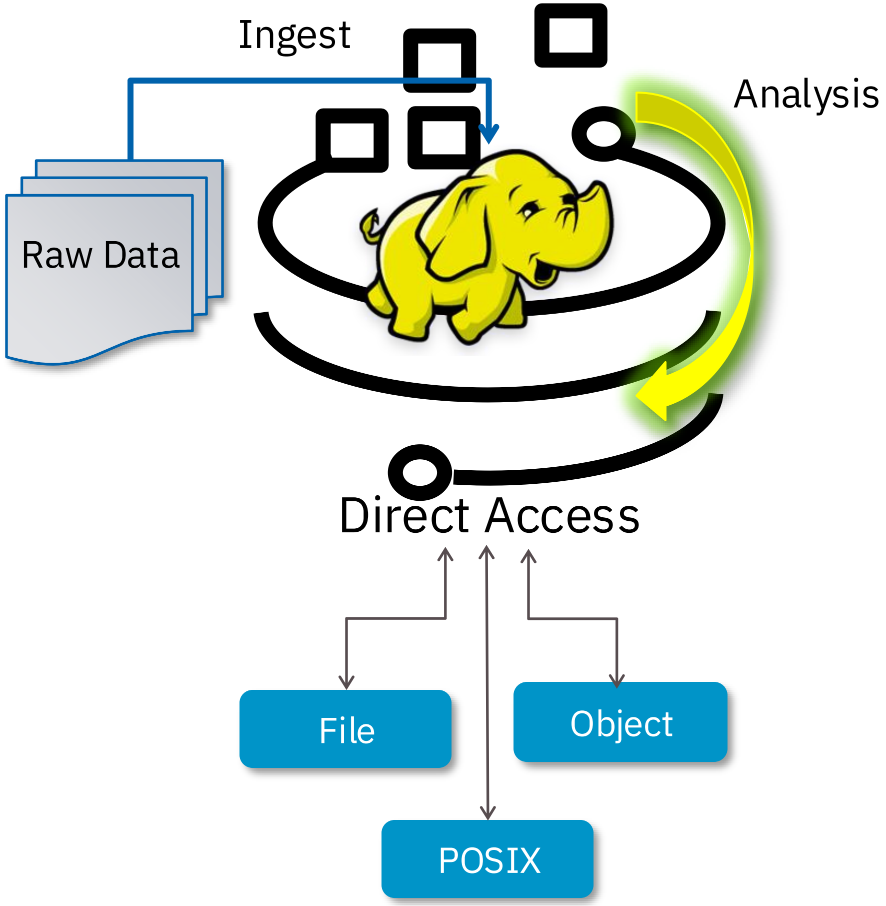
A couple important ways to access IBM Storage Scale



Big Data and Analytics – Storage Scale's HDFS Transparency

- Map/Reduce on either shared or "shared-nothing" storage
  - Shared storage allows uses far less storage than traditional HDFS 3-way replication.
- No need to wait for data transfer between storage systems – access by most appropriate method (POSIX, HDFS, S3) without copying.
- Scale can be a single "Data Lake" for <u>all</u> applications
- Archive and Analysis happen in-place – Immediately share results

All this, without rewriting existing Hadoop or Spark applications



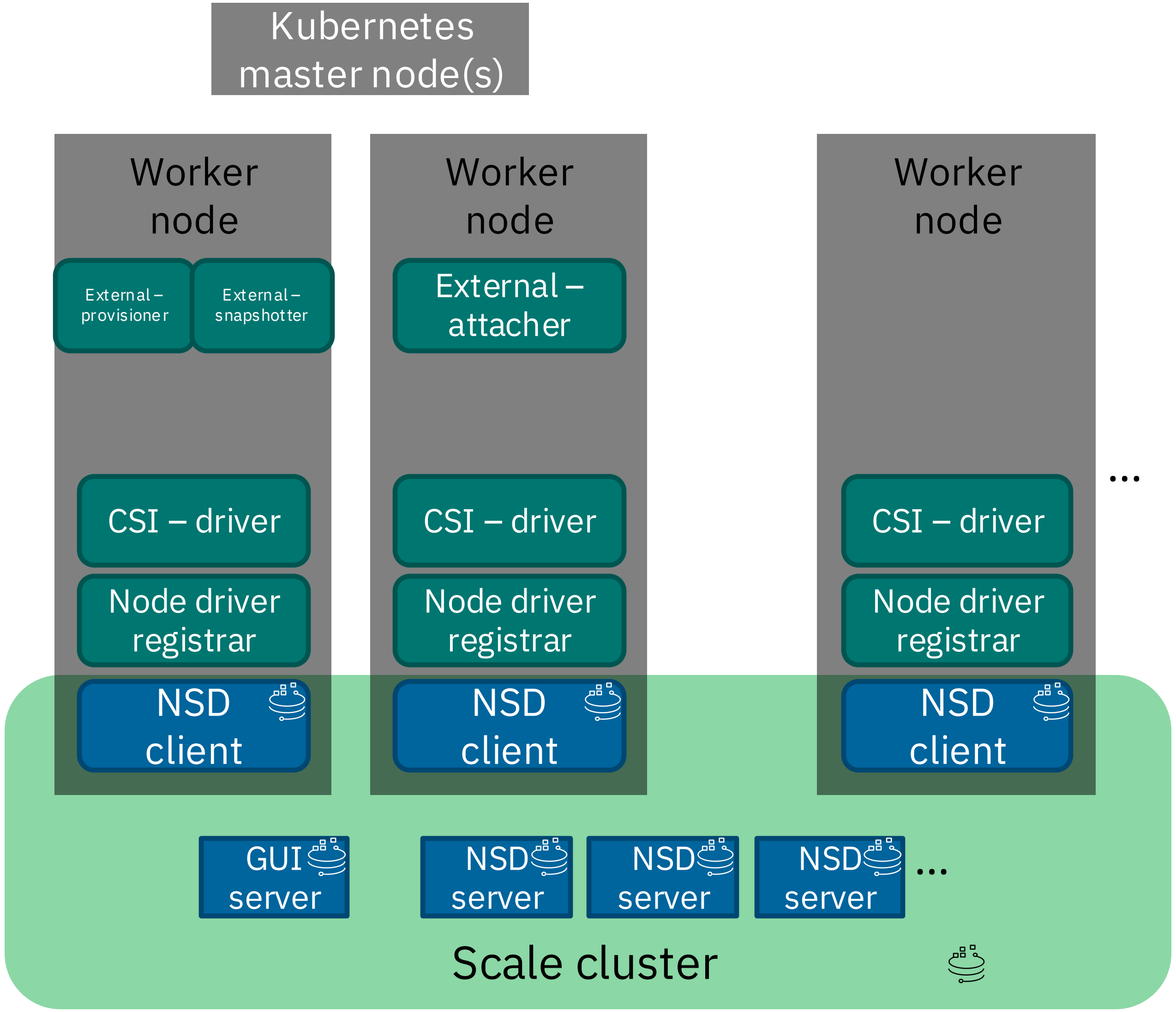
Container Storage Interface (CSI) driver

The Scale Container Storage Interface (CSI) driver lets Kubernetes (K8s) containers use Scale storage for persistent volumes.

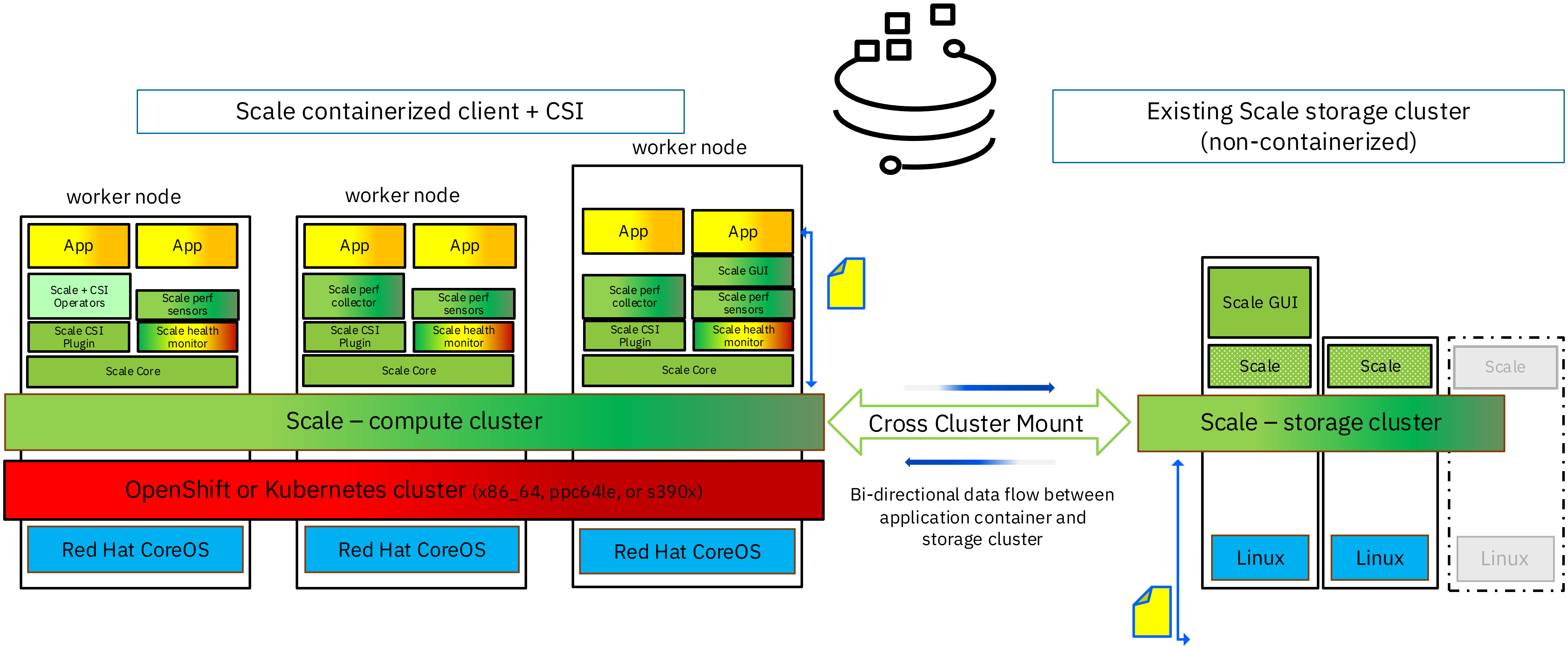
- Both RWX (ReadWriteMany) and RWO (ReadWriteOnce)
- Both static and dynamic provisioning
- Volumes may be directories or filesets

The driver uses the REST management API provided by the Scale GUI backend.

Scale Container Native Storage Access (CNSA) runs the Scale NSD clients as K8s pods in an OpenShift environment.



Container Native Storage Access (CNSA)

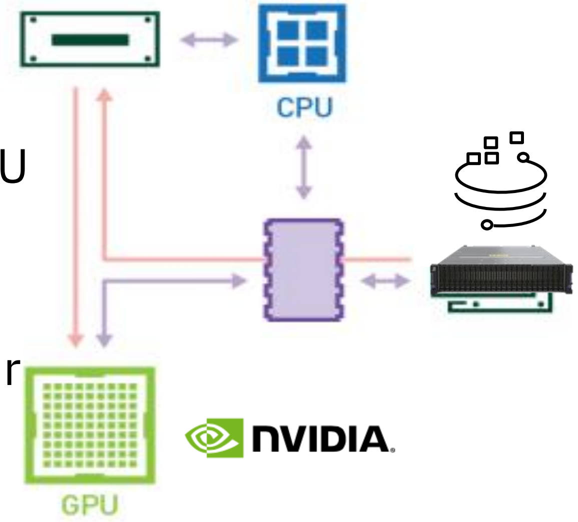


GPU Direct Storage (GDS) – Scale with NVIDIA GPU

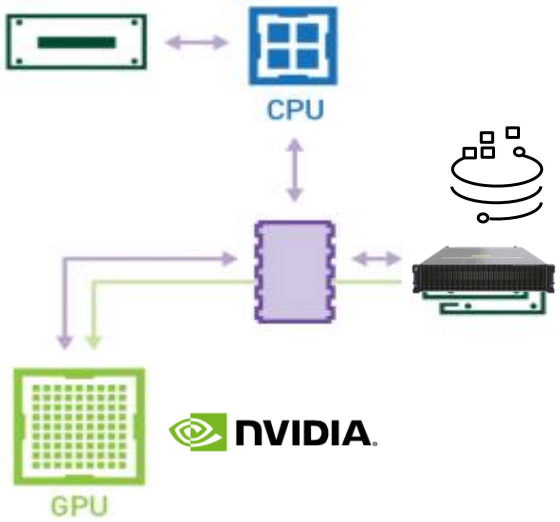
GPUDirect Storage (GDS) enables an NVIDIA CUDA developer to:

- Use direct memory access (DMA) between GPU memory and storage
- Bypass the CPU and system memory
- Reduce latency, increase bandwidth, and lower CPU utilization for a specific read from storage

GPUDirect leverages RDMA (over either InfiniBand or RoCE).



CPU in the middle



GPUdirect gives direct link between GPU and storage

* Requires CUDA, provided by NVIDIA, to be installed on the Storage Scale compute node

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

Storage Scale

Beside the file system: Remote mounts, caching, and replication

… for working with data stored elsewhere

“Multi-cluster” – Remote file system mounts

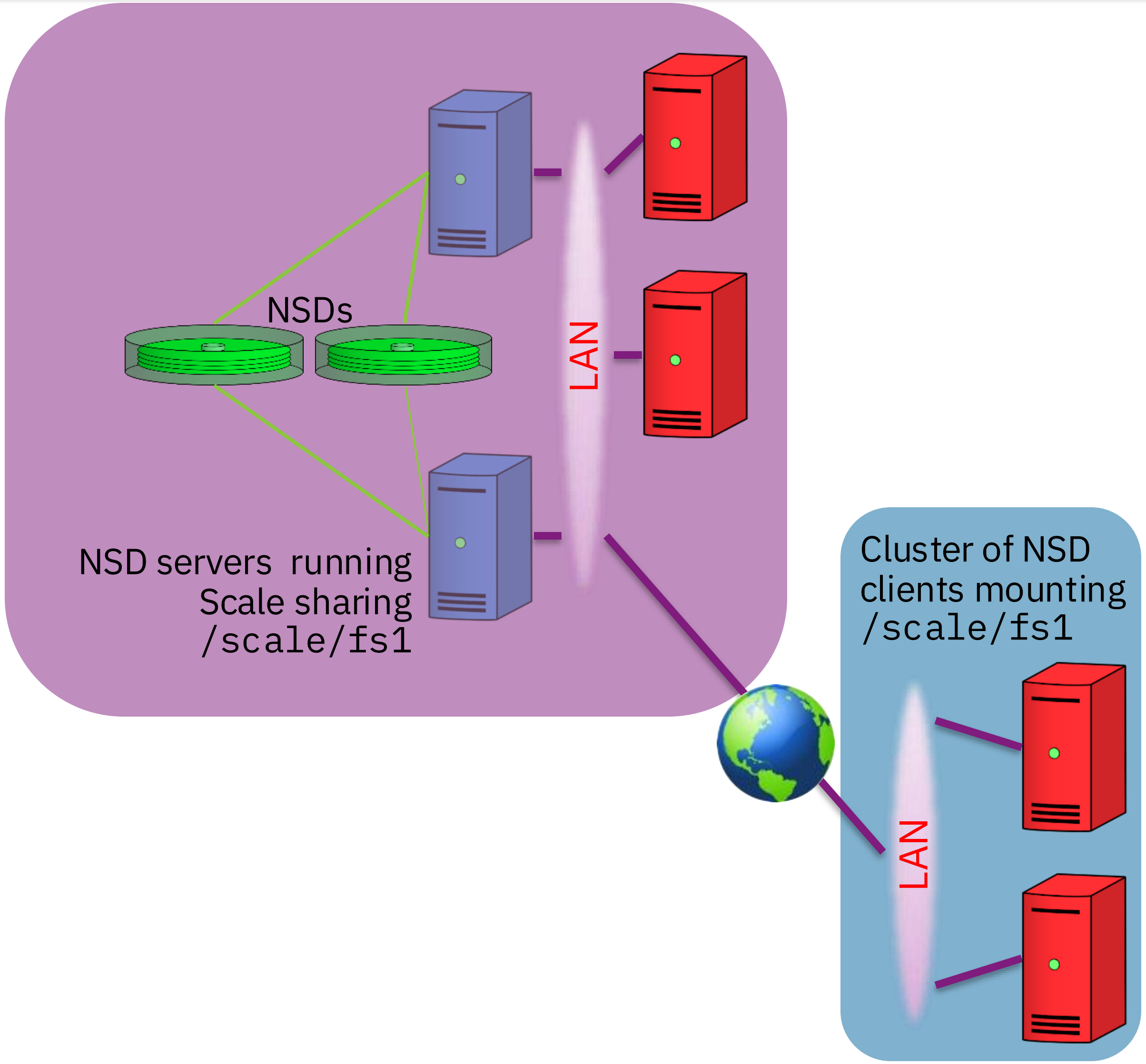
Scale clusters can mount file systems deployed on other clusters.

- It is common to have multiple Scale clusters in a single data center, but it is also possible to more a remote cluster at a distance.

Multi-cluster remote mounts can use either the NSD protocol over the LAN, or directly access NSDs over through a SAN.

- But in all cases, LAN connectivity is required.

Likely problem with *long-distance* remote mounts: latency! But this is typically not an issue within a single data center.



Overview of AFM caching

AFM **cache filesets** cache data from a target or **home**, using several modes:

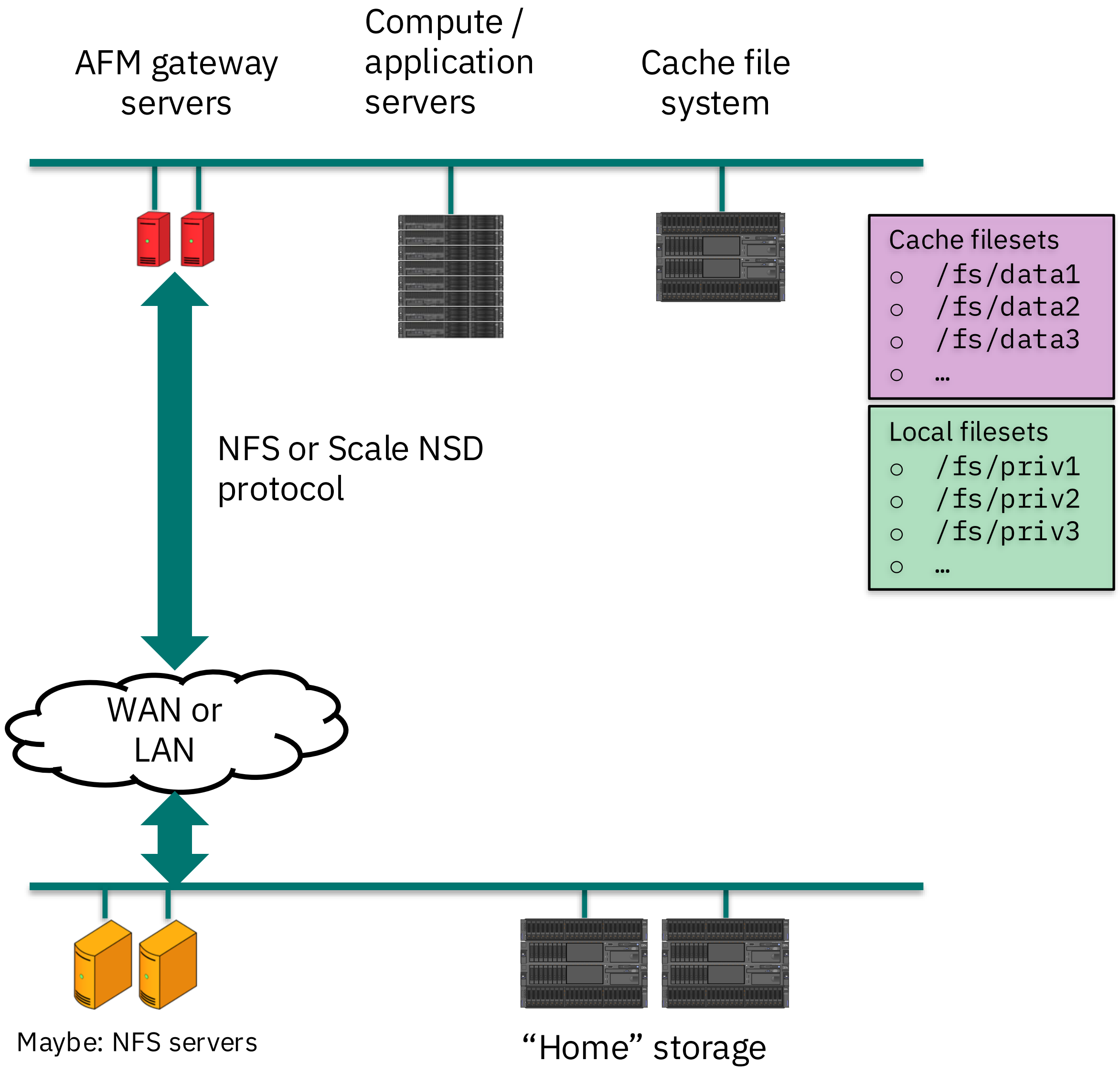
- **Read-only** – A read-only cache of the home
- **Local update** – Read-only cache, but local changes allowed
- **Single writer** – Only the single cache; it alone may update home
- **Independent writers** – multiple caches may update same home

Gateway servers maintain the freshness of caches and send updates to the home. They queue updates even if the link to home is down.

A **home** is a directory tree being cached by AFM somewhere else. The home is unaware of the cache and need not even be in Scale.

The transport is either the native Scale NSD transport, or NFS.

Cache revalidation is on demand, and the granularity of updates is the file system block.



Central data storage with AFM caches at branch offices

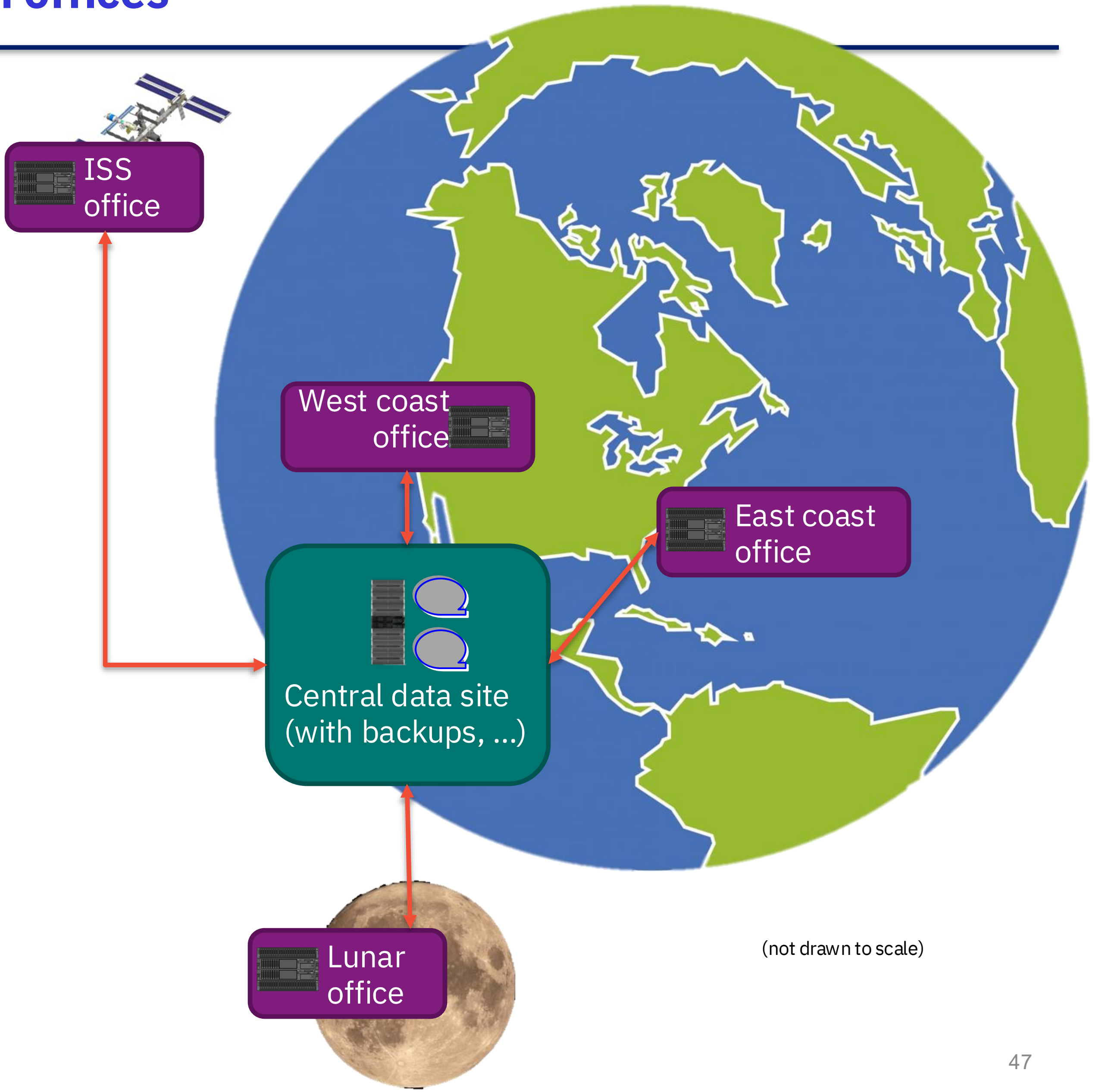
Central data center contains “home” storage.

- Backups, archives, etc. are handled at this single central site.

Branch offices have local caches.

- Caches can be small but fast.
- AFM masks much of the latency and unreliability of WAN links.
- Central site is the “source of truth”, containing the “master copy” of all data, so loss of a cache is not itself catastrophic.

Edge computing often uses a similar approach.



(not drawn to scale)

Creating a global namespace

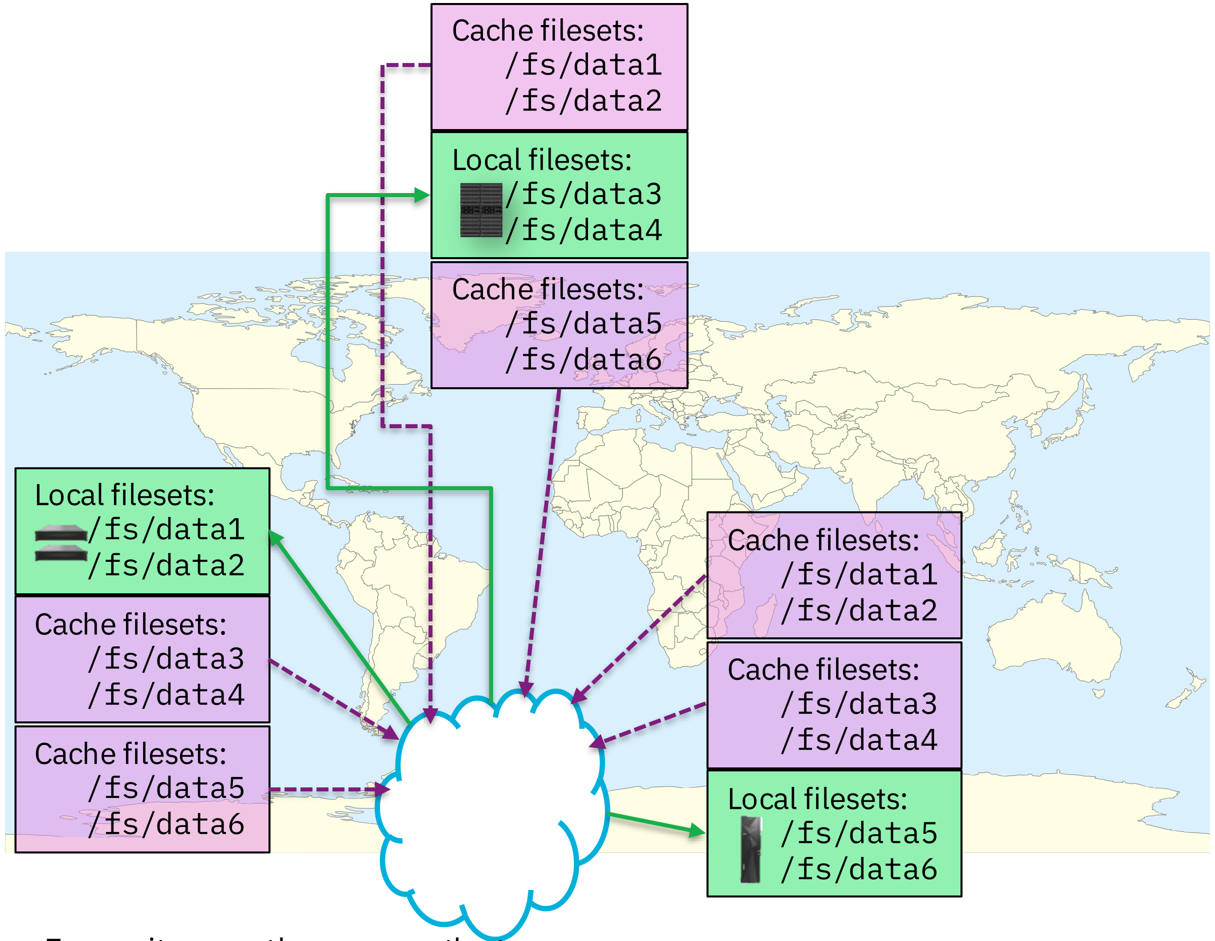
In this use case, there is no central site serving as the source of truth for all other sites.

- Rather, each site is the source of truth of some portion of the namespace but needs access to all parts of the namespace.

Each site is an independent Scale cluster with several filesets in the file system(s):

- Local filesets are standard filesets for that portion of the name space served as the source a truth.
- Otherwise, AFM cache filesets point to the appropriate "home" serving as the corresponding source of truth.

Each site can have an identical view of the entire namespace but dividing up the responsibilities of ownership.



Cache filesets: /fs/data1 /fs/data2
Local filesets: /fs/data3 /fs/data4
Cache filesets: /fs/data5 /fs/data6
Local filesets: /fs/data1 /fs/data2
Cache filesets: /fs/data3 /fs/data4
Cache filesets: /fs/data5 /fs/data6
Cache filesets: /fs/data1 /fs/data2
Cache filesets: /fs/data3 /fs/data4
Local filesets: /fs/data5 /fs/data6

Every site uses the same paths to access files, through:

```
/fs/data1     /fs/data3     /fs/data5
/fs/data2     /fs/data4     /fs/data6
```

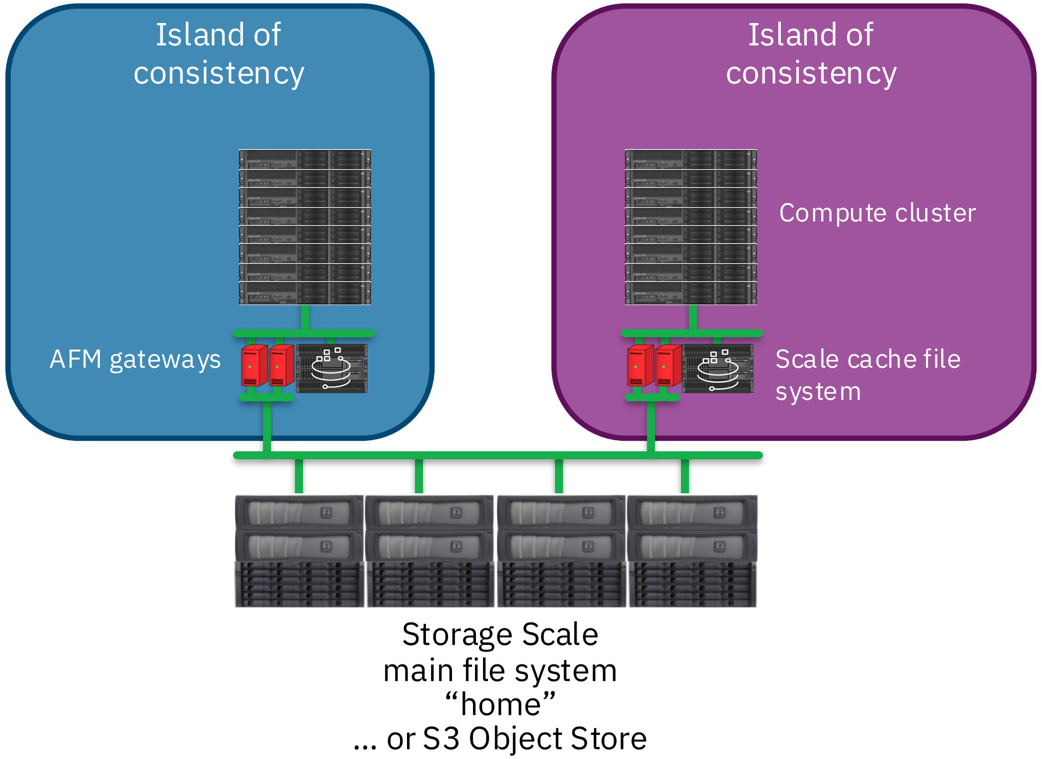
High-performance tier using AFM

A single "main file system" may be made available to multiple compute clusters, each a separate Scale cluster.

- A cluster's cache file system need only be large enough for the actual working set size, and perhaps some local storage, e.g., scratch space.
- The AFM caches function as "burst buffers" to accelerate storage speeds for the local compute cluster.
- Different compute clusters may use different interconnect technologies, e.g., different InfiniBand generations.

Each compute cluster and corresponding AFM cache is a separate "island of consistency"

- Within a single island, Scale lock management works across all nodes.
- Between islands, or between an island and the "home", there is no lock management.



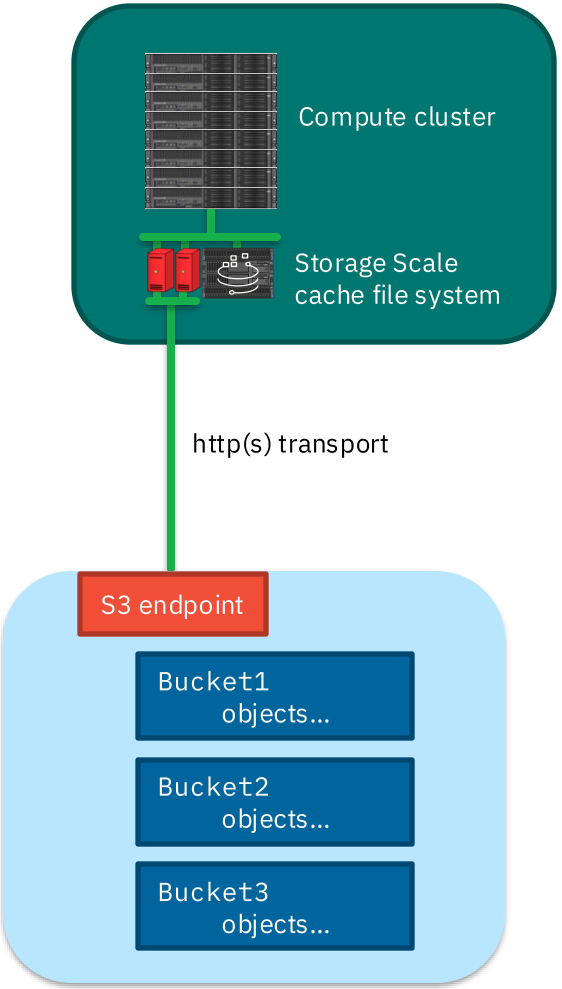
Archiving to cloud storage using AFM

AFM filesets act as caches of cloud buckets.

Files written to a cache will be automatically written to the cloud bucket.

- This can either happen automatically, or upon request.

Additional (or new) caches can be made of the same buckets – making those file available.



AFM-based Active/Passive Asynchronous DR

AFM-DR turns caching sideways!

The AFM Single Write cache model is extended to support a writeable **primary** fileset associated with a read-only **secondary** fileset.

Data can be trucked to preload the secondary fileset. (With other AFM modes, cache must start empty, copying data from home as it is fetched.)

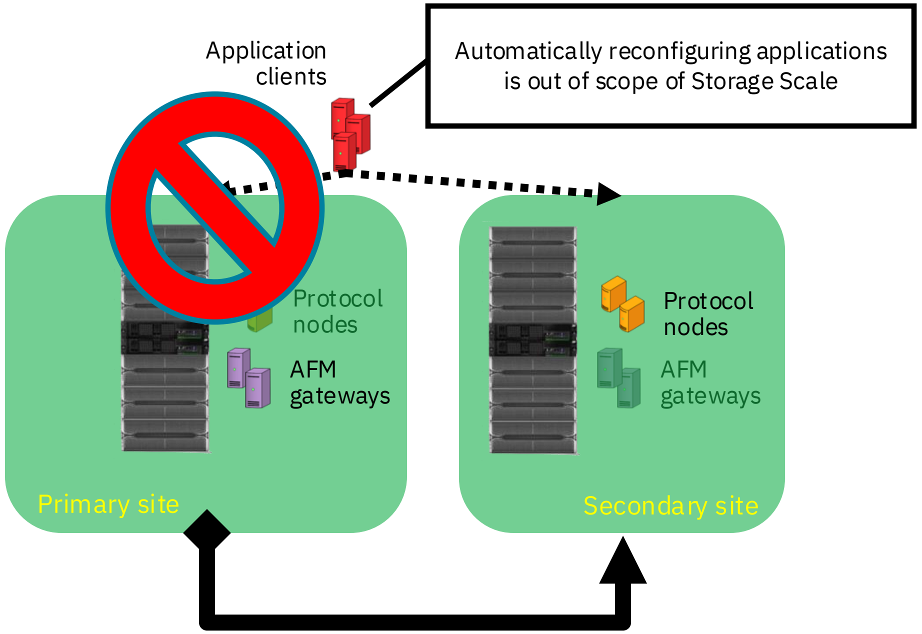
Data written to the primary fileset is constantly being replicated to the secondary fileset, asynchronously.

- *Optionally*, a regular “peer snapshot” can be taken at a configurable interval, based on an RPO.
- When promoting the secondary to be the acting primary, it can be rolled back to this snapshot.
- Mostly useful when there is risk out-of-order updates leading to a corrupted acting primary.

Upon failover, the secondary site becomes the acting primary site (and the fileset becomes writeable).

Upon failback, the original primary is updated with all the changes made to the acting primary (original secondary). Then the secondary is converted back to a true secondary.

If necessary, a new primary or secondary site can be established.



Storage Scale

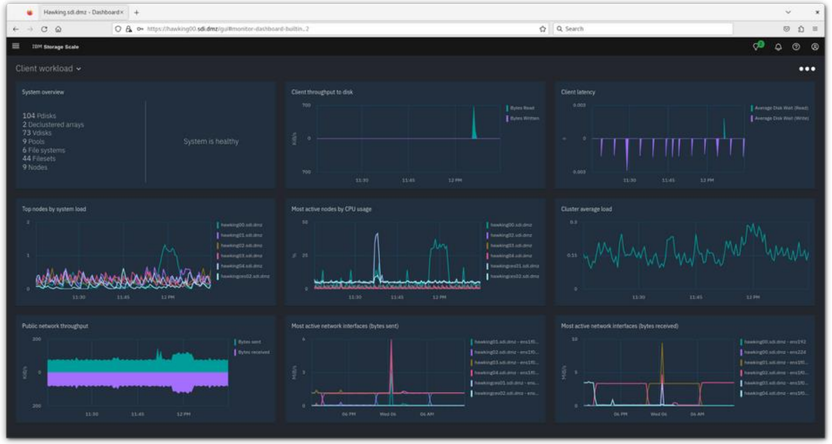
Introspection: Management, monitoring, and auditing

… to keep things running smoothly

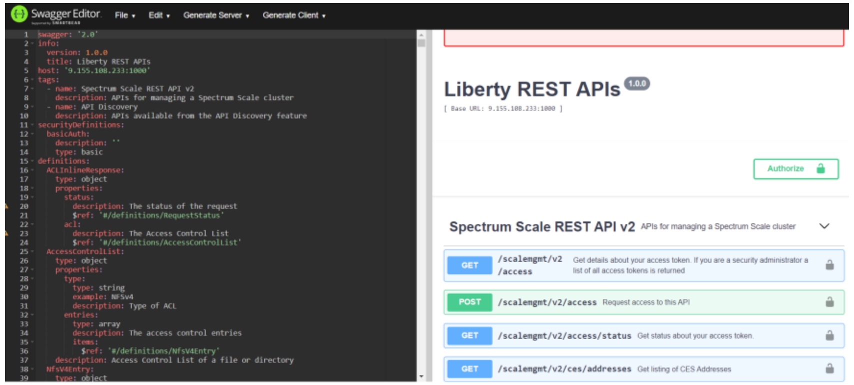
Managing a Scale cluster

There are multiple ways to manage a Scale cluster:

- The traditional **CLI** uses consistent naming and syntax and is suitable for both scripting and regular systems administration.
- The Scale **GUI** provides tools to manage a cluster such as creating filesets, exported shares, and snapshots.
- The **REST API** enables creating advanced tools for managing a Scale cluster.
  - The GUI is itself a client of the REST API.
- The brand-new **Native REST API** provides secure role-based management of Scale nodes and resources, along with a new **`scalectl`** CLI that non-root users can use to manage those aspects of the Scale cluster consistent with their role.



GUI dashboard



```
swagger: '2.0'
info:
  version: 1.0.0
  title: Liberty REST APIs
host: '9.155.108.233:1000'
tags:
  - name: Spectrum Scale REST API v2
    description: APIs for managing a Spectrum Scale cluster
  - name: API Discovery
    description: APIs available from the API Discovery feature
securityDefinitions:
  basicAuth:
    description: ''
    type: basic
definitions:
  ACLInlineResponse:
    type: object
    properties:
      status:
        description: The status of the request
        $ref: '#/definitions/RequestStatus'
      acl:
        description: The Access Control List
        $ref: '#/definitions/AccessControlList'
  AccessControlList:
    type: object
    properties:
      type:
        type: string
        example: NFSv4
        description: Type of ACL
      entries:
        type: array
        description: The access control entries
        items:
          $ref: '#/definitions/NfsV4Entry'
    description: Access Control List of a file or directory
  NfsV4Entry:
    type: object
```

Liberty REST APIs 1.0.0

[Base URL: 9.155.108.233:1000]

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

GET `/scalemgmt/v2/access` Get details about your access token. If you are a security administrator a list of all access tokens is returned

POST `/scalemgmt/v2/access` Request access to this API

GET `/scalemgmt/v2/access/status` Get status about your access token.

GET `/scalemgmt/v2/ces/addresses` Get listing of CES Addresses

REST API
(accessed through Swagger)

Monitoring Scale’s health and performance

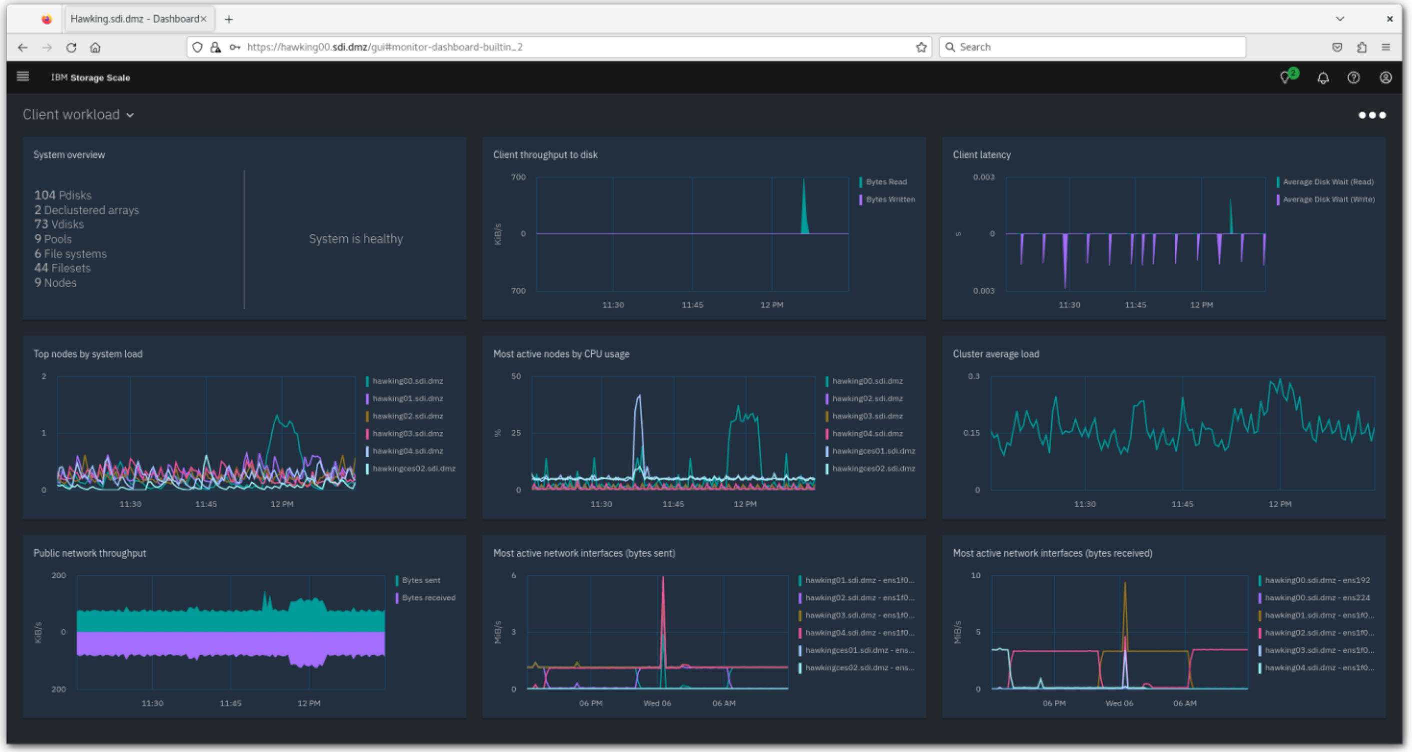
The Storage Scale GUI supports monitoring health and performance.

- The performance monitoring system maintains history.

REST API exposes performance and health information to other consumers.

Performance and health monitoring framework exposed to the CLI with `mmperfmon` and `mmhealth`

“Call home” facility assists with proactively collecting data for problem determination.



Hawking.sdi.dmz - Dashboard

https://hawking00.sdi.dmz/gui#monitor-dashboard-builtin_2

IBM Storage Scale

Client workload

System overview

104 Pdisks
2 Declustered arrays
73 Vdisks
9 Pools
6 File systems
44 Filesets
9 Nodes

System is healthy

Search
Client throughput to disk
700
0
700
KiB/s
11:30
11:45
12 PM
Bytes Read
Bytes Written
Client latency
0.003
0
0.003
s
Average Disk Wait (Read)
Average Disk Wait (Write)
Top nodes by system load
2
1
0
hawking00.sdi.dmz
hawking01.sdi.dmz
hawking02.sdi.dmz
hawking03.sdi.dmz
hawking04.sdi.dmz
hawkingces02.sdi.dmz
Most active nodes by CPU usage
50
25
0
%
hawking00.sdi.dmz
hawking02.sdi.dmz
hawking03.sdi.dmz
hawking04.sdi.dmz
hawkingces01.sdi.dmz
hawkingces02.sdi.dmz
Cluster average load
0.3
0.15
0
Public network throughput
200
0
200
KiB/s
Bytes sent
Bytes received
Most active network interfaces (bytes sent)
6
3
0
MiB/s
06 PM
Wed 06
06 AM
hawking01.sdi.dmz - ens1f0...
hawking02.sdi.dmz - ens1f0...
hawking03.sdi.dmz - ens1f0...
hawking04.sdi.dmz - ens1f0...
hawkingces01.sdi.dmz - ens...
hawkingces02.sdi.dmz - ens...
Most active network interfaces (bytes received)
10
5
0
MiB/s
hawking00.sdi.dmz - ens192
hawking00.sdi.dmz - ens224
hawking01.sdi.dmz - ens1f0...
hawking02.sdi.dmz - ens1f0...
hawking03.sdi.dmz - ens1f0...
hawking04.sdi.dmz - ens1f0...

Clustered Watch Folders (DME, DE, ECE)

The clustered watch folder facility monitors file system, filesets, and “inode spaces” for selected events, as “watches”.

- Somewhat like the Linux inotify, except...
- Events and associated information (JSON-formatted) are written to a Kafka topic.
- Each cluster node is a Kafka producer, so must have network access to the Kafka brokers.
- Kafka consumers can react to file system events.

The **mmwatch** command is used to configure and examine watches.

- Up to 25 watches can be established.

Features and limitations

- NFS (Ganesha), SMB, and native POSIX supported.
- Only available on Linux systems. (SELinux is supported.)

File access event	Description
IN_ACCESS	A file was accessed (read or ran).
IN_ATTRIB	Metadata was changed (for example, **chmod**, **chown**, **setxattr** etc.).
IN_CLOSE_NOWRITE	A file or folder that was not opened for writing was closed.
IN_CLOSE_WRITE	A file that was opened for writing was closed.
IN_CREATE	A file or folder was created in a watched folder.
IN_DELETE	A file or folder was deleted from a watched folder.
IN_DELETE_SELF	A watched file or folder was deleted.
IN_IGNORED	Event that is triggered when a file system is unmounted. Always follows the IN_UNMOUNT event.
IN_ISDIR	A directory is listed.
IN_MODIFY	A file was modified (for example, write or truncate).
IN_MOVE_SELF	A folder that was being watched was moved.
IN_MOVED_FROM	A file within the watched folder was renamed.
IN_MOVED_TO	A file was moved or renamed to this folder.
IN_OPEN	A file or folder was opened.
IN_UNMOUNT	A file system that is being watched is unmounted.

File Audit Logging

Captures the most common types of file activity:

- `open`, `close`, `delete`, `rename`, POSIX permission changes, ACL changes, etc., configurable with `mmaudit`.
- Doesn’t capture internal operations (e.g., restripe).

Events are captured within the Storage Scale `mmfsd` daemon, representing attributes of filesystem action at that point.

Interfaces to IBM Guardium and Varonis.

- Each file system enabled has a dedicated fileset where the audit logs will go.
  - Default option is `.audit_log` at the root of the file system.
- Audit log files are nested within `/FSNAME/.audit_log/topic/year/month/date/*`
- Log files are append-only with default retention of 365 days.

```
{"LWE_JSON": "0.0.1", "path": "/newfs/1Kfile2.restore", "oldPath": null,
"clusterName": "pardie.cluster", "nodeName": "c6f2bc3n10", "nfsClientIp": "",
"fsName": "newfs", "event": "OPEN", "inode": "26626", "openFlags": "32962",
"poolName": "sp1", "fileSize": "0", "ownerUserId": "0", "ownerGroupId": "0",
"atime": "2017-10-25_12:36:22-0400", "ctime": "2017-10-25_12:36:22-0400",
"eventTime": "2017-10-25_12:36:22-0400", "clientUserId": "0", "clientGroupId":
"0", "processId": "10437", "permissions": "200100644", "acls": "u::rwc, g::r,
o::r, ", "xattrs": null }
```

Example audit log entry

Storage Scale

What is a Scale System?

Or maybe you’ve heard it called ESS?

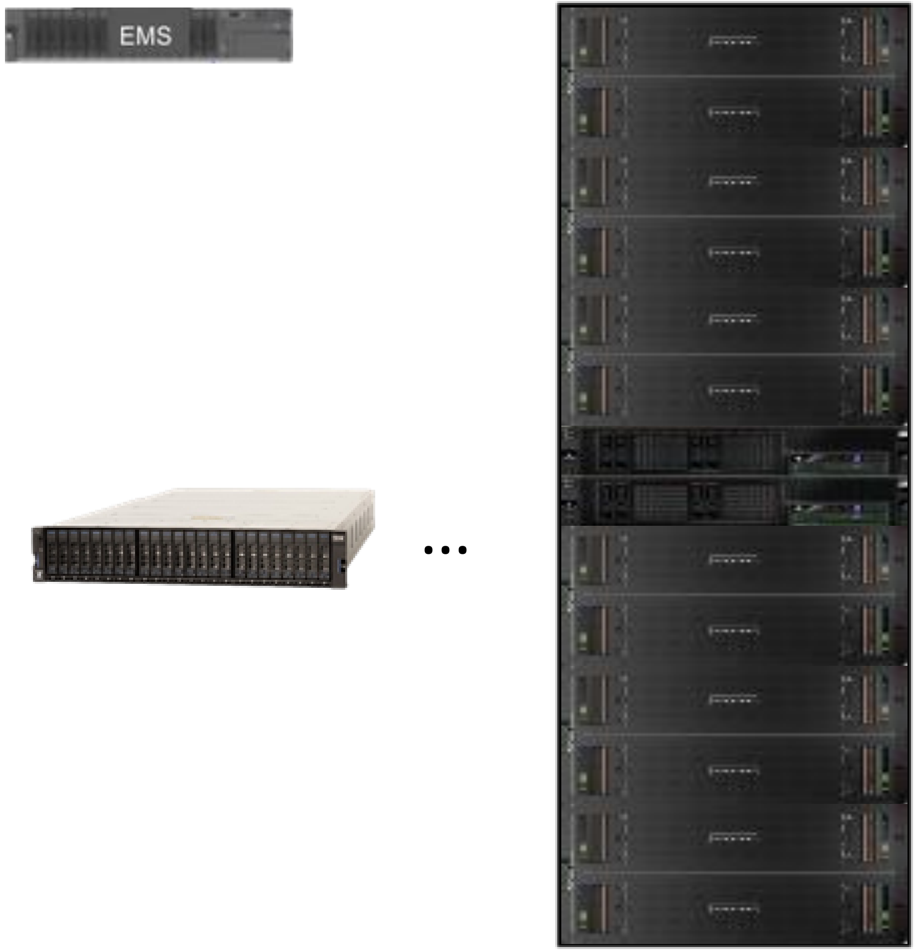
IBM Scale System (formerly ESS)

The IBM Scale System (SSS) 6000 and ESS 3500 are the latest generation of integrated and tested IBM-provided **NSD server building block** solutions for Scale

- Fully validated IBM hardware and software stack
- Pre-assembled, pre-configured and installed
- Scale + Storage Scale RAID + Scale System GUI
- Scale System-aware performance, monitoring, installation, and upgrade
- Containerized Scale deployment based on Ansible playbooks, simplifying install and updates.

Scale System mitigates risks and makes it quicker to deploy and grow a Scale cluster

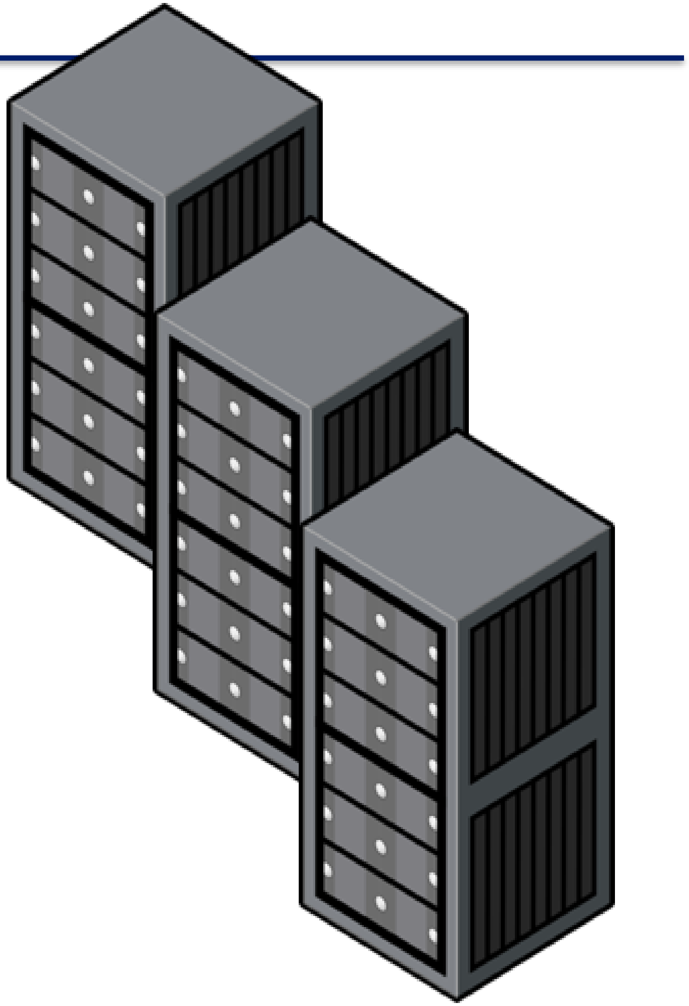
Erasure Code Edition (ECE) is NOT build-your-own Scale System, even though both are based on Storage Scale RAID.



What is Storage Scale Erasure Code Edition (ECE) ?

This is Scale for storage rich nodes using “Storage Scale RAID” with Data Management Edition (DME)

- Scale running in storage rich servers connected to each other with a high-speed network infrastructure
- Bring your own hardware (as long as it meets requirements)
- Provided Storage devices can be HDD, SSD, NVMe or a mixture
- Features of an enterprise storage controller all in software

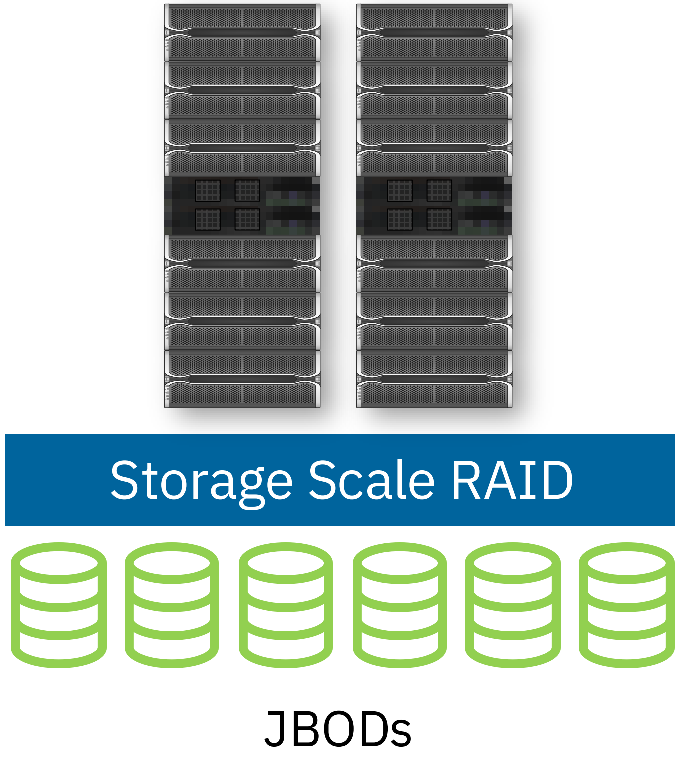


This is similar in *concept* to Storage Scale System. However, Scale ECE is built using commodity storage-rich servers, without shared storage.

Declustered software RAID

IBM **Storage Scale RAID** is a *software* implementation of “declustered” or “**distributed RAID**”:

- Extremely fast rebuild after a disk failure, with minimal impact on performance
- Very strong data integrity checks
  - Additional erasure codes, such as 8+3p
  - Error detection codes enable detecting track errors and dropped writes
- Consistent performance from 0 – 99% utilization or 1 to many jobs in parallel
- All NSDs are wide-striped over many internal disks.



Storage Scale RAID is currently available only with Storage Scale System (IBM’s reference architecture) and Erasure Code Edition.

YouTube: IBM Storage Scale RAID for petabyte storage: https://www.youtube.com/watch?v=2g5rx4gP6yU

Storage Scale RAID – Data Integrity Enhancements

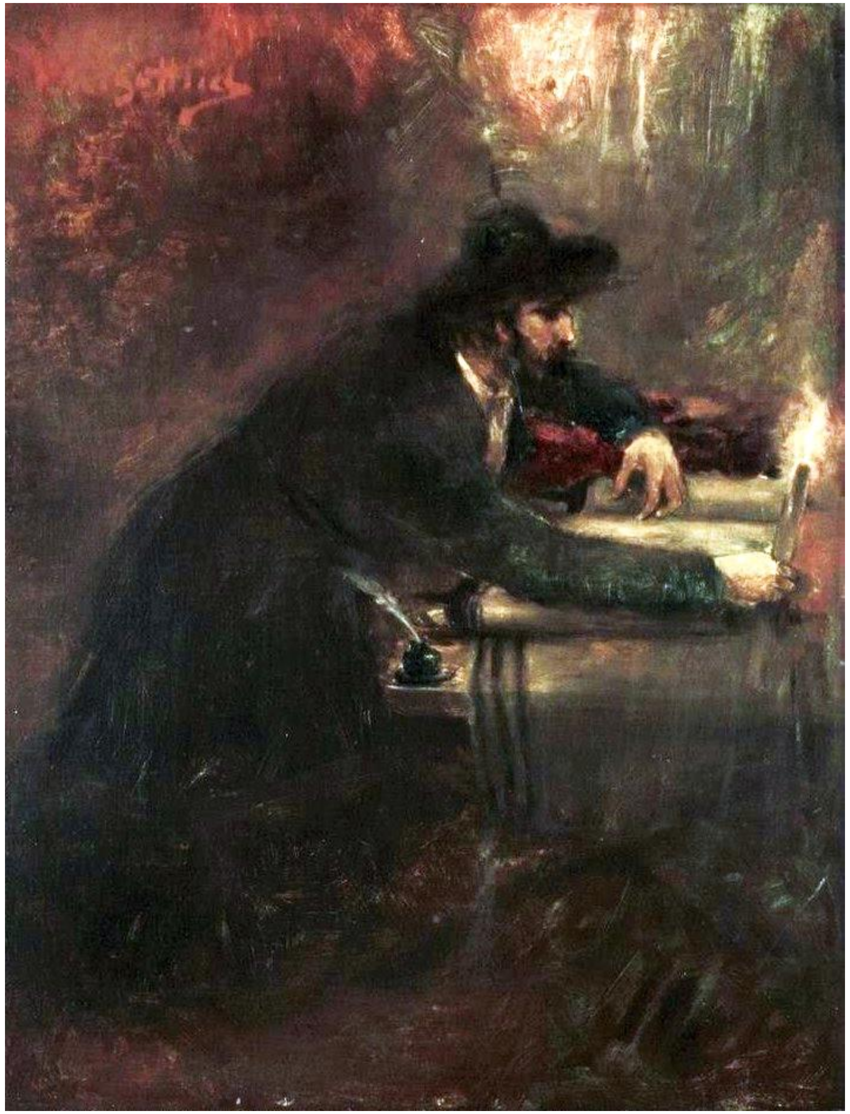
End-to-end *checksum provides superior protection to current hardware-based RAID arrays*

- Checksums maintained on disk and in memory and are transmitted to/from client
- Eliminates soft/latent read errors
- Eliminates silent dropped writes

Protection against lost writes eliminates additional costs to deploy mirroring alternatives

Advanced "Disk Hospital" proactively diagnoses and reduces potential issues, expediting repair actions

- Remote call-home capability for faster problem resolution



Storage Scale

The best Global Data Platform

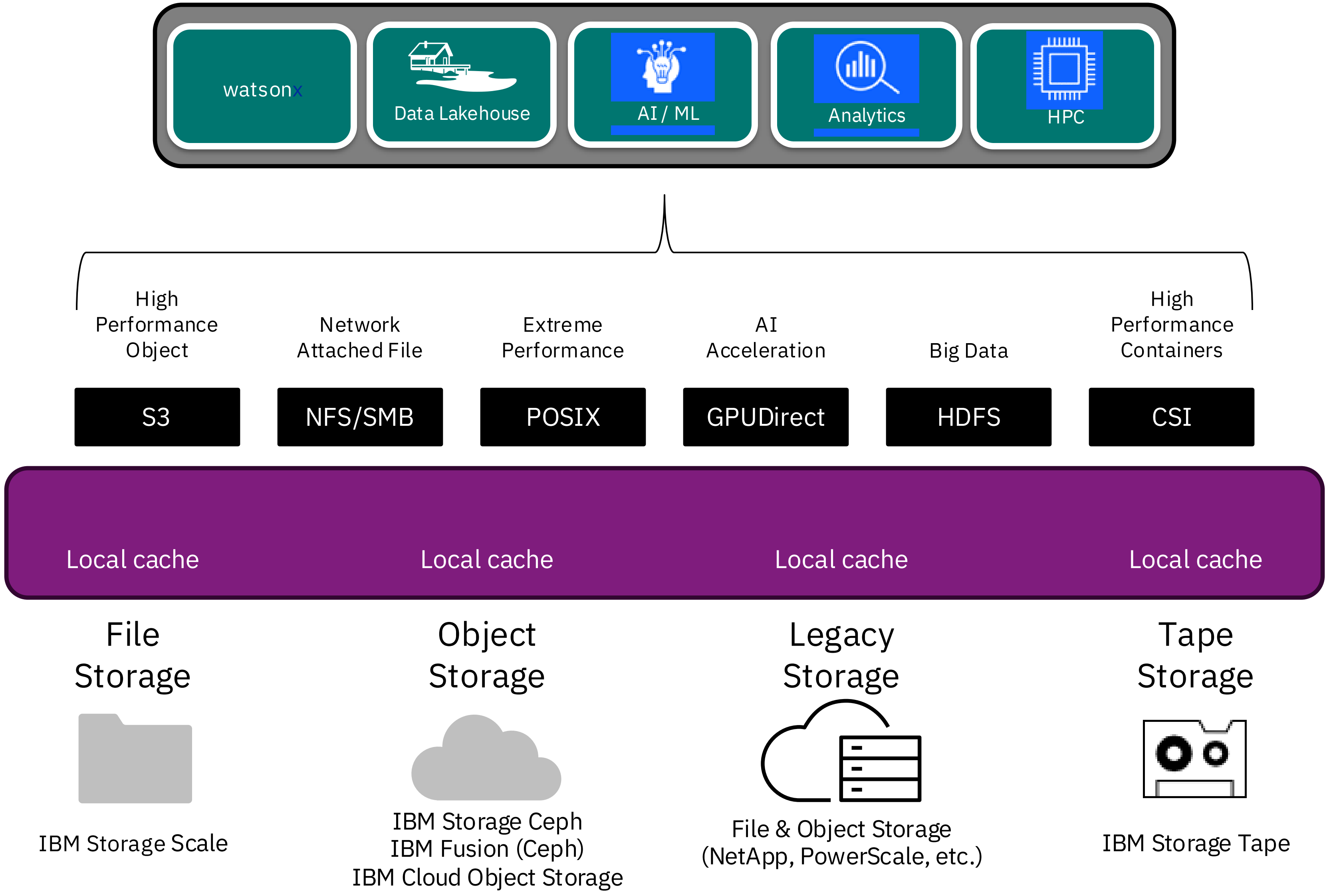
… based on the best General Parallel File System

Storage Scale: Still the best General Parallel File System!
Storage Scale: Still the only true Global Data Platform!

Data Access Services

Data Acceleration

Data Abstraction



watsonx

Data Lakehouse

AI / ML

Analytics

HPC

High Performance Object

Network Attached File

Extreme Performance

AI Acceleration

Big Data

High Performance Containers

S3

NFS/SMB

POSIX

GPUDirect

HDFS

CSI

Local cache

Local cache

Local cache

Local cache

File Storage

Object Storage

Legacy Storage

Tape Storage

IBM Storage Scale

IBM Storage Ceph
IBM Fusion (Ceph)
IBM Cloud Object Storage

File & Object Storage
(NetApp, PowerScale, etc.)

IBM Storage Tape

Data Assurance Services

Data Administration Services

Data Aware Services

Storage Scale

Further resources

... and thank you for listening!

Further resources

IBM Documentation

- for Storage Scale – https://www.ibm.com/docs/en/storage-scale
- for Storage Scale System – https://www.ibm.com/docs/en/ess

Storage Scale FAQ – https://www.ibm.com/docs/en/STXKQY/gpfsclustersfaq.html

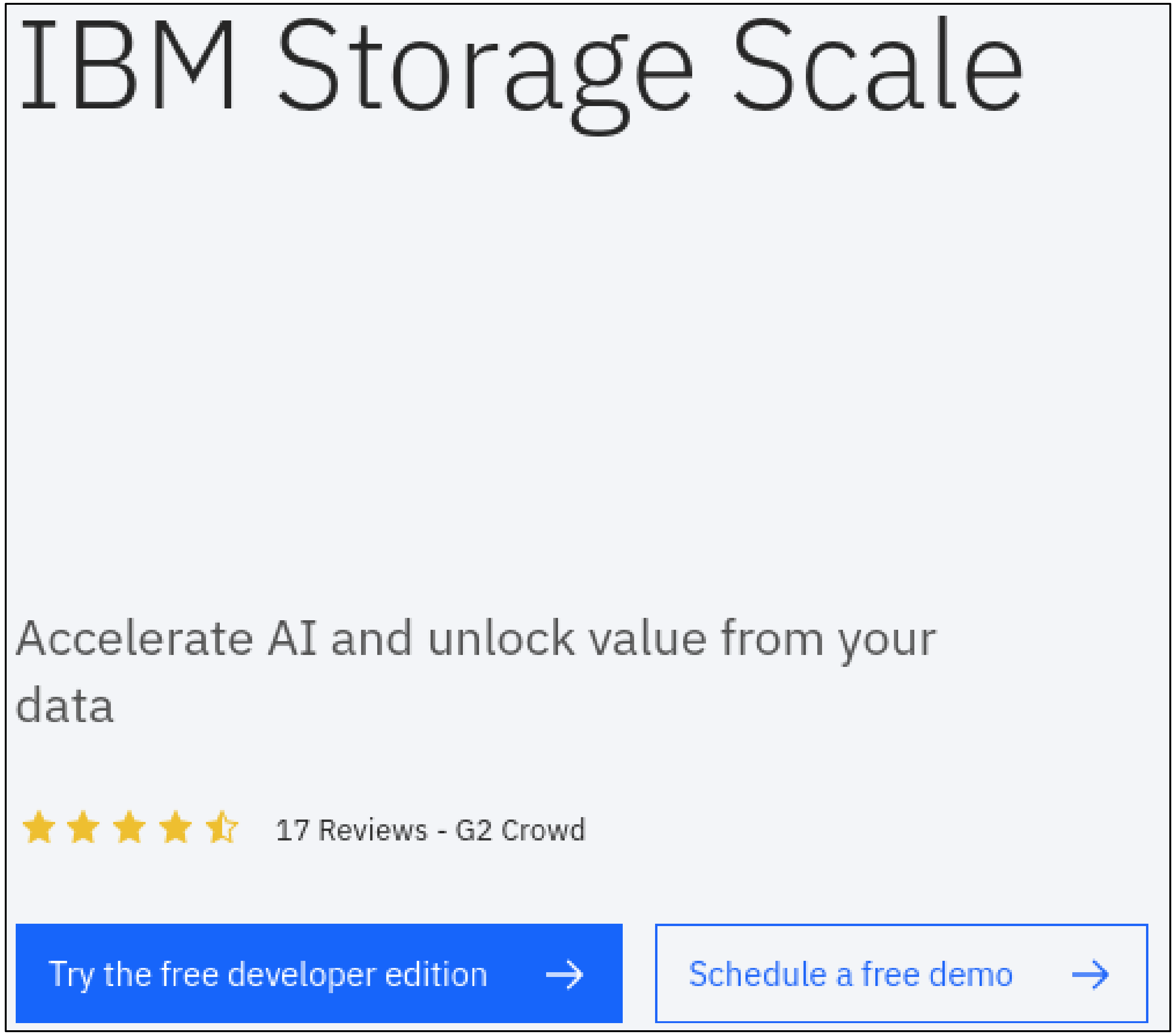
Storage Scale User Group – https://storagescale.org/

Storage Scale Resource Collection (a Box folder) – https://ibm.biz/Scale-Resource-Collection

IBM Storage Scale Developer Edition

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

- Fully functional!



IBM Storage Scale

Accelerate AI and unlock value from your data

★★★★☆ 17 Reviews - G2 Crowd

Scale User Group

The Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Storage Scale.

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

See our web page for upcoming events and presentations of past events. Join our conversation via mail and Slack.

www.storagescale.org

Accelerate with ATG Technical Webinar Series

The Free IBM ATG – Technical Series continues in 2025…..

Advanced Technology Group experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with ATG webinars, Clients, Business Partners and IBMers are welcome to send an email request to **accelerate-join@hursley.ibm.com.**

Important Links to Bookmark:

Accelerate with ATG - Click here to access the Accelerate with ATG webinar schedule for 2025, view presentation materials, and watch past replays dating back two years. **https://ibm.biz/BdSUFN**

ATG MediaCenter Channel - This channel offers a wealth of additional videos covering a wide range of storage topics, including IBM Flash, DS8, Tape, Ceph, Fusion, Cyber Resiliency, Cloud Object Storage, and more. **https://ibm.biz/BdfEgQ**

ATG Publisher Site - Learn more about the IBM Advanced Technology Group Storage team by clicking here! Dive deeper into who we are and who we serve! **https://ibm.biz/BdGSD3** **(For internal IBM only)**

Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code **5151 0447**
or

Direct link https://www.menti.com/alhsf3bgvxu6
or

QR Code