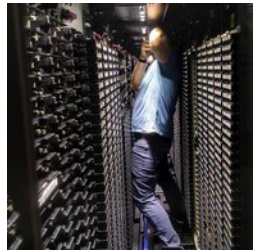


# Data Lifecycle Management - IBM

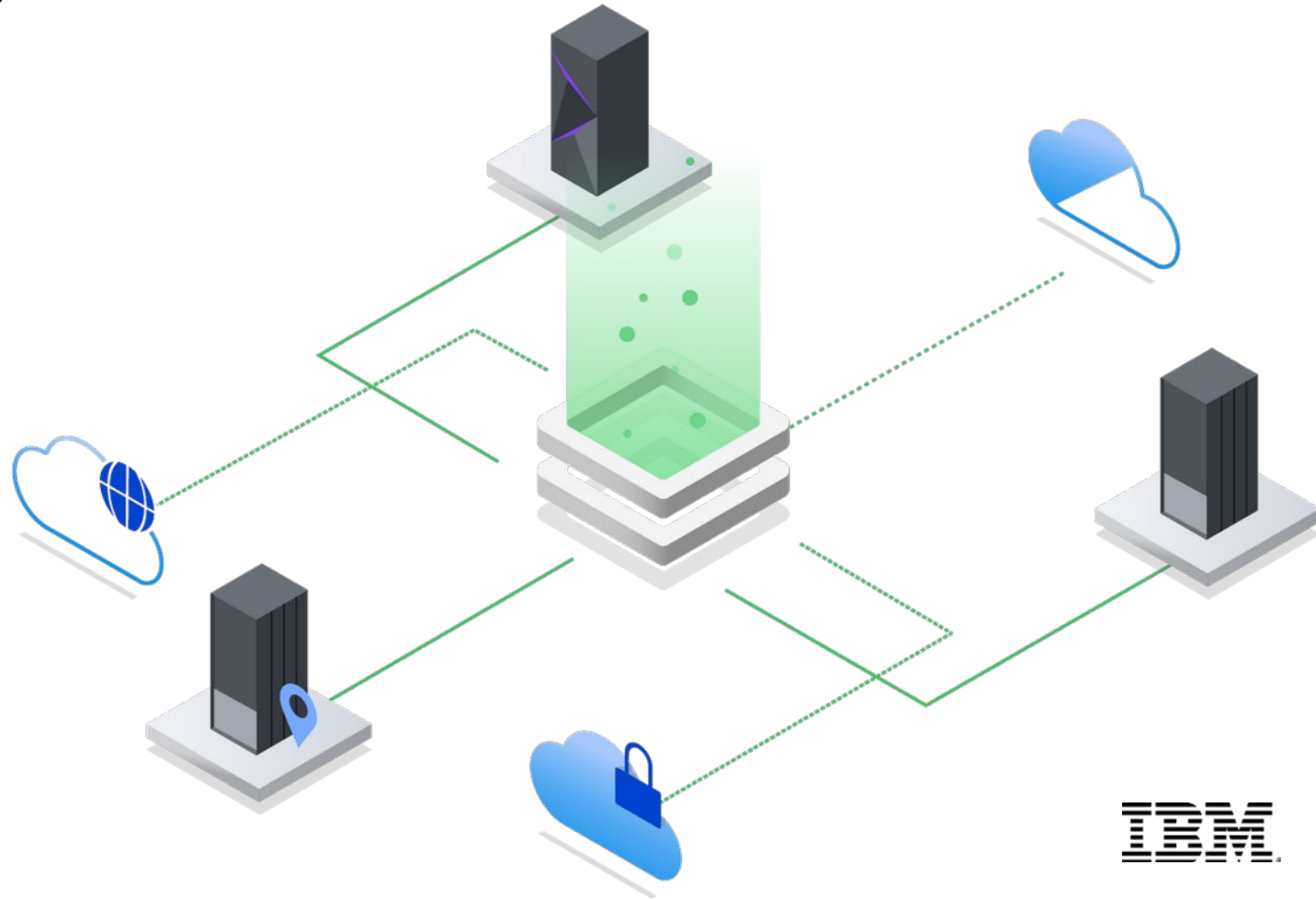
## Storage Ceph, S3 Deep Archive and other archive solutions



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



**Ben Rawlins**  
Tape Applications Engineer





**Christopher Maestas** ✓ • You  
CTO, Data and AI Storage Solutions  
1mo • 🌐



It was my pleasure to visit how IBM Storage technology is able to meet our clients demands in Chile. We have a healthy roadmap to move with them from today into the future! [#IBMStorage](#) [#GPFS](#) [#Scale](#) [#Tape](#) [#TS4500](#)



**Copernicus LAC Chile**

1mo • 🌐

Ayer recibimos en CopernicusLAC Chile a [Christopher Maestas](#), Director de **Tecnología** para soluciones de almacenamiento de datos e Inteligencia Artificial de [IBM](#). La visita se enmarcó en el proceso de... more

Show translation



Dave McDonnell and 133 others

4 comments • 2 reposts

# STORAGE for AI MATTERS

The fastest servers in the world are the world's  
slowest servers if they are waiting for data  
ALL GPUs & CPUs WAIT AT THE SAME SPEED



# Storage for Data and AI Customer Challenges



- Escalating Storage and GPU costs

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



- AI Data is Distributed

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



- Data is a Critical Business Asset

- Storage Solution:
  - ✓ Resilient to cyber threats
  - ✓ Quick recovery time
  - ✓ Highly available

# IBM Storage Scale's Global Data Platform solves the AI data platform challenges



## AI Data Challenges:

- Data bottlenecks
- Data integration
- Data silos
- Data quality
- Data availability



**Accelerate AI Results**  
Scalable Parallel Performance

**Solves the data bottleneck**



**Access the Data for AI**  
Single source of data – Global Data Platform

**Solves data integration**



**Abstraction to Optimize AI Economics**  
Data lineage preservation

**Solves data coexistence,  
no additional silos!**



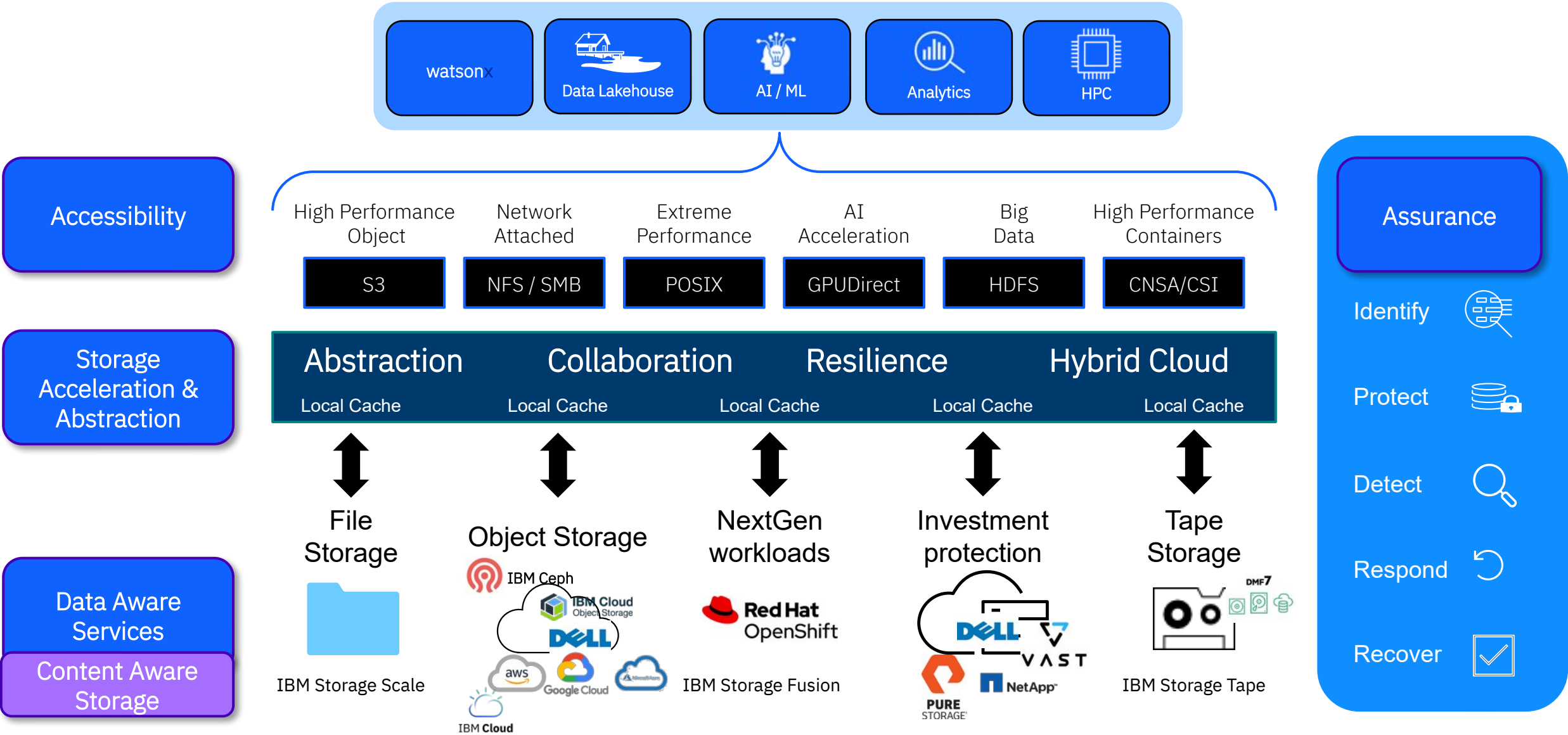
**Assurance to Secure, Govern, Protect Data**  
NIST framework compliant

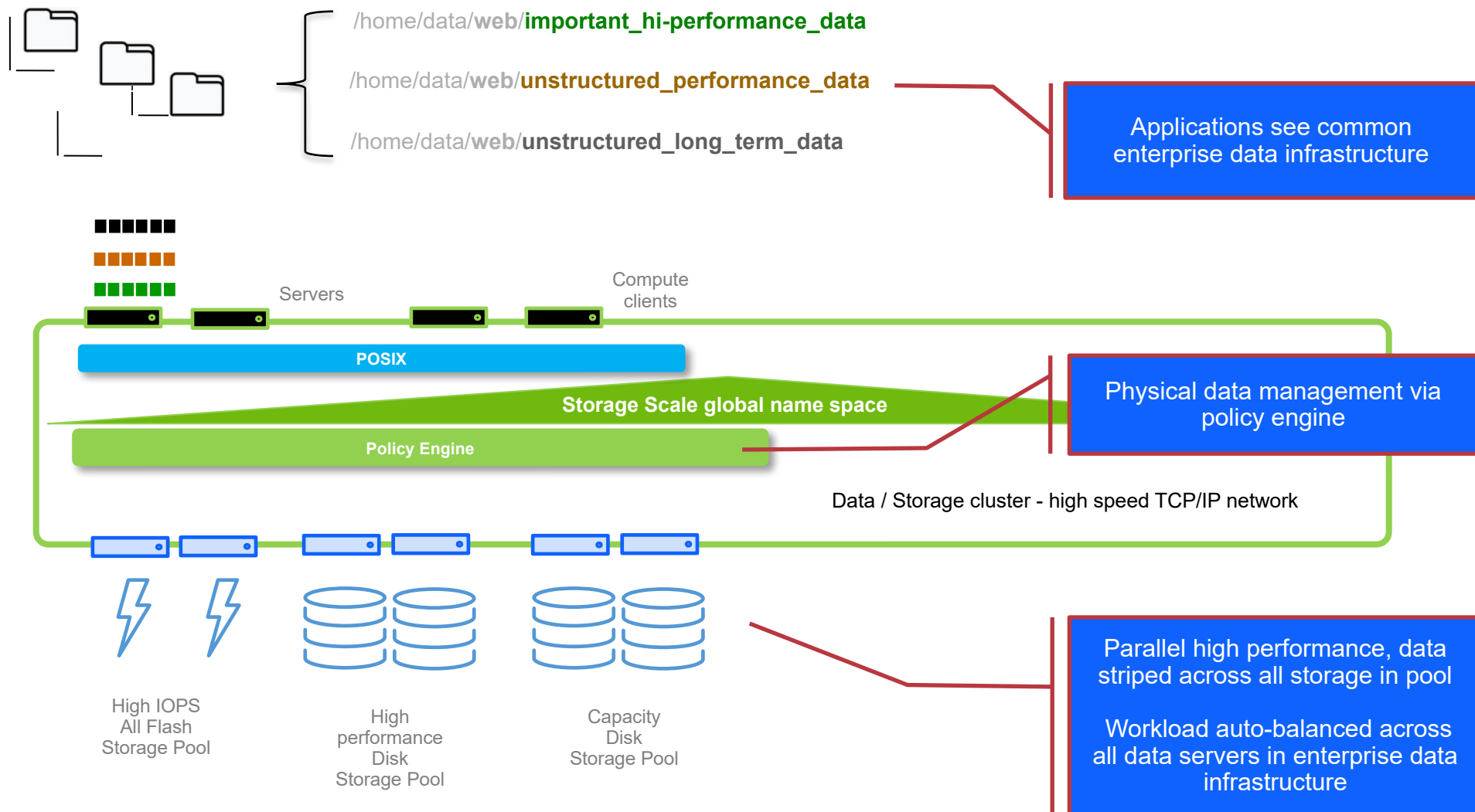
**Solves data availability**

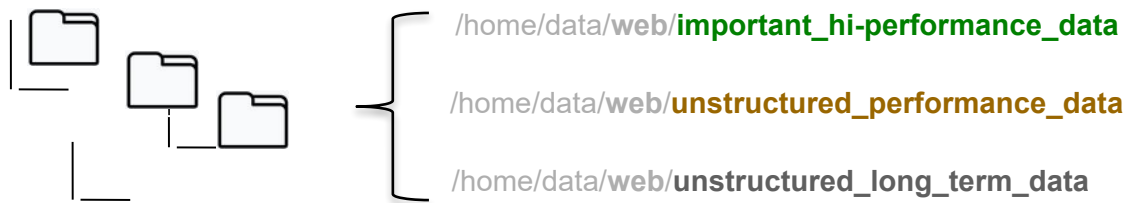
- a Global Data Platform -

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

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

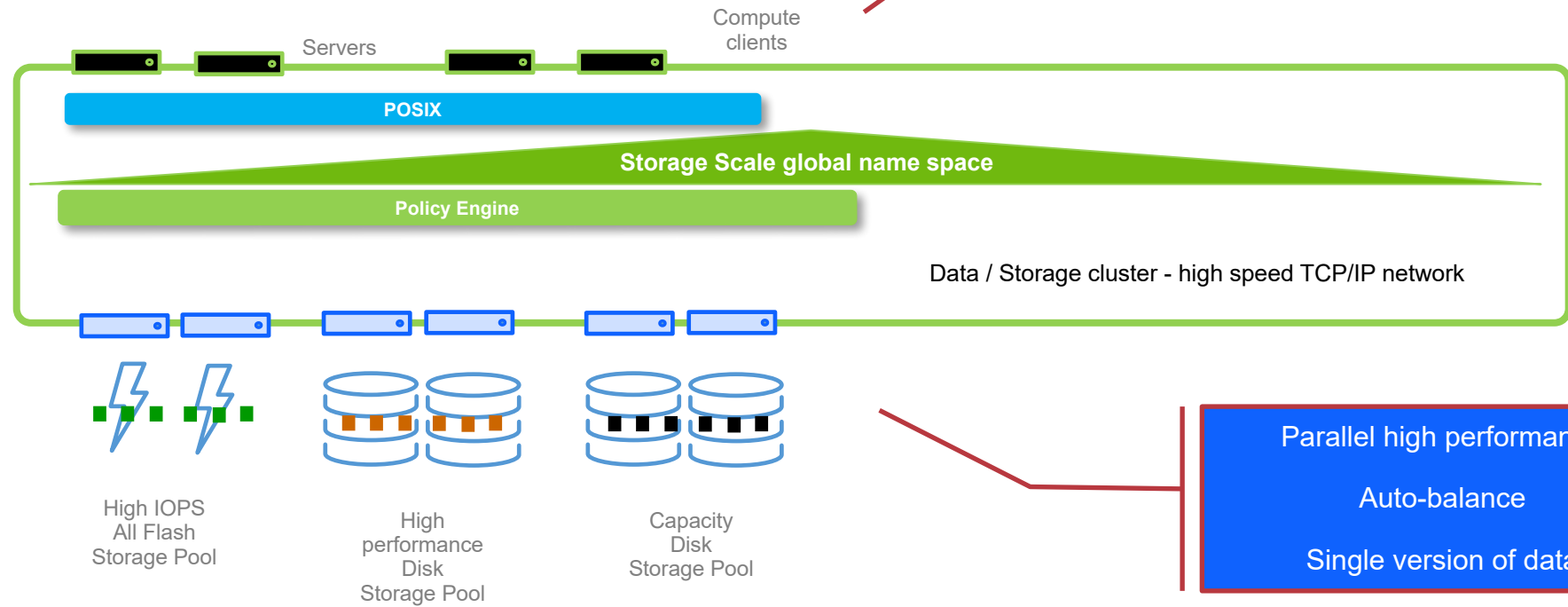






Single version of data flows to other applications / servers from common enterprise data infrastructure

Read/write with full data integrity across multiple users

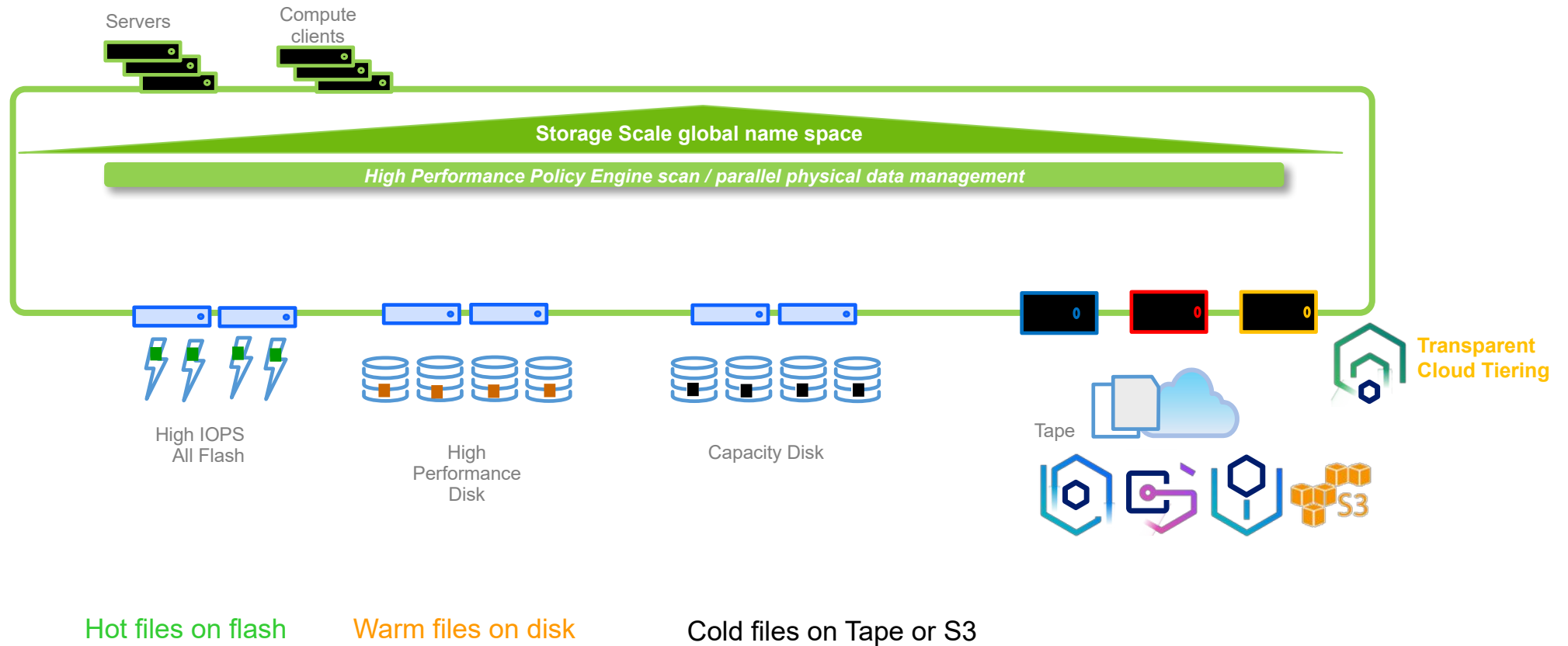


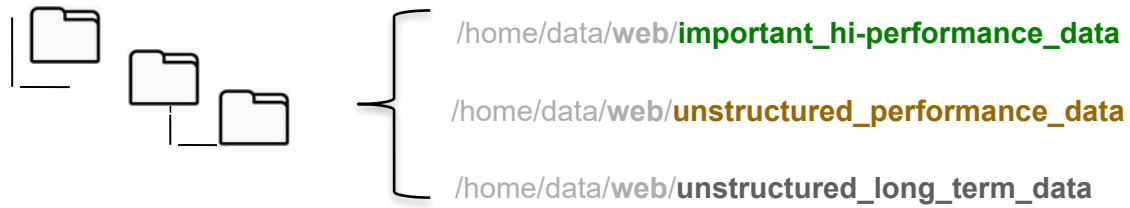
Parallel high performance

Auto-balance

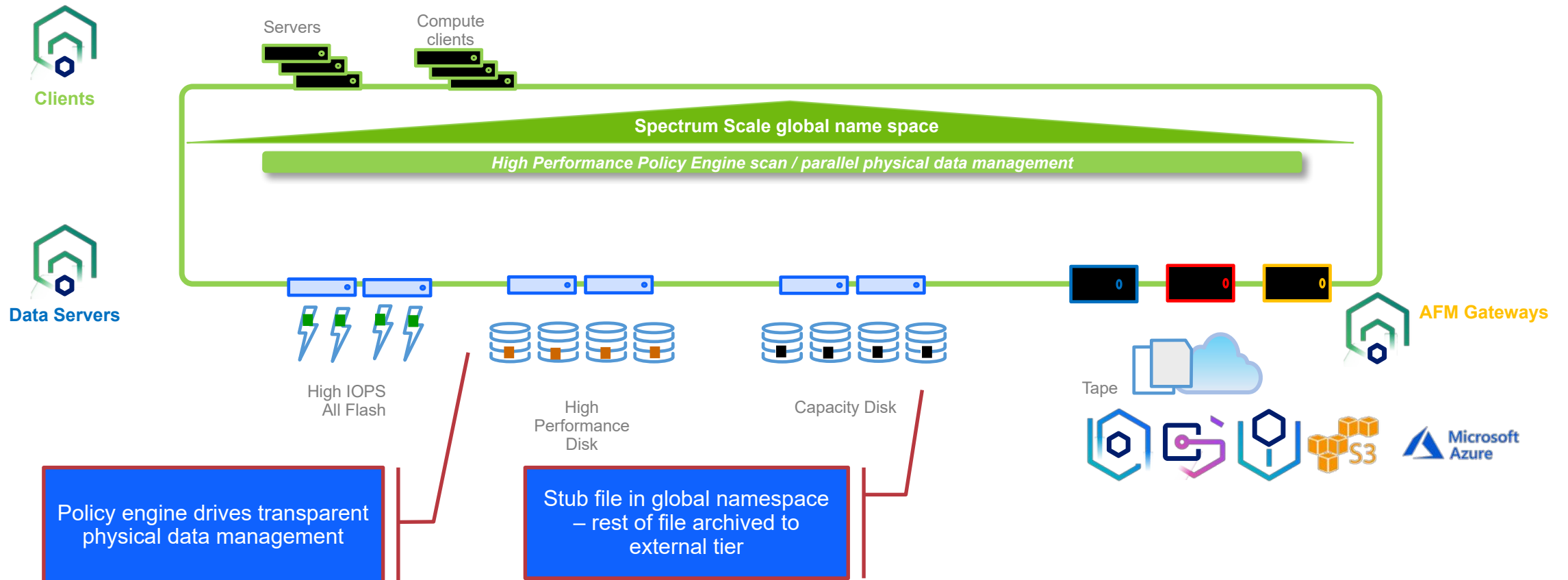
Single version of data

90% of data is never touched after 90 days - when data gets **COLDER** move it to a lower cost storage tier



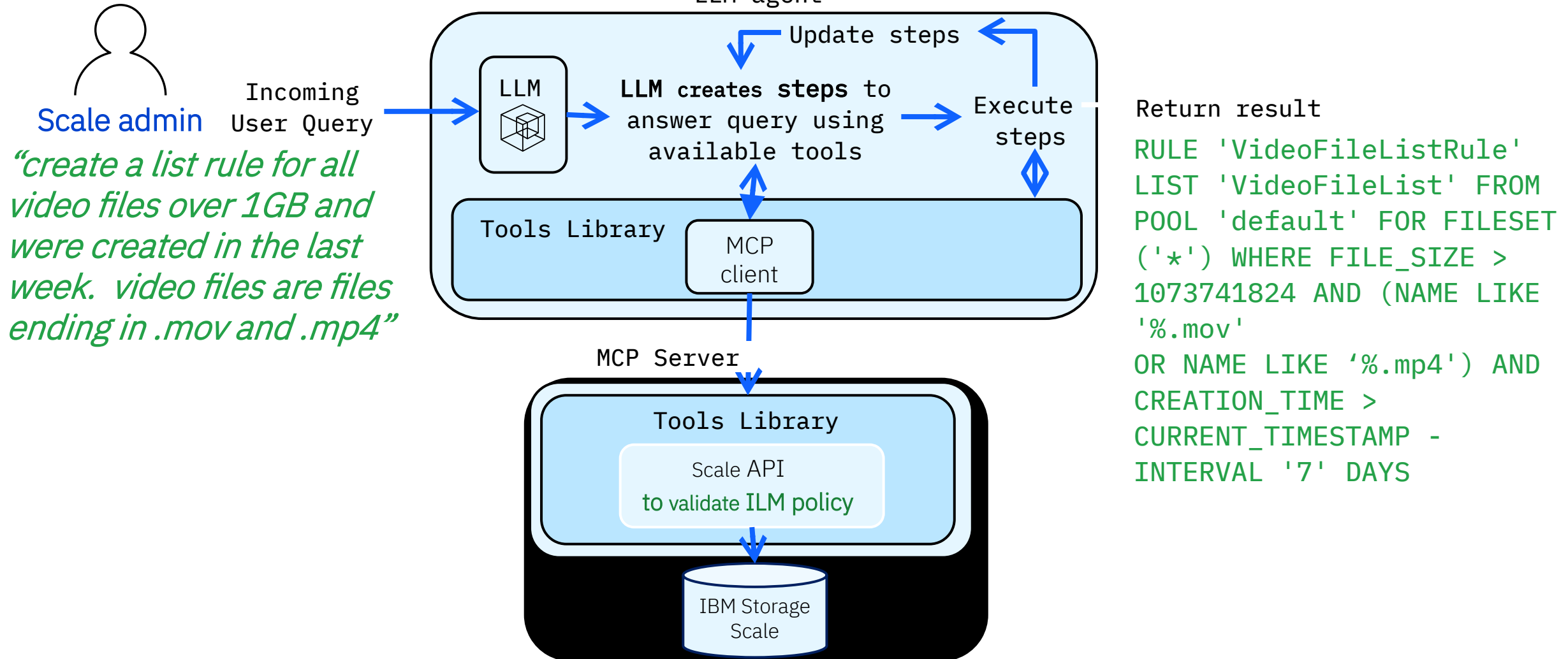


Applications always see common enterprise data infrastructure

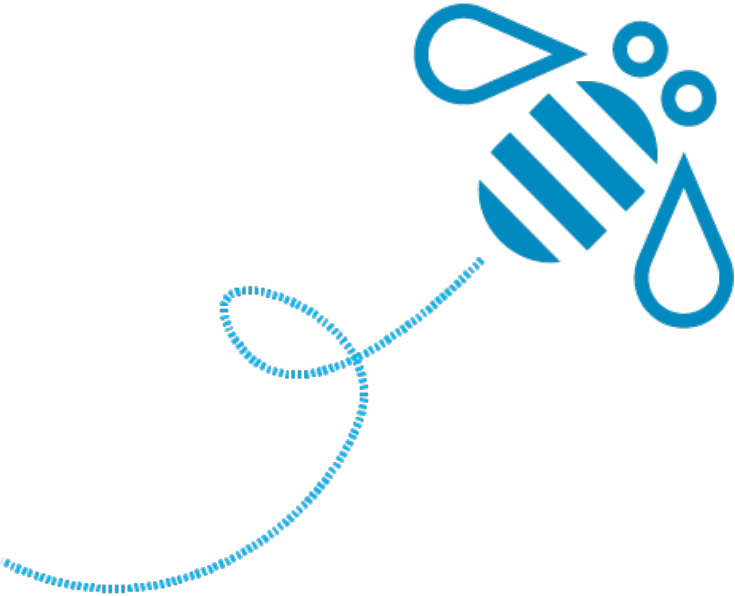
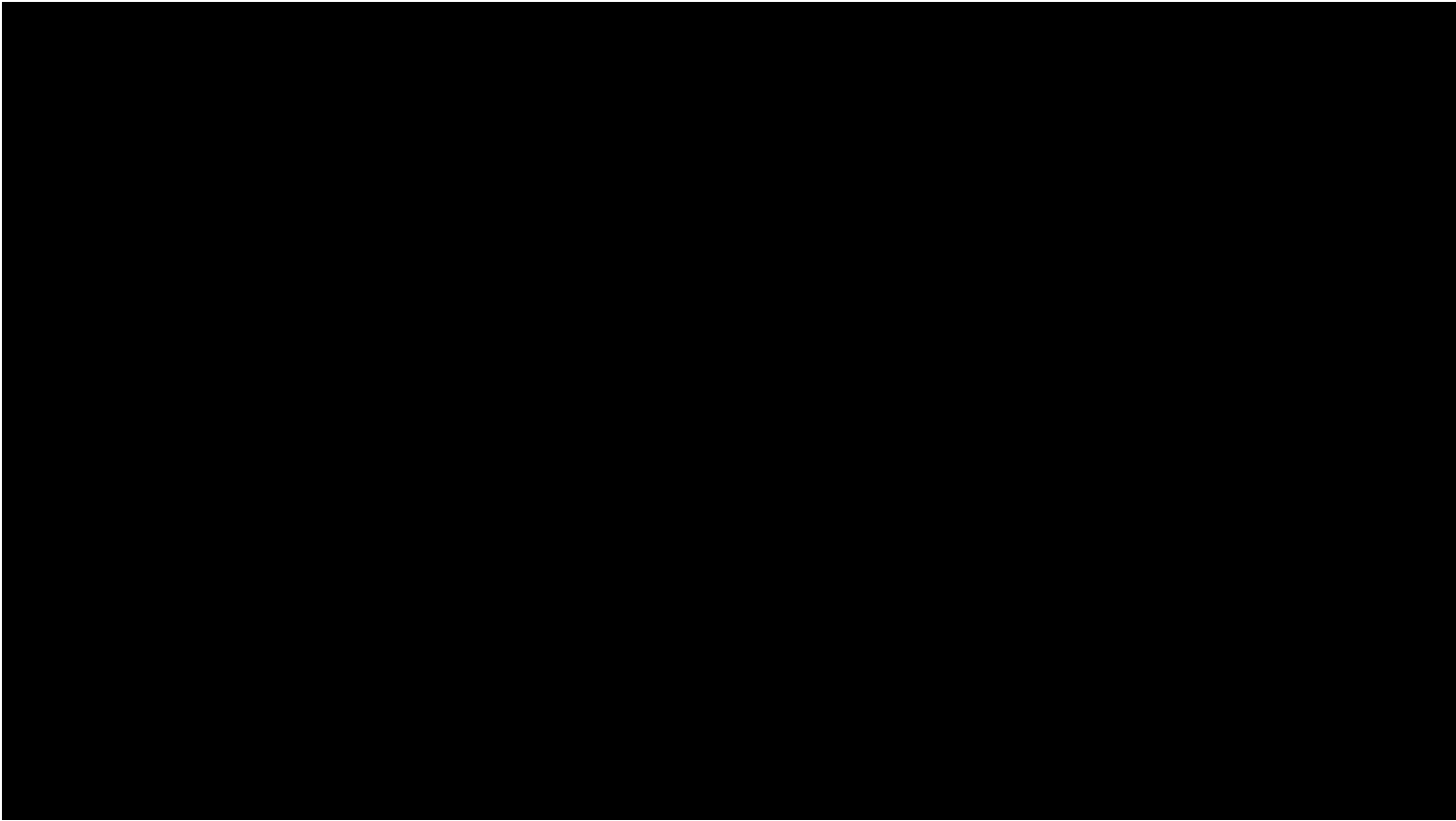


## – Policy Creation

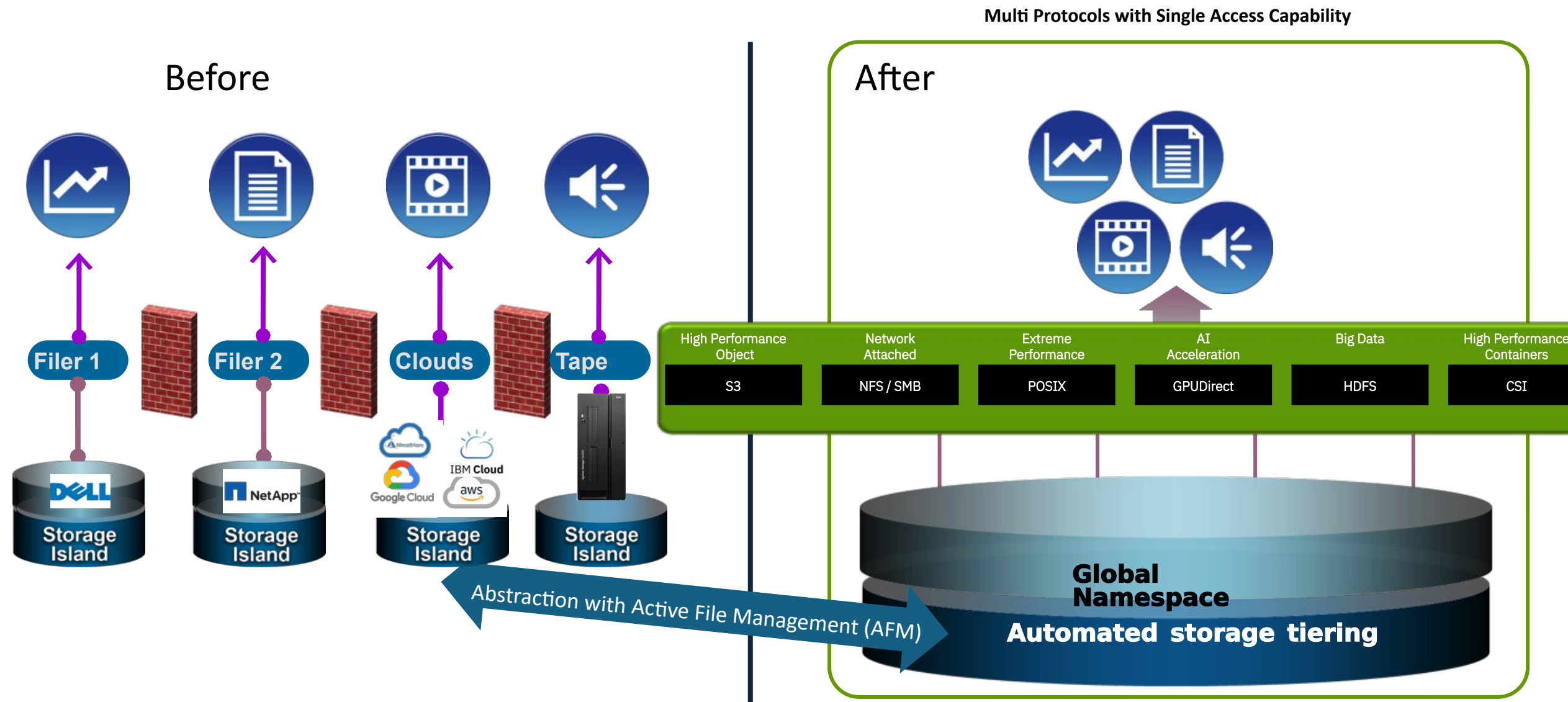
*Reduce the level of expertise required by Scale admins by use of an AI agent to create ILM policy*



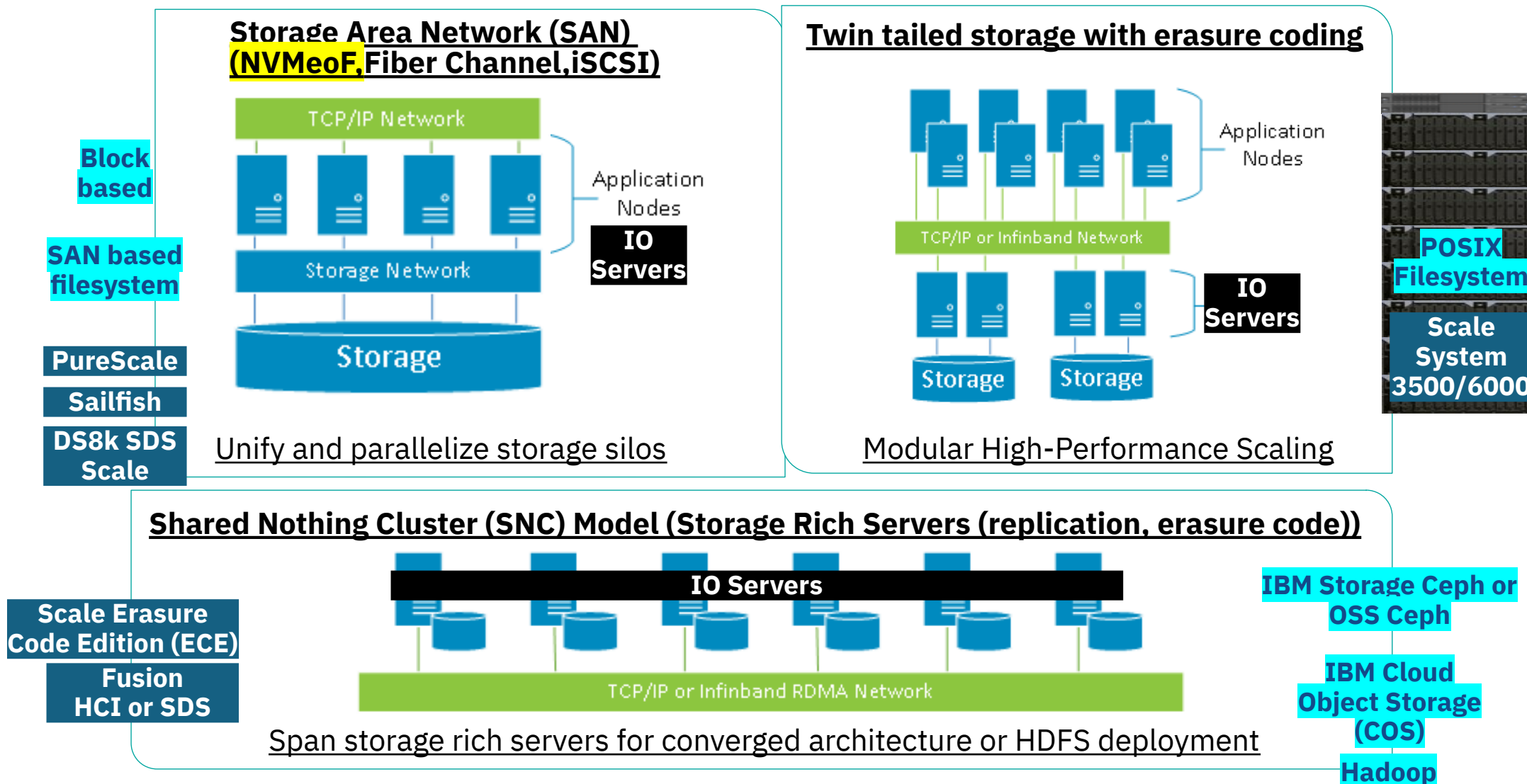
**Demo: Usage Case – Storage Scale GUI - ILM policy**



# Abstract your Storage with Unified Access

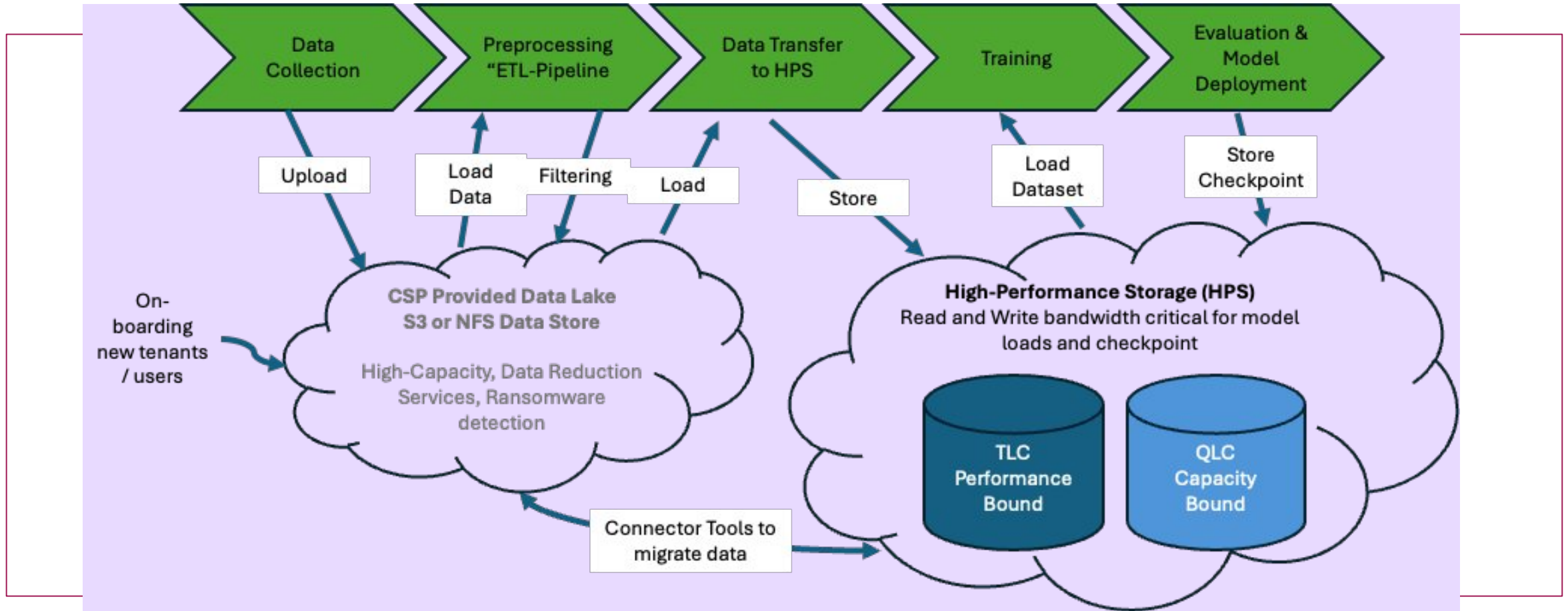


# Scale Deployment model comparison



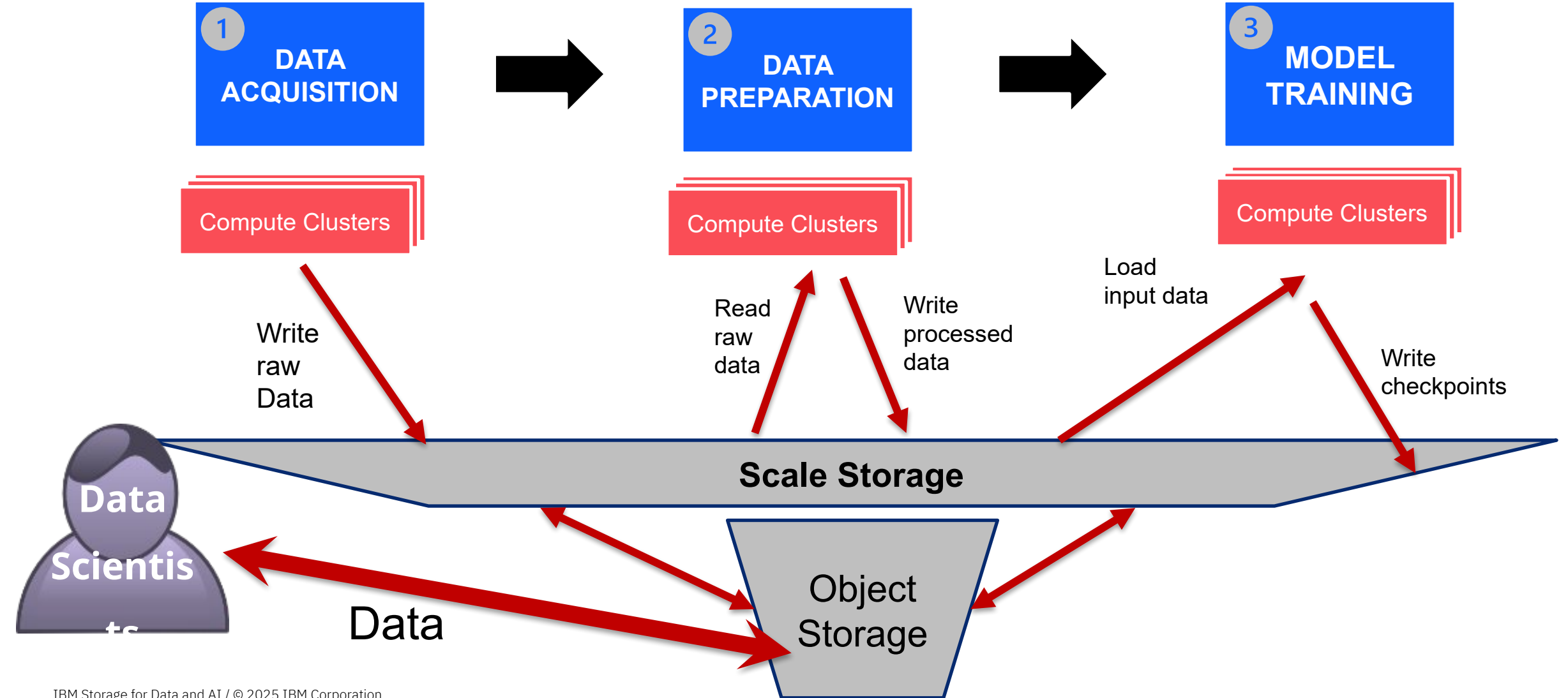
# High-Performance Storage

Offered by Certified Storage Partners



Secure, Multi-tenancy, Highly-Available, and Resilient

# Data Access in HPC and AI workflows @ IBM



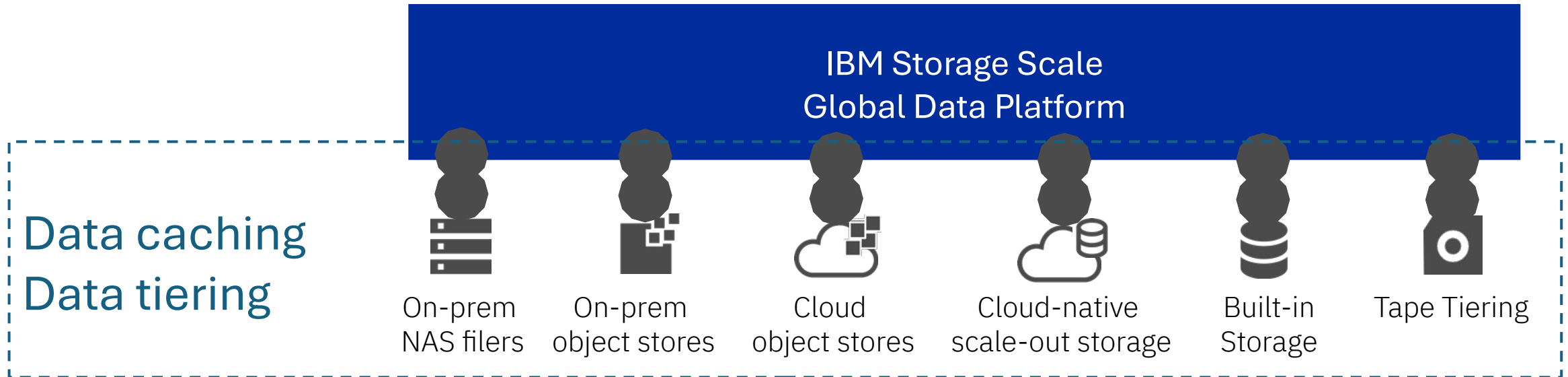
# Abstraction - Data caching and tiering

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

Policy-based Information Lifecycle Management (ILM)

Hierarchical Storage Management (HSM)

Active File Management (AFM)



# Data aware cost optimization

## Powerful policy engine

- Information Lifecycle Management
- Fast metadata 'scanning' and data movement
- Could automate data migration based on threshold

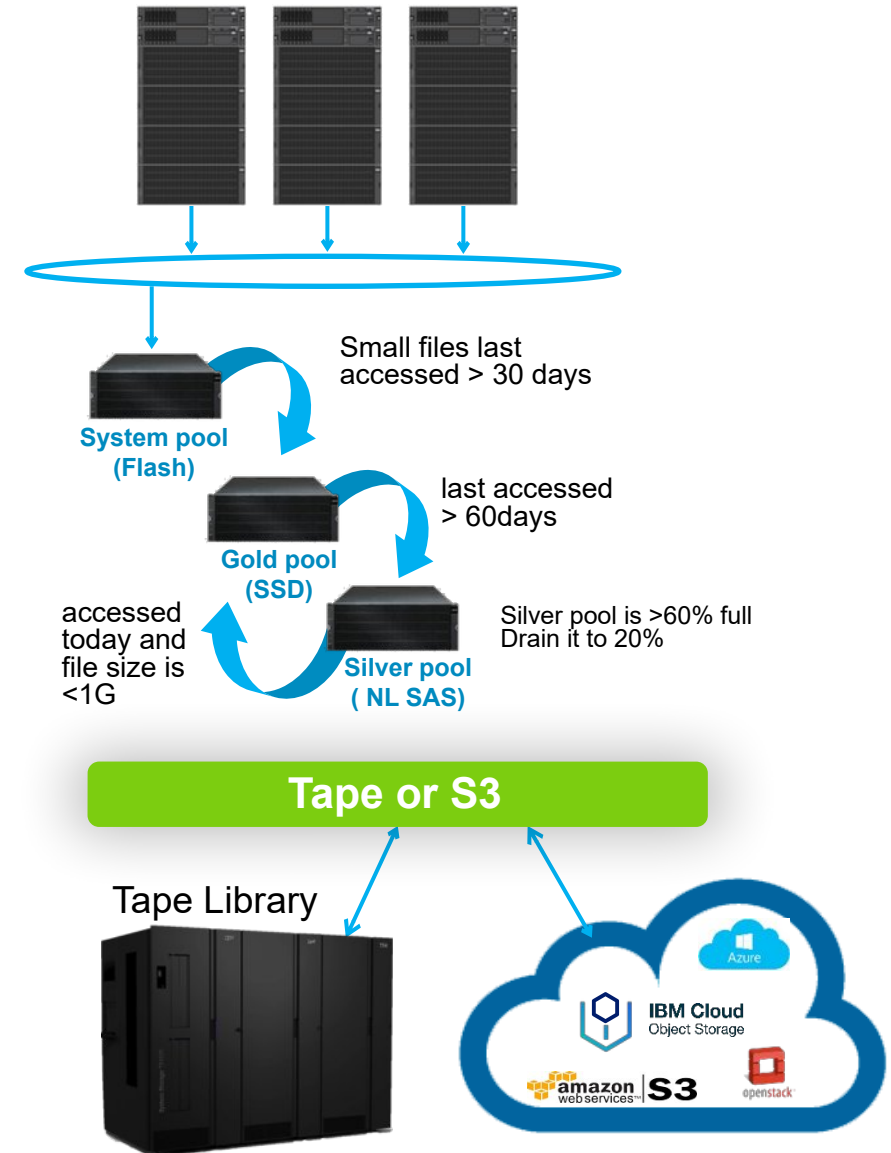
## Users not affected by data migration

- Single namespace

## Integrated with Tape, S3 or NFS endpoints

- DMAPI
- Active File Management (AFM)

Example: Online storage reaches 90% full then move all 1 GB or larger files that are 60 days old to offline to free up space



# Data aware performance optimization

## Storage pools

A collection of disks or arrays with similar properties that are managed together as a group.

## File placement policies

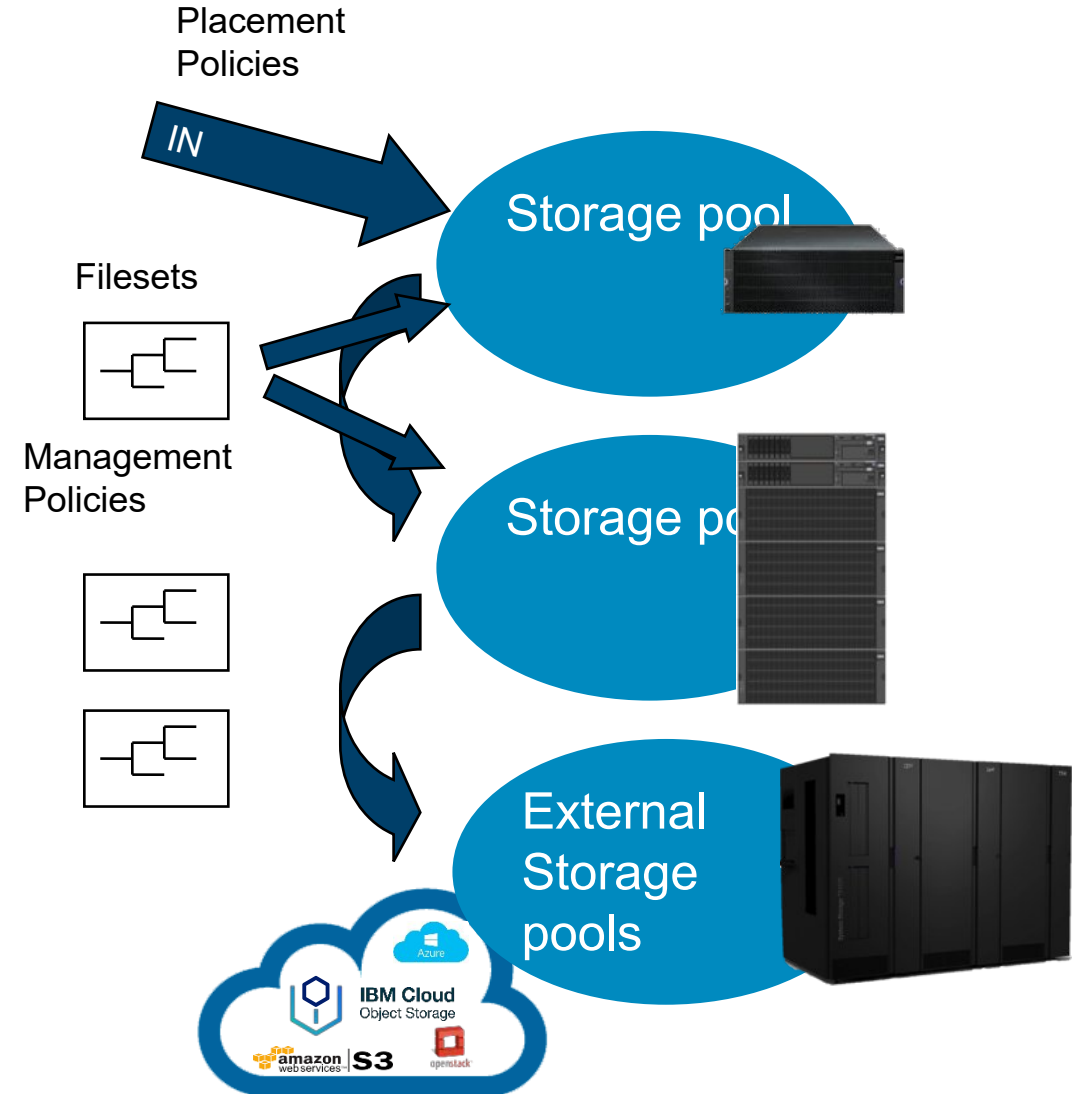
Determines where the file data is placed on creation.

## File management policies

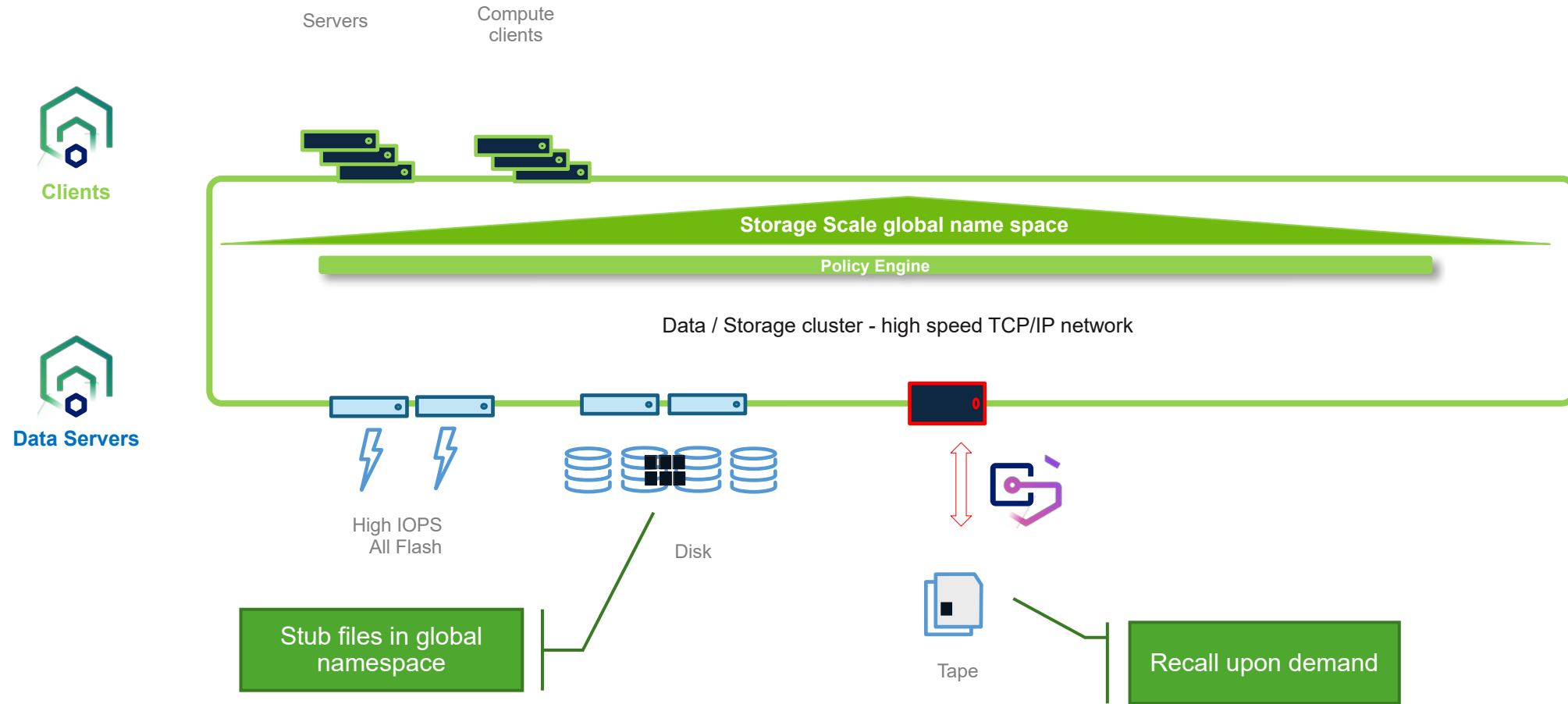
Migrates or deletes file based on business rules.  
Can look at file heat

## Filesets

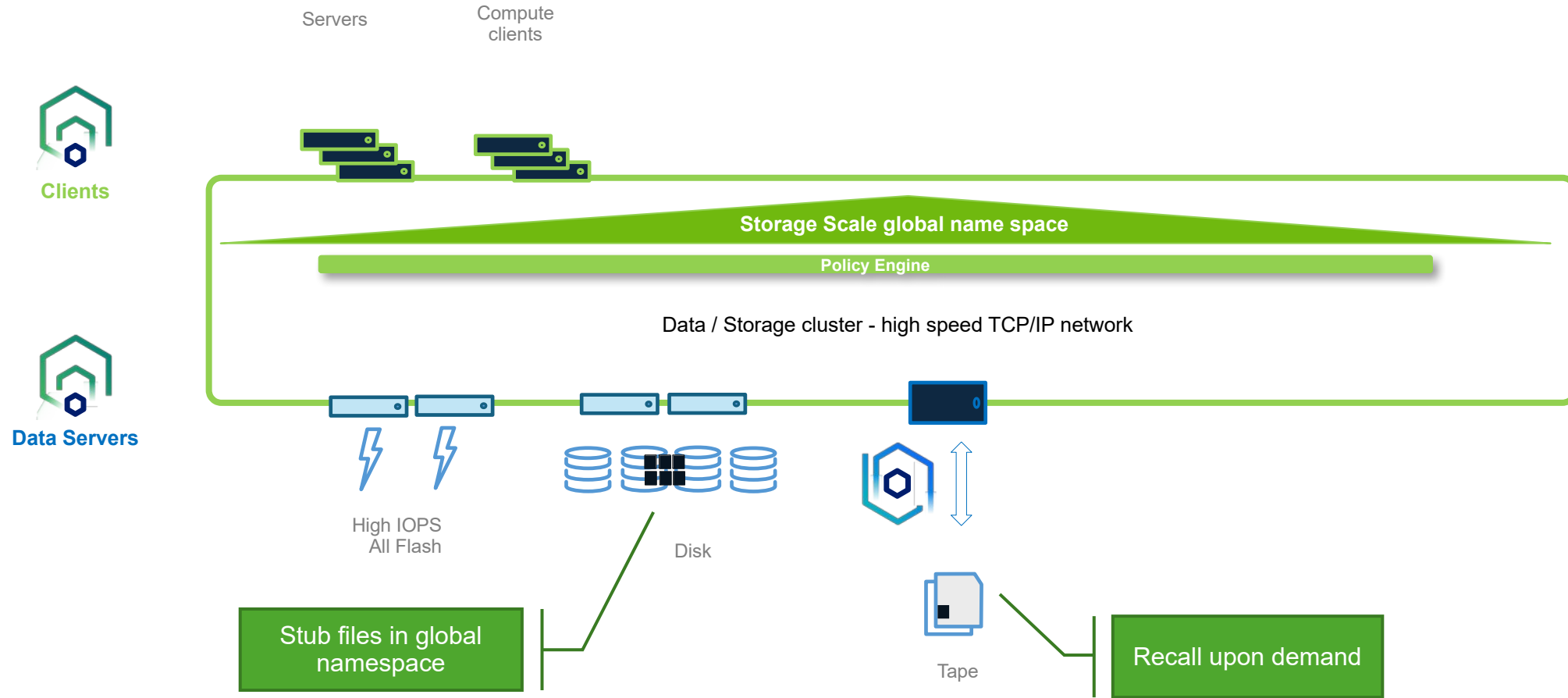
Logical subtrees within a file system that act as metadata containers for files.



## Archive with Storage Archive

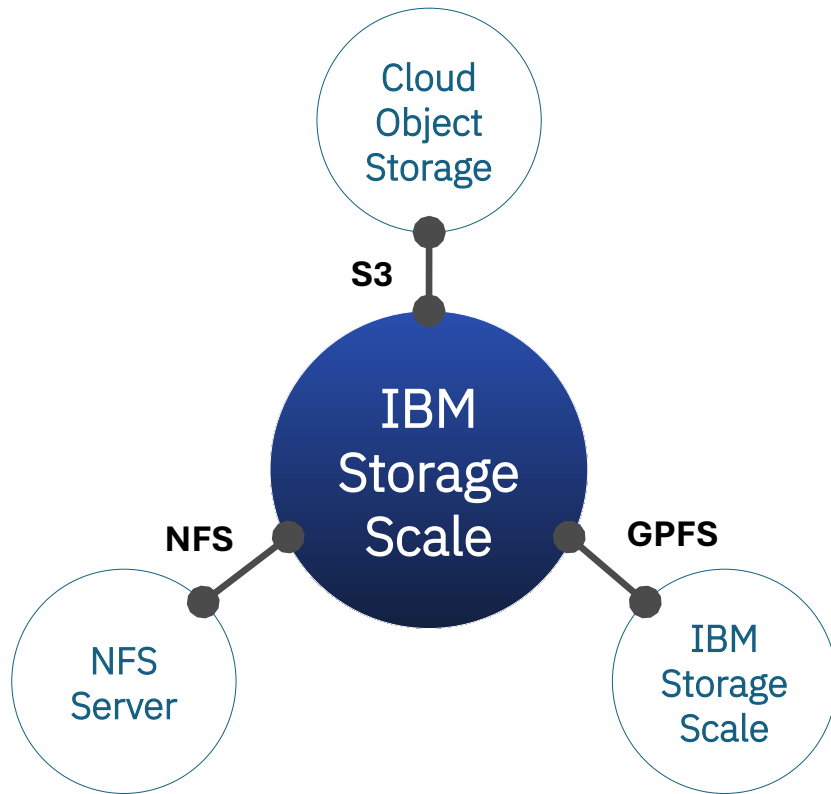


# Utilize Storage Scale to contain Storage Protect Storage Pools

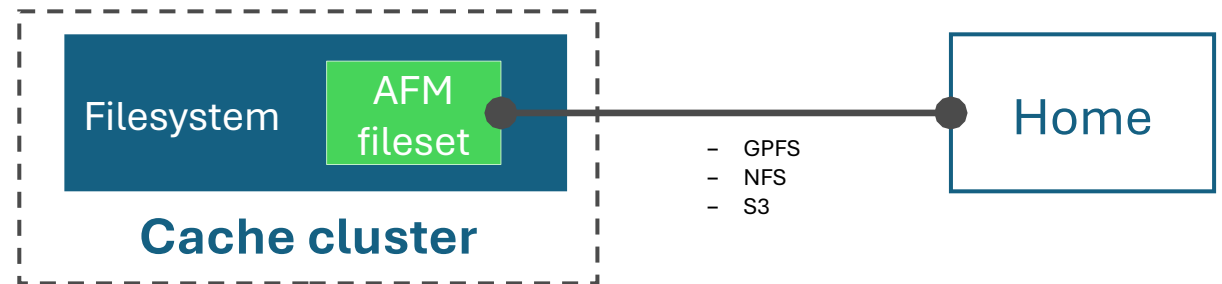


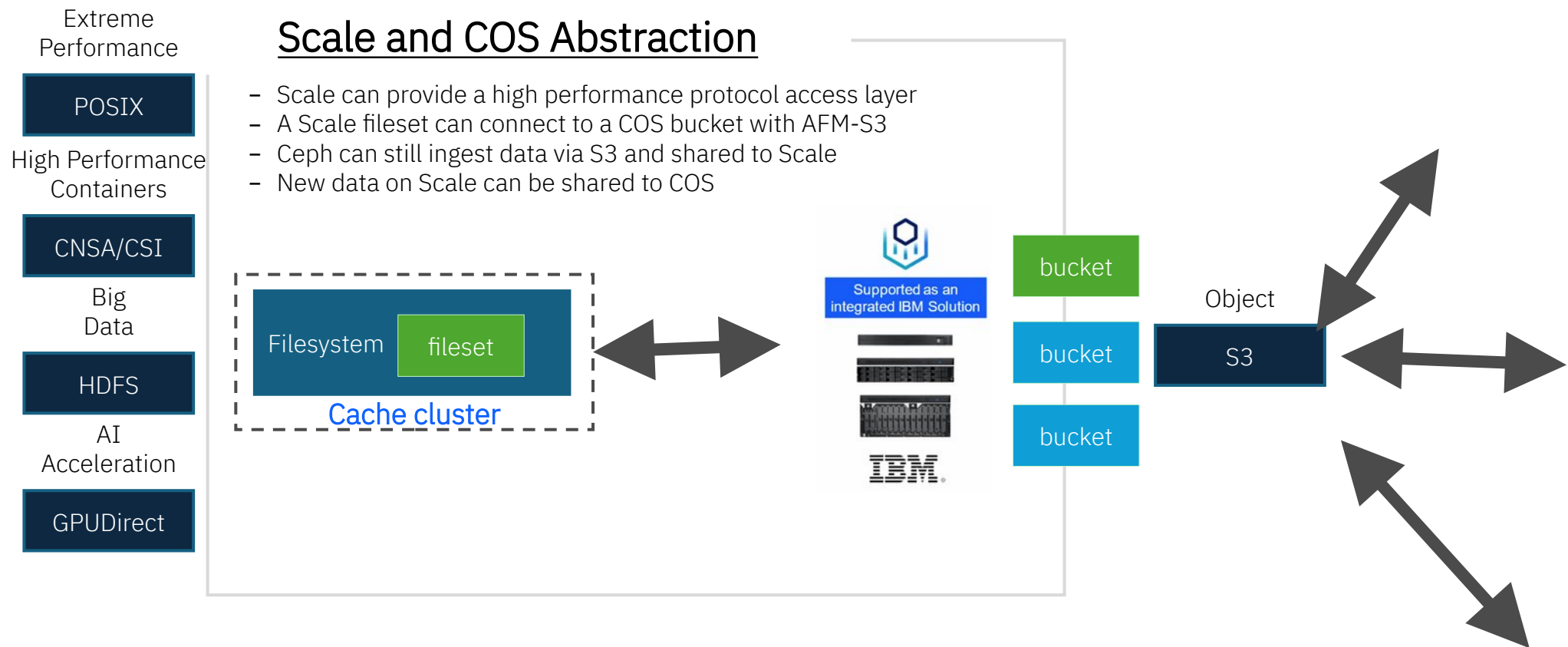
# Active File Management (AFM) overview

Active File Management enables caching data across other data sources.

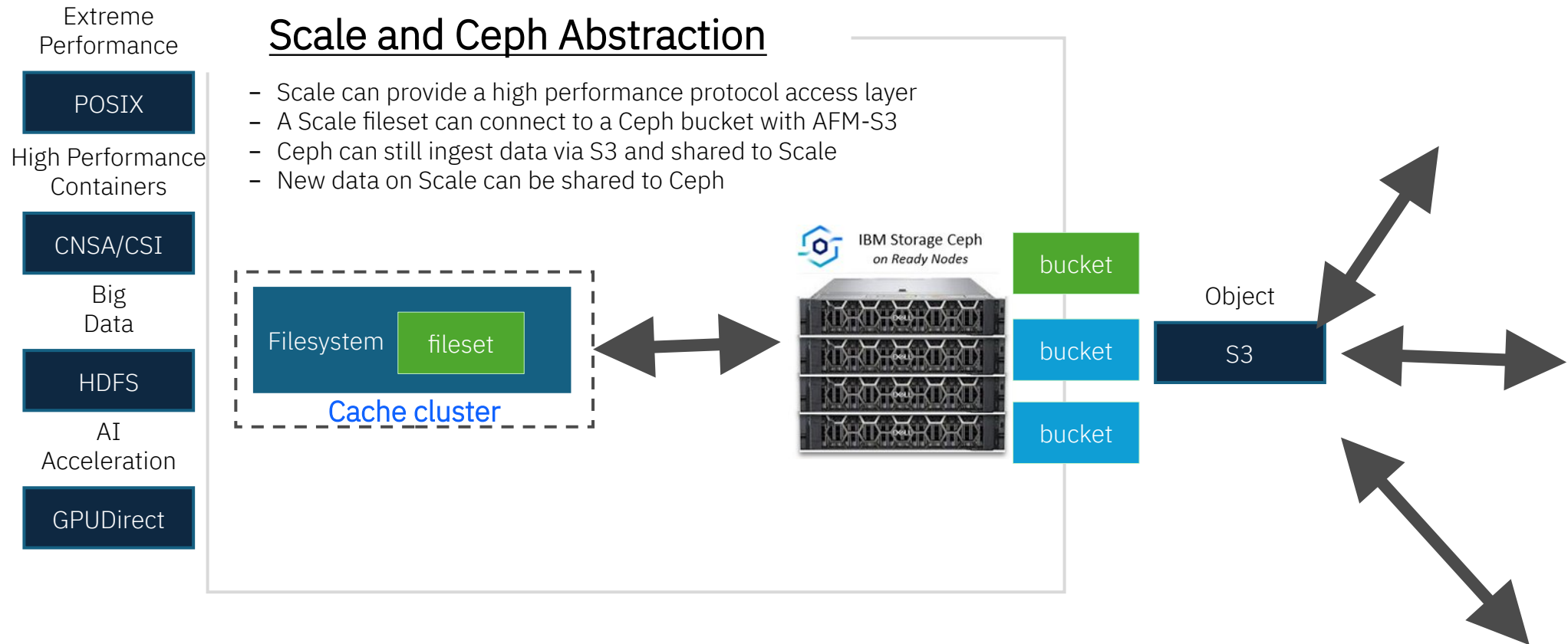


- Each AFM fileset has a distinct set of AFM attributes.
- An IBM Storage Scale cluster that contains AFM filesets is called a cache cluster.
- A cache cluster has a relationship with another remote site called the home, where either the cache or the home can be the data source or destination.
- A cache cluster must be an IBM Storage Scale Cluster.
- A home can be an IBM Storage Scale, NFS server and Object Storage.





Easily push and pull data to and from an Object repository  
Can have multiple filesets to multiple buckets



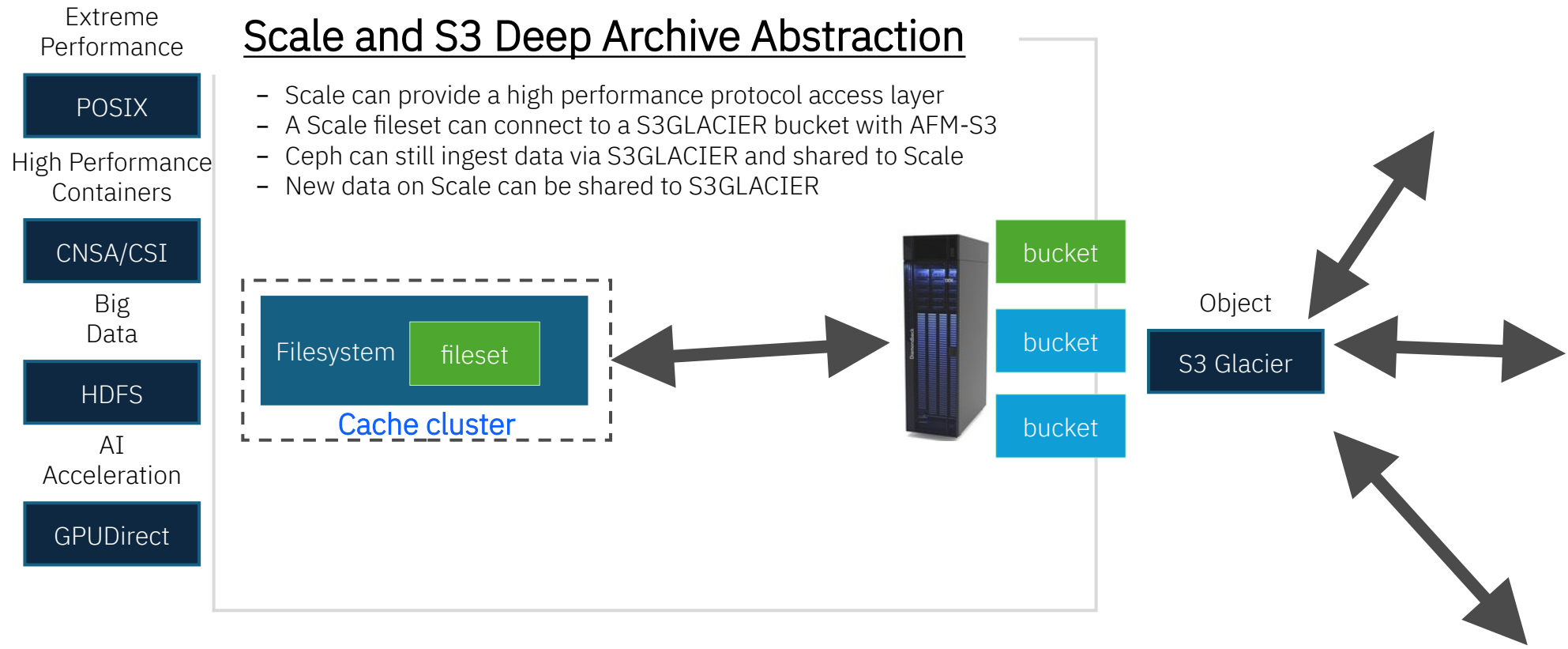
Easily push and pull data to and from an Object repository  
Can have multiple filesets to multiple buckets

# IBM Deep Archive

A low-cost, on-premises cloud providing secure and durable storage for data archiving and online backup.

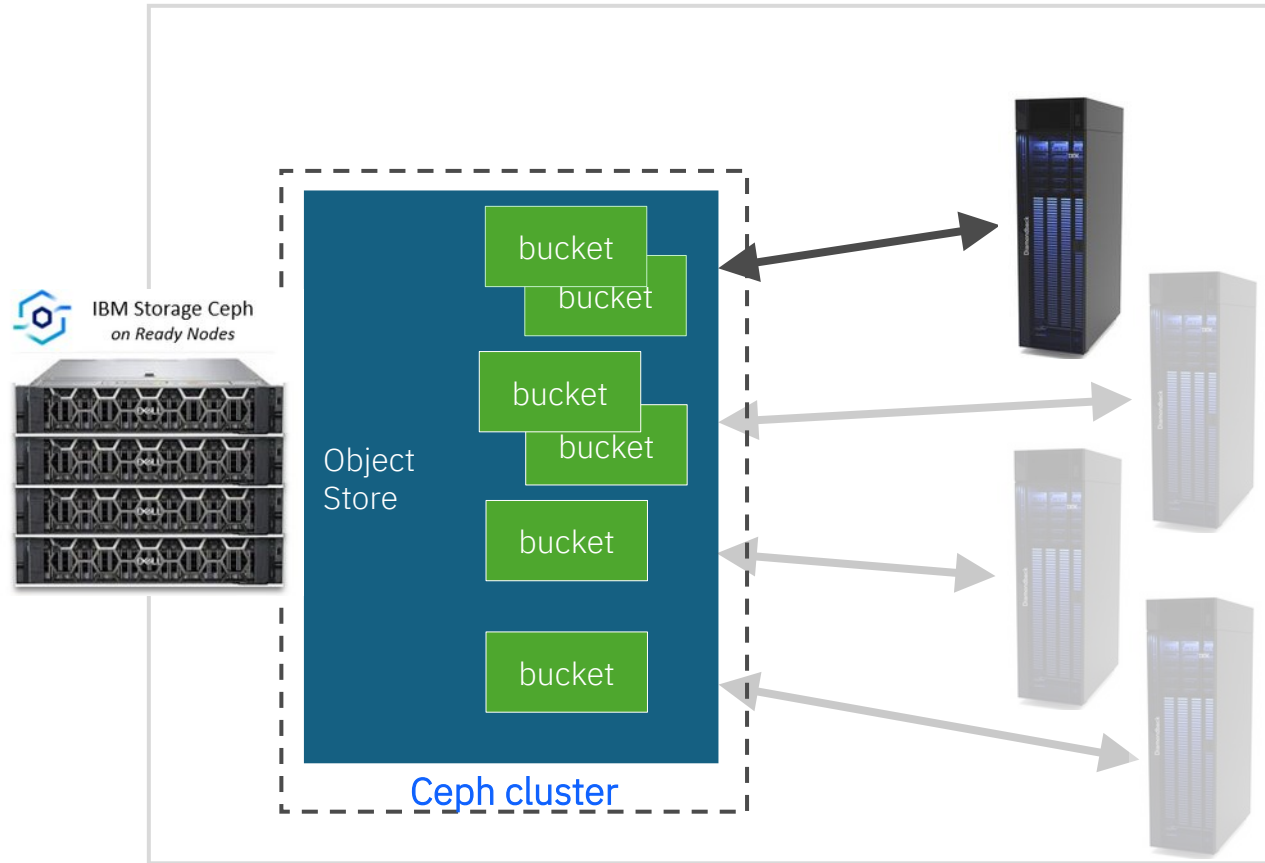
- S3 Glacier compliant interface
- Install-and-go deployment
- Simple administration
- Up to 61PB object storage
- Up to 13.2TB per hour data rate





Easily push and pull data to and from a GLACIER Object repository  
Can have multiple filesets to multiple buckets

# IBM Storage Ceph S3 Support for Storage Deep Archive

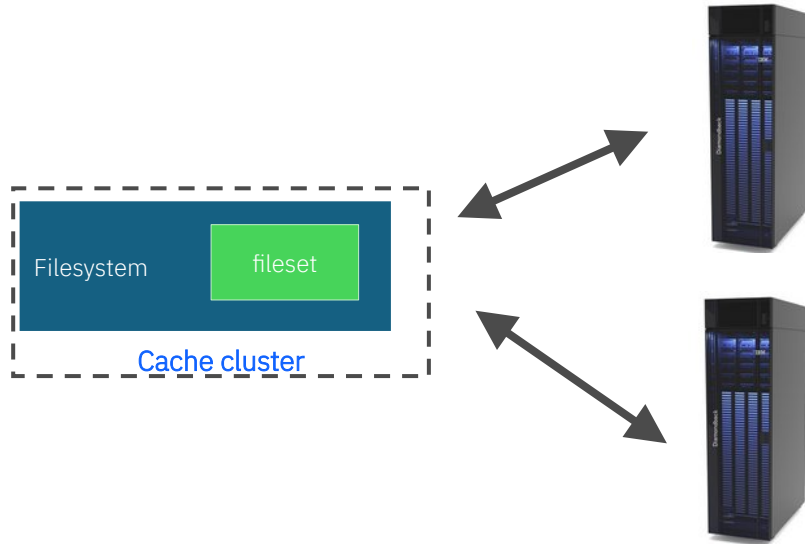


Scale the number of buckets and/or Storage Deep Archives

# IBM Storage Ceph S3 Support for Storage Deep Archive Replication

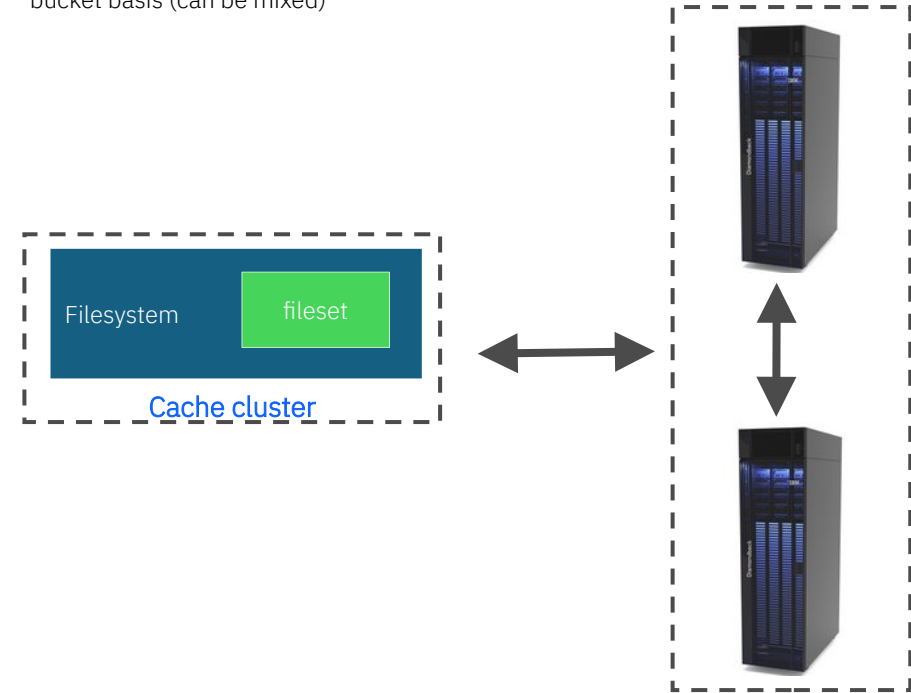
## Multi-target replication

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



## Single Target Deep Archive Replication

- Cache can push to single target which is replicated internally on a bucket-by-bucket basis (can be mixed)



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

# IBM Storage Ceph RGW

Object Retrieval AWS Glacier S3 tiers

Object Storage  
Class Cloud tier



Amazon AWS  
Glacier S3 tier



Object Storage  
Tape S3 tier



Object  
retrieval



IBM Storage Ceph  
RADOS Gateway  
RGW



IBM Storage Ceph  
Cluster

## Functionality

Enables clients for seamless retrieval of objects, that earlier have been transitioned to external locations

In example AWS Glacier, amongst other capabilities, like Tape S3 storage tier-IBM Deep Archive.

Multiple enhancements have been established in addition to AWS Glacier support.

## Business value

Having the option to retrieve data from public cloud or tape resources allows organizations to manage data more effectively.

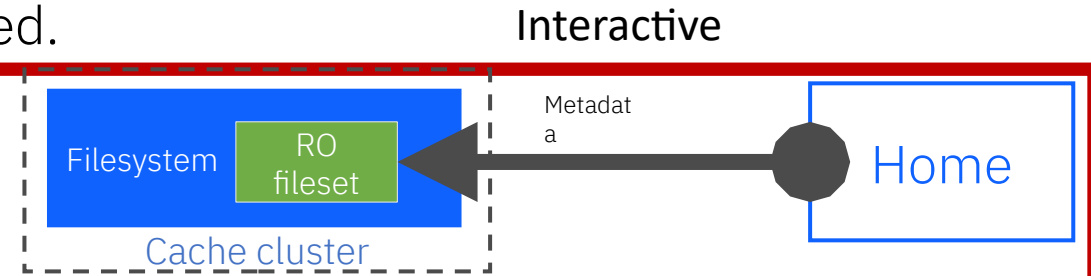
Transitioned Object data remains to be available.

The functionality allows businesses to optimize their data management strategies without requiring additional management utilities.

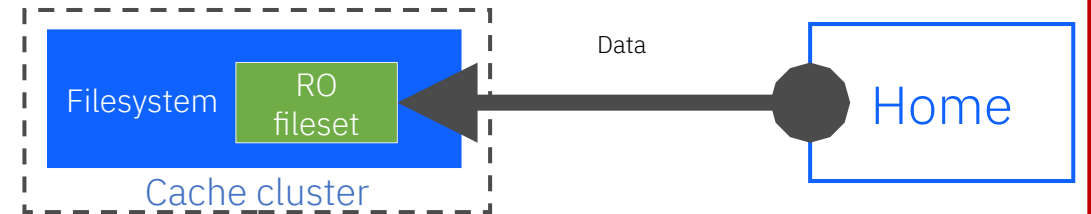
# AFM Propagation

The metadata and data of the file is propagated when needed.

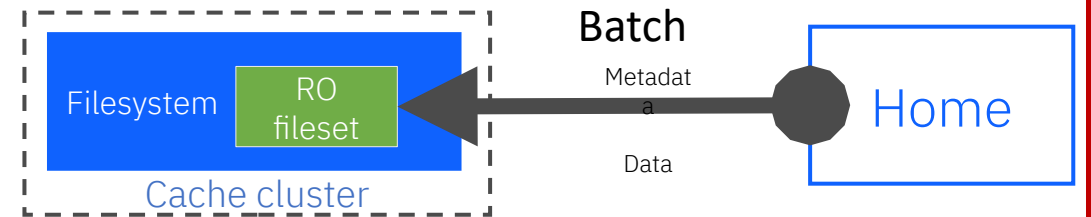
- Run the command, "ls" on the cache.  
Metadata is propagated from the home to the cache.



- Run the command, "cat" or "head" on the cache,  
Data is propagated from the home to the cache.



- Run the Scale command, "prefetch" on the cache,  
All metadata and data is propagated to the cache.



- Run the Scale command, "evict" on the cache  
Cached data is cleared in the cache.



# AFM Caching Mode

AFM has four caching modes.

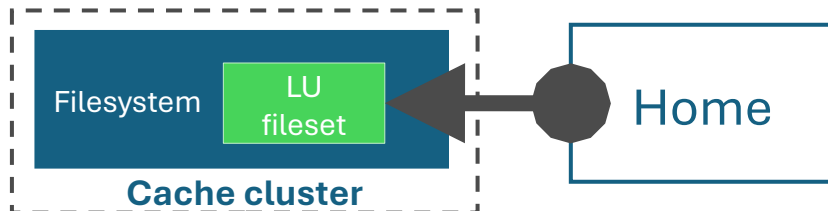
## Read only (RO) mode

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



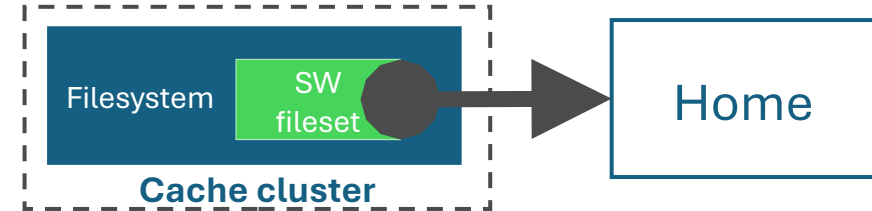
## Local update (LU) mode - “Hotel California”

- Data in the cache can be read and written.
- Data which is created or modified in the cache is never updated by home.
- Data source is the home and data destination is the cache.



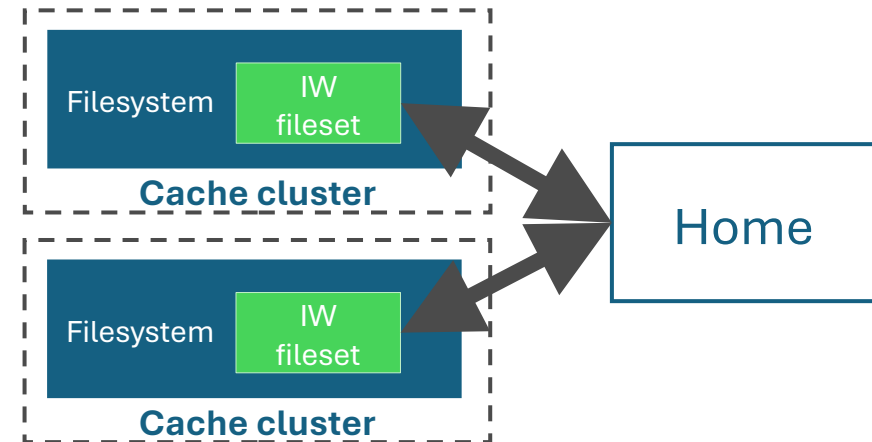
## Single writer (SW) mode

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



## Independent writer (IW) mode

- Each cache reads from home and updates to the home independently of each other.
- Updates are propagated to the home in an asynchronous and can be delayed due to network.



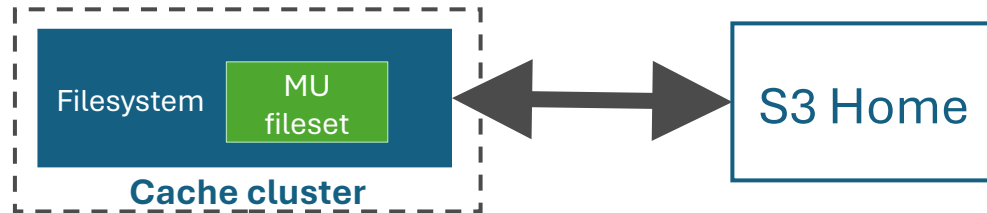
# AFM S3 Caching Mode

Support of CAS (Content Aware Storage)

Async Notifications to detect changes at object store instead of polling.

## Manual Update (MU) mode

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

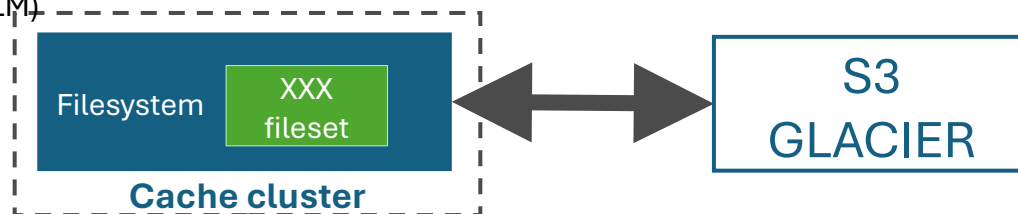


**mmafmcconfig --storage-class GLACIER**

**mmafmcconfig --lazy-migrate**

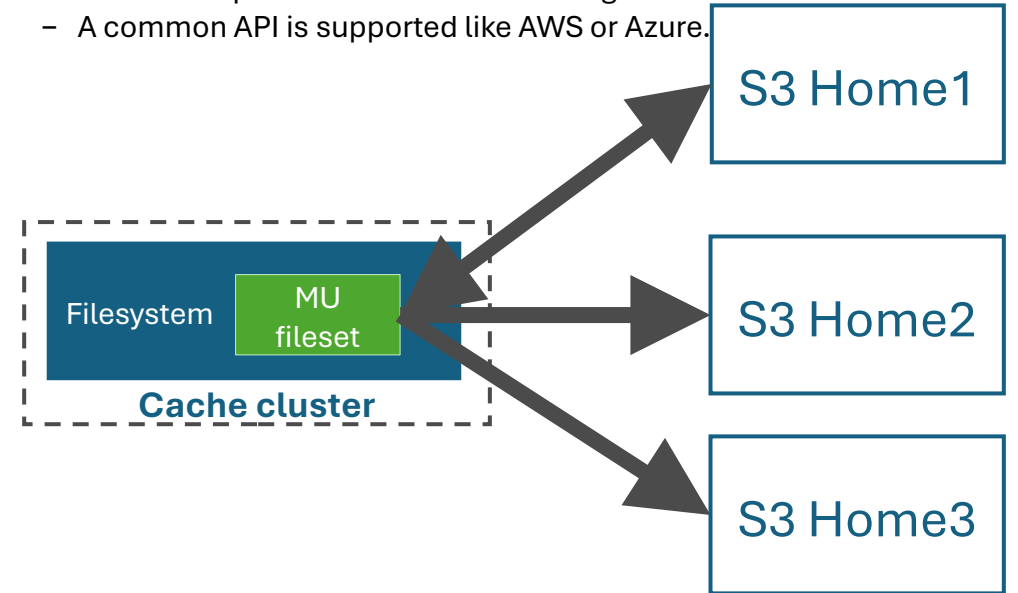
## Support of S3 Glacier

- Data in the cache or in the S3 GLACIER object store.
- Data can be moved bi-directional via Integrated Life Cycle Management (ILM).



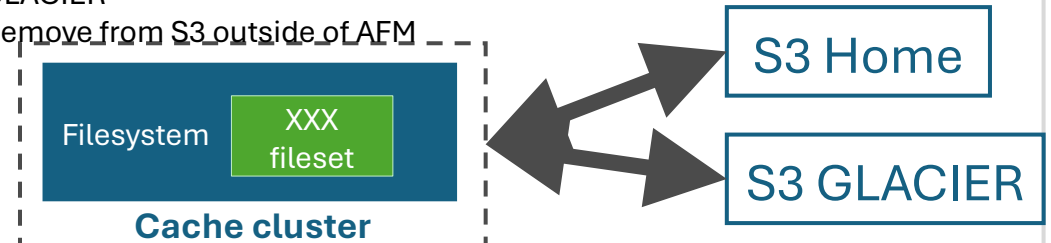
## Multi-target replication

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

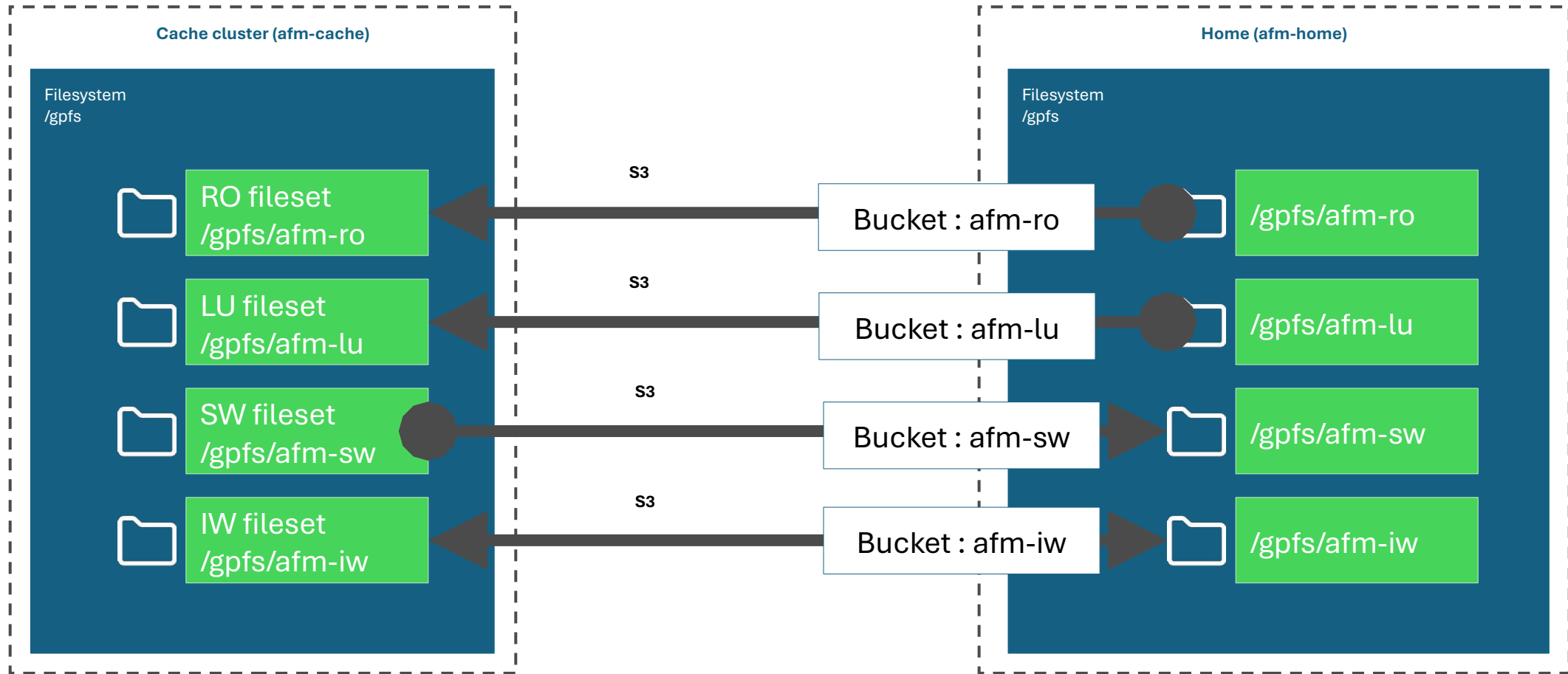


## Free up S3 but keep S3 GLACIER data

- Place data on S3, then S3 Glacier and remove from S3 but keep S3 GLACIER
- Remove from S3 outside of AFM



# AFM Demo environment



# AFM Demo

