Active File Management (AFM)

Spectrum Scale Strategy Days 2017

Achim Christ · achim.christ@de.ibm.com
Karl Schulz · karl.schulz@csi-online.de

Table of contents

AFM Overview and Concepts – “Stretched” Cluster, Multi-cluster, AFM

Use Case 1: **Branch office**

Use Case 2: **Data ingest**

Use Case 3: **Disaster protection**

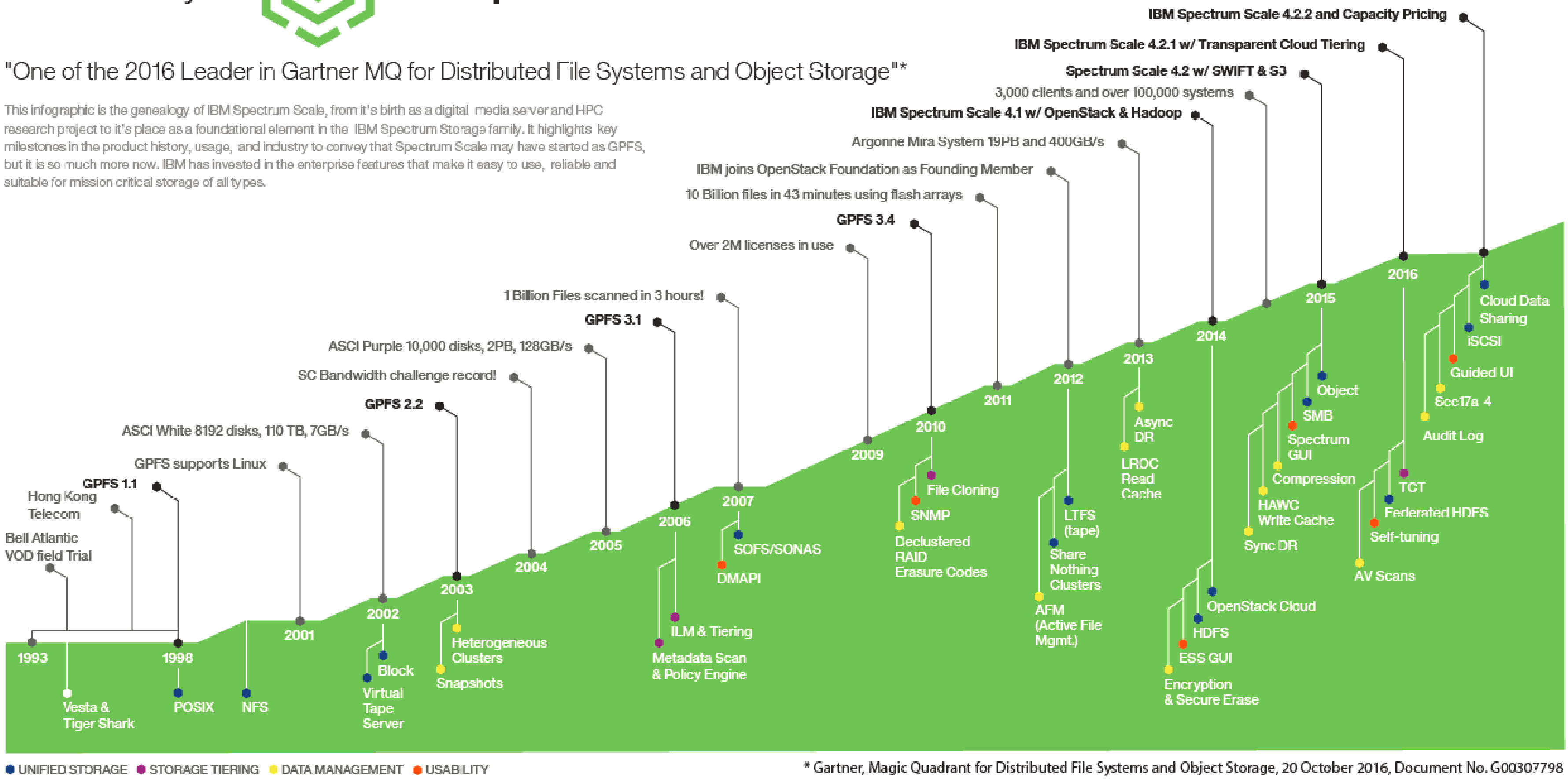
Use Case 4: **Migration**

AFM Overview and Concepts

The History of IBM **Spectrum** Scale

"One of the 2016 Leader in Gartner MQ for Distributed File Systems and Object Storage"*

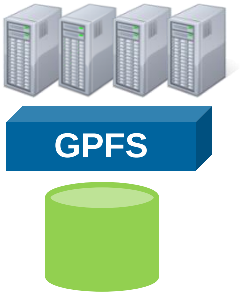
This infographic is the genealogy of IBM Spectrum Scale, from it's birth as a digital media server and HPC research project to it's place as a foundational element in the IBM Spectrum Storage family. It highlights key milestones in the product history, usage, and industry to convey that Spectrum Scale may have started as GPFS, but it is so much more now. IBM has invested in the enterprise features that make it easy to use, reliable and suitable for mission critical storage of all types.



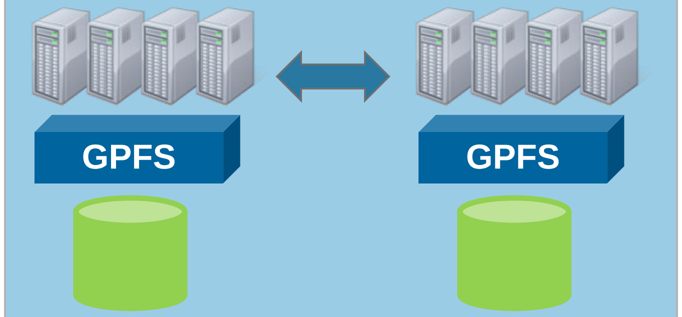
IBM Spectrum Scale 4.2.2 and Capacity Pricing
IBM Spectrum Scale 4.2.1 w/ Transparent Cloud Tiering
Spectrum Scale 4.2 w/ SWIFT & S3
3,000 clients and over 100,000 systems
IBM Spectrum Scale 4.1 w/ OpenStack & Hadoop
Argonne Mira System 19PB and 400GB/s
IBM joins OpenStack Foundation as Founding Member
10 Billion files in 43 minutes using flash arrays
GPFS 3.4
Over 2M licenses in use
1 Billion Files scanned in 3 hours!
GPFS 3.1
ASCI Purple 10,000 disks, 2PB, 128GB/s
SC Bandwidth challenge record!
GPFS 2.2
ASCI White 8192 disks, 110 TB, 7GB/s
GPFS supports Linux
GPFS 1.1
Hong Kong Telecom
Bell Atlantic VOD field Trial
1993
Vesta & Tiger Shark
1998
POSIX
NFS
2001
2002
Block
Virtual Tape Server
2003
Heterogeneous Clusters
Snapshots
2004
2005
2006
ILM & Tiering
Metadata Scan & Policy Engine
2007
SOFS/SONAS
DMAPI
2009
2010
File Cloning
SNMP
Declustered RAID Erasure Codes
2011
2012
LTFS (tape)
Share Nothing Clusters
AFM (Active File Mgmt.)
2013
Async DR
LROC Read Cache
2014
OpenStack Cloud
HDFS
ESS GUI
Encryption & Secure Erase
2015
Object
SMB
Spectrum GUI
Compression
HAWC Write Cache
Sync DR
2016
TCT
Federated HDFS
Self-tuning
AV Scans
Cloud Data Sharing
iSCSI
Guided UI
Sec17a-4
Audit Log
● UNIFIED STORAGE ● STORAGE TIERING ● DATA MANAGEMENT ● USABILITY

* Gartner, Magic Quadrant for Distributed File Systems and Object Storage, 20 October 2016, Document No. G00307798

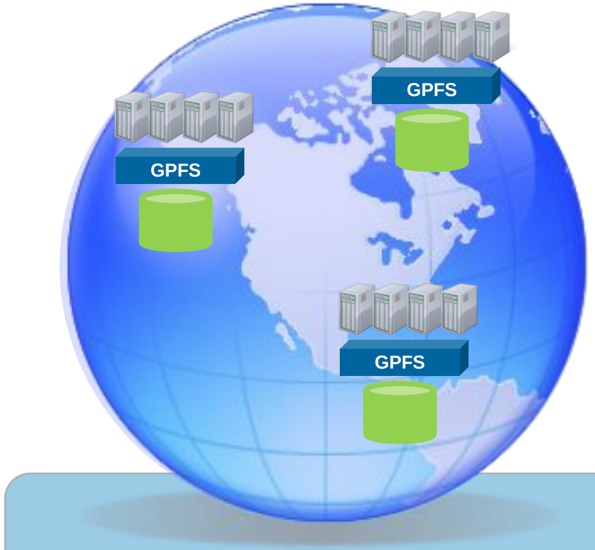
Spectrum Scale evolution



GPFS introduced concurrent file system access from multiple nodes



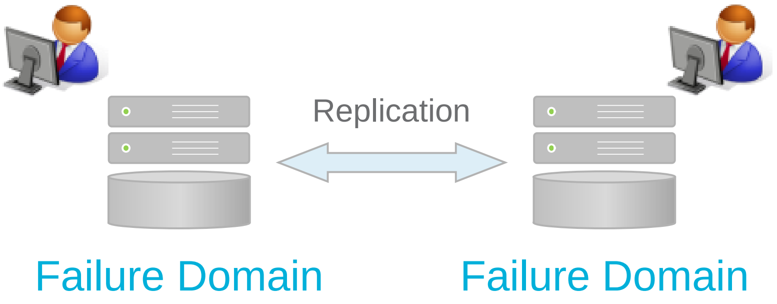
Multi-cluster expands the global namespace by connecting multiple sites



Active File Management takes global namespace truly global by automatically managing asynchronous replication of data

1993 2005 2012

Spectrum Scale “Stretched” Cluster

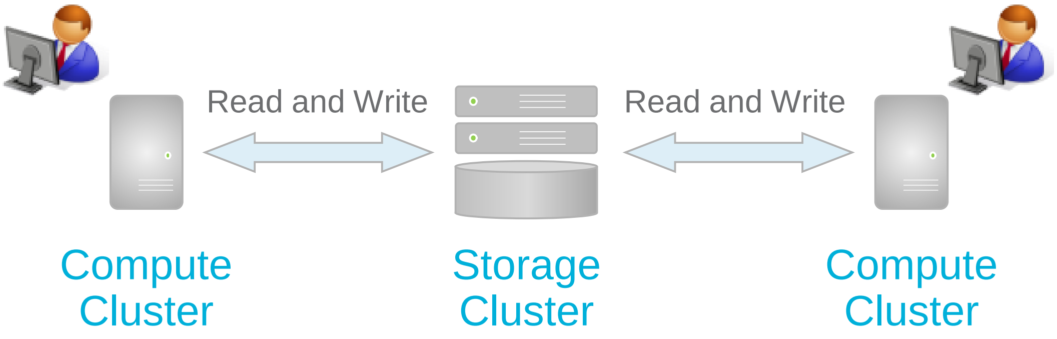


- Single Spectrum Scale cluster spans sites (failure domains)
  - Single administrative domain
  - Synchronous operation, consistent locking

- Synchronous replication between sites based on NSD Failure Groups
  - Distinct Failure Groups indicate resources which could fail simultaneously
  - (Optional) replication based on different Failure Groups
  - Default replication factor for filesystem, overridden via policy

- Supports idea of high-availability
  - Failure of individual Failure Group compensated by GPFS
  - Details and recovery steps: Whitepaper | Wiki

Spectrum Scale Multi-cluster



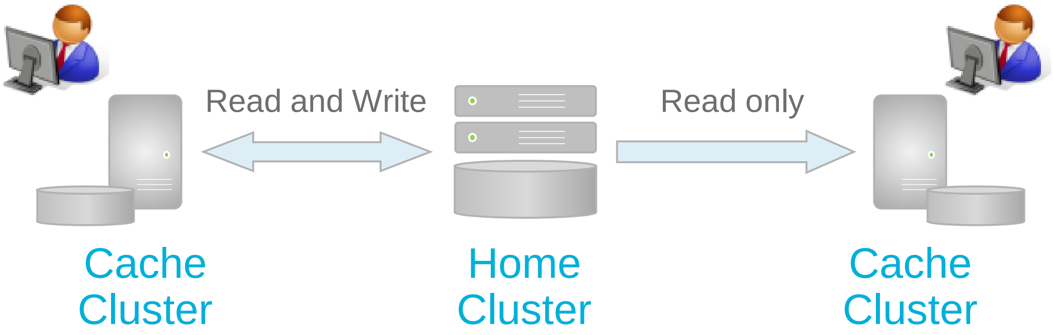
- Independent Spectrum Scale clusters
  - Separate administrative domains
  - Synchronous operation, consistent locking

- Single storage cluster owns NSDs, one or many remote clusters mount file system(s)
  - Cross-cluster mount: [Knowledge Center]
  - Facilitates parallelism to optimize performance

- Unavailability of storage cluster affects all remote clusters (single data copy)

- Supports idea of multi-tenancy (clusters can be authorized for individual file systems)

Spectrum Scale Active File Management (AFM)



- Geographically dispersed systems share their files over WAN
  - Low bandwidth, expensive lines, temporarily unavailable, etc.

- Requirements
  - All users & applications should “see” their files stored somewhere else
  - Files should only be transferred when accessed and then cached locally

- Solution
  - Represent namespace (home) in remote system (cache) without transferring files
  - Transfer files when required and cache them, users work on local copy
  - Use parallelism to minimize transfer times
  - Asynchronous operation, no consistent locking

Active File Management (AFM) architecture

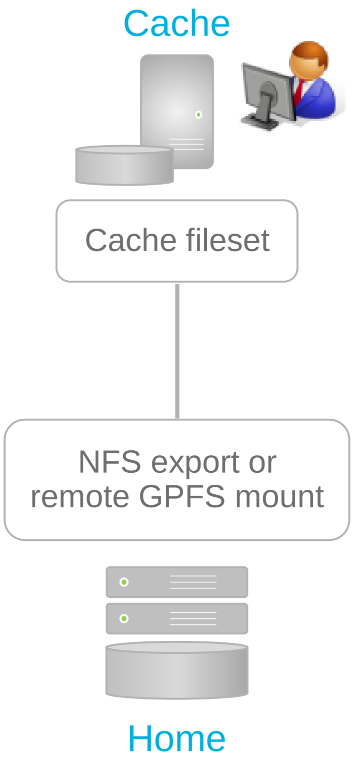
- AFM uses home-cache model
  - Single home provides primary storage of data which is exported
  - Exported data is cached in local GPFS file system

- Home can be NFS export or remotely mounted GPFS cluster
  - Only GPFS-based home file systems support ACLs, EA, and sparse files (irrespectively of NFS or GPFS protocol)

- GPFS cache presents home export in a fileset
  - One cache fileset can cache one home export
  - One cache server can cache multiple home exports (one fileset each)

- Different modes supported, can be combined per fileset

- AFM supported on AIX and Linux, gateway nodes must be Linux

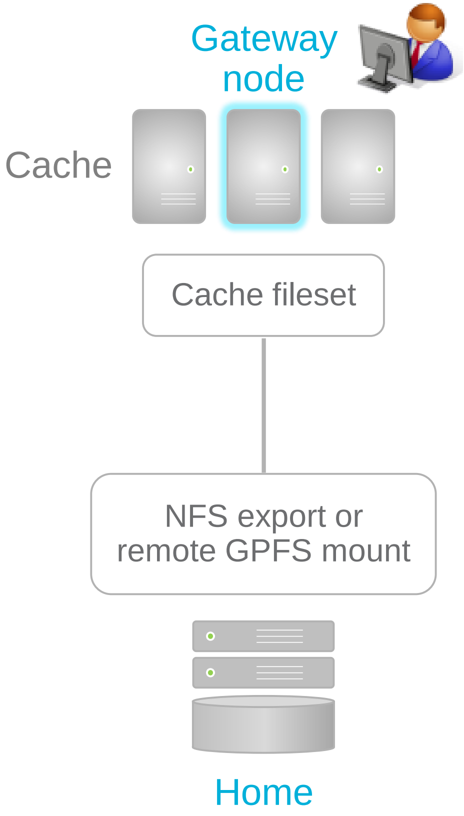


AFM gateway nodes

- Gateway node on cache manages communication with home
  - At least one GPFS cache node must be assigned as gateway node
  - Gateway node must be network connected to home server
  - Multiple gateway nodes can be used for redundancy and parallel I/O
  - Each cache fileset has gateway node (metadata server) in cache cluster

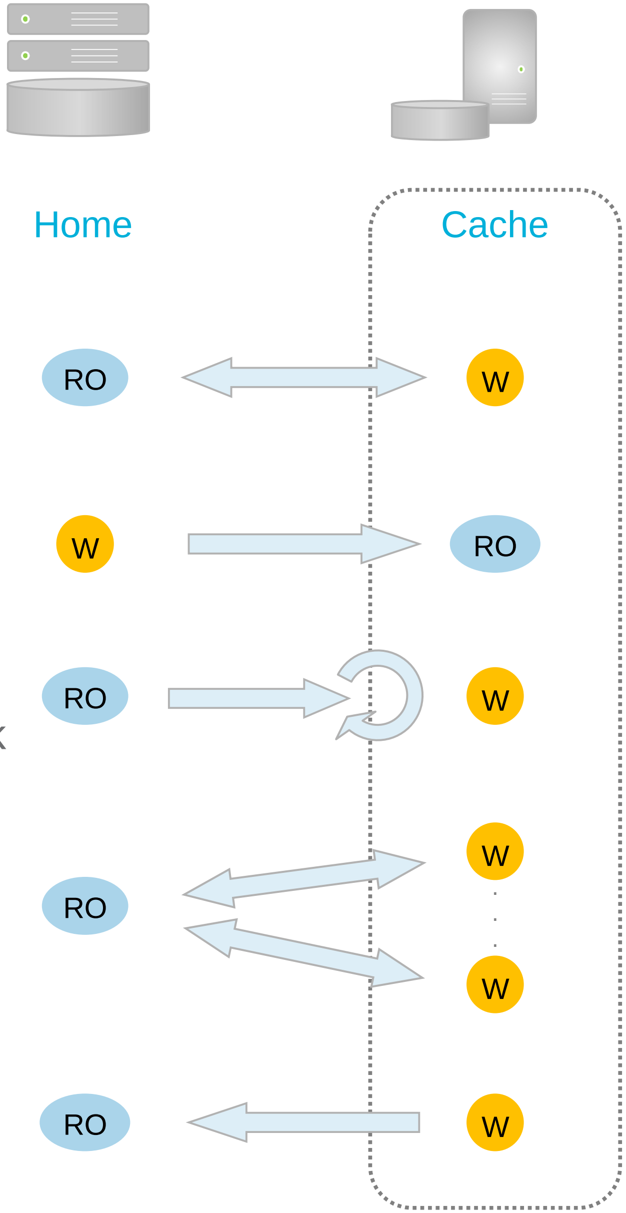
- Gateway nodes are setup during AFM configuration
  - Node role, recommendation to use dedicated resources
  - Command `mmchnode --gateway`

- Gateway nodes must be Linux, require server license



AFM cache modes

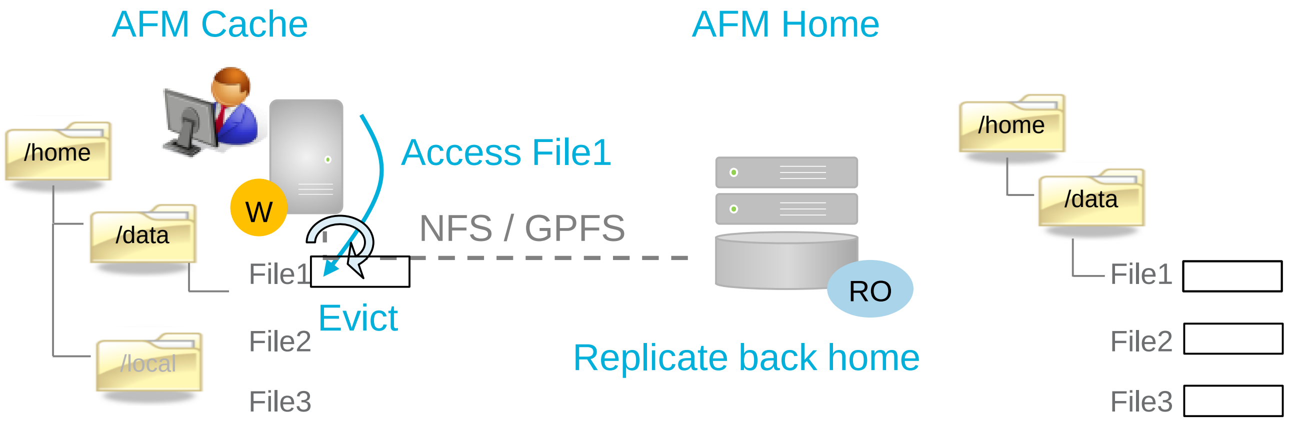
- Single Writer
  - Only cache can write data. Home can’t change. Other peer caches have to be setup in RO mode
- Read Only
  - Cache can only read data, no data change allowed
- Local Updates
  - Data is cached from home like in SW mode
  - Once data is changed on cache it is not replicated back to home (stays local)
- Independent Writer
  - Multiple caches pointing to the same home
  - No file locking or write ordering from cache to home
- Primary / Secondary
  - Similar to SW, but no caching
  - All files created on primary are replicated to secondary



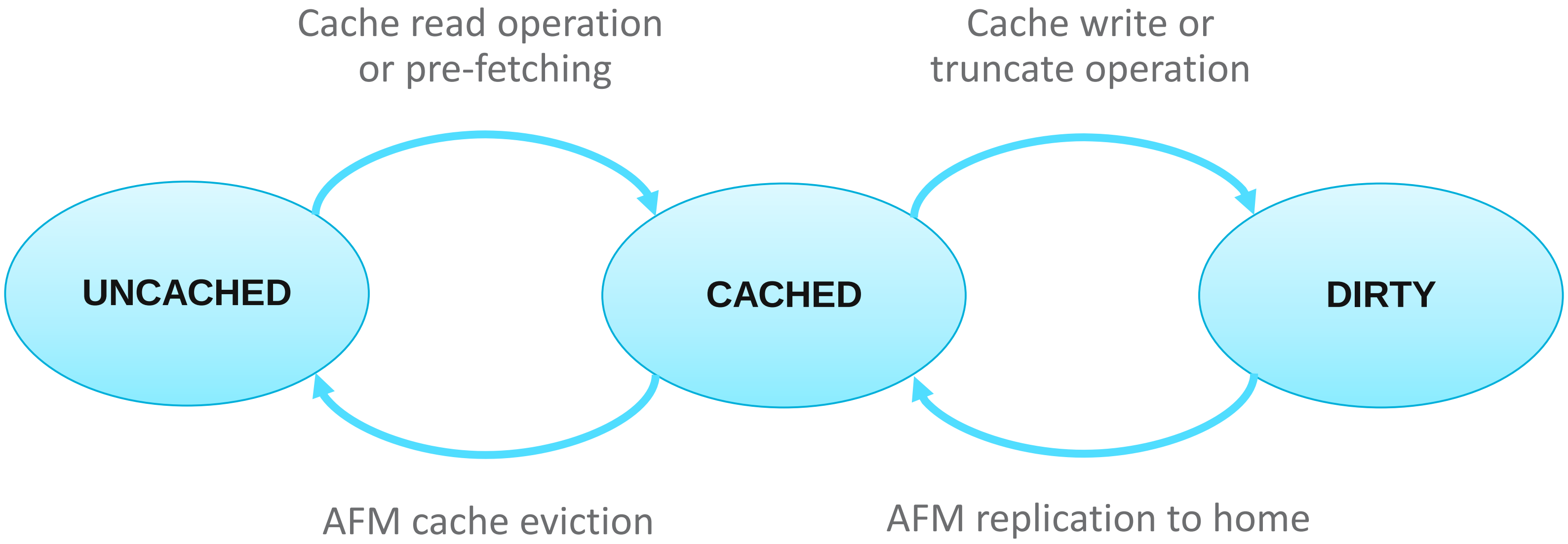
Use Case 1: Branch office

AFM caching use case (single writer)

- Requires GPFS cache system, home can be any NFS storage
  - Complex ACLs are not maintained with non-GPFS home
- File stubs (inodes) are created in cache upon AFM setup
- Data is being copied from home (fetched) upon file access in cache
  - Files can be pre-fetched based on policies to improve performance
- Changed files on cache are replicated back to home
- When file comes to rest it can be evicted from cache based on thresholds
  - File remains visible in cache but does not consume space



AFM file states in SW cache



Cache read operation or pre-fetching

Cache write or truncate operation

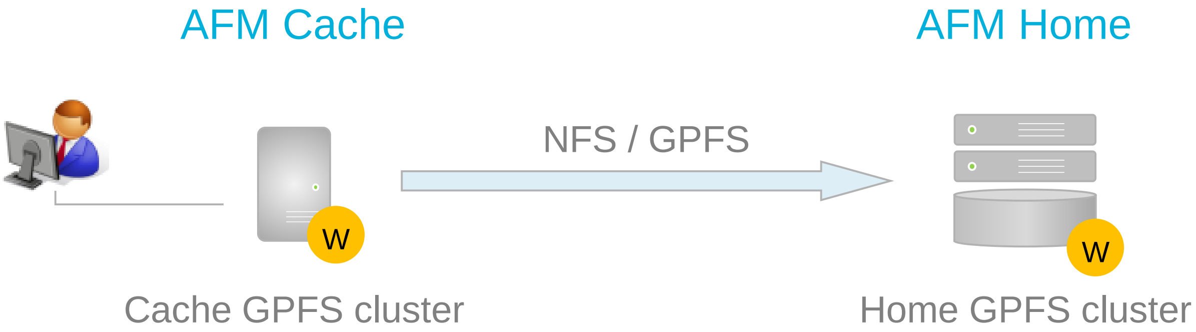
AFM cache eviction

AFM replication to home

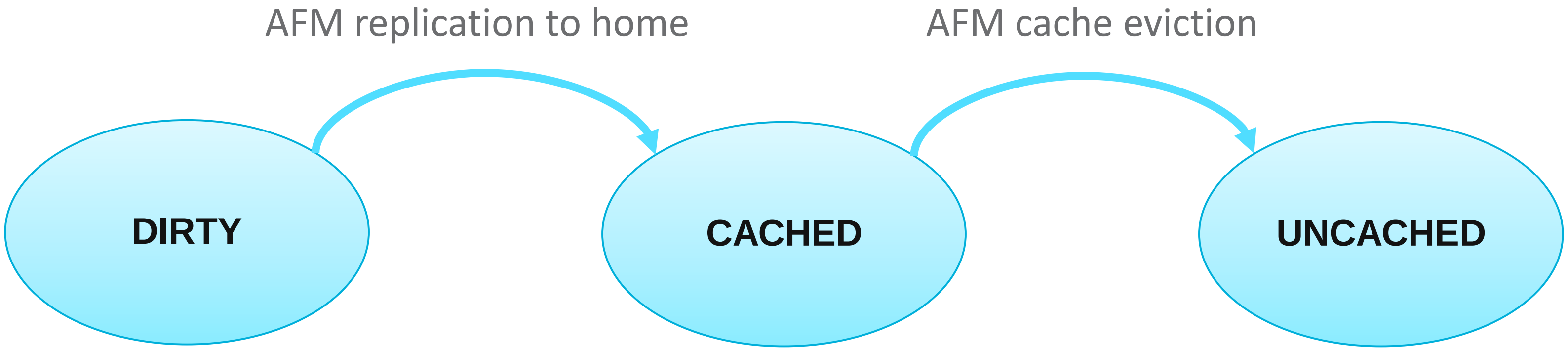
Use Case 2: Data ingest

AFM data ingest use case (independent writer)

- Asynchronous replication between two GPFS clusters
  - Data is generated at cache site, transferred to home site
- Data is transferred to home GPFS system automatically using AFM
- Further processing of data on home system
  - Workflow prevents write conflicts for files
  - Data not modified on cache after creation
  - Centralized analysis, indexing, backup, etc. on home
- Cache eviction frees up storage on cache system for new data



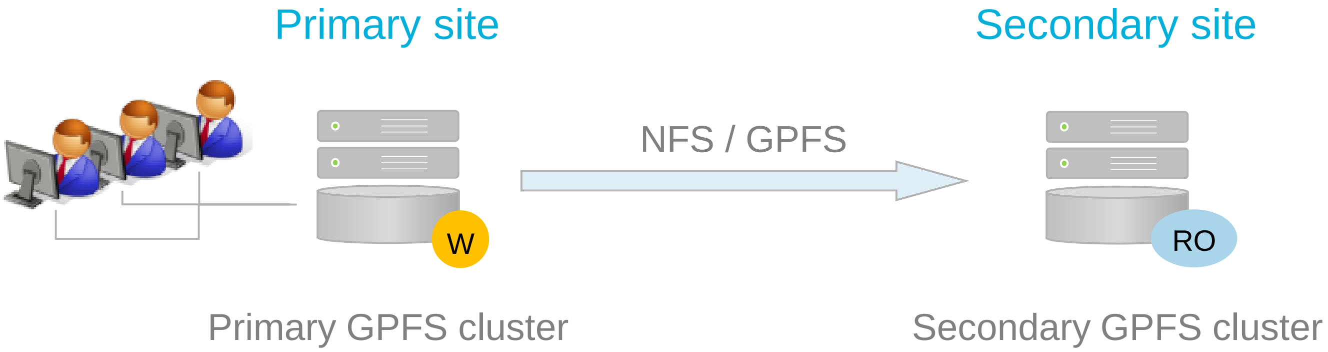
AFM file states in IW cache



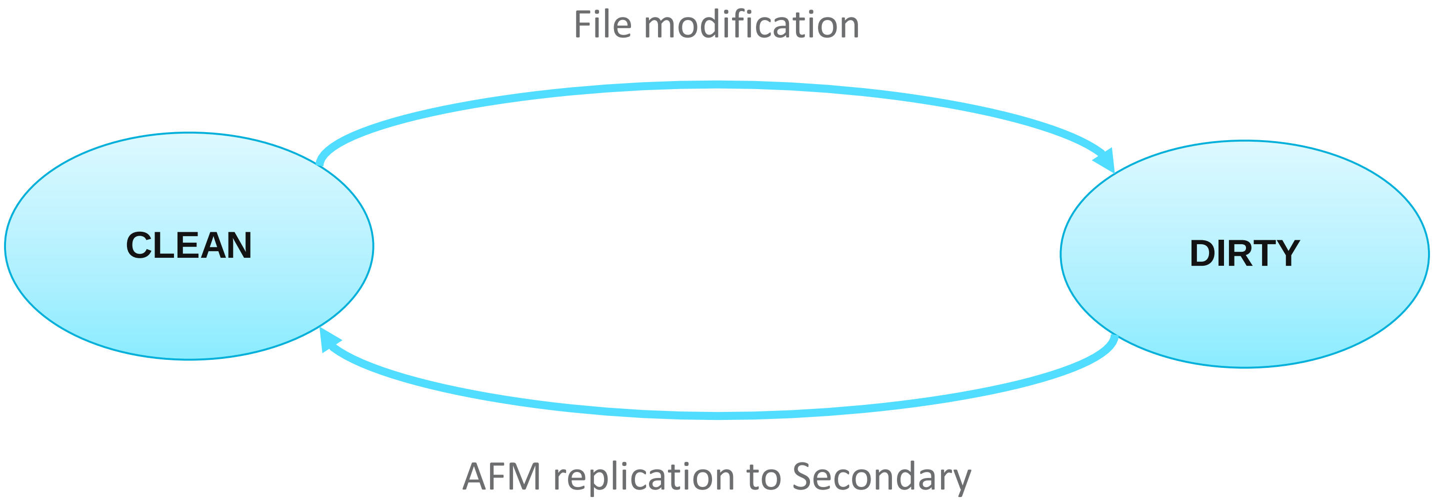
Use Case 3: Disaster protection

AFM Disaster Recovery (primary, secondary)

- Asynchronous replication between two GPFS clusters
  - Cache site is primary and is used for I/O, home site is secondary
  - Peer snapshots are used to provide Recovery Point Objective (RPO)
- When primary is down secondary can be made new primary during failover
- When old primary is back online it can be made primary again during failback
  - Requires delta to be copied from home to cache
- New secondary can be defined if old secondary fails



AFM file states in DR primary



File modification

CLEAN

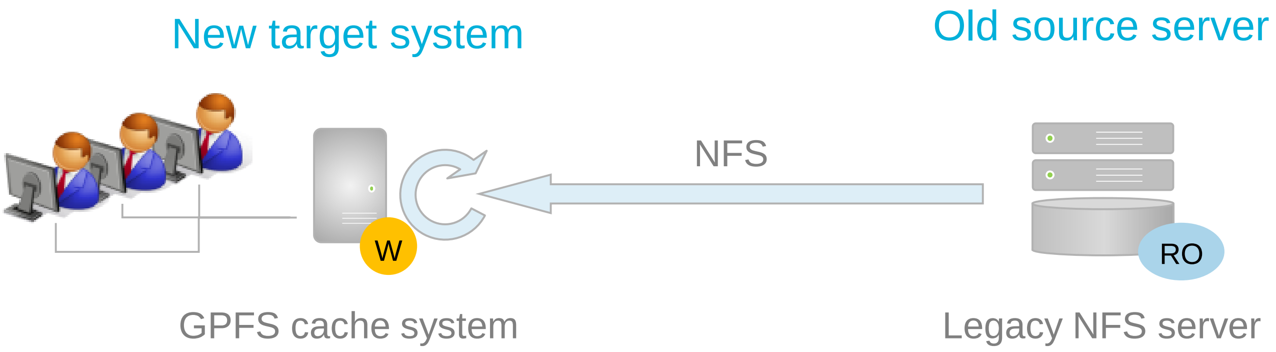
DIRTY

AFM replication to Secondary

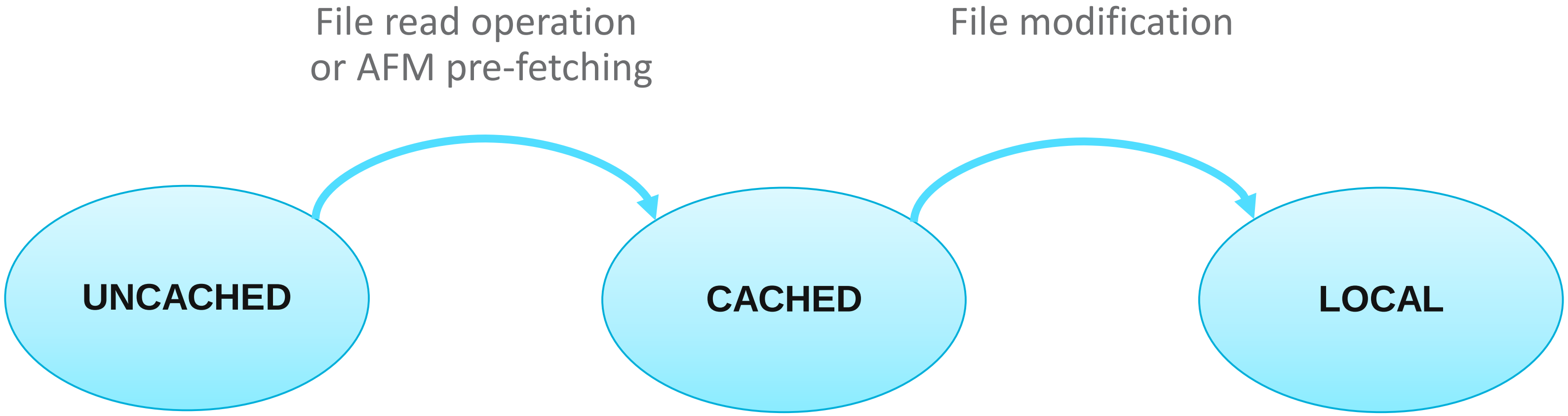
Use Case 4: Migration

AFM migration use case (local update)

- Migrate files from old NFS server to new GPFS system
- Cache system “sees“ all files from old NFS server after establishing AFM relationship
  - Cache is configured in LU mode
  - Home provides NFS export(s)
  - Files can be pre-fetched (migrated) based on results of policy scans
- Switch over when sufficient files are pre-fetched
  - Uncached files accessed on cache are transferred from home
  - Files changed on cache are not replicated back



AFM file states in LU cache



File read operation
or AFM pre-fetching

File modification

UNCACHED

CACHED

LOCAL

Active File Management (AFM)

Spectrum Scale Strategy Days 2017

Achim Christ · achim.christ@de.ibm.com
Karl Schulz · karl.schulz@csi-online.de

Kernkompetenzen

Strak

Interieur
Exterieur
Grauzone
Poly-Modeling

Karosserie

Leichtbau
Fügetechnik
Karosseriearchitektur
Multimaterialbauweise
Simulation

Interieur

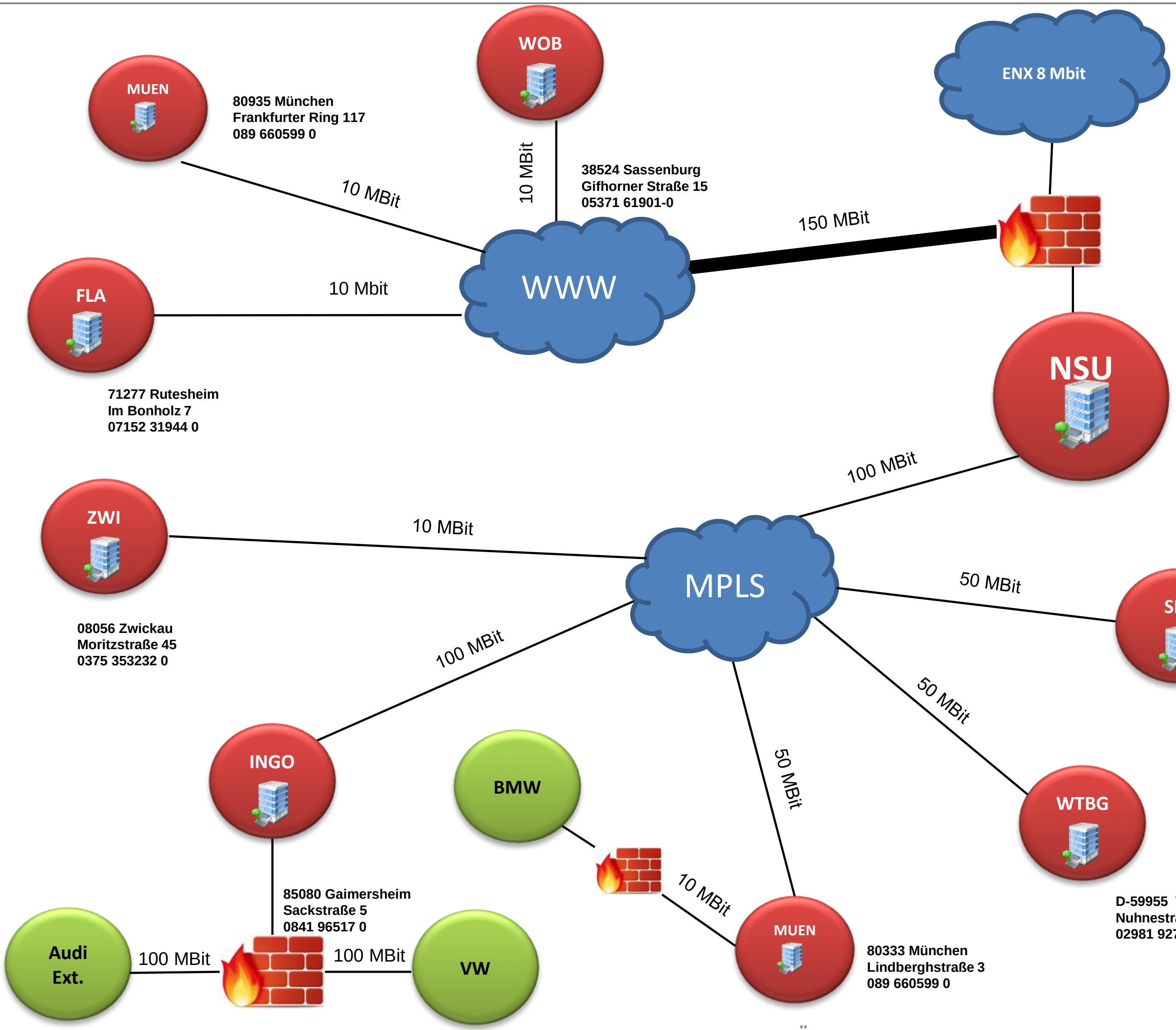
Fahrzeugsicherheit
Industrialisierung
Emotionen
Leichtbau
Simulation

Exterieur

Fahrzeugsicherheit
Leichtbau
Material
Kinematik
Simulation

- Innovationsstudio ▪ csi akademie
- Design Thinking ▪ Technisches Consulting
- Prozessoptimierung ▪ Benchmarkhalle
- Toolentwicklung ▪ Methodenentwicklung

Netzplan - csi Gruppe



80935 München
Frankfurter Ring 117
089 660599 0

38524 Sassenburg
Gifhorner Straße 15
05371 61901-0

71277 Rutesheim
Im Bonholz 7
07152 31944 0

74172 Neckarsulm
Robert-Mayer-Straße 10
07132 9326-0

08056 Zwickau
Moritzstraße 45
0375 353232 0

71063 Sindelfingen
Mercedesstraße 19
07031 61186 0

85080 Gaimersheim
Sackstraße 5
0841 96517 0

D-59955 Winterberg
Nuhnestraße 2
02981 9272 0

80333 München
Lindberghstraße 3
089 660599 0

- Spectrum Scale seit 2010
- Linux HSM / Spectrum Protect
- Filesystemgröße 6 TB – 800 TB
- Clients per SMB / NFS
- File-basierte Arbeitsweise

-> Catia, Icem, Pamcrash,
Moldflow, MS Office…

Herausforderung:

Verbesserung der Zusammenarbeit zwischen den Standorten

- AFM seit Juni 2016 produktiv
- Multiple Writer Modus
- Cache Parameter je Fileset justierbar
- Beliebige Filesets zwischen den Standorten

Alternativen?

...