GPFS for advanced users

Disaster Recovery using GPFS

Outline

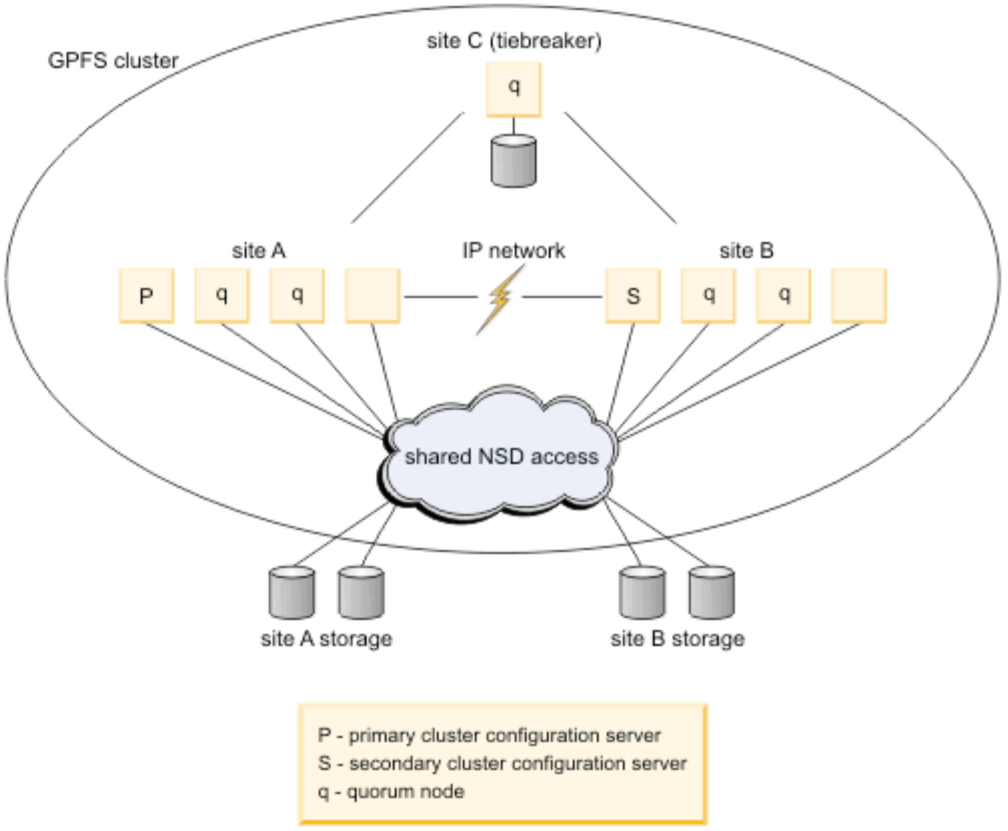
- Disaster recovery solution using GPFS replication
- The GPFS mmfsctl command
- Examples

HA features of GPFS

- HA against catastrophic HW failures using
  - **Replication** of the file system's data at a geographically -separated site → data availability in the event of a total failure of the primary (production) site
  - **Snapshot** allows a backup process to run concurrently with user updates → assures consistency of the data used for backup
  - **AFM** enables sharing data across unreliable or high latency networks.

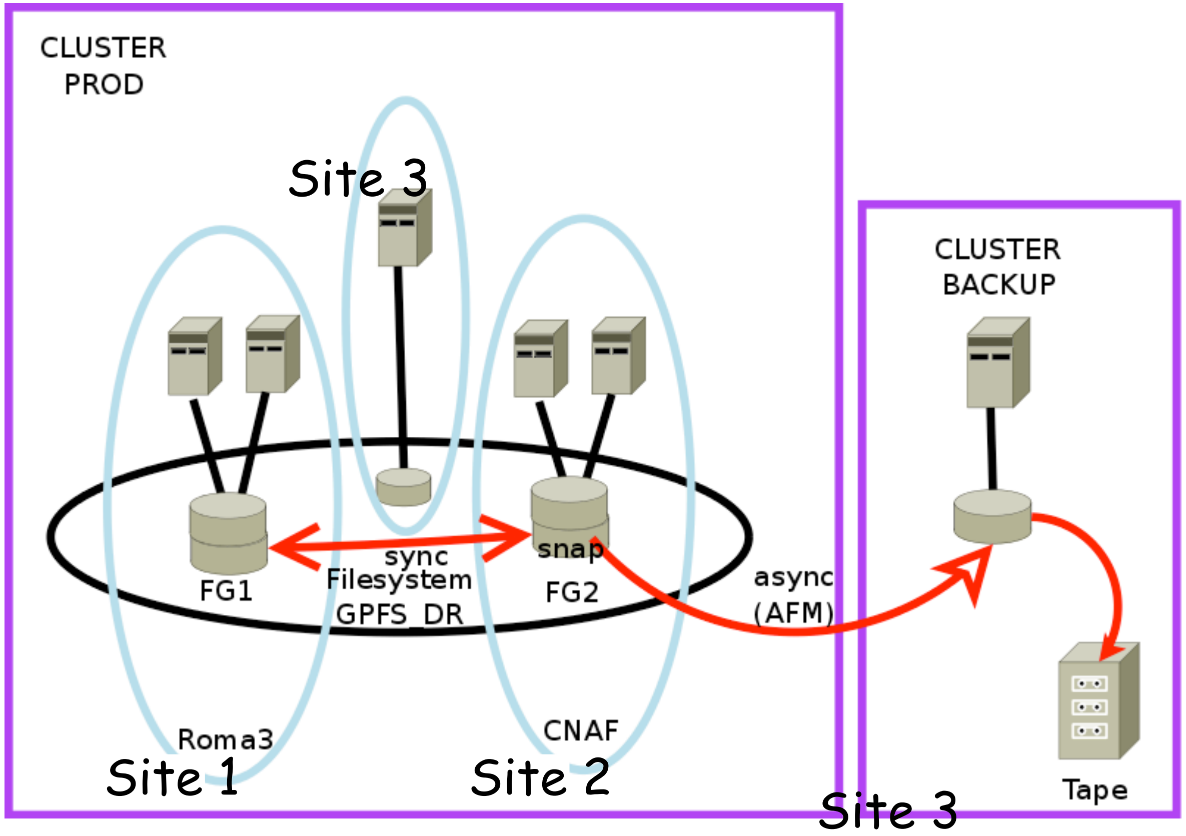
Synchronous mirroring using GPFS Replication

- Data and metadata replication of GPFS can be used to implement synchronous mirroring between a pair of geographically separate sites



Replication + Snapshot + AFM = Complete Solution

- 3 or 4 geo separated sites
  - 2 sites in close vicinity to compose HA cluster
  - 1 tie breaker site
    - Keeping also FS descriptor and cluster configuration
  - 1 backup site
    - Can coincide con tie breaker



- Backup can be done from a Snapshot copied to backup site via AFM
  - Backup window = time to stop/sync/start application
  - All data transferred to backup site in background (asynchronously)
  - Backup will be kept in 4 copies (2 on disk in prod cluster, 1 on disk and 1 on tape in backup cluster

Failure scenarios

failures	effects	actions	downtime
Disk on site1	Switching to access disk remotely from site 2	non	0
WAN network connection to site 1	no access to data, application crashes or hangs	ensure that application is not running on site 1, restart application on site2	t1
Site1 failure		restart application on site2	t1
Site3 (tiebreaker) failure	non	non	0
site2 and site3 failure	No access to data, file system down application crashes or hangs	reconfigure quorum nodes, restart application	t1+ 1min

t1 = time to restart application

Production site failure

After a production site failure, no administrative intervention is required.

- GPFS detects the failure and reacts to it as follows:
- The failed nodes are marked as down.
- The failed disks are marked as unavailable.
- The application continues running at the surviving site.

After prod site recovery

Perform the following steps:

- Restart GPFS on all nodes at the recovered site:
  - mmstartup –a
- Bring the recovered disks online:
  - mmchdisk … start

Failure on the Production site and the Tiebreaker site

GPFS loses quorum and the file system is unmounted after the failure occurs.

The administrator initiates the manual takeover procedure:

- Relaxes the node quorum:
  - mmchnode --nonquorum –N …
- Relaxes the file system descriptor quorum:
  - mmfsctl … exclude

mmfsctl command

implements disaster recovery functionality:

mmfsctl Device {suspend | resume}

- suspends file-system I/O and flushes the GPFS cache to ensure the integrity of the FlashCopy image

mmfsctl Device syncFSconfig {-n RemoteNodesFile | -C RemoteCluster} [-S SpecFile]

- use this command to synchronizes the file system's configuration state between peer recovery clusters

mmfsctl Device {exclude | include} {-d DiskList | -F DiskFile | -G FailureGroup}

- Use this command for minority takeover in Active-Active replicated configurations. It tells GPFS to exclude the specified disks or failure groups from the file system descriptor quorum