DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

Lecture on Parallel Filesystems

GPFS

Jacques-Charles Lafoucriere

ENSIIE| 2018

GPFS

General Parallel File System

- IBM // FS
- 1992: Tiger Shark (multimedia FS)
- 1998: renamed GPFS
- 2015: renamed Spectrum Scale

- Support Linux, AIX and Windows

GPFS Architecture

Shared Disk File System

- All nodes must see all the storage
- Distributed locking
- Native block device use

Data Organization

- Files are divided into equal size blocks
- Blocks are placed on different disks in a round-robin fashion

GPFS Architecture (Cont.)

Namespace Organization

- Use of extensible hashing
- The block containing the directory entry for a particular name can be found by applying a hash function to the name

MetaData Consistency

- Records all metadata updates that affect file system consistency in a journal or write-ahead log

GPFS Architecture (Cont.)

Locking

- Distributed Locking
- Allows greater parallelism than centralized management as long as different nodes operate on different pieces of data/metadata
- The global lock manager coordinates locks between local lock managers by handing out lock tokens
- GLM delegates token managements to LLM

GPFS Components

Cluster

- A number of nodes and network shared disks (NSDs) for management purposes

Node

- Any server that has the Spectrum Scale product installed with direct storage access or network access to another node

Network Shared Disk (NSD)

- For global device naming and data access in a cluster
- Allow sharing of host connected storage

GPFS Components (cont.)

Cluster manager

- The node which
  - Monitors disk leases
  - Detects failures and manages recovery from node failure within the cluster
  - Determines whether or not a quorum of nodes exists to allow the Spectrum Scale daemon to start and for file system usage to continue
  - Handle configuration information
  - Selects the file system manager node

GPFS Components (cont.)

File system manager

- Maintains the availability information for the disks in the file system
- Manages file system configuration
- Manages disk space allocation
- Manages quota configuration
- Handle security services

Metanode

- Handles metadata, also referred to as “directory block updates”

Application node

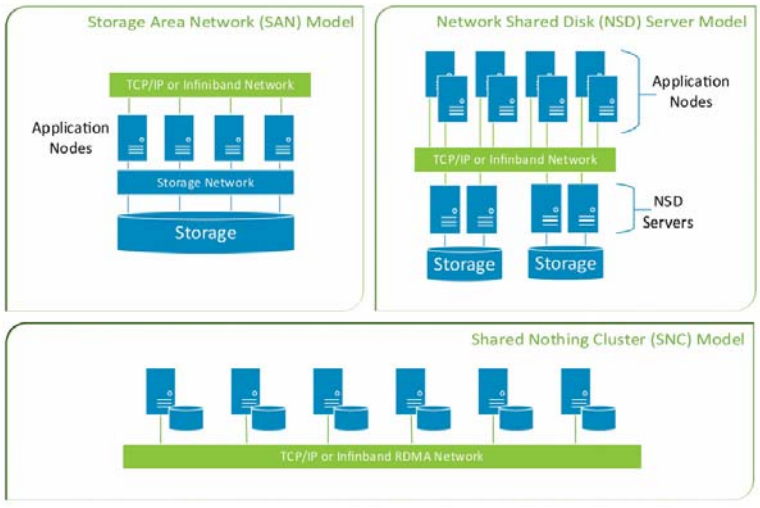
- Mounts a Spectrum Scale file system and runs a user application that accesses the file system

GPFS Components (cont.)

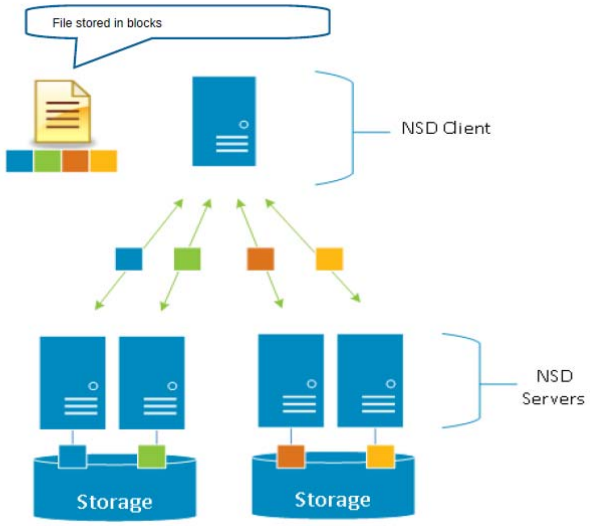
Quorum nodes

- Nodes participating in maintaining the Spectrum Scale cluster active
  - Node quorum
    - the cluster is maintained online when most of quorum nodes are available
  - Node quorum with tiebreaker disks
    - the cluster is online when one quorum node is up and it has access to the tiebreaker disks

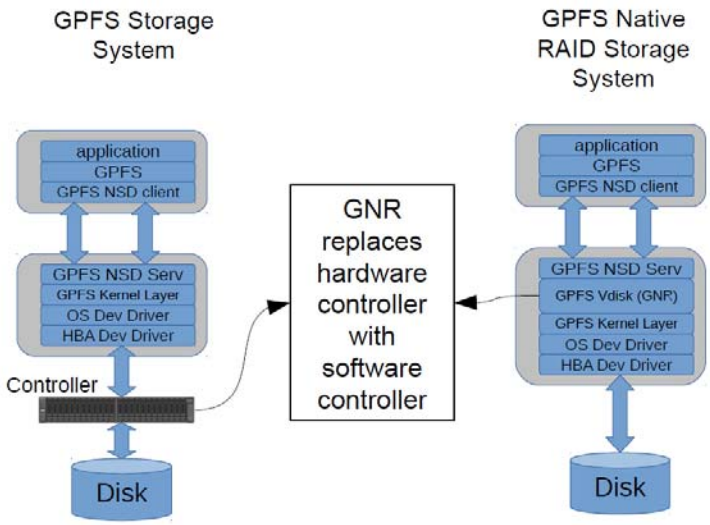
GPFS Cluster Modes



GPFS Parallel Architecture



GPFS Software RAID



GPFS Network Layer

TCP/IP based

- Bounding
- Routing

RDMA support

- High Speed

GPFS Supported Feature

- ILM
  - HSM
- Fileset
  - A subtree of a file system namespace that in many respects behaves like an independent file system (quotas, snapshot)
  - Independent or dependent (inode space)
- Quotas
- ACL
- Extended Attributes
- Immutable or AppendOnly files
- Disks Pools
- Snapshots
- Policies
  - Used to assign files to specific storage pools

GPFS Connectors

Provide Storage Service Based on GPFS

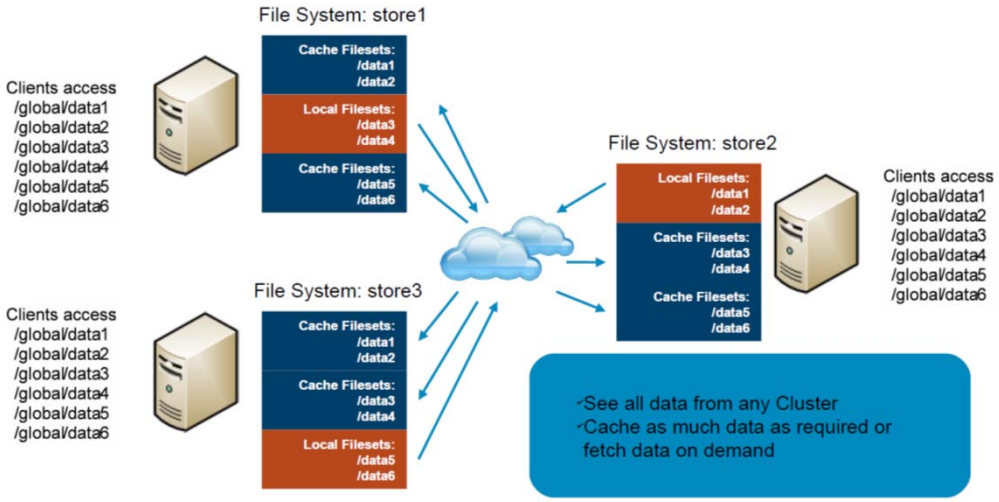
- NFS
  - To Lan or Work Stations
- Hadoop
  - For Map/Reduce workload
- iSCSI
  - For Virtual Machines or diskless nodes
- SMB
  - For Windows world
- OpenStack Swift
  - For Cloud type workload

Active File Management

Objective

- Allow multiple distributed namespace synchronization
- Scalable, high-performance, file system caching feature
- Capable of masking wide area network (WAN) latencies
- Allow automation of the control of data location and data copies, independently of the total number of objects in the namespace

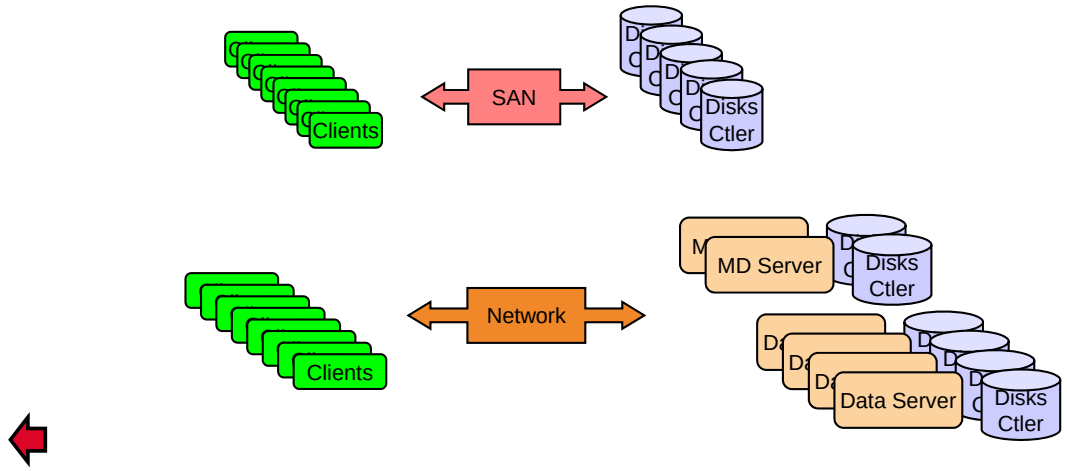
AFM (Cont.)



Thank you for your attention

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM-Ile de France | 91297 Bruyères-le-Châtel Cedex
T. +33 (0)1 69 26 40 00 | F. +33 (0)1 69 26 70 86
Direction des applications militaires
Département sciences de la simulation et de l'information
Service informatique scientifique et réseaux

Shared Disk vs Client/server



Network Shared Disks