Grid Computing

# IBM LoadLeveler and IBM GPFS Multicluster
## *Unicore Summit*


Europe Design Center for On Demand Business

***Solution Architect***

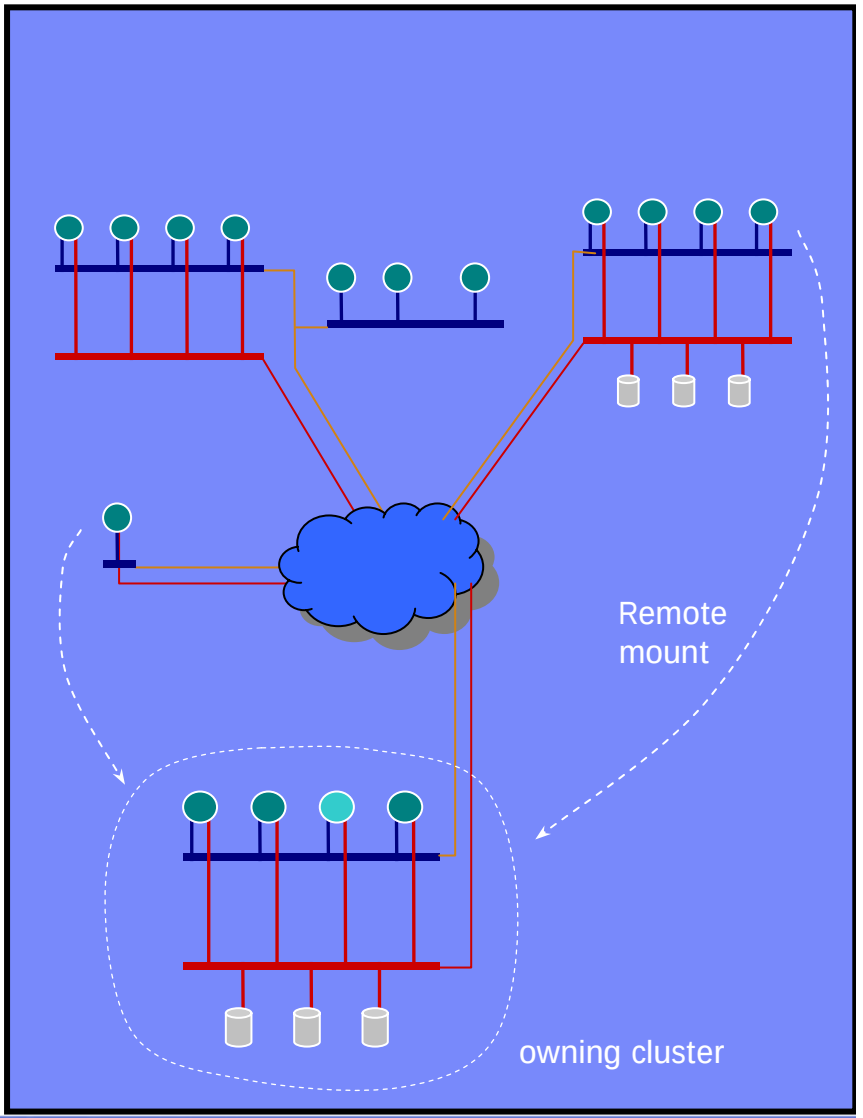
***Jean-Yves Girard***

# Agenda

- GPFS and LoadLeveler Multi-Cluster
- DEISA Project
- Implementation details

# GPFS 2.3

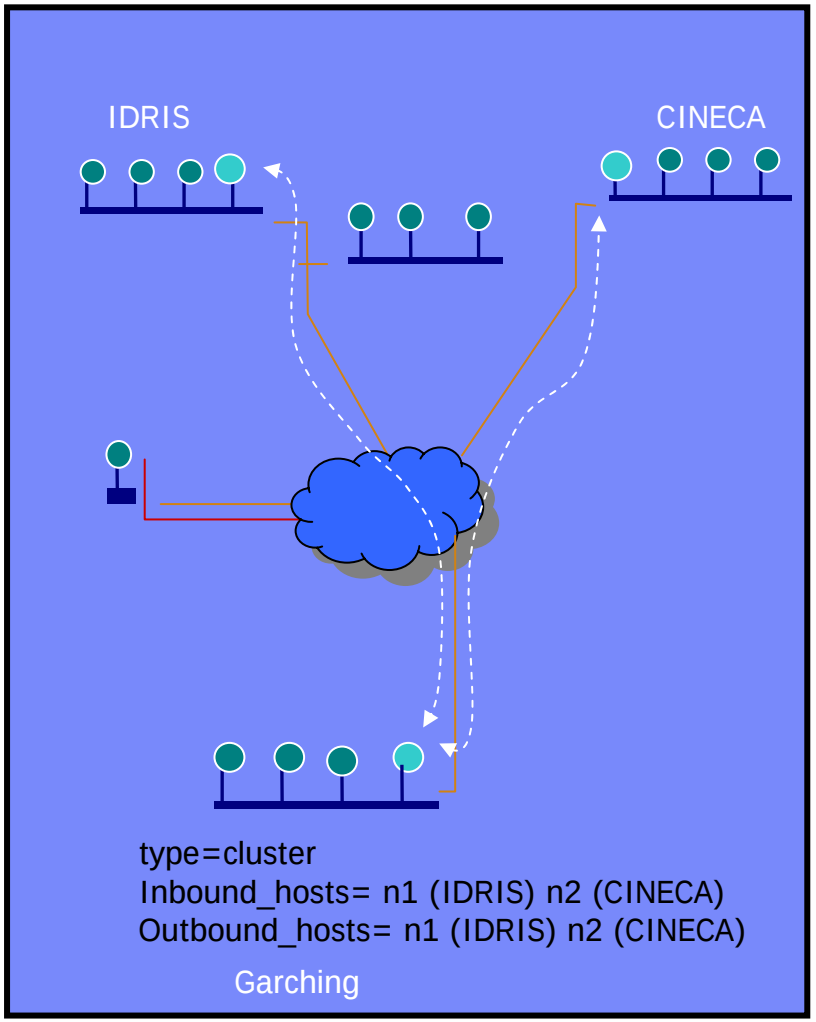
- Each FS belongs to one “owning” cluster, responsible for
  - administration
  - lock management
  - recovery
- “Remote” nodes can mount FS and
  - request locks
  - access data & metadata directly over the SAN
  - do not participate in quorum or administration.
- Disk access is through NSD or an external disk facility (FC/IP, iSCSI …)
- Cluster manager detects node failures and drives recovery
- Global configuration data published via local configuration data.
- Security: disk access controlled by SAN; non-SAN data and controls through SSL
- UID mapping for file access control on remote nodes
- Scaling by limiting protocol traffic to “active” nodes
- Failure detection using GPFS disk leasing/heartbeating outside that set of machines.



# LoadLeveler 3.3 extensions

- Provide users the capability to submit jobs to more than one LoadLeveler clusters easily. This includes accessing a Linux cluster from an AIX cluster, and vice versa.
- Facilitate workload balancing across multiple clusters
- Address scalability issue as the size of clusters exceeds the currently supported limit
- Security

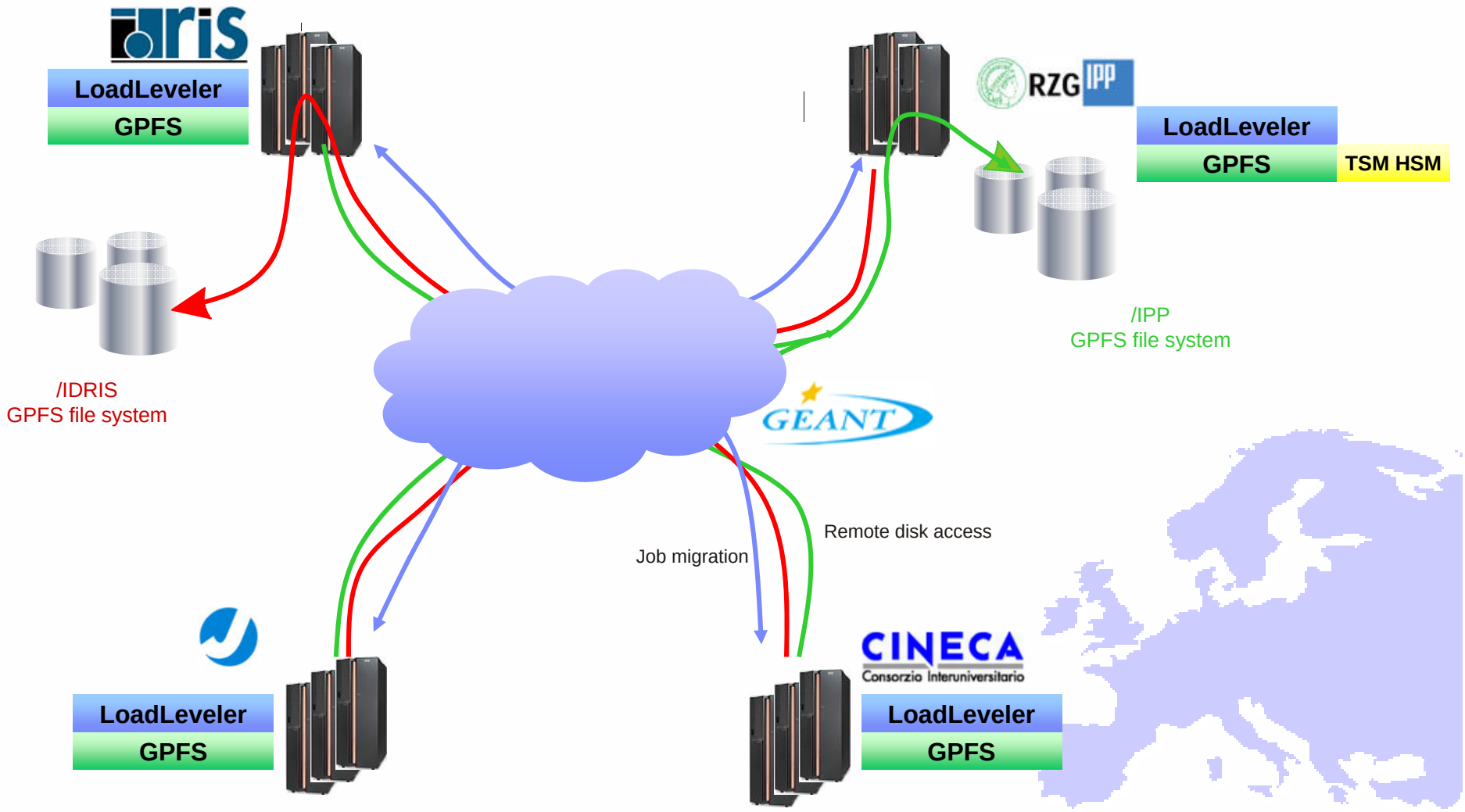
- Preemption
- Advanced reservation
- Processor/Memory Scheduling Affinity on Power4/5 MCMs



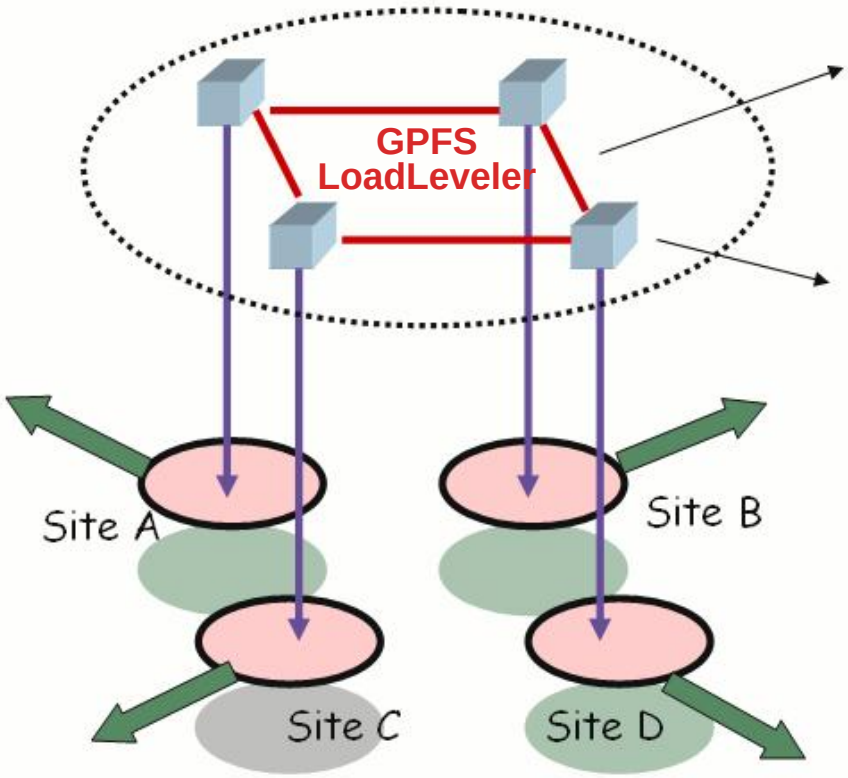
# Agenda

- GPFS and LoadLeveler Multi-Cluster
- DEISA Project
- Implementation details

# LL/GPFS MC Architecture for DEISA



# Extending LL/GPFS MC Grid




**NFS v4**

**Globus**

**Unicore**

**Dedicated bandwidth network.**
**GEANT, RENATER, DFN, GARR, …**

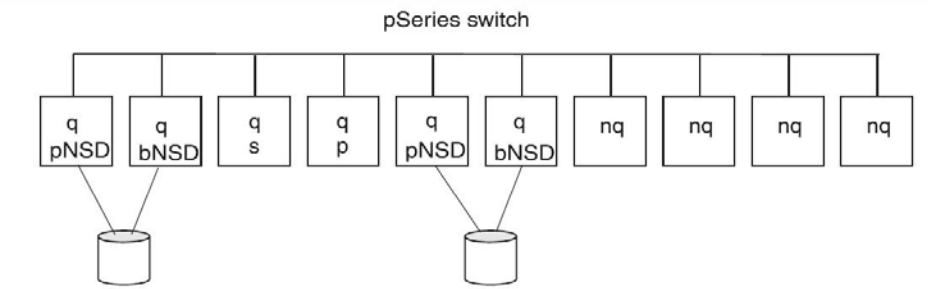
**National supercomputing platforms:**
**IDRIS - France**
**JULICH - Germany**
**GARCHING - Germany**
**CINECA - Italy**
**…**
**SARA – The Netherlands**
**CSC - Finland**

**Extended Grid services :**
**Portals, Web-like services, …**
**Interfacing the core platform to other virtual organizations.**
**Hiding complex environments from end users**

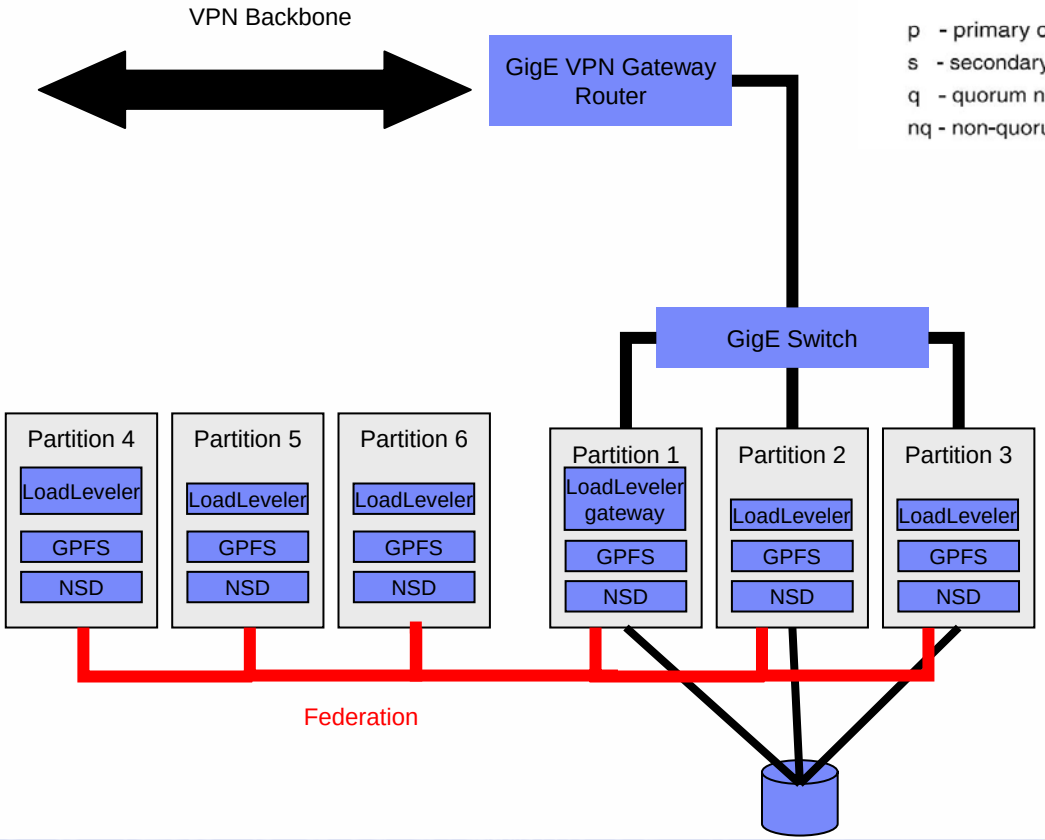
# Agenda

- GPFS and LoadLeveler Multi-Cluster
- DEISA Project
- Implementation details

# GPFS Operational Model

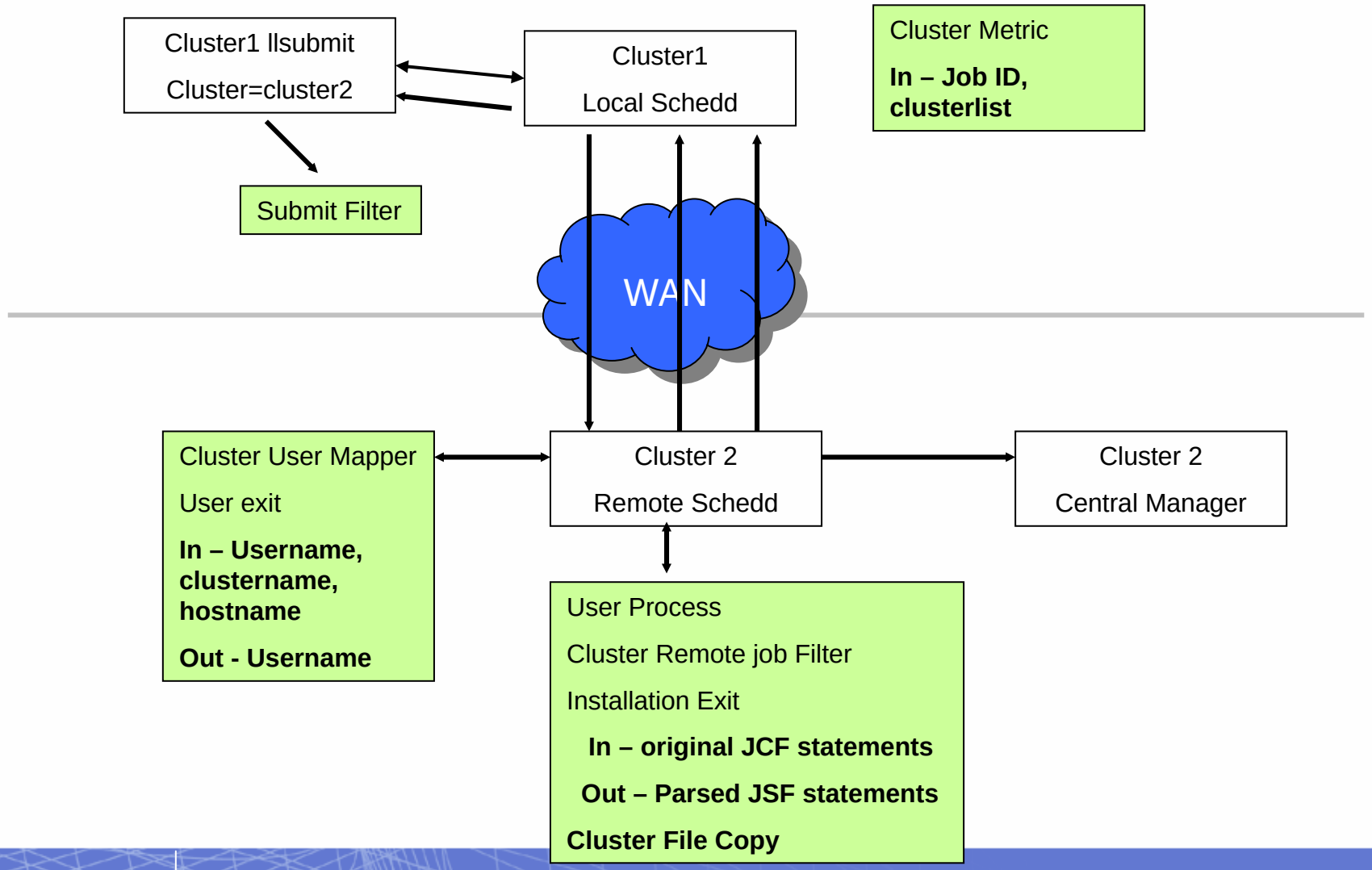



p - primary cluster configuration server
s - secondary cluster configuration server
q - quorum node
nq - non-quorum node
pNSD - primary NSD server
bNSD - backup NSD server




Declare the remote cluster
**mmremotecluster** add dream.ppd.pok.ibm.com -n c5n97g
Update contact nodes
**mmremotecluster** update dream.ppd.pok.ibm.com -n c5n97g,c5n98g
Display remote cluster information
**mmremotecluster** show
*dream.ppd.pok.ibm.com c5n97g,c5n98g*
Declare remote file system
**mmremotefs** add fs1 -f gpfs1 –C dream.ppd.pok.ibm.com -T /fs1
Display remote file system information
**mmremotefs** show
*fs1 gpfs1 dream.ppd.pok.ibm.com*
Display remote cluster information
**mmremotecluster** show
*dream.ppd.pok.ibm.com c5n97g,c5n98g*
*fs1 gpfs1*
Mount the remote GPFS file system
**mount** /fs1

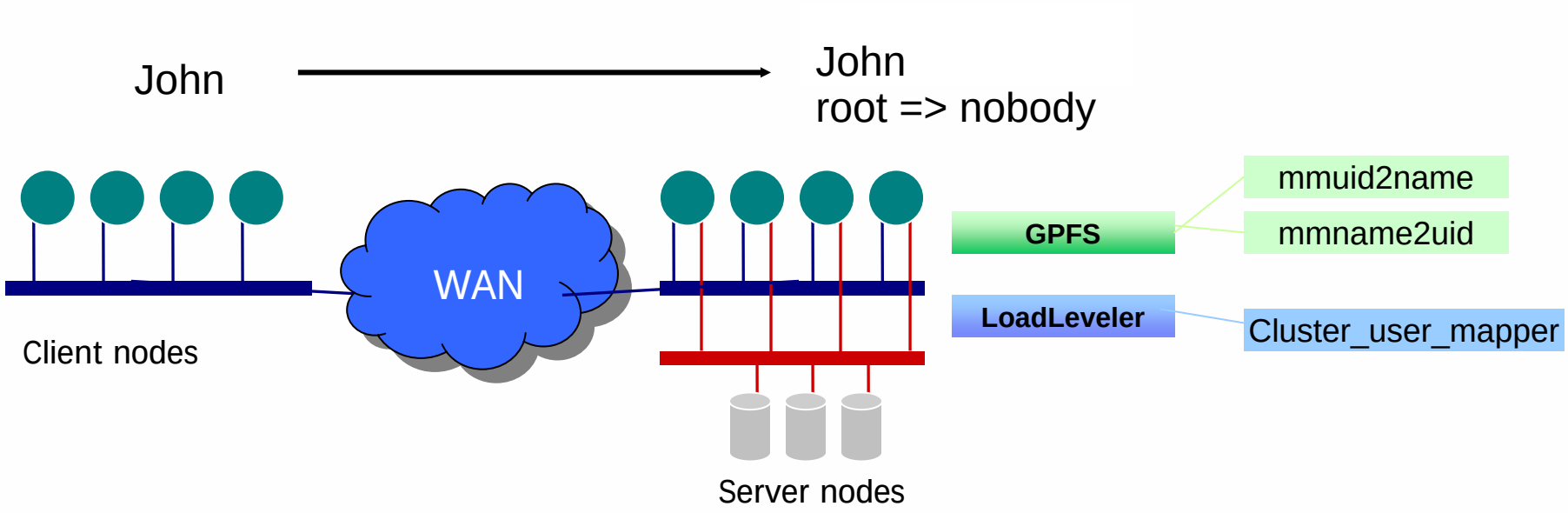
# LoadLeveler Multi-Cluster exits



# Examples

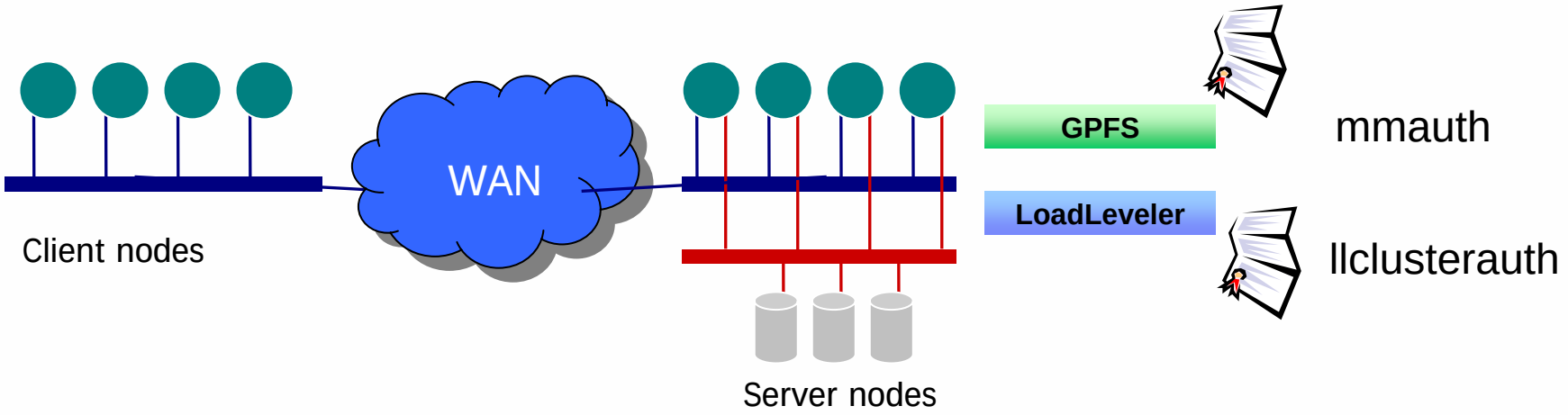
- This example moves the idle job silver.11 from the local cluster to the remote cluster cluster1:
  - `llmovejob -C cluster1 -j silver.11`

# User Management




- GPFS calls external plug-in to map client UID to global name, global name to file system UID/GID
- GPFS caches mapping so subsequent checks are fast
- http://www-1.ibm.com/servers/eserver/clusters/whitepapers/uid_gpfs.html
- mmchconfig enableUIDRemap=yes

# Security




- Each site controls and manages access to its resources
  - Certificates authenticates remote cluster access
- Root id cannot be mapped by LoadLeveler
- GPFS maps root to nobody via mmuid2name and mmname2uid scripts
- Access Control List
- Local administrators can blocks incoming jobs and decide if a job must be migrated

# GPFS Howto

On Cluster_A

1. Generate public/private key pair
mmauth genkey
creates public key file with default name id_rsa.pub
start GPFS daemons after this command!
2. Enable authorization
mmchconfig cipherList=AUTHONLY

3. Sysadm gives following file to Cluster_B
/var/mmfs/ssl/id_rsa.pub
rename as cluster_A.pub

7. Authorize Cluster_B to mount FS owned by Cluster_
mmauth add cluster_B **-k cluster_B.pub**

On Cluster_B

4. Generate public/private key pair
mmauth genkey

creates public key file with default name id_rsa.pub
start GPFS daemons after this command!
5. Enable authorization
mmchconfig cipherList=AUTHONLY

6. Sysadm gives following file to Cluster_A
/var/mmfs/ssl/id_rsa.pub
rename as cluster_B.pub

8. Define cluster name, contact nodes and public key for cluster_A
mmremotecluster add cluster_A -n nsd_A1,nsd_A2,nsd_A3,nsd_A4 **-k Cluster_A.pub**

9. Identify the FS to be accessed on cluster_A
mmremotefs add /dev/fsAonB -f /dev/fsA -C Cluster_A -T /dev/fsAonB
10. mount FS locally
mount /fsAonBc

# LL Howto

1. Create the SSL authorization keys by invoking the llclusterauth command with the -k option on all gateway nodes. Result: LoadLeveler creates a public key, a private key, and a security certificate for each gateway node.
2. Distribute the public keys to gateways on other secure clusters. This is done by exchanging the public keys found in /var/load/ssl/id_rsa.pub file with the other clusters you wish to communicate with.
3. Copy the public keys of the clusters you wish to communicate with into the authorized_keys directory on your inbound schedd nodes. (for AIX, /var/loadl/ssl/authorized_keys v for Linux, /var/opt/LoadL/ssl/authorized_key ). The authorization key files can be named anything within the authorized_keys directory.
4. Define the cluster stanzas within the LoadLeveler administration file, using the multicluster_security = SSL keyword. Define the keyword ssl_cipher_list if a specific OpenSSL cipher encryption method is desired. Use secure_schedd_port to define the port number to be used for secure inbound transactions to the cluster.
5. Notify LoadLeveler daemons by issuing the llctl command with the recycle keyword. Otherwise, LoadLeveler will not process the modifications you made to the administration file.
6. Configure firewalls to accept connections to the secure_schedd_port numbers you defined in the administration file.


# ibm.com

**Jean-Yves Girard**

SGB Solution Architect
Europe Design Center
for On Demand Business
girardjy@fr.ibm.com
+ 33 4 67 34 45 39