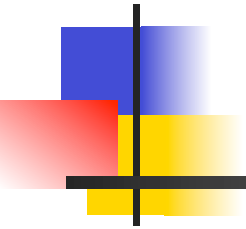
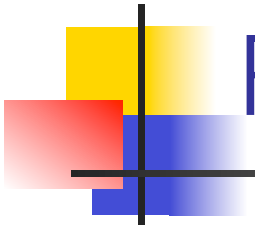


Remote cluster

Alessandro Brunengo INFN-Genova

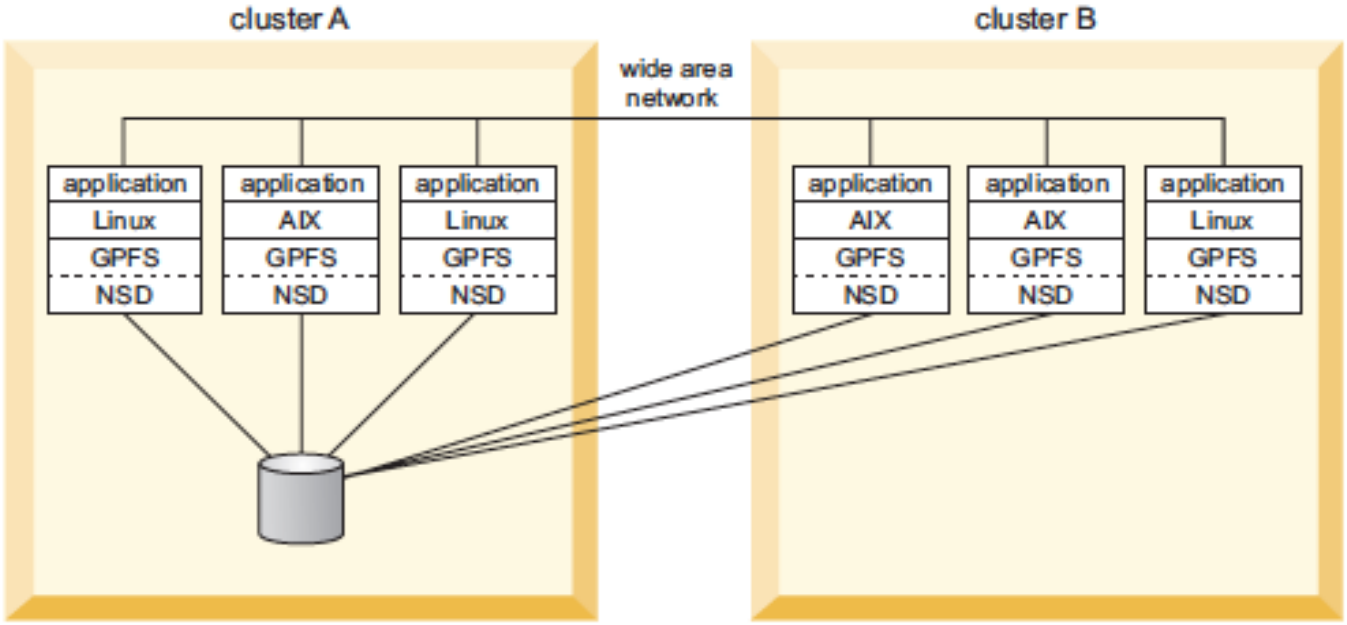


Remote cluster export

- Il file system GPFS creato su un cluster e' accessibile implicitamente a tutti i nodi del cluster
- GPFS permette di rendere un file system accessibile anche ai membri di un cluster GPFS diverso
- Prerequisito e' che ogni nodo del cluster che vuole montare il file system deve avere connettivita' TCP verso tutti i nodi del cluster proprietario del file system
  - la comunicazione avviene tra i demoni GPFS: non c'e' comunicazione di management via ssh
  - se due cluster A e B montano un file system del cluster C, i nodi di A e B non devono comunicare tra loro

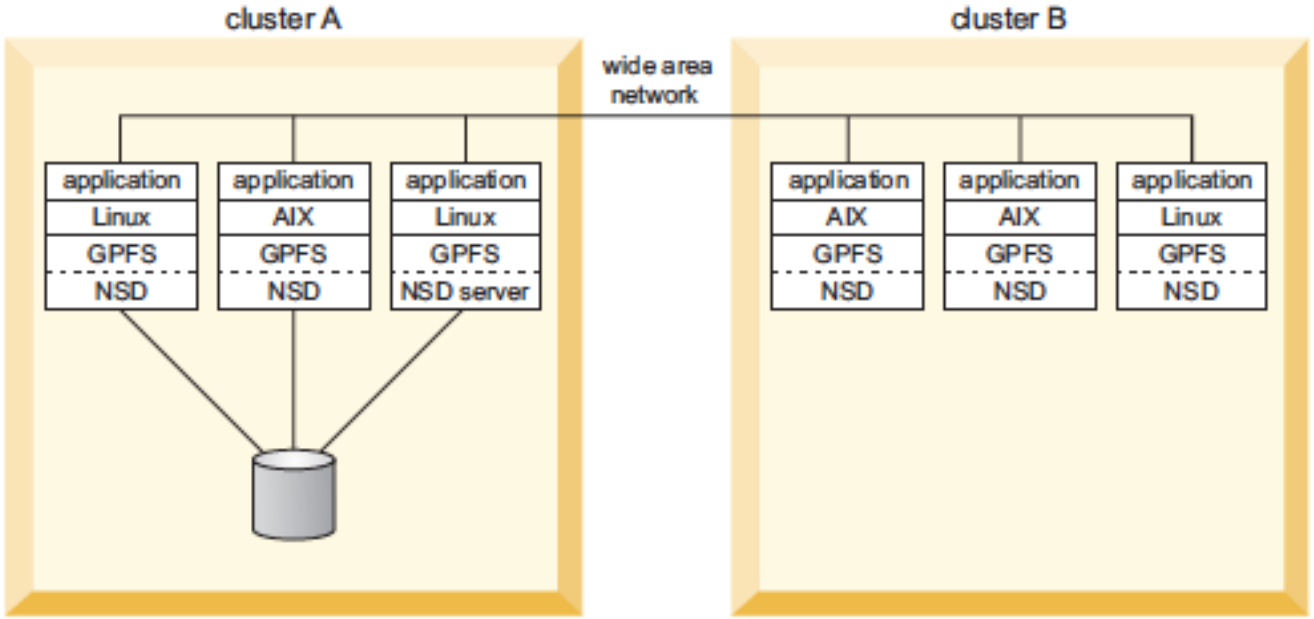
Accesso diretto agli NSD

Se i client del cluster remoto vedono gli NSD del file system tramite una SAN, l'accesso ai NSD e' diretto



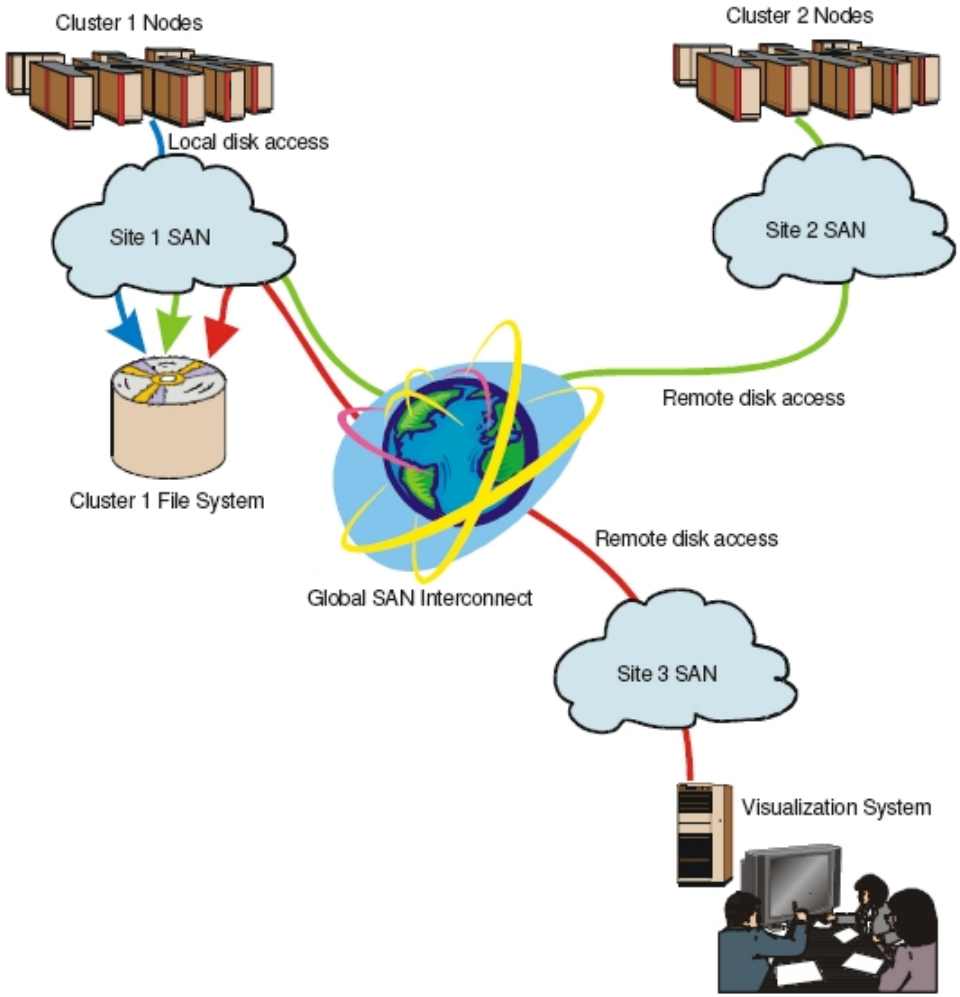
Accesso tramite NSD server

In mancanza di visibilita' diretta, gli NSD sono visti tramite gli NSD server



GPFS Multi-Cluster Feature
The Big Picture

- Problem: nodes outside the cluster need access to GPFS files
- Solution: allow nodes outside the cluster to natively (i.e., no NFS) mount the file system
  - "Home" cluster responsible for admin, managing locking, recovery, etc.
  - Separately administered remote nodes have limited status
    - Can request locks and other metadata operations
    - Can do I/O to file system disks over global SAN
    - Are trusted to enforce access control, map user Ids, ...
- Uses:
  - High-speed data ingestion, postprocessing (e.g. visualization)
  - Sharing data among clusters
  - Separate data and compute sites (Grid)
  - Forming multiple clusters into a "supercluster" for grand challenge problems
- Scaling: max supported GPFS cluster size



GPFS Multi-Cluster
Example

Mount a GPFS file system from Cluster_A onto Cluster_B

On Cluster_A

1. Generate public/private key pair
   `mmauth genkey new`
   COMMENTS
   - key pair is placed in /var/mmfs/ssl
   - public key default file name `id_rsa.pub`
2. Enable authorization
   `mmauth update . -l AUTHONLY`
3. Sysadm gives following file to Cluster_B
   `/var/mmfs/ssl/id_rsa.pub`
   COMMET: rename as `cluster_A.pub`
7. Authorize Cluster_B to mount file systems owned by Cluster_A
   `mmauth add cluster_B -k cluster_B.pub`
8. Authorize Cluster_B to mount a particular FS owned by Cluster_A
   `mmauth grant cluster_B -f /dev/fsA`

On Cluster_B

4. Generate public/private key pair
   `mmauth genkey`
   COMMENTS
   - key pair is placed in /var/mmfs/ssl
   - public key default file name `id_rsa.pub`
5. Enable authorization
   `mmauth update . -l AUTHONLY`
6. Sysadm gives following file to Cluster_A
   `/var/mmfs/ssl/id_rsa.pub`
   COMMENT: rename as `cluster_B.pub`

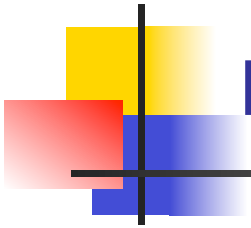
9. Define cluster name, contact nodes and public key for cluster_A
   `mmremotecluster add cluster_A -n nsd_A1,nsd_A2 -k Cluster_A.pub`
10. Identify the FS to be accessed on cluster_A
    `mmremotefs add /dev/fsAonB -f /dev/fsA -C Cluster_A -T /fsAonB`
11. mount FS locally
    `mmmount /dev/fsAonB`

Communication Between Clusters

All nodes in both clusters must have TCP/IP connectivity between each other; this is used only for "daemon to daemon" (*n.b.*, mmfsd) communication via the GPFS protocol. This does not allow remote shell (e.g., ssh or rsh) access between nodes; remote users can **not** access the home cluster by this mechanism. OpenSSL guarantees secure communications.

Contact Nodes

The contact nodes are used only when a remote cluster first tries to access the home cluster; one of them sends configuration information to the remote cluster after which there is no further communication. It is recommended that the primary and backup cluster manager be used as the contact nodes.



mmauth

- # mmauth update . -l AUTHONLY
  - il comando configura il cluster per utilizzare il livello di encription specificato (authonly: encription per l'autorizzazione all'accesso)
  - quando si passa da un livello non sicuro a sicuro e viceversa, **il cluster deve essere down**
- # mmauth delete <remote-cluster>
  - elimina un cluster (e la sua chiave) tra quelli precedentemente autorizzati
- # mmauth grant <remote-cluster>
  - opzione per specificare solo un file system (**-f**)
  - opzione per specificare accesso rw o ro (**-a**)
  - opzione per specificare root-squash (**-r [uid:gid | no]**)
- # mmauth deny <remote-cluster> [-f <device>]
  - rimuove una autorizzazione concessa
- # mmauth show
  - visualizza i cluster a cui si e' dato accesso a file system locali, ed i file system relativi

Problema con libreria ssl

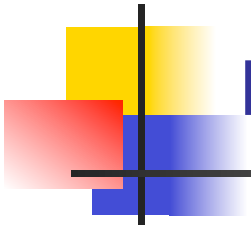
- Si puo' verificare un problema con libreria SSL
  - GPFS non trova libssl.so.4
  - RHEL5/6 troppo up-to-date: hanno libssl.so.1.0.0
- Si deve modificare un parametro di configurazione undocumented:

  mmchconfig openssllLibName=libssl.so.1.0.0

- E' meglio aggiungere il nome della libreria attuale a quelli preconfigurati; per vedere l'attuale valore (a cluster su):

  ```
  # mmdiag --config | grep openssllLibName
                    openssllLibName libssl.so:libssl.so.0:libssl.so.4
  ```

- Per aggiungere il nome in piu':
  - mmchconfig openssllLibName=<old string>:<new file name>



mmremote*

- # mmremotecluster add | update | delete | show
  - gestisce i cluster remoti da cui si desidera montare un file system
- # mmremotefs add <device>
  - crea un “device” corrispondente al file system remoto sui nodi del cluster locale
  - permette di specificare le seguenti opzioni identiche a quelle del file system locale:
    - il mount point (**-T <dir>**), omogeneo sul cluster locale
    - quando montare (**-A yes | no | automount**)
    - opzioni del mount (**-o <MountOptions>**)
    - priority del mount (**--mountPriority <prio>**)
- # mmremotefs update modifica le opzioni
- # mmremotefs delete rimuove un remote file system

Subnets in multi-cluster

- E' possibile configurare il parametro subnets anche per definire il path di accesso a nodi di cluster remoti:

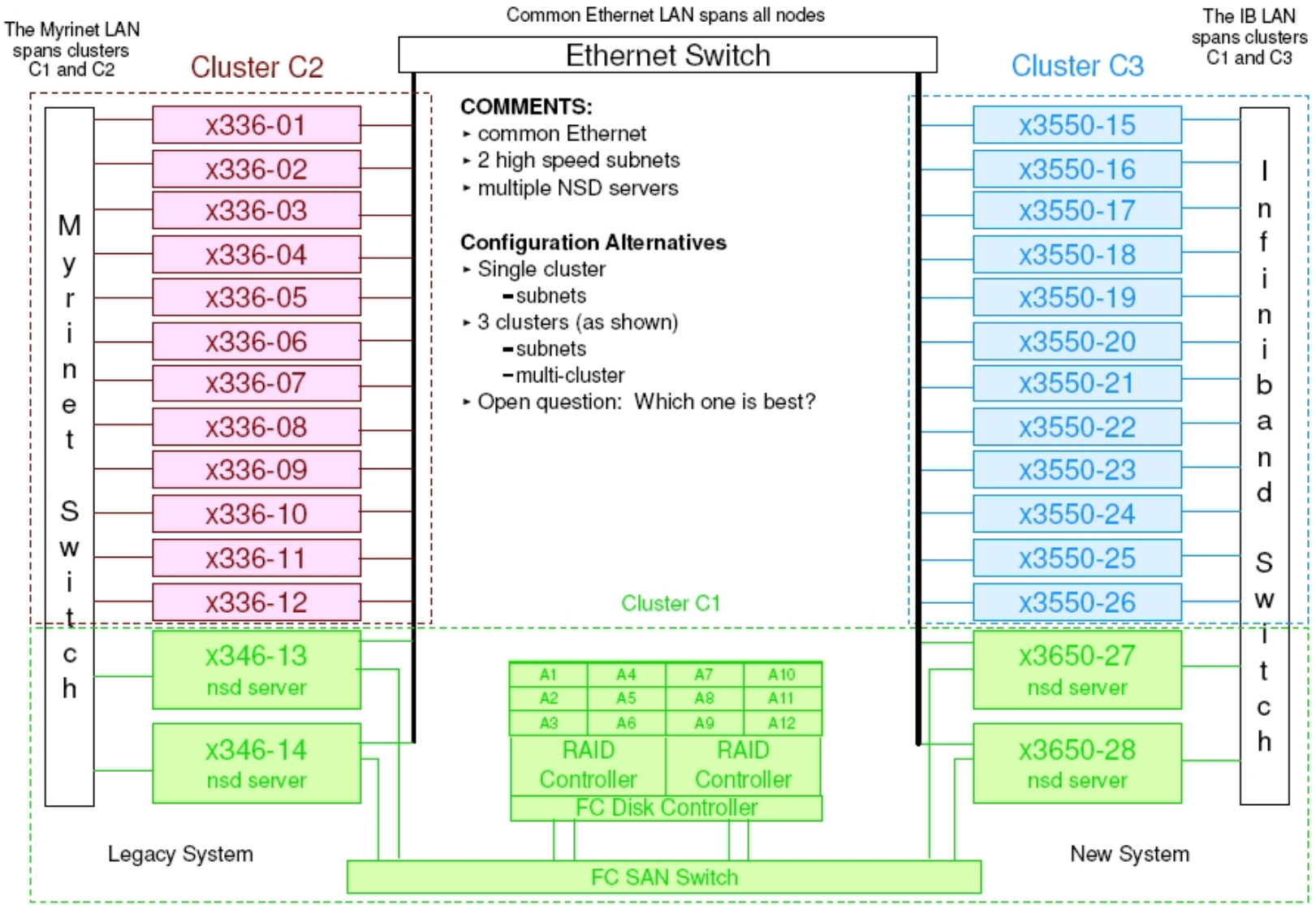
  # mmchconfig subnets="<ip-net1>/cl1 <ip-net2>/cl2"

  questo comando induce ad utlilzzare preferenzialmente la rete <ip-net1> per connettersi ai nodi dei cluster cl1, e <ip-net2> per connettersi ai nodi di cl2
- se si configurano reti IP private per remote cluster, GPFS assume che siano reti disgiunte
- per operare tra cluster remoti su una rete privata connessa, la subnet va configurata specificando i due cluster name assieme:

  # mmchconfig subnets="<ip-priv_net>/cl1;cl2"

Subnet *vs.* Multi-Cluster

Combining Subnets with Multi-Clusters to Support Multiple Fabrics



Modifica delle chiavi SSL

- E' possibile modificare le chiavi SSL di un cluster in modo da non influire sulle connessioni attive tra cluster
  - # mmauth genkey new
    questo genera una nuova coppia di chiavi, mantenendo la vecchia coppia ancora usabile
  - il file **/var/mmfs/ssl/id_rsa.pub** che contiene la parte pubblica della nuova chiave va copiata sul cluster remoto
  - sul cluster remoto va eseguito il comando

    # mmremotecluster update <cname> -k <file>
    se la chiave e' modificata su chi esporta il file system, o

    # mmauth update <cname> -k <file>
    se la chiave e' modificata su chi importa il file system
  - infine si rende la nuova chiave come unica chiave valida con il comando
    # mmauth genkey commit

Da tenere presente...

- Un file system e' amministrato unicamente dal cluster in cui e' stato creato
  - le uniche operazioni eseguibili da un cluster che monta in file system remoto sono
    - accedere ai dati in read/write secondo gli accessi
    - eseguire comandi di **visualizzazione**: mmlsfs, mmlsdisk, mmlsmount, mmdf
- Ogni cluster e' gestito in modo indipendente, quindi modifiche di configurazione non sono propagate automaticamente tra due cluster
  - se il cluster proprietario del file system cancella o rinomina il file system, il cluster remoto non potra' piu' montarlo
    - ci deve essere coordinamento e update delle informazioni in mmauth e mmremotefs
- Update al database dei cluster remoti deve essere eseguito se
  - cambia il nome del cluster remoto
  - vengono rimossi i contact nodes del cluster remoto
  - cambiano le chiavi SSL (visto prima)
- Possono insorgere problemi se il parametro **maxblocksize** dei due cluster non e' uguale
  - in generale non sara' possibile montare un file system remoto con block size maggiore della maxblocksize del cluster che vuole montare
  - si deve modificare il parametro in modo opportuno (**richiede cluster restart**)