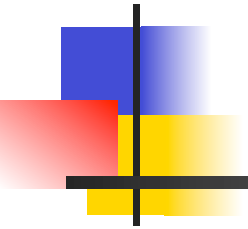
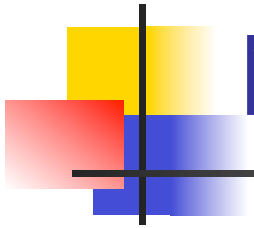


Clustered NFS

Alessandro Brunengo INFN-Genova



NFS

NFS export di file system GPFS

- Un file system GPFS puo' essere esportato via NFS da uno o piu' nodi del cluster
- Gli accessi al file system via NFS e via GPFS coesistono senza problemi
  - NFS basa sul timestamp dei metadati la consistenza della sua cache
    - questo richiede **sincronizzazione tra i nodi** che accedono al file system sia via NFS che via GPFS

Considerazioni per l'export via NFS

- Sul server NFS il file system deve essere montato via GPFS **prima** di essere esportato via NFS
  - il file system deve essere montato **automaticamente** allo startup di GPFS
  - e' utile inserire il comando exportfs –ra nel file (da creare) **/var/mmfs/etc/mmfsup** (script eseguito dopo che GPFS e' partito ed i file system sono stati montati)

Considerazioni per l'export via NFS

- per kernel 2.6 o sup, per esportare file system GPFS, in /etc/exports deve essere specificato il parametro fsid = <num>:

```
/gpfs/dir1      cluster1(rw,fsid=745)
```

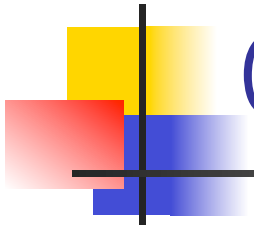
- fsid deve essere unico per ciascun file system
- fsid non deve cambiare nemmeno al reboot
- se la stessa directory e' esportata da piu' NFS server, l'fsid deve essere lo stesso su tutti i nodi che lo esportano
- se si esportano due subtree diversi dello stesso file system GPFS, i due fsid devono essere differenti
  - ATTENTI alla documentazione errata in "*A Guide to the IBM Clustered Network File System*"

Considerazioni per l'export via NFS

- Il nodo che agisce come NFS server puo' avere benefici nell'incremento della cache GPFS
  - maxFileToCache = 1000
  - maxStatCache = 50000
    - attenzione al token manager per le grosse installazioni
- Se il file system GPFS deve essere esportato via NFS v4, il file system deve essere configurato opportunamente:
  - deve supportare il deny-write open lock (-D nfs4)
    - si DEVE usare –D posix per l'export via NFS v3
  - deve supportare gli ACL di tipo NFS v4 (-k nfs4 o –k all)

Considerazioni per l'export via NFS

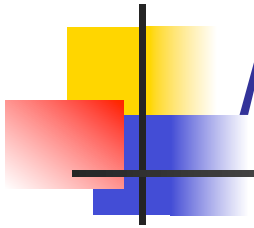
- Si deve valutare l'utilizzo di opzioni che penalizzano le prestazioni
  - write sincroni: -o sync,no_wdelay
  - disabilitare caching dei metadati sull'NFS client: -o noac
- nfsd puo' allocare file, ed impedire l'umount del file system GPFS
  - si deve fare shutdown di nfs prima di eseguire mmumount del file system GPFS sull'NFS server



CNFS

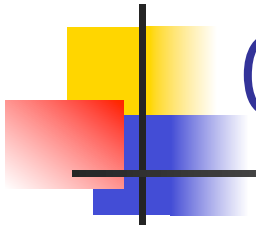
Clustered NFS

- GPFS supporta l'export via NFS realizzato tramite un subset di nodi del cluster che operano in modo da fornire un servizio ad alta affidabilita (Clustered NFS)
  - Il protocollo di trasporto e' NFS (ordinario): non ci sono requisiti sui client che montano il file system
  - I nodi che partecipano all'export sono designati come **CNFS member nodes**
  - Il loro insieme e' indicato come **CNFS cluster**
- I membri del CNFS cluster devono esportare gli stessi volumi (stesso file /etc/exports)
- Attualmente CNFS e' supportato solo su piattaforma Linux (non RHEL6!)



Alta affidabilita'

- In caso di failure (o di shutdown) di un membro del cluster CNFS, l'infrastruttura di cluster GPFS sottostante viene utilizzata per implementare un meccanismo di failover:
  - i nodi del cluster CNFS si accorgono della failure del server
  - un altro nodo del cluster CNFS si fara' carico delle attivita' pendenti del nodo in failure senza interruzione di servizio
  - la gestione del failover e' completamente trasparente ai client



CNFS

- I meccanismi su cui si basa CNFS sono
  - **NFS monitoring**: ogni nodo del cluster CNFS esegue una utility che controlla lo stato di NFS, GPFS e della rete. In caso di malfunzionamenti l'utility innesca la procedura di NFS failover
  - **NFS failover**: il meccanismo di NFS failover e' eseguito tra i nodi del cluster CNFS in aggiunta ai meccanismi del GPFS recovery
    - identifica un nodo del cluster CNFS per il takeover
    - trasferisce il carico NFS del nodo indisponibile sul nodo selezionato
    - il meccanismo di failover si basa sull'*IP address failover*
    - supporta il recovery degli NFS lock
  - **IP address failover**: il cluster CNFS necessita di un set di indirizzi IP (usualmente uno per ogni nodo) in aggiunta a quelli principali
    - l'IP address failover consiste nella migrazione di questi indirizzi tra i nodi del cluster

Network setup per CNFS

- Ad ogni nodo del cluster CNFS deve essere assegnato un indirizzo IP secondario
  - l'indirizzo puo' essere reale o virtuale (alias)
    - se e' un alias, non deve essere configurato
    - altrimenti deve essere configurato staticamente
  - l'interfaccia di rete relativa non deve partire al boot
- L'insieme di tali indirizzi deve essere registrato nel DNS, con un alias
  - l'alias sara' il nome utilizzato dai client per specificare il server NFS da cui montare
  - l'alias permette di realizzare
    - load balancing via DNS
    - una configurazione (lato client) indipendente dai nodi effettivamente configurati nel cluster CNFS

CNFS: failover

- CNFS failover nel dettaglio:
  - l'NFS monitor utility identifica un problema relativo ad NFS
  - l'NFS monitor utility ferma l'NFS serving e ferma (kill) il daemon GPFS.
  - Il cluster GPFS identifica il node failure e, come parte del recovery, tutti i nodi del cluster CNFS entrano in stato di grace period (fermano tutte le richieste di lock dei client)
  - il cluster GPFS completa il recovery rilasciando tutti i lock posseduti dal nodo che e' in failure
  - il cluster CNFS sposta tutti i lock posseduti dal nodo in failure su un nuovo nodo ed invoca l'NFS recovery
  - Il cluster CNFS opera l'IP address takeover (inclusa emissione di gratuitous ARP)
  - i nodi del cluster CNFS notificano ai client di iniziare un lock reclamation, secondo il protocollo NFS
  - alla fine del grace period le operazioni ricominciano normalmente

CNFS: prerequisiti

- linux 2.6 kernel
  - distribuzioni supportate: RHEL 4, 5 e 6, SLES 9 e 10
- OS patches (fino a RHEL5.4 esclusa)
  - patch del ***lockd*** daemon per la propagazione delle informazioni di lock al file system sottostante
  - patch per permettere a ***statd*** di inviare ai client reclaim messages provenienti da un indirizzo IP specifico del server (necessario in conseguenza dell'IP failover)
    - richiede l'utilizzo di opportune versioni di util-linux o nfs-utils
- Verificare sulla documentazione in base alla distribuzione ed alla release

Setup di CNFS (I)

- Definire una directory per i file shared del cluster CNFS
  - # mmchconfig cnfsSharedRoot=*directory*
  - <dir> e' il path ad una directory su file system GPFS che:
    - non sia esportato via NFS (preferibilmente un piccolo file system dedicato)
    - sia montato automaticamente allo startup di GPFS (-A yes)
  - GPFS deve essere down sui nodi ? (documentazione ambigua: pagg. 12 e 148 di *Administration and Programming Reference 3.5*): provate!
- Configurare i file system da esportare in /etc/exports dei nodi del cluster CNFS
  - la configurazione **deve essere uguale per tutti**
- nfsd non deve essere configurato per partire al boot (GPFS si incarica di fare start/stop)

Setup di CNFS (II)

- Aggiungere i nodi al cluster CNFS
  - # mmchnode --cnfs-interface=<nfs_ip> -N <node>
  - <nfs_ip> e' l'indirizzo virtuale (puo' essere piu' d'uno)
  - <node> e' il nome del nodo da aggiungere al cluster CNFS
- Definire parametri di configurazione
  - # mmchconfig cnfsMountdPort=<port> (la porta a cui risponde rpc.mountd)
  - # mmchconfig cnfsNFSDprocs=<nproc> (numero processi NFS, default 32)
  - # mmchconfig cnfsVIP=<aliasDNSname> (inutile)

Failover groups

- E' possibile forzare la selezione del nodo subentrante in funzione del nodo andato in failure
- Questo e' implementato attraverso i CNFS group id:
  - # mmchnode --cnfs-groupid=<nn> -N <node>
  - <nn>: CNFS group ID (numero di due cifre)
  - <node>: il nodo del cluster a cui assegnare il CNFS group id
- Quando un nodo fallisce, la selezione del nodo a cui assegnare il carico viene determinata:
  - uno dei nodi con lo stesso group id
  - uno dei nodi con group id nella stessa decina
    - ad esempio, se fallisce un nodo con group id = 15, si cerca un nodo
      - con group id = 15
      - con group id tra 10 e 19
  - in mancanza di nodi con group id nella stessa decina il failover fallisce
- Per default tutti i nodi hanno CNFS group id = 0 (tutti possono subentrare a tutti)

CNFS management

- La configurazione del cluster CNFS viene visualizzata tramite il comando:

```
# mmlscluster --cnfs

GPFS cluster information
========================
  GPFS cluster name:         grid-ge.grid03
  GPFS cluster id:           13965265603593503612

Cluster NFS global parameters
-----------------------------
  Shared root directory:/mnt/atlas2_mnt/HA
  Virtual IP address:     (undefined)
  rpc.mountd port number:               (undefined)
  nfsd threads:                         32
  Reboot on failure enabled:            yes
  CNFS monitor enabled:                 yes

 Node  Daemon node name    IP address       CNFS state  group  CNFS IP address list
-------------------------------------------------------------------------------
   15   cnfs1.ge.infn.it   193.206.151.4     enabled     0      193.206.151.5
   16   cnfs2.ge.infn.it   193.206.151.6     enabled     0      193.206.151.7
```

CNFS management

- E' possibile abilitare e disabilitare nodi del cluster CNFS:

  # mmchnode --cnfs-enable -N <node-list>
  # mmchnode --cnfs-disable -N <node-list>

- Per rimuovere un nodo dal cluster CNFS:

  # mmchnode --cnfs-interface=DELETE

- Disabilitare un nodo significa escluderlo dai meccanismi di failover
  - l'export ordinario via NFS non e' coinvolto dal comando: il server NFS e le connessioni restano attivi
  - per disabilitare o rimuovere anche l'export NFS senza dare disservizio, si deve prima fermare GPFS sul nodo tramite mmshutdown
    - questo innesca la procedura di failover
- Per la rimozione vale lo stesso discorso
  - il nodo deve anche essere rimosso dal round-robin del DNS