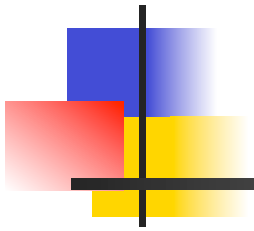
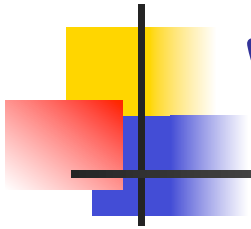


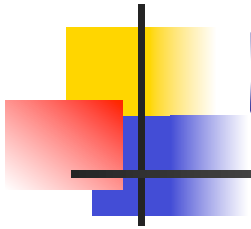
Disk Management

Alessandro Brunengo INFN-Genova



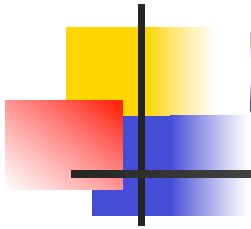
Visualizzare i dischi del cluster

- **mmlsnsd**
  - Visualizza tutti i **dischi inizializzati** (NSD) del cluster
  - Permette di visualizzare le caratteristiche di configurazione degli NSD (**NSD server**, **file system** di appartenenza)
- **mmlsdisk**
  - Visualizza lo **stato corrente** dei dischi appartenenti ad un file system (**status**, **availability**, NSD server **attualmente** usato per l'accesso, **tipo di utilizzo**, **storage pool**)



Disk availability

- Lo stato di un disco e' la combinazione del **disk status** e **disk availability**
- La disk availability segnala se GPFS e' in grado di scrivere o leggere dal disco
  - **up**: condizione di normalita': il disco e' usato per operazioni r/w
  - **down**: il disco non e' utilizzato per operazioni di I/O
    - condizione automatica in conseguenza di errori di I/O ripetuti
    - e' una condizione permanente: si deve usare mmchdisk per modificarla
  - **recovering**: condizione transitoria tra lo stato down e lo stato up: GPFS sta' verificando il contenuto del disco prima di renderlo disponibile; e' permessa la scrittura, non la lettura
  - **unrecovered**: GPFS non ha potuto completare l'operazione di recovering



Disk status

- Il disk status controlla il **data placement** e la migrazione dei dati
  - **ready**: il disco e' in condizioni normali, ed usato per dati e/o metadati secondo la sua configurazione
  - **suspended**: il dati sul disco possono essere letti o aggiornati, ma non viene allocato nuovo spazio
  - **being emptied**: condizione transitoria per un disco in attesa di completare la sua rimozione
  - **replacing**: condizione transitoria per un disco mentre e' in corso la sua sostituzione
  - **replacement**: condizione transitoria per un disco che sostituisce un altro disco
- GPFS alloca spazio solo su dischi **ready** o **replacement**

Modificare l'availability di un disco

- **# mmchdisk <device> {stop|start} -d "<disk-descr>"**
  - **stop**: il disco viene messo "**down**" e non deve piu' essere utilizzato da GPFS
    - avviene anche automaticamente
    - il restart di GPFS non modifica lo stato "**down**"
    - non puo' essere utilizzato nemmeno per **mmfsck**
  - **start**: fa ripartire il disco "**down**"
    - il disco diviene "**recovering**", quindi "**up**" o "**unrecovered**"
    - il disco "**unrecovered**" puo' essere utilizzato per operare un **mmfsck**

Modificare lo status del disco

- **# mmchdisk <device> {suspend|resume} -d "<disk-descr>"**
  - **suspend**: istruisce GPFS di non allocare piu' spazio su questo disco
    - generalmente prima di rimuovere un disco: in questa condizione si puo' migrare via tutti i dati dal disco e rimuoverlo
    - lo status "suspended" **non viene modificato** nemmeno da GPFS restart: **solo manualmente**
  - **resume**: istruisce GPFS di rimettere in stato "ready" un disco precedentemente messo in "suspended"
    - il disco torna ad essere pienamente disponibile (se l'availability lo permette)

Modificare disk usage e failure group

- **mmchdisk** puo' essere utilizzato per modificare **disk usage** e **failure group** di appartenenza del disco:

  # mmchdisk <device> change -d "<disk-descr>"

  <disk-descr> e' una stringa del tipo:

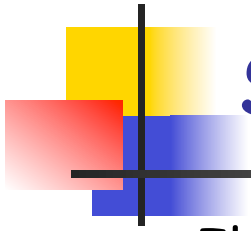
  DiskName:::DiskUsage:FailureGroup:::

  che specifica i parametri desiderati
- Questo comando **non sposta dati**: se si modifica il failure group si deve eseguire un **mmrestripefs** per rimettere a posto le cose

mmrestripefs

- **mmrestripefs** viene utilizzato per migrare dati in funzione dello stato dei dischi e della replica di dati e metadati
  - **b: ribilancia tutti i dati e metadati su dischi non suspended (da eseguire dopo aggiunta/rimozione di dischi dal file system)**
  - **m: migra tutti i dati e metadati che non esistono altrove da dischi suspended e fa restripe**
    - **quelli replicati altrove non li sposta**
  - **r: sposta tutti i dati e metadati da dischi suspended e fa restripe**
    - **ricrea le repliche, ed alla fine il disco e' vuoto**
    - **come -m se non c'e' replica di dati e metadati**
  - **p: rimette a posto la collocazione dei dati di file ill placed (cioe' con blocchi nello storage pool errato)**
  - **R: modifica il replica setting di tutti i file al default del file system e crea o rimuove le repliche secondo la necessita'**
  - **b implica r (o m) e p**
- E' una operazione che genera molto I/O

Aggiunta di un disco al file system



- Il nuovo disco deve essere inizializzato come NSD (mmcrnsd), definendo failure group, disk usage e storage pool di appartenenza
  - i dischi inizializzati ma non ancora assegnati ad un file system possono essere visti tramite il comando **mmlsnsd –F**
- Il nuovo disco puo' essere aggiunto al file system con il comando

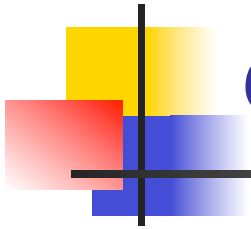
  # mmadddisk Device {"DiskDesc"| -F StanzaFile} [-a] [-r]

  dove DiskDesc e' una stringa del tipo:

  DiskName:::DiskUsage:FailureGroup::StoragePool

  - l'opzione –r richiede che il file system **ribilanci il suo contenuto** su tutti i dischi, compreso il nuovo (operazione I/O intensive)
  - l'opzione –a indica che il comando non deve aspettare la fine del ribilanciamento per ritornare

Considerazioni sulla aggiunta di un disco al file system



- E' opportuno mantenere **omogenea** la dimensione dei dischi all'interno di uno storage pool
- E' opportuno che la stripe size del volume RAID sia comunque un **sottomultiplo** della block size del file system
  - GPFS esegue singole I/O delle dimensioni della block size
  - se la block size non e' multiplo della stripe size avremo una inefficienza RAID level

Rimozione di un disco dal file system

- E' possibile rimuovere a caldo un disco dal file system:

  # mmdeldisk <device> "<disk-name>"

- I dati contenuti sul disco vengono **automaticamente migrati** su altri dischi dello stesso pool
  - usa mmdf per verificare la disponibilita' di spazio
- Si possono usare le opzioni (-b, -m, -r) di **mmrestripefs**
- In caso di fallimento, il disco rimane in stato suspended
  - si puo' rieseguire la rimozione dopo aver sistemato la causa dell'errore
- Se c'e' un disco disponibile, si puo' usare **mmrpldisk** (sostituisce il disco copiando il suo contenuto sull'altro)