Introduzione a GlusterFS

Paolo Veronesi - (INFN CNAF)

Agenda

- Perchè un file system distribuito
- Descrizione di GlusterFS

Alta Affidabilita' di specifiche aree dati

- Con particolare riguardo alla gestione di una infrastruttura basata su openstack, si evidenzia la necessità di dotare tale infrastruttura di specifiche aree dati altamente affidabili e che permettano un certo numero di operazioni da parte dei gestori dell'infrastuttura.
- Per esempio, per quel che riguarda il servizio di provisioning di VM, è molto utile configurare l'infrastruttura per permettere la live migration di VM tra compute node.
  - Si definisce come "*live migration*" l'operazione attraverso la quale è possibile migrare una macchina virtuale dal server fisico sul quale è in esecuzione (sorgente) verso un server fisico differente (destinazione). Per permettere questa operazione, è necessario che l'area dati che ospita le macchine virtuali sia condivisa tra il server sorgente e il server destinazione. Questa area dati è quindi caratterizzata da alto I/O e dal fatto che deve essere condivisa tra più server.
-

GlusterFS (1/2)

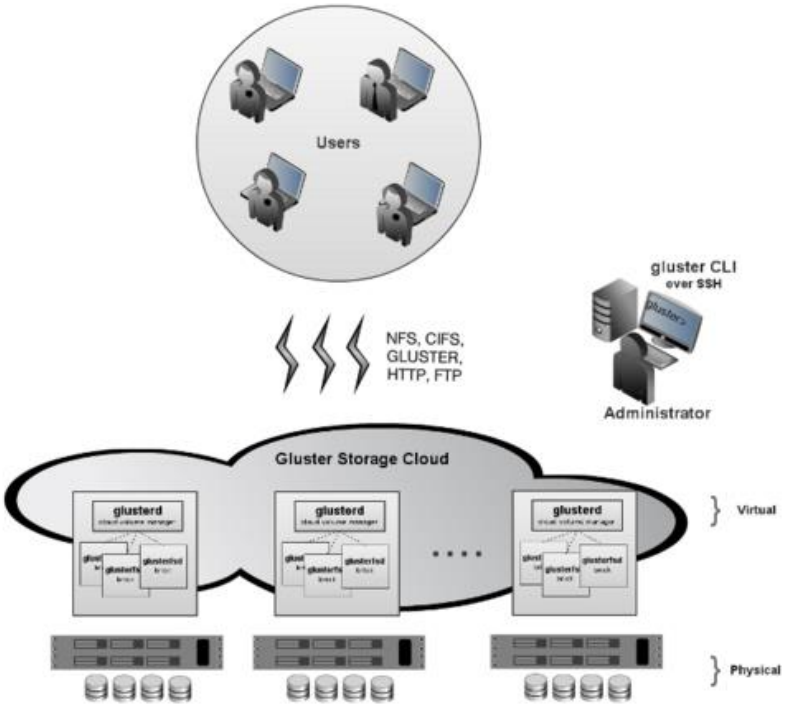
- **GlusterFS** è un **file system distribuito** e scalabile orizzontalmente, la cui capacità può essere dinamicamente espansa mediante l'aggiunta di nuovi nodi. È un prodotto open source .
  - Un file system LOCALE rende disponibile I file (cioe' spazio logico contiguo) ai processi del sistema stesso
  - Un file system DISTRIBUITO e' una architettura che rende disponibili I file a sistemi remoti e ne rende possibile la condivisione tra piu' applicazioni che operano su sistemi diversi
  - SAN (Storage Area Network):
    - I device fisici sono collegati as un sistema dedicato che serve dei dispositivi logici chiamati LUN risultato di aggregazione o partizione dei device fisici
    - Tipicamente le SAN offrono dei servizi aggiuntivi per la gestione del Mass Storage permettendo sulle singole LUN o su interi RAID group operazioni di:
      - Cloning
      - Mirroring
      - Snapshot
      - Backup/restore
      - Disaster recovery
- GlusterFS è in grado di arrivare a gestire fino a diversi Petabyte, migliaia di client e diverse aree dati (storage) organizzandole in blocchi che rende accessibili su Infiniband RDMA (remote direct memory access, fibra ottica) o attraverso connessioni TCP/IP.

GlusterFS (2/2)

Le risorse vengono rese disponibili sotto un unico punto di condivisione; tali risorse possono essere montate dai client mediante tre diversi protocolli:

- CIFS;
- NFS;
- il client nativo Gluster.
- **Libgfapi**

Nel file system Gluster, il termine con cui si identifica una risorsa condivisa è ***volume***, ossia un insieme logico di blocchi (***bricks***), dove per blocco si intende una directory esportata da un server compreso nel pool degli storage fidati (***trusted storage pool***).

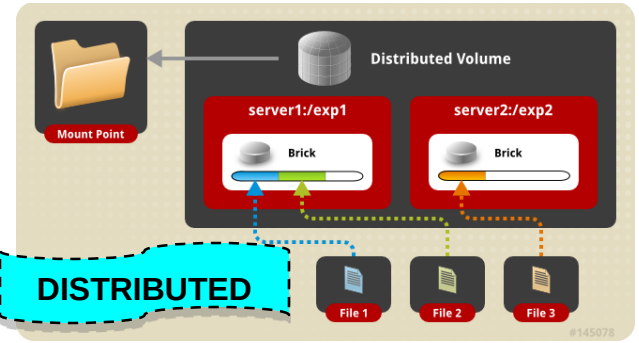


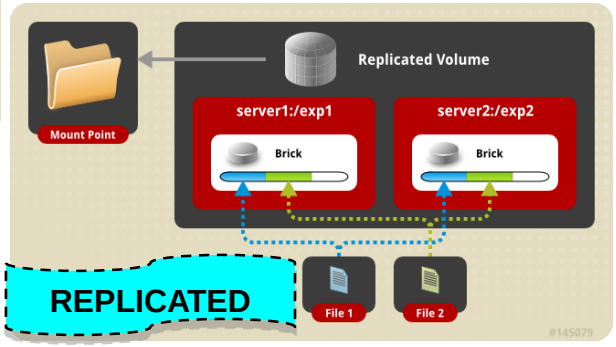
- Esistono diverse tipologie di volume, e principali sono le seguenti:
- ***Distributed***: distribuisce i file all'interno dei brick del volume;
- ***Replicated***: replica i file nei brick del volume;
- ***Striped***: blocchi di dati (stripes) vengono registrati nei brick del volume;
- ***Distributed striped***: distribuisce i file nei blocchi di dati (stripes) presenti nei brick del volume;
- ***Distributed replicated***: distribuisce i file nelle repliche presenti nei brick del volume;

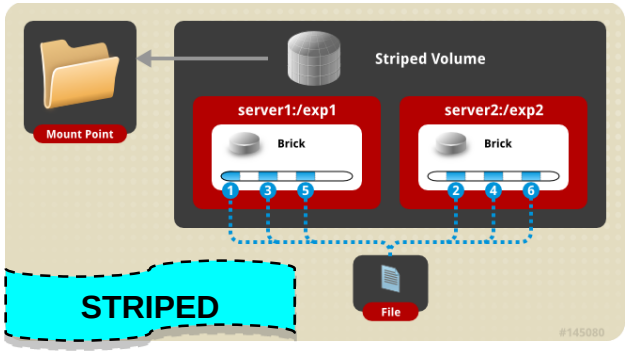
Volumi in GlusterFS (1/2)

Le diverse tipologie di volume configurabili sono le seguenti:

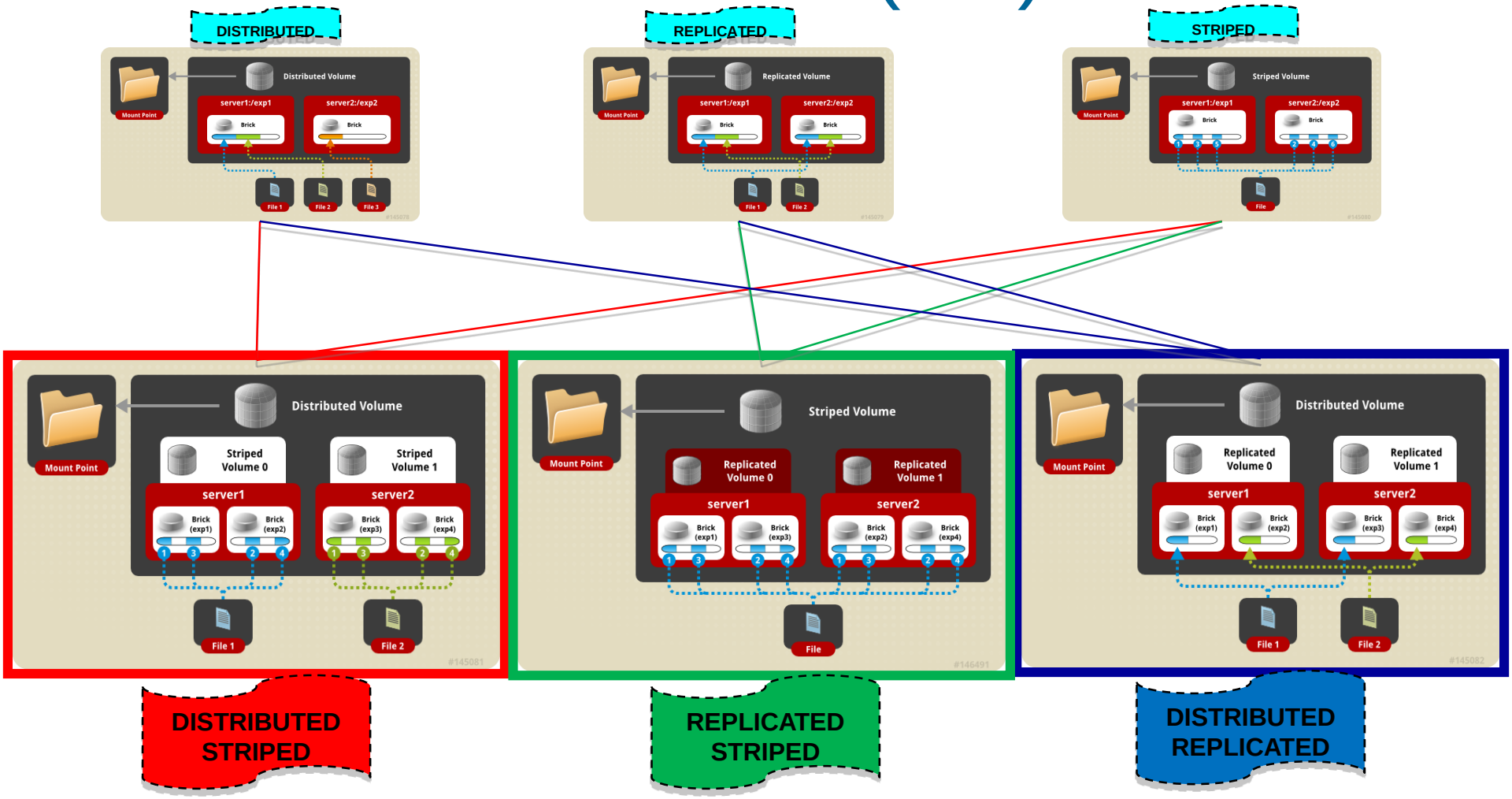
- *Distributed*: distribuisce i file all'interno dei brick del volume;
- *Replicated*: replica i file nei brick del volume;
- *Striped*: blocchi di dati (stripes) vengono registrati nei brick del volume;







Volumi GlusterFS (2/2)



Distributed striped: distribuisce i file nei blocchi di dati (stripes) presenti nei brick del volume;
Distributed replicated: distribuisce i file nelle repliche presenti nei brick del volume;
Replicated Striped: *replica i blocchi di dati (stripes) nei brick*

GlusterFS con Qemu, Nova, Cinder

- **Libgfapi** is a POSIX-like C library shipped along with GlusterFS, which allows to access Gluster's volumes without passing through its FUSE client. This integration brings in some benefits but, the most relevant ones are:
  - Performance improvements by removing FUSE's overhead.
  - Reduce the number of steps required to get to GlusterFS
- There's no special configuration needed to use this, as long as you have QEMU >=1.3 and GlusterFS>=3.4 you should be fine. This is an example of what you can do:

- `qemu-img create gluster://$GLUSTER_HOST/$GLUSTER_VOLUME/images 5G`

GlusterFS con Qemu, Nova, Cinder

- # NOVA COMPUTE
- # Libvirt handlers for remote volumes.
- `volume_drivers=nova.virt.libvirt.volume.LibvirtGlusterfsVolumeDriver`
- # Directory where the glusterfs volume is mounted on the compute node
- `glusterfs_mount_point_base=$state_path/mnt`
- # CINDER
- `volume_driver=cinder.volume.drivers.glusterfs.GlusterfsDriver`
- `glusterfs_shares_config=/etc/cinder/shares.conf`
  - `xx.yy.zz.aa:/cindervol1 -o backupvolfile-server=xx.yy.zz.bb`