GlusterFS

A Scale-out Software Defined Storage

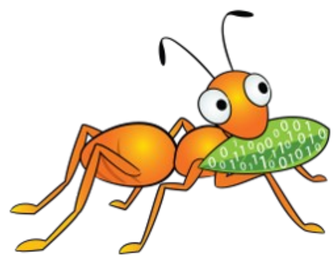
Rajesh Joseph
Poornima Gurusiddaiah

usenix LISA15

November 8–13, 2015 | Washington, D.C.
www.usenix.org/lisa15 #lisa15

Note

- This holds good for 3.7 version of GlusterFS, other version might have variations
- Commands shown here work on CentOS, other distributions might have different command or options
- At the right corner of the slides, there is a link to the live demo



GlusterFS Installation

- Installation via Repo
  - Download latest repo file from download.gluster.org

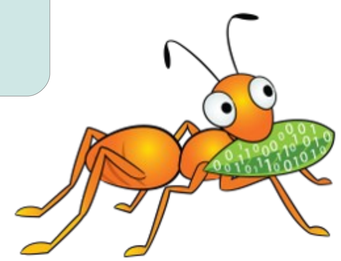
```
wget -P /etc/yum.repos.d
http://download.gluster.org/pub/gluster/glusterfs/LATEST/CentOS/epel-7/x86_64/
```

  - Install GlusterFS

```
yum install glusterfs-server
```

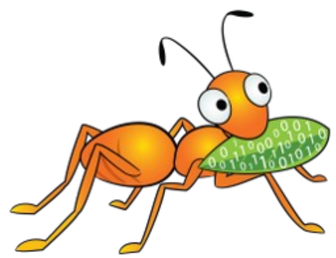
- Installation via RPM
  - Download latest gluster RPMs from download.gluster.org

```
http://download.gluster.org/pub/gluster/glusterfs/LATEST/CentOS/epel-7/x86_64/
```

GlusterFS Packages

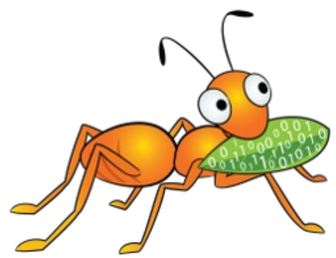
- **GlusterFS Server Packages**
  - glusterfs
  - glusterfs-server
  - glusterfs-api
  - glusterfs-cli
  - glusterfs-libs
- **GlusterFS Client Packages**
  - glusterfs
  - glusterfs-client-xlators
  - glusterfs-libs
  - glusterfs-fuse
- **GlusterFS Feature Packages**
  - glusterfs-extra-xlators
  - glusterfs-ganesha
  - glusterfs-geo-replication
  - glusterfs-rdma
- **GlusterFS Devel Packages**
  - glusterfs-debuginfo
  - glusterfs-devel
  - glusterfs-api-devel



Ports used by GlusterFS

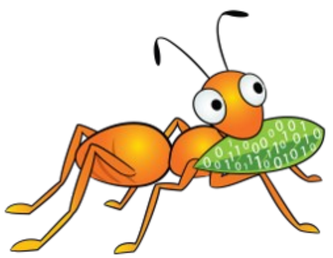
- UDP Ports
  - 111 – RPC
  - 963 – NFS lock manager (NLM)

- TCP Ports
  - 22 – For sshd used by geo-replication
  - 111 – RPC
  - 139 – netbios service
  - 445 – CIFS protocol
  - 965 – NLM



Ports used by GlusterFS

- TCP Ports
  - 2049 – NFS exports
  - 4379 – CTDB
  - 24007 – GlusterFS Daemon (Management)
  - 24008 – GlusterFS Daemon (RDMA port for Management)
  - 24009 – Each brick of every volume on the node (GlusterFS version < 3.4)
  - 49152 – Each brick of every volume on the node (GlusterFS version >= 3.4)
  - 38465-38467 – GlusterFS NFS service
  - 38468 – NFS Lock Manager (NLM)
  - 38469 – NFS ACL Support



Starting Gluster Server

- Gluster server/service can be started by the following command

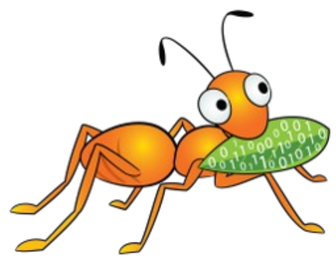
```
# systemctl start glusterd
```

- Gluster server should be started on all the nodes
- To automatically start GlusterFS on node start use chkconfig command

```
# systemctl enable glusterd
```

or

```
# chkconfig glusterd on
```

Setting up Trusted Storage Pool

- Use gluster peer probe command to include a new Node to the Trusted Storage Pool

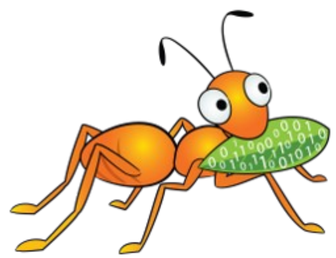
```
# gluster peer probe <Node IP/Hostname of new Node>
```

- Removing Node from the Trusted Storage Pool

```
# gluster peer detach <Node IP/Hostname>
```

- Verify the peer probe/detach succeeded by executing the following command on all the nodes

```
# gluster peer status
```

Creating Bricks

- Create thinly provisioned volume (dm-thin)
  - Create Physical Volume (PV)

```
# pvcreate /dev/sdb
```

  - Create Volume Group (VG) from the PV

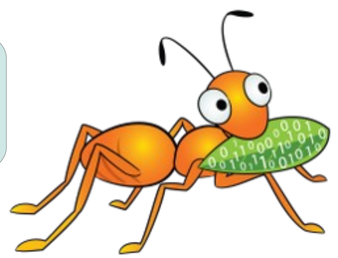
```
# vgcreate vgname1 /dev/sdb
```

  - Create Thin Pool

```
# lvcreate -L 2T –poolmetadatasize 16G -T vgname1/thinpoolname1
```

  - Create Thinly provisioned Logical Volume (LV)

```
# lvcreate -V 1T -T vgname1/thinpoolname1 -n lvname1
```

Creating Bricks

- Create

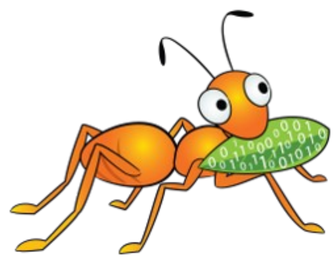
```
# mkfs.xfs -i size=512 /dev/mapper/vgname1-lvname1
```

- Mount

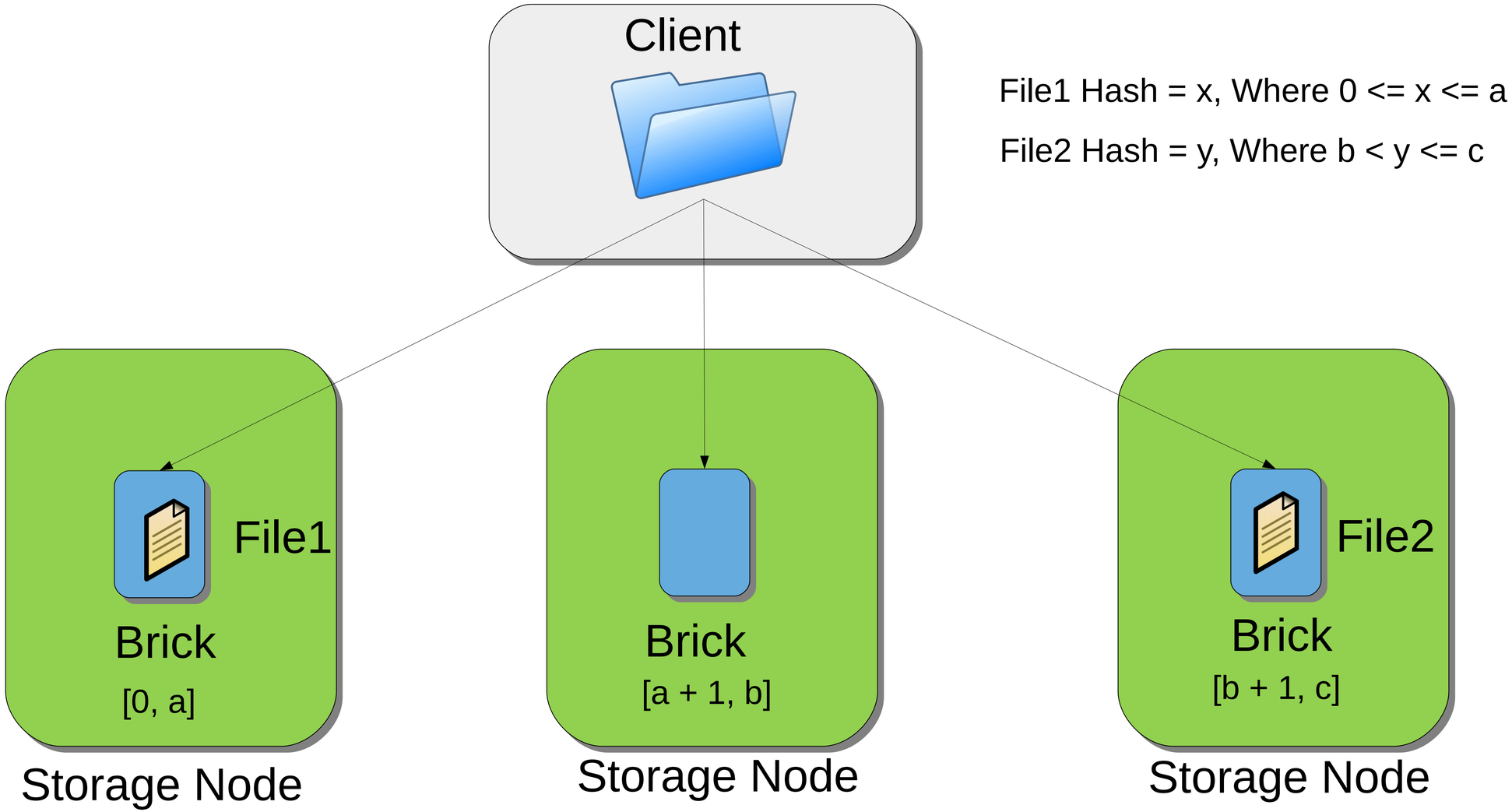
```
# mount  /dev/mapper/vgname1-lvname1 /mnt/brick1
```

- And use it

```
# mkdir /mnt/brick1/data
```

Distribute Volume



Creating Volumes - Distribute

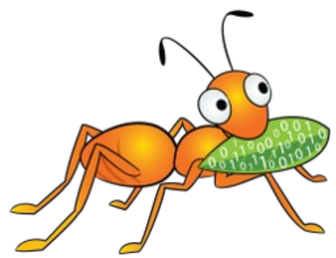
Demo

- Distributed volumes distributes files throughout the bricks in the volume

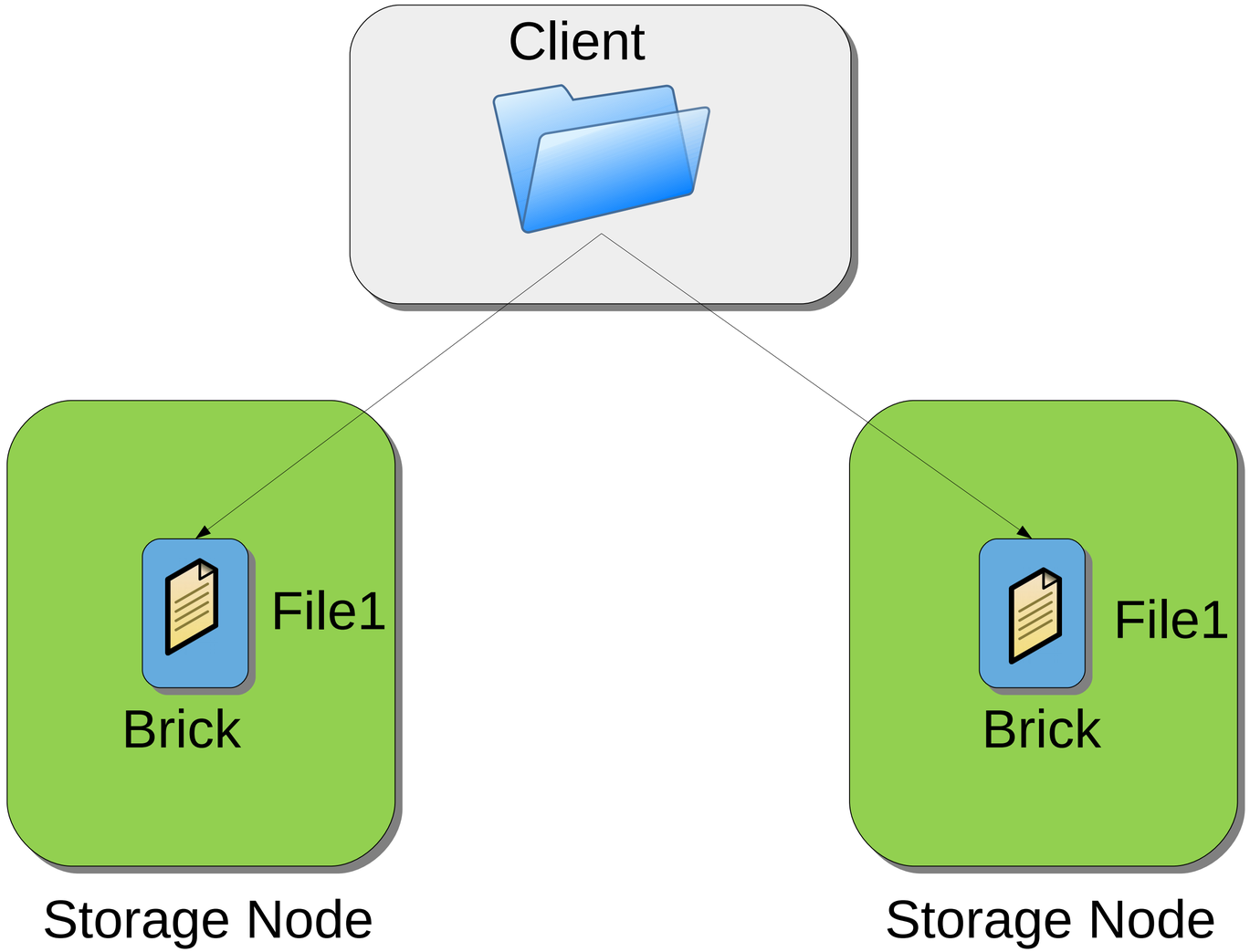
```
# gluster volume create <volume name> [transport <tcp|rdma|
tcp,rdma>] <Node IP/hostname>:<brick path>.... [force]

e.g.
# gluster volume create dist_vol host1:/mnt/brick1/data
host2:/mnt/brick1/data
```

- Its advised to provide a nested directory in the brick mount point as the brick directory
- If transport type is not specified 'tcp' is used as default



Replicate Volume



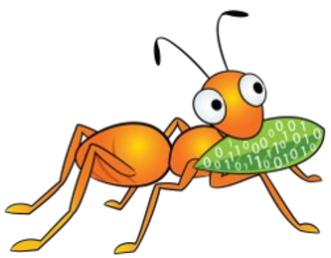
Creating Volumes - Replicate

- Replicated volumes provides file replication across n (replica) bricks

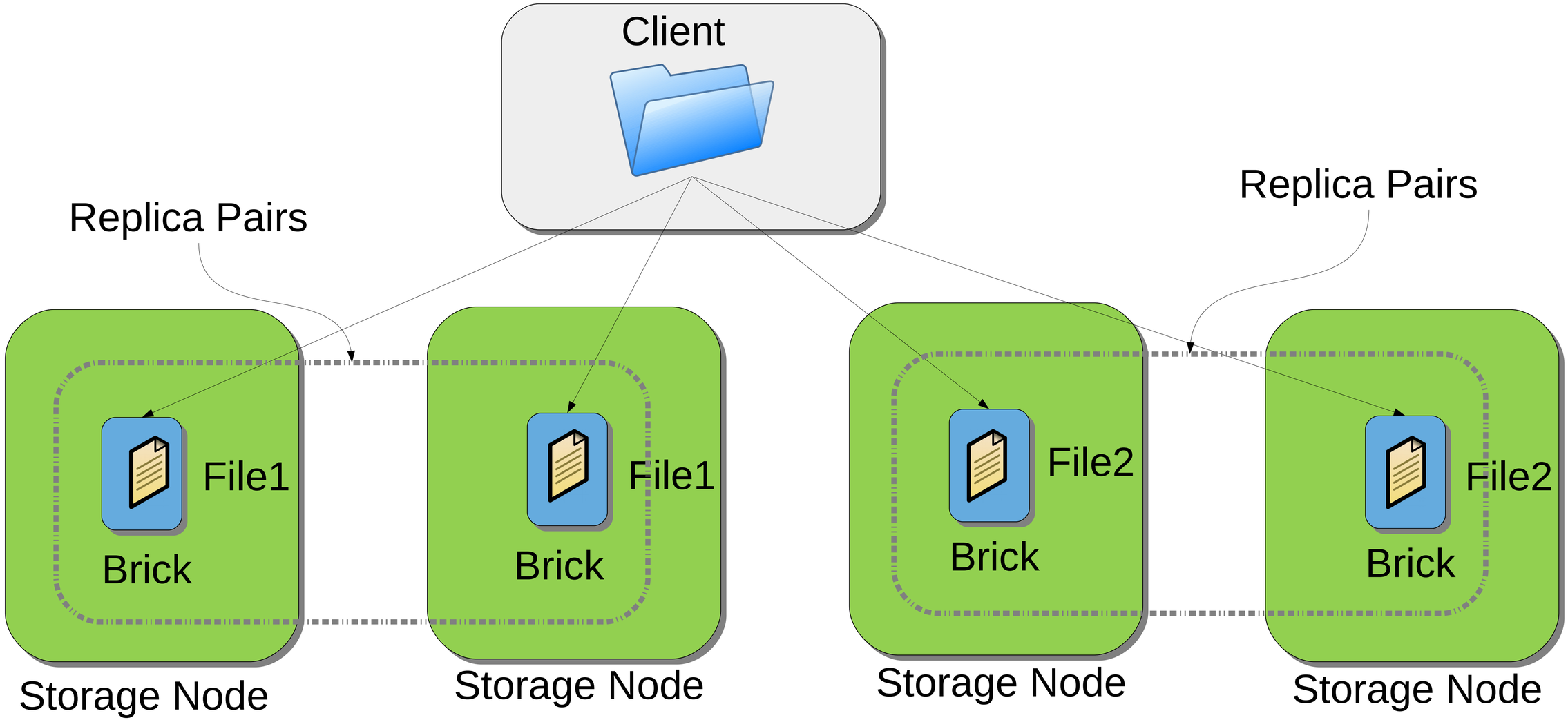
```
# gluster volume create <volume name> [replica <COUNT>] [transport
<tcp|rdma|tcp,rdma>] <Node IP/hostname>:<brick path>.... [force]

e.g.
# gluster volume create repl_vol replica 3 host1:/mnt/brick1/data
host2:/mnt/brick1/data host3:/mnt/brick1/data
```

- Number of bricks must be a multiple of the replica count
- It is advised to have bricks in different servers
- The replication is synchronous in nature, hence it is not advised to combine a brick in different geo location as it may reduce the performance drastically



Distribute Replicate Volume



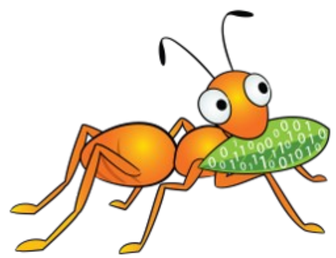
Creating Volumes – Distribute Replicate

- Distributed replicated volumes distributes files across replicated bricks in the volume

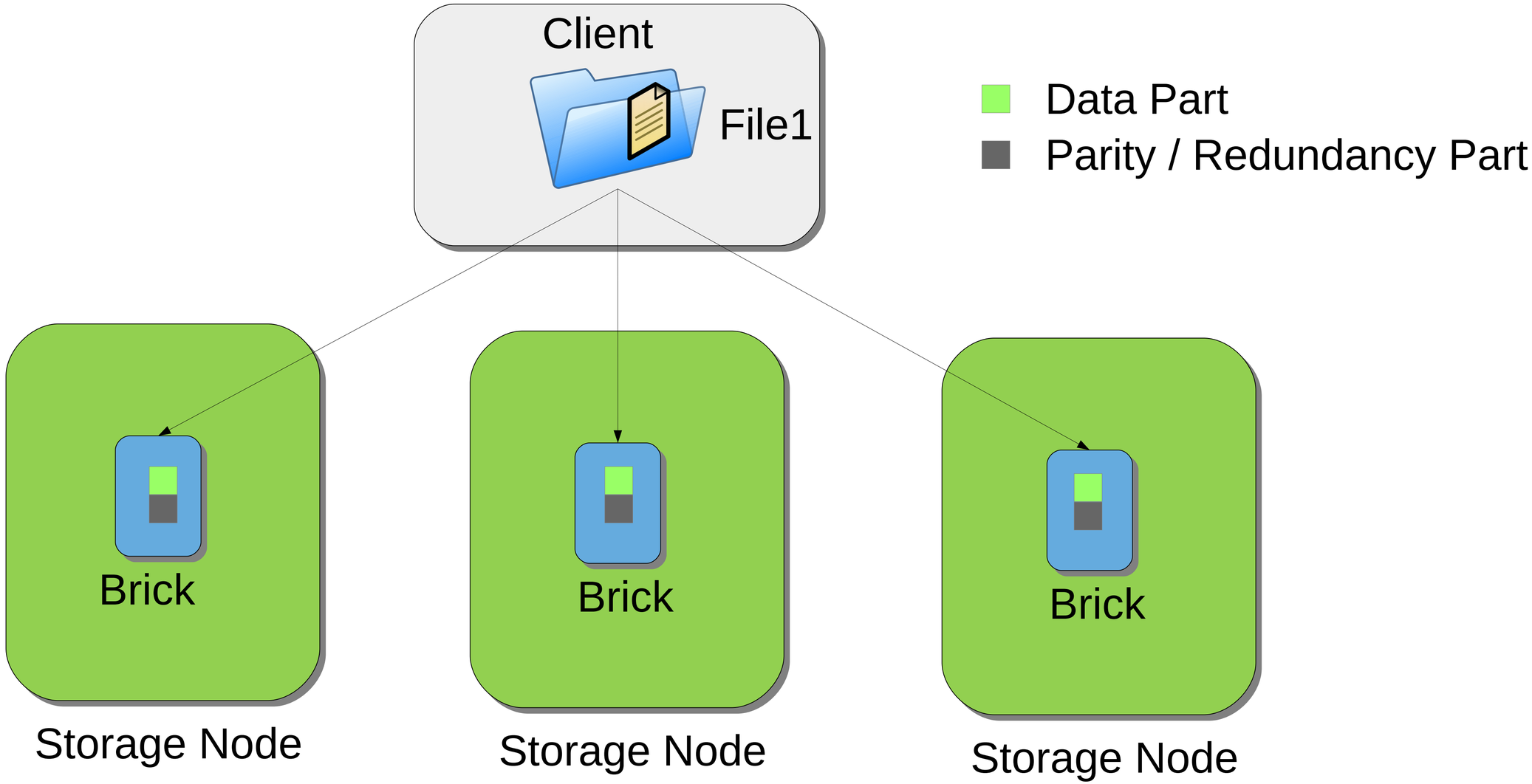
```
# gluster volume create <volume name> [replica <COUNT>] [transport
<tcp|rdma|tcp,rdma>] <Node IP/hostname>:<brick path>.... [force]

e.g.
# gluster volume create repl_vol replica 3 host1:/mnt/brick1/data
host2:/mnt/brick1/data host3:/mnt/brick1/data host1:/mnt/brick2/data
host2:/mnt/brick2/data host3:/mnt/brick2/data
```

- Number of bricks must be a multiple of the replica count.
- Brick order decides replica set and distribution set



Disperse Volume



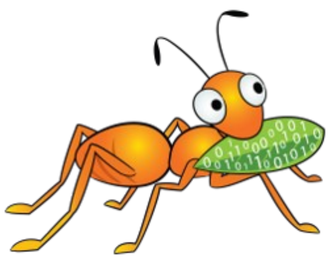
Creating Volumes – Disperse

- Dispersed volumes are based on erasure codes, providing space-efficient protection against disk or server failures

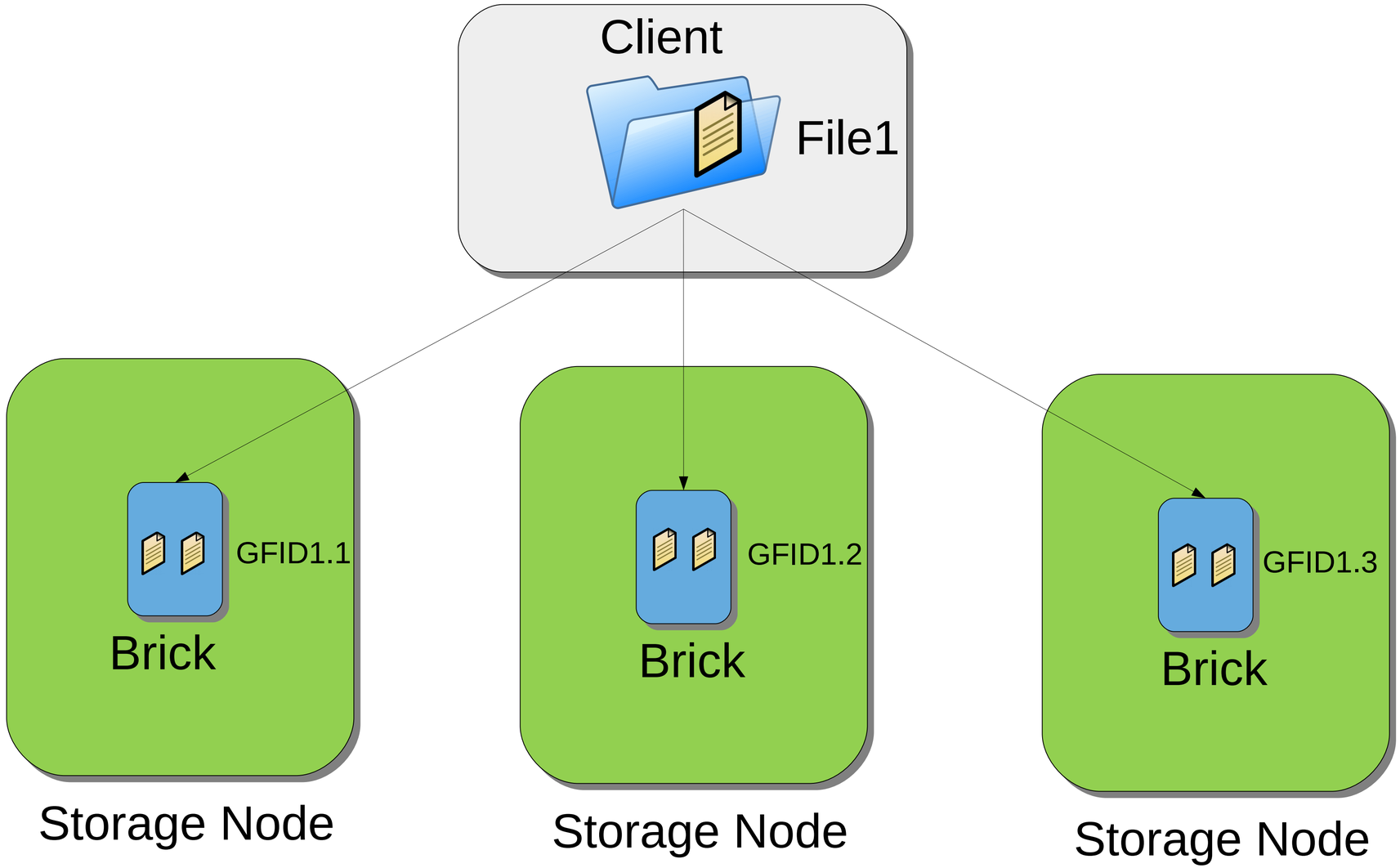
```
# gluster volume create <volume name> [disperse COUNT] [disperse-
data COUNT] [redundancy COUNT] [transport tcp|rdma|tcp,rdma] <Node
IP/hostname>:<brick path>.... [force]
```

- The data protection offered by erasure coding can be represented as n = k + m
  - n = total number of bricks, disperse count
  - k = total number of data bricks, disperse-data count
  - m = number of brick failure that can be tolerated, redundancy count
- Any two counts need to be specified while creating volume

Eg: 6 = 4 + 2 i.e. a 10MB file is split into 6 2.5MB chunks and stored in all 6 bricks(=15MB) but can withstand failure of 2 bricks



Sharded Volume

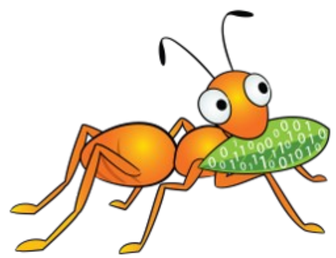


Creating Volumes – Sharded

- Sharded volume is similar to striped volume
- Unlike other volume types shard is a volume option which can be set on any volume

```
# gluster volume set <volume name> features.shard on
```

- To disable sharding it is advisable to create a new volume without sharding and copy out contents of this volume into the new volume
- This feature is disabled by default, and is beta in 3.7.4 release

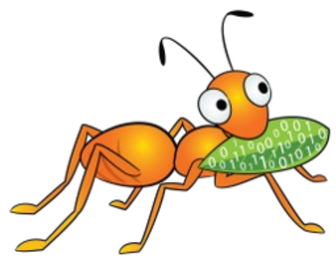


Starting Volumes

- Volumes must be started before they can be mounted
- Use the following command to start volume

```
# gluster volume start <volname>

e.g.
# gluster volume start dist_vol
```

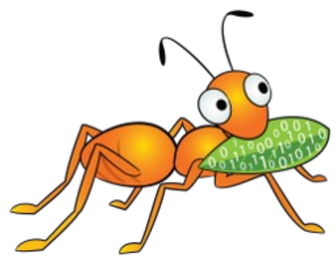
Configuring Volume Options

- Current volume options

```
# gluster volume info
```

- Volume options can be configured using the following command

```
# gluster volume set <volname> <option> <value>
```

Expanding Volume

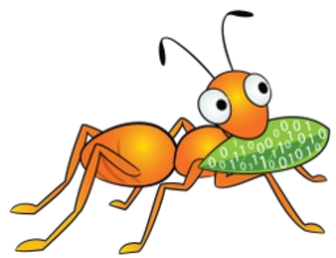
- Volume can be expanded when the cluster is online and available
- Add Node to the Trusted Storage Pool

```
# gluster peer probe <IP/hostname>
```

- Add bricks to the volume

```
# gluster volume add-brick <VOLNAME> <Node IP/hostname>:<brick path>....
```

- In case of replicate, the bricks count should be multiple of replica count



Expanding Volume

- To change the replica count, following command needs to be executed

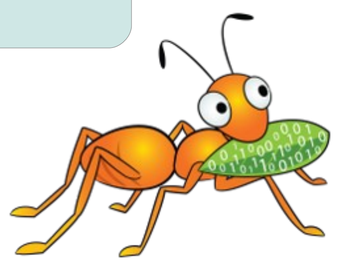
```
# gluster volume add-brick replica <new count> <VOLNAME> <Node IP/hostname>:<brick path>...
```

- Number of replica bricks to be added must be equal to the number of distribute sub-volumes
  - e.g change replica 2 distribute 3, to replica 3 distribute 3 for volume dist-repl

```
# gluster volume dist-repl replica 3 host1:/brick1/brick1 host2:/brick1/brick1
host3:/brick1/brick1
```

- Rebalance the bricks

```
# gluster volume rebalance <volname> <start | status | stop>
```



Shrinking Volume

- Remove a brick using the following command

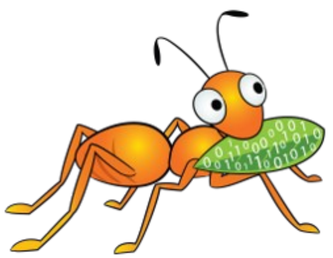
```
# gluster volume remove-brick <volname> BRICK start [force]
```

- You can view the status of the remove brick operation using the following command

```
# gluster volume remove-brick <volname> BRICK status
```

- After status shows complete run the following command to remove brick

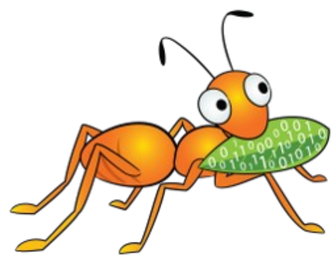
```
# gluster volume remove-brick <volname> BRICK commit
```

Volume Self Healing

- In Replicate volume when an offline bricks comes online the updates on the online brick needs to be synced to this brick – Self Healing
- File is healed by
  - Self-Heal daemon (SHD)
  - On-access
  - On-demand
- SHD automatically initiates heal every 10 minutes

```
# gluster volume set <volname> cluster.self-heal-deamon <on | off>
```



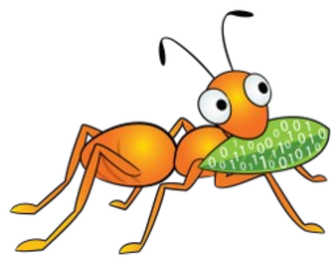
Volume Self Healing

- On-demand healing can by done by

```
# gluster volume heal <volname>
# gluster volume heal <volname> full
# gluster volume heal <volname> info
```

- To enable/disable healing when file is accessed from the mount point

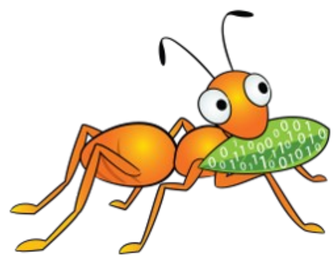
```
# gluster volume set <volname> cluster.data-self-heal off
# gluster volume heal <volname> cluster.entry-self-heal off
# gluster volume heal <volname> cluster.metadata-self-heal off
```

Volume Self Healing

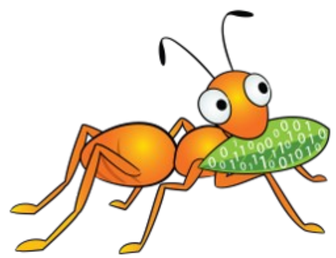
- In Replicate volume when an offline bricks comes online the updates on the online brick needs to be synced to this brick – Self Healing
- File is healed by
  - Self-Heal daemon (SHD)
  - On-access
  - On-demand
- SHD automatically initiates heal every 10 minutes

```
# gluster volume set <volname> cluster.self-heal-deamon <on | off>
```



Accessing Data

- Volume can be mounted on local file-system
- Following protocols supported for accessing volume
  - GlusterFS Native client
    - Filesystem in Userspace (FUSE)
  - NFS
    - NFS Ganesha
    - Gluster NFSv3
  - SMB / CIFS



GlusterFS Native Client

Demo

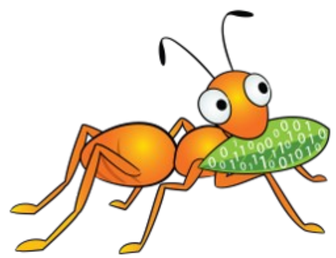
- Client machines should install GlusterFS client packages
- Mount the started GlusterFS volume

```
# mount -t glusterfs host1:/dist-vol /mnt/glusterfs
```

- Use any Node from Trusted Storage Pool to mount
- Use /etc/fstab for automatic mount

```
e.g. to mount dist-vol append following to /etc/fstab

host1:/dist-vol /mnt/glusterfs glusterfs defaults,_netdev,transport=tcp 0 0
```

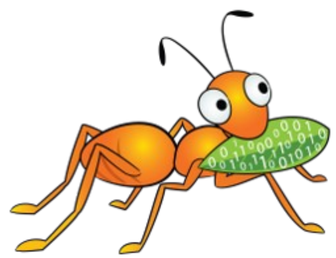



NFS Client

- Install NFS client packages
- Mount the started GlusterFS volume via NFS

```
# mount -t nfs -o vers=3 host1:/dist-vol /mnt/glusterfs
```

- Gluster NFS supports only version 3
- Use /etc/fstab for automatic mount

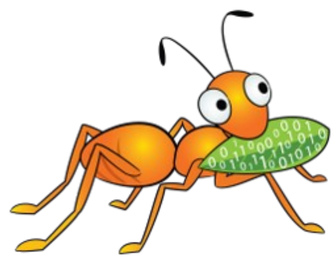


SMB Client

Demo

- For high availability and lock synchronization SMB uses CTDB
- Install CTDB and GlusterFS Samba packages
- GlusterFS Samba pacakges can be downloaded from

```
http://download.gluster.org/pub/gluster/glusterfs/samba/
```

CTDB Setup

Demo

- Create n-way replicated CTDB volume
  - n – Number of nodes that will be used as samba server

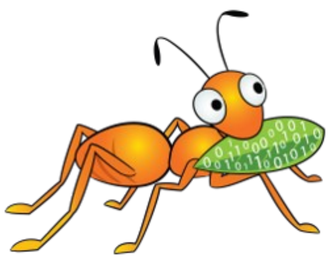
```
# gluster volume create ctdb replica 4 host1:/mnt/brick1/ctdb
host2:/mnt/brick1/ctdb host3:/mnt/brick1/ctdb host4:/mnt/brick1/ctdb
```

- Replace META=all to META=ctdb in the below files on all the nodes

```
/var/lib/glusterd/hooks/1/start/post/S29CTDBsetup.sh
/var/lib/glusterd/hooks/1/stop/pre/S29CTDB-teardown.sh
```

- Start the ctdb volume

```
# gluster volume start ctdb
```



CTDB Setup

Demo

- On volume start following entries are created in /etc/samba/smb.conf

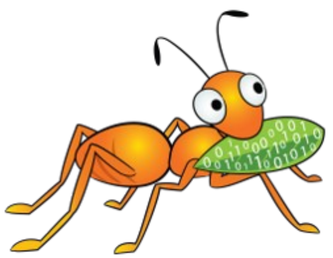
```
clustering = yes
idmap backend = tdb2
```

- CTDB configuration files are stored on all the nodes used as Samba server

```
/etc/sysconfig/ctdb
```

- Create /etc/ctdb/nodes file on all the nodes that is used by Samba server

```
192.168.8.100
192.168.8.101
192.168.8.102
192.168.8.103
```

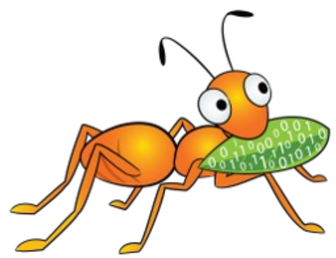
CTDB Setup

Demo

- For IP failover create /etc/ctdb/public_addresses file on all the nodes
- Add virtual IPs that CTDB should create in this file

```
<Virtual IP>/<routing prefix><node interface>

e.g.
192.168.1.20/24 eth0
192.168.1.21/24 eth0
```

Sharing Volumes over Samba

Demo

- Set following options to gluster volume

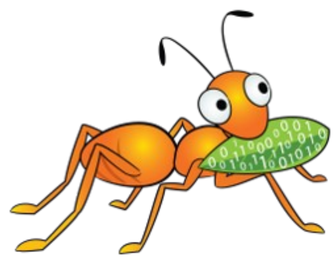
```
# gluster volume set <volname> stat-prefetch off
# gluster volume set <volname> server.allow-insecure on
```

- Edit /etc/glusterfs/glusterd.vol in each Node and add the following

```
option rpc-auth-allow-insecure on
```

- Restart glusterd service on each Node
- Set following options to gluster volume

```
# gluster volume set <volname> storage.batch-fsync-delay-usec 0
```

Sharing Volumes over Samba

Demo

- On GlusterFS volume start following entry will be added to /etc/samba/smb.conf

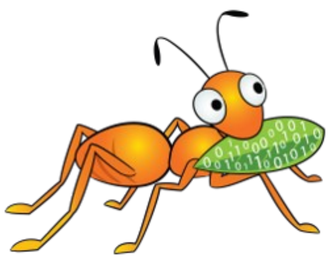
```
[gluster-VOLNAME]
comment = For samba share of volume VOLNAME
vfs objects = glusterfs
glusterfs:volume = VOLNAME
glusterfs:logfile = /var/log/samba/VOLNAME.log
glusterfs:loglevel = 7
path = /
read only = no
guest ok = yes
```

- Start SMBD

```
# systemctl start smb
```

- Specify the SMB password. This password is used during the SMB mount

```
# smbpasswd -a username
```



Mounting Volumes using SMB

Demo

- Mount from Windows system

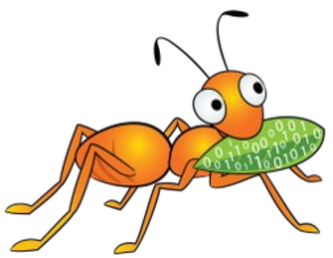
```
# net use <drive letter> \\<virtual IP>\gluster-VOLNAME

e.g.
# net use Z: \\192.168.1.20\gluster-dist-vol
```

- Mount from Linux system

```
# mount -t cifs \\<virtual IP>\gluster-VOLNAME /mnt/cifs

e.g.
# mount -t cifs \\192.168.1.20\gluster-dist-vol /mnt/cifs
```

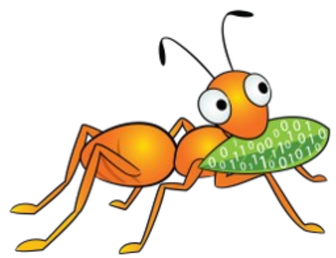



Troubleshooting

- Log files
  - Following command will give you log file location

```
# gluster –print-logdir
```

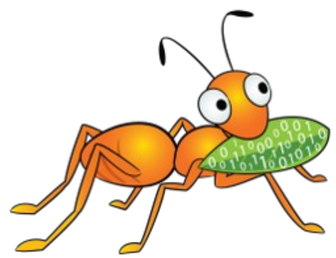
  - Log dir will contain logs for each GlusterFS process
    - glusterd - `/var/log/glusterfs/etc-glusterfs-glusterd.vol.log`
    - Bricks - `/var/log/glusterfs/bricks/<path extraction of brick path>.log`
    - Cli - `/var/log/glusterfs/cmd_history.log`
    - Rebalance - `/var/log/glusterfs/VOLNAME-rebalance.log`
    - Self-Heal Daemon (SHD) - `/var/log/glusterfs/glustershd.log`
    - Quota - `/var/log/glusterfs/quotad.log`



Troubleshooting

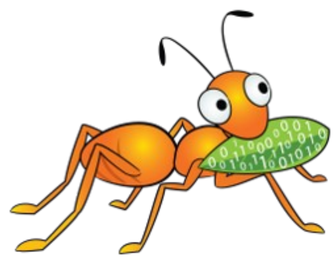
- Log files
  - Log dir will contain logs for each GlusterFS process
    - NFS - `/var/log/glusterfs/nfs.log`
    - Samba - `/var/log/samba/glusterfs-VOLNAME-<ClientIP>.log`
    - NFS-Ganesha - `/var/log/nfs-ganesha.log`
    - Fuse Mount - `/var/log/glusterfs/<mountpoint path extraction>.log`
    - Geo-replication - `/var/log/glusterfs/geo-replication/<master>`
- Volume status

```
# gluster volume status [volname]
```

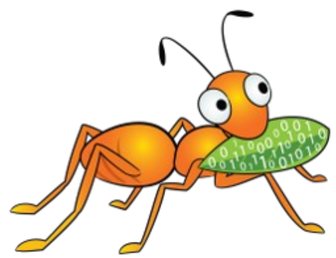
Troubleshooting

- Connectivity issues
  - Check network connectivity
  - Check all necessary GlusterFS processes are running
  - Check Firewall rules



Troubleshooting – Split Brain

- Is a scenario where in a replicate volume GlusterFS is not in a position to determine the correct copy of file
- Three different types of split-brain
  - Data split-brain
  - Metadata split-brain
  - Entry split-brain
- The only way to resolve split-brains is by manually inspecting the file contents from the backend and deciding which is the true copy



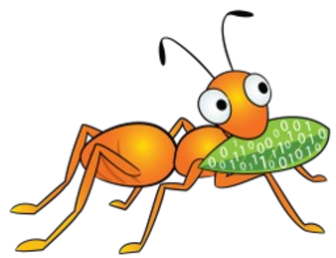
Troubleshooting – Preventing Split Brain

- Configuring Server-Side Quorum
  - Number of server failures that the trusted storage pool can sustain
  - Server quorum can be by volume option

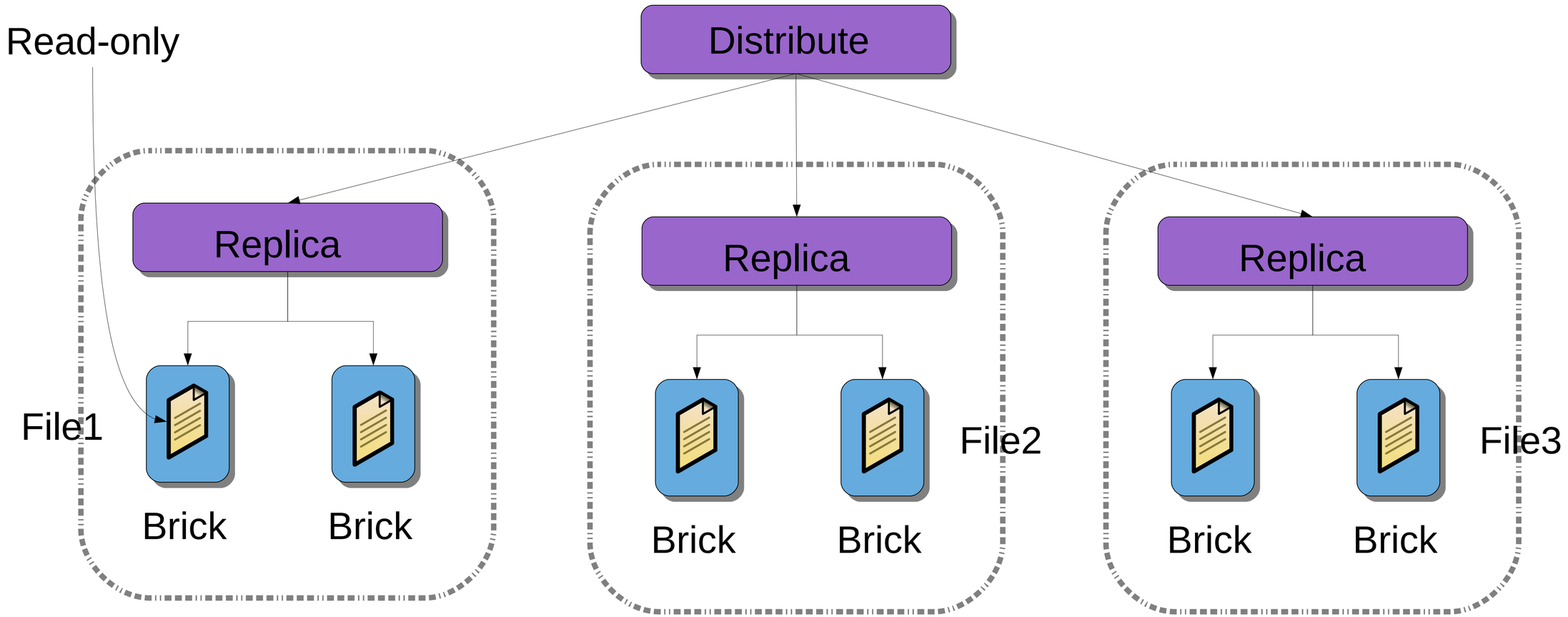
```
# gluster volume set all cluster.server-quorum-ratio <Percentage>

e.g.
# gluster volume set all cluster.server-quorum-ratio 51%
```

  - All bricks on the node are brought down in case server-side quorum is not met



Client-side Quorum



Troubleshooting – Preventing Split Brain

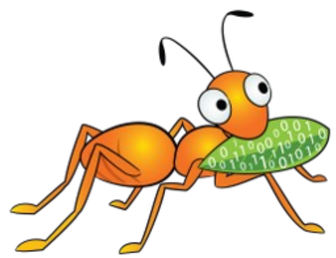
- Configuring Client-Side Quorum
  - Determines number of bricks that must be up for allowing data modification
  - Files will become read-only in case of quorum failure
  - Two types of client-side quorum

```
# gluster volume set all cluster.quorum-type <fixed | auto>
```

    - Fixed – fixed number of bricks should be up

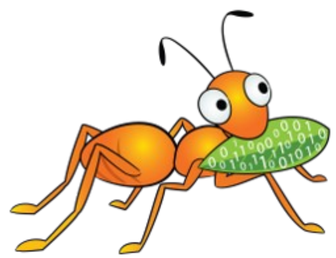
```
# gluster volume set all cluster.quorum-count <count>
```

    - Auto – Quorum conditions are determined by GlusterFS



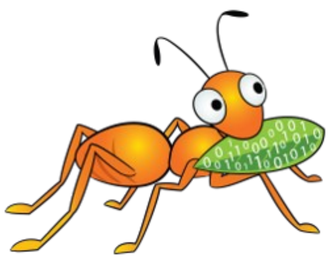
Troubleshooting – Preventing Split Brain

- Configuring Client-Side Quorum
  - Auto quorum type
    - At least n/2 brick needs to be up, where is n is the replica count
    - If n is even and exactly n/2 bricks are up then first brick of the replica set should be up



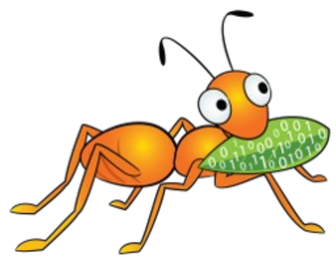
Community

- IRC channels:
  - #gluster – For any gluster usage or related discussions
  - #gluster-dev – For any gluster development related discussions
  - #gluster-meeting – To attend the weekly meeting and bug triage
- Mailing lists:
  - gluster-users@gluster.org - For any user queries or related discussions
  - gluster-devel@gluster.org - For any gluster development related queries/discussions



References

- www.gluster.org
- https://gluster.readthedocs.org/en/latest/
- https://github.com/gluster/gluster-tutorial



Thanks