By Marc

Published: 2009-03-04 18:21

High-Availability Storage Cluster With GlusterFS On Ubuntu1. Introduction

Original article: **http://blogama.org**

In this tutorial I will show you how to install GlusterFS in a scalable way to create a storage cluster, starting with 2 servers on Ubuntu 8.04 LTS server. Files will be replicated and splitted accross all servers which is some sort of RAID 10 (raid 1 with < 4 servers). With 4 servers that have each 100GB hard drive, total storage will be 200GB and if one server fails, the data will still be intact and files on the failed server will be replicated on another working server.

GlusterFS is a clustered file-system capable of scaling to several peta-bytes. It aggregates various storage bricks over Infiniband RDMA or TCP/IP interconnect into one large parallel network file system. Storage bricks can be made of any commodity hardware such as x86-64 server with SATA-II RAID and Infiniband HBA.

2. Installation

First you need to install some software:

```
sudo su
```

```
apt-get install sshfs build-essential flex bison byacc vim wget
```

Now we need to install fuse from source:

```
cd /root/

wget http://europe.gluster.org/glusterfs/fuse/fuse-2.7.4glfs11.tar.gz

tar -zxvf fuse-2.7.4glfs11.tar.gz

cd /root/fuse-2.7.4glfs11
```

Next we compile fuse:

```
./configure

make && make install
```

Next we will install GlusterFS:Get the same exact version, otherwise there is good chances it wont work. I tried with 2.0.0rc1 and 1.3.12 and there was some issues (1.4.0rc7 works fine).

```
cd /root/

wget http://ftp.gluster.com/pub/gluster/glusterfs/2.0/LATEST/glusterfs-2.0.0rc2.tar.gz

tar -zxvf glusterfs-2.0.0rc2.tar.gz

cd /root/glusterfs-2.0.0rc2/
```

Take a minute break and compile:

```
./configure

make && make install
```

For some reasons, libraries are going in the wrong directory so we need to (if someone has a clean fix to this please post it!):

```
cp /usr/local/lib/* -R /usr/lib/
```

Next we create some folders that will be used later on:

```
mkdir /mnt/glusterfs

mkdir /data/

mkdir /data/export

mkdir /data/export-ns

mkdir /etc/glusterfs/
```

3. Servers configuration

Before you go further, you need to know that GlusterFS works in a client/server way. What we will do is to make our servers both client and server for GlusterFS.

Lets start with the server configuration file ON ALL SERVERS:

```
vi /etc/glusterfs/glusterfs-server.vol
```

and make it look like this:

```
# file: /etc/glusterfs/glusterfs-server.vol

volume posix
  type storage/posix
  option directory /data/export
end-volume

volume locks
  type features/locks
  subvolumes posix
end-volume

volume brick
  type performance/io-threads
  option thread-count 8
  subvolumes locks
end-volume

volume posix-ns
  type storage/posix
  option directory /data/export-ns
```

```
end-volume

volume locks-ns
  type features/locks
  subvolumes posix-ns
end-volume

volume brick-ns
  type performance/io-threads
  option thread-count 8
  subvolumes locks-ns
end-volume

volume server
  type protocol/server
  option transport-type tcp
  option auth.addr.brick.allow *
  option auth.addr.brick-ns.allow *
  subvolumes brick brick-ns
end-volume
```

Now do:

```
glusterfsd -f /etc/glusterfs/glusterfs-server.vol
```

to start the server daemon.

4. Clients configuration

In these example files, I will use the following hosts:

```
server1 : 192.168.0.1
server2 : 192.168.0.2
server3 : 192.168.0.3
server4 : 192.168.0.4
```

[...]

Now we edit the client configuration file ON ALL SERVERS (because servers are clients as well in this howto):

```
vi /etc/glusterfs/glusterfs-client.vol
```

2 servers configuration (sort of RAID1)

```
### Add client feature and attach to remote subvolume of server1
volume brick1
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.1    # IP address of the remote brick
 option remote-subvolume brick       # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server2
volume brick2
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.2     # IP address of the remote brick
 option remote-subvolume brick       # name of the remote volume
end-volume
```

```
### The file index on server1
volume brick1-ns
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.1    # IP address of the remote brick
 option remote-subvolume brick-ns        # name of the remote volume
end-volume

### The file index on server2
volume brick2-ns
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.2      # IP address of the remote brick
 option remote-subvolume brick-ns        # name of the remote volume
end-volume

#The replicated volume with data
volume afr1
 type cluster/afr
 subvolumes brick1 brick2
end-volume

#The replicated volume with indexes
volume afr-ns
 type cluster/afr
 subvolumes brick1-ns brick2-ns
end-volume

#The unification of all afr volumes (used for > 2 servers)
volume unify
  type cluster/unify
  option scheduler rr # round robin
  option namespace afr-ns
  subvolumes afr1
end-volume

```

4 servers configuration (sort of RAID10)

```
### Add client feature and attach to remote subvolume of server1
volume brick1
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.1    # IP address of the remote brick
 option remote-subvolume brick       # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server2
volume brick2
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.2      # IP address of the remote brick
 option remote-subvolume brick       # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server3
volume brick3
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.3      # IP address of the remote brick
 option remote-subvolume brick       # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server4
volume brick4
```

```
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.4      # IP address of the remote brick
 option remote-subvolume brick        # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server1
volume brick1-ns
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.1    # IP address of the remote brick
 option remote-subvolume brick-ns        # name of the remote volume
end-volume

### Add client feature and attach to remote subvolume of server2
volume brick2-ns
 type protocol/client
 option transport-type tcp/client
 option remote-host 192.168.0.2      # IP address of the remote brick
 option remote-subvolume brick-ns        # name of the remote volume
end-volume

volume afr1
 type cluster/afr
 subvolumes brick1 brick4
end-volume

volume afr2
 type cluster/afr
 subvolumes brick2 brick3
end-volume

volume afr-ns
```

```
 type cluster/afr
 subvolumes brick1-ns brick2-ns
end-volume

volume unify
  type cluster/unify
  option scheduler rr # round robin
  option namespace afr-ns
  subvolumes afr1 afr2
end-volume
```

So on and so forth... For configuration over 4 servers, simply add brick volumes 2 by two, replicate them and dont forget to put them in the "unify" volume.

Now mount the GlusterFS on all servers in the cluster:

```
glusterfs -f /etc/glusterfs/glusterfs-client.vol /mnt/glusterfs
```

5. Testing

Once you mounted the GlusterFS to `/mnt/glusterfs` you can start copying files and see what is happening. Below are my tests on 4 servers. Everything works as it should, files in `/data/export` only show in 2 out of 4 server and everything is there under `/mnt/glusterfs` and `/data/export-ns`:

```
server 1 (ls -la /data/export)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 budget.ods
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
```

```
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server 4 (ls -la /data/export)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 budget.ods
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server 2 (ls -la /data/export)
-rwxr--r-- 1 marc marc 135793 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc 112640 2008-11-17 21:41 cv.doc
-rwxrwxrwx 1 marc marc  13546 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc  25088 2006-07-03 17:07 menulaurentien.doc
-rwxr--r-- 1 marc marc  33734 2009-02-06 12:58 opera6.htm

server 3 (ls -la /data/export)
-rwxr--r-- 1 marc marc 135793 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc 112640 2008-11-17 21:41 cv.doc
-rwxrwxrwx 1 marc marc  13546 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc  25088 2006-07-03 17:07 menulaurentien.doc
-rwxr--r-- 1 marc marc  33734 2009-02-06 12:58 opera6.htm

server x (ls -la /mnt/glusterfs)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxr--r-- 1 marc marc 135793 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 budget.ods
-rwxrwxrwx 1 marc marc 112640 2008-11-17 21:41 cv.doc
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13546 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc  25088 2006-07-03 17:07 menulaurentien.doc
```

```
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
-rwxr--r-- 1 marc marc  33734 2009-02-06 12:58 opera6.htm
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server 1 (ls -la /data/export-ns)
-rwxrwxrwx 1 marc marc    0 2007-09-14 14:14 6-instructions2.pdf
-rwxr--r-- 1 marc marc    0 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc    0 2008-12-18 11:54 budget.ods
-rwxrwxrwx 1 marc marc    0 2008-11-17 21:41 cv.doc
-rwxr--r-- 1 marc marc    0 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc    0 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc    0 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc    0 2006-07-03 17:07 menulaurentien.doc
-rwxrwxrwx 1 marc marc    0 2008-04-02 18:48 odometre.pdf
-rwxr--r-- 1 marc marc    0 2009-02-06 12:58 opera6.htm
-rwxrwxrwx 1 marc marc    0 2008-05-23 19:42 Thumbs.db


server 2 (ls -la /data/export-ns)
-rwxrwxrwx 1 marc marc    0 2007-09-14 14:14 6-instructions2.pdf
-rwxr--r-- 1 marc marc    0 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc    0 2008-12-18 11:54 budget.ods
-rwxrwxrwx 1 marc marc    0 2008-11-17 21:41 cv.doc
-rwxr--r-- 1 marc marc    0 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc    0 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc    0 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc    0 2006-07-03 17:07 menulaurentien.doc
-rwxrwxrwx 1 marc marc    0 2008-04-02 18:48 odometre.pdf
-rwxr--r-- 1 marc marc    0 2009-02-06 12:58 opera6.htm
-rwxrwxrwx 1 marc marc    0 2008-05-23 19:42 Thumbs.db
```

Now let's say we want to test how redundant is the setup. Lets reboot server1 and create new files while its down:

```
> /mnt/glusterfs/testfile

> /mnt/glusterfs/testfile2

> /mnt/glusterfs/testfile3

> /mnt/glusterfs/testfile4
```

Once server1 is back, let's check file consistency:

```
server 1 (ls -la /data/export)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 b4udget.ods
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server 4 (ls -la /data/export)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 budget.ods
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
-rw-r--r-- 1 root root      0 2009-02-19 11:32 testfile
-rw-r--r-- 1 root root      0 2009-02-19 11:32 testfile3
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server 1 (ls -la /data/export-ns)
```

```
-rwxrwxrwx 1 marc marc    0 2007-09-14 14:14 6-instructions2.pdf
-rwxr--r-- 1 marc marc    0 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc    0 2008-12-18 11:54 budget.ods
-rwxrwxrwx 1 marc marc    0 2008-11-17 21:41 cv.doc
-rwxr--r-- 1 marc marc    0 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc    0 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc    0 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc    0 2006-07-03 17:07 menulaurentien.doc
-rwxrwxrwx 1 marc marc    0 2008-04-02 18:48 odometre.pdf
-rwxr--r-- 1 marc marc    0 2009-02-06 12:58 opera6.htm
-rwxrwxrwx 1 marc marc    0 2008-05-23 19:42 Thumbs.db
```

Oups, we have an inconstency here. To fix that, gluster documentation says missing files have to be read. So let's do this simple command to read all files:

```
ls -lR /mnt/glusterfs/
```

Now, let's check what we have on server1:

```
server1 (ls -la /data/export)
-rwxrwxrwx 1 marc marc 215663 2007-09-14 14:14 6-instructions2.pdf
-rwxrwxrwx 1 marc marc   2256 2008-12-18 11:54 budget.ods
-rwxr--r-- 1 marc marc  21281 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc  13308 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc 196375 2008-04-02 18:48 odometre.pdf
-rw-r--r-- 1 root root      0 2009-02-19 11:32 testfile
-rw-r--r-- 1 root root      0 2009-02-19 11:32 testfile3
-rwxrwxrwx 1 marc marc   5632 2008-05-23 19:42 Thumbs.db

server1 (ls -la /data/export-ns)
-rwxrwxrwx 1 marc marc    0 2007-09-14 14:14 6-instructions2.pdf
```

```
-rwxr--r-- 1 marc marc    0 2009-02-02 15:26 bookmarks.html
-rwxrwxrwx 1 marc marc    0 2008-12-18 11:54 budget.ods
-rwxrwxrwx 1 marc marc    0 2008-11-17 21:41 cv.doc
-rwxr--r-- 1 marc marc    0 2009-02-18 16:45 cv_nouveau.docx
-rwxrwxrwx 1 marc marc    0 2007-09-11 15:43 cv.odt
-rwxrwxrwx 1 marc marc    0 2009-01-26 10:49 cv.pdf
-rwxrwxrwx 1 marc marc    0 2006-07-03 17:07 menulaurentien.doc
-rwxrwxrwx 1 marc marc    0 2008-04-02 18:48 odometre.pdf
-rwxr--r-- 1 marc marc    0 2009-02-06 12:58 opera6.htm
-rw-r--r-- 1 root root    0 2009-02-19 11:29 testfile
-rw-r--r-- 1 root root    0 2009-02-19 11:29 testfile2
-rw-r--r-- 1 root root    0 2009-02-19 11:29 testfile3
-rw-r--r-- 1 root root    0 2009-02-19 11:29 testfile4
-rwxrwxrwx 1 marc marc    0 2008-05-23 19:42 Thumbs.db
```

Now everything is as it should be.

6. Conclusion

GlusterFS has a lot of potential. What you saw here is a small portion of what GlusterFS can do. As I said in the first page, this setup was not tested on a live webserver and very little testing was done. If you plan to put this on a live server and test this setup in depth, please share your experience in the forums or simply post a comment on this page. Also, it would be very interesting if someone can post benchmarks to see how well it scale.

Further reading : **http://www.gluster.org**