实验环境：
操作系统：CentOS 6.5 x86
GlusterFS:3.5.2
主机名将统一使用短域名

主机名称	主机IP	主机角色
gfs1.bj1.haodf.net	10.1.6.200	存储节点1
gfs2.bj1.haodf.net	10.1.6.201	存储节点2
gfs3.bj1.haodf.net	10.1.6.202	存储备用节点
opc5.bj1.haodf.net	10.1.6.114	客户端1
opc6.bj1.haodf.net	10.1.6.115	客户端2

一、安装
1、配置yum源
vim /etc/yum.repos.d/glusterfs.repo

```
[gluster]
name=glusterfs
baseurl=http://download.gluster.org/pub/gluster/glusterfs/3.5/3.5.2/CentOS/epel-6.5/x86_64/
enable=1
gpgcheck=0
```

清除yum缓存
yum clean all

2、安装
服务器端安装：
yum -y install glusterfs glusterfs-server gluster-fuse

安装客户端：
yum install glusterfs glusterfs-fuse

二、集群操作
1、添加集群
集群操作只需要在其中一台服务器上操作即可，步骤如下：

```
[root@gfs1 ~]# gluster
gluster> pool list
UUID                                    Hostname        State
0cb874cf-fa65-4d6c-a654-950cfbb8ff7c    localhost       Connected
```

查看目前集群情况

```
gluster> peer probe gfs2.bj1
peer probe: success.
gluster> pool list
UUID                                    Hostname        State
178cb6c9-e335-4ad9-a2e4-0244c6694ca2    gfs2.bj1        Connected
0cb874cf-fa65-4d6c-a654-950cfbb8ff7c    localhost       Connected
```

添加gfs2.bj1 到集群

2、volume操作

```
gluster> volume list
No volumes present in cluster
gluster> volume create v1 replica 2 gfs1.bj1:/Data/gfs/v1 gfs2.bj1:/Data/gfs/v1
volume create: v1: success: please start the volume to access data
gluster> volume list
v1
```

创建一个v1的复制卷(理解为raid 1)

```
gluster> volume start v1
volume start: v1: success
gluster> volume status v1
Status of volume: v1
Gluster process                                 Port    Online  Pid
------------------------------------------------------------------------------
Brick gfs1.bj1:/Data/gfs/v1                     49152   Y       26128
Brick gfs2.bj1:/Data/gfs/v1                     49152   Y       5131
NFS Server on localhost                         N/A     N       N/A
Self-heal Daemon on localhost                   N/A     Y       26146
NFS Server on gfs2.bj1                          2049    Y       5145
Self-heal Daemon on gfs2.bj1                    N/A     Y       5149

Task Status of Volume v1
------------------------------------------------------------------------------
There are no active volume tasks
```

启动volume v1

3、客户端挂载

```
[root@opc5 ~]# mount -t glusterfs gfs1.bj1:/v1 /v1/
[root@opc5 ~]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda2        20G  1.3G   18G   7% /
tmpfs            16G     0   16G   0% /dev/shm
/dev/sda7       859G   61G  755G   8% /Data
/dev/sda1       194M   29M  155M  16% /boot
/dev/sda5       2.0G   68M  1.9G   4% /tmp
/dev/sda3        20G  811M   18G   5% /var
gfs1.bj1:/v1    902G  200M  856G   1% /v1
```

2台客户端使用同样的命令进行挂载操作。

三、数据安全
分别测试小文件大文件

```
[root@opc5 ~]# cd /v1
[root@opc5 v1]# ls
[root@opc5 v1]# echo gfstest > t1.txt
[root@opc5 v1]# dd if=/dev/zero of=1G bs=1M count=1024
1024+0 records in
1024+0 records out
1073741824 bytes (1.1 GB) copied, 18.1286 s, 59.2 MB/s
```

创建一个小文本文件和一个1G的大文件

1、GlusterFS 单台数据损坏

```
[root@gfs1 ~]# cd /Data/gfs/v1/
[root@gfs1 v1]# ls
1G  t1.txt
[root@gfs1 v1]# rm t1.txt
rm: remove regular file `t1.txt'? y
[root@gfs1 v1]# ls
1G
[root@gfs1 v1]#
```

模拟删除gfs1.bj1存储节点的数据

```
[root@opc5 v1]# ls
1G  t1.txt
[root@opc5 v1]# cat t1.txt
gfstest
```

从客户端查看文件，没有影响数据的访问

等发生触发条件的情况下，此被删除的文件将会从gfs2.bj1节点重新同步过来。触发条件：重新上传文件、手动触发等。

```
gluster> volume heal v1 full
Launching heal operation to perform full self heal on volume v1 has been successful
Use heal info commands to check status
gluster>
```

手动触发数据修复

```
[root@gfs1 v1]# ls
1G
[root@gfs1 v1]# ls
1G  t1.txt
[root@gfs1 v1]#
```

小文件t1.txt已经同步

```
[root@gfs1 v1]# rm 1G
rm: remove regular file `1G'? y
[root@gfs1 v1]# ls
t1.txt
[root@gfs1 v1]#
```

删除大文件1G，使用手动触发的方式让GlusterFS触发同步

```
[root@opc5 v1]# md5sum 1G
cd573cfaace07e7949bc0c46028904ff  1G
[root@opc5 v1]# md5sum 1G
cd573cfaace07e7949bc0c46028904ff  1G
```

gfs1.bj1删除大文件1G后，客户端访问依然没受到影响

```
[root@gfs1 v1]# ls
t1.txt
[root@gfs1 v1]# ls
1G  t1.txt
[root@gfs1 v1]# md5sum 1G
cd573cfaace07e7949bc0c46028904ff  1G
[root@gfs1 v1]#
```

gfs1.bj1上被破坏的大文件1G已经被自动修复

2、GlusterFS 单台服务器故障

```
[root@opc5 v1]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda2        20G  1.3G   18G   7% /
tmpfs            16G     0   16G   0% /dev/shm
/dev/sda7       859G   61G  755G   8% /Data
/dev/sda1       194M   29M  155M  16% /boot
/dev/sda5       2.0G   68M  1.9G   4% /tmp
/dev/sda3        20G  812M   18G   5% /var
gfs1.bj1:/v1    902G  1.2G  855G   1% /v1
```

```
[root@opc6 v1]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda2        20G  1.3G   18G   7% /
tmpfs            32G     0   32G   0% /dev/shm
/dev/sda7       1.8T  291M  1.7T   1% /Data
/dev/sda1       194M   29M  155M  16% /boot
/dev/sda5       2.0G   68M  1.9G   4% /tmp
/dev/sda3        20G  881M   18G   5% /var
gfs1.bj1:/v1    902G  1.2G  855G   1% /v1
```

现在opc5 opc6 这2台客户端都是通过gfs1.bj1挂载到本地。现在模拟gfs1.bj1这个存储节点故障宕机的场景。

```
[root@gfs1 v1]# halt

Broadcast message from huangyi@gfs1.bj1.haodf.net
        (/dev/pts/0) at 15:42 ...

The system is going down for halt NOW!
[root@gfs1 v1]# Connection to gfs1 closed by remote host.
Connection to gfs1 closed.
```

```
[root@opc5 v1]# ping gfs1.bj1
PING gfs1.bj1.haodf.net (10.1.6.200) 56(84) bytes of data.
From 10.1.6.114 icmp_seq=17 Destination Host Unreachable
From 10.1.6.114 icmp_seq=18 Destination Host Unreachable
From 10.1.6.114 icmp_seq=20 Destination Host Unreachable
From 10.1.6.114 icmp_seq=21 Destination Host Unreachable
^C
--- gfs1.bj1.haodf.net ping statistics ---
21 packets transmitted, 0 received, +4 errors, 100% packet loss, time 20035ms
pipe 2
```

存储节点gfs1.bj1已经被关机

```
[root@opc5 v1]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda2        20G  1.3G   18G   7% /
tmpfs            16G     0   16G   0% /dev/shm
/dev/sda7       859G   61G  755G   8% /Data
/dev/sda1       194M   29M  155M  16% /boot
/dev/sda5       2.0G   68M  1.9G   4% /tmp
/dev/sda3        20G  812M   18G   5% /var
gfs1.bj1:/v1    902G  1.2G  855G   1% /v1
[root@opc5 v1]# cat /v1/t1.txt
gfstest
[root@opc5 v1]# md5sum /v1/1G
cd573cfaace07e7949bc0c46028904ff  /v1/1G
```

```
[root@opc5 v1]# echo 'gfs1.bj1 is down' >> /v1/t1.txt
[root@opc5 v1]# cat /v1/t1.txt
gfstest
gfs1.bj1 is down
[root@opc5 v1]#
```

```
[root@opc6 v1]# cat /v1/t1.txt
gfstest
gfs1.bj1 is down
[root@opc6 v1]# md5sum /v1/1G
cd573cfaace07e7949bc0c46028904ff  /v1/1G
```

客户端依然能正常访问数据

3、GlusterFS故障替换

当其中一个存储节点彻底毁坏，需要使用备用节点顶替。

切换到另外一台存储节点gfs2.bj1

```
gluster> pool list
UUID                                    Hostname        State
0cb874cf-fa65-4d6c-a654-950cfbb8ff7c    10.1.6.200      Disconnected
178cb6c9-e335-4ad9-a2e4-0244c6694ca2    localhost       Connected
```

这里能看见，gfs1.bj1(10.1.6.200)已经失去联系。

```
gluster> peer probe gfs3.bj1
peer probe: success.
gluster> pool list
UUID                                    Hostname        State
0cb874cf-fa65-4d6c-a654-950cfbb8ff7c    10.1.6.200      Disconnected
73404649-0d0a-4c3f-9da0-242663dfde48    gfs3.bj1        Connected
178cb6c9-e335-4ad9-a2e4-0244c6694ca2    localhost       Connected
gluster>
```

把备用节点gfs3.bj1加入到集群

```
gluster> volume replace-brick v1 gfs1.bj1:/Data/gfs/v1 gfs3.bj1:/Data/gfs/v1 start force
gluster> volume status
gluster> quit
[root@gfs2 ~]# /etc/init.d/glusterd  restart
Starting glusterd:                                         [  OK  ]
[root@gfs2 ~]# gluster
gluster> volume status
Status of volume: v1
Gluster process                                 Port    Online  Pid
------------------------------------------------------------------------------
Brick gfs2.bj1:/Data/gfs/v1                     49152   Y       5131
NFS Server on localhost                         2049    Y       5325
Self-heal Daemon on localhost                   N/A     Y       5336
NFS Server on gfs3.bj1                          2049    Y       6653
Self-heal Daemon on gfs3.bj1                    N/A     Y       6652

Task Status of Volume v1
------------------------------------------------------------------------------
Task                 : Replace brick
ID                   : 9a984a1a-10aa-4d9c-a219-5543742c4eeb
Source Brick         : gfs1.bj1:/Data/gfs/v1
Destination Brick    : gfs3.bj1:/Data/gfs/v1
Status               : completed

gluster>
```

```
gluster> volume status
Status of volume: v1
Gluster process                                 Port    Online  Pid
------------------------------------------------------------------------------
Brick gfs3.bj1:/Data/gfs/v1                     49153   Y       6697
Brick gfs2.bj1:/Data/gfs/v1                     49152   Y       5131
NFS Server on localhost                         2049    Y       5350
Self-heal Daemon on localhost                   N/A     Y       5354
NFS Server on gfs3.bj1                          2049    Y       6704
Self-heal Daemon on gfs3.bj1                    N/A     Y       6708

Task Status of Volume v1
------------------------------------------------------------------------------
There are no active volume tasks

gluster>
```

gf3.bj1已经加入volume v1

```
[root@gfs3 glusterd]# ls /Data/gfs/v1/
[root@gfs3 glusterd]#
```

gfs3.bj1此时数据还没有被同步,需要达到触发条件。

```
[root@opc6 ~]# mount -t glusterfs gfs3.bj1:/v1 /v1/
[root@opc6 ~]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda2        20G  1.3G   18G   7% /
tmpfs            32G     0   32G   0% /dev/shm
/dev/sda7       1.8T  291M  1.7T   1% /Data
/dev/sda1       194M   29M  155M  16% /boot
/dev/sda5       2.0G   68M  1.9G   4% /tmp
/dev/sda3        20G  882M   18G   5% /var
gfs3.bj1:/v1     98G  1.2G   92G   2% /v1
[root@opc6 ~]# cat /v1/
1G      t1.txt
[root@opc6 ~]# cat /v1/t1.txt
gfstest
gfs1.bj1 is down
gfs3.bj1 up
[root@opc6 ~]# echo 'gf3test' >> /v1/t1.txt
[root@opc6 ~]# cat /v1/t1.txt
gfstest
gfs1.bj1 is down
gfs3.bj1 up
gf3test
[root@opc6 ~]#
```

客户端已经可以重新挂载心的存储节点gfs3.bj1。

四、性能

性能随存储模式的不同而不同。