第四十八章：GlusterFS 分布式存储

一、GlusterFS 概述；
二、GlusterFS 存储架构；
三、GlusterFS 工作原理；
四、GlusterFS 卷的类型；
五、案例：搭建 Gluster 分布式文件系统；

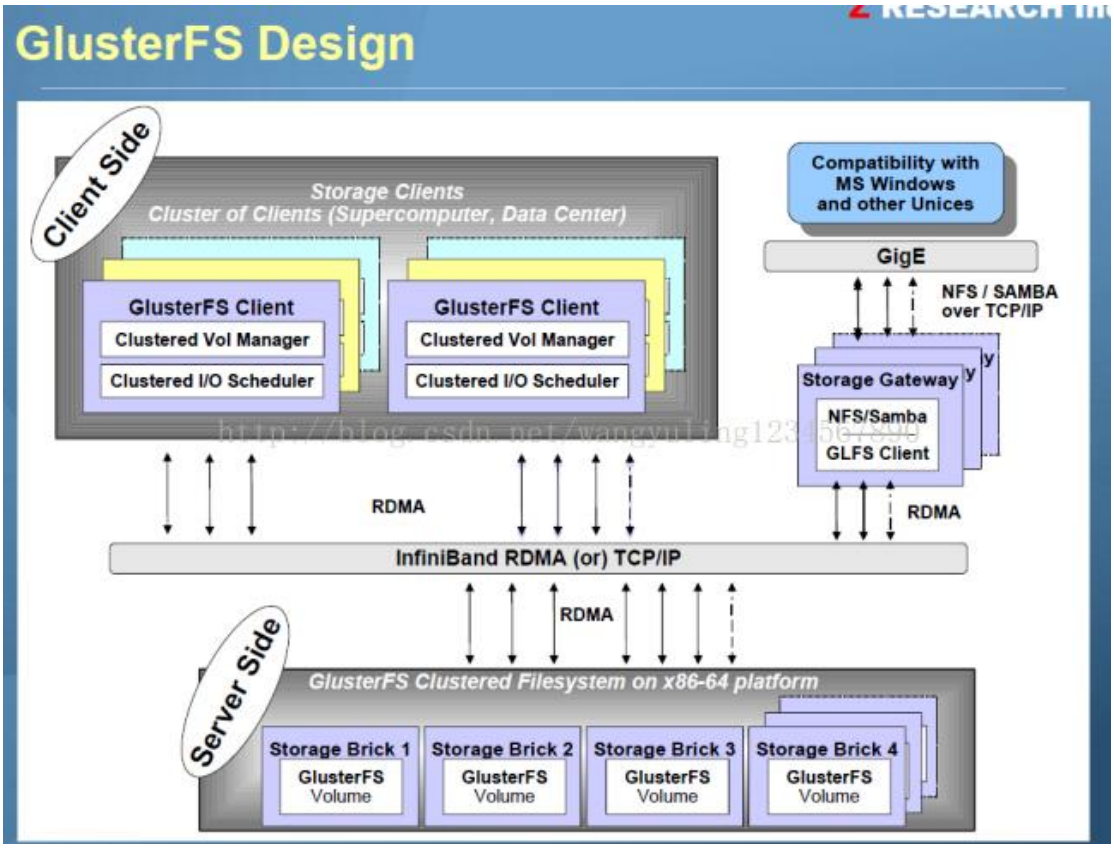
一、GlusterFS 概述；

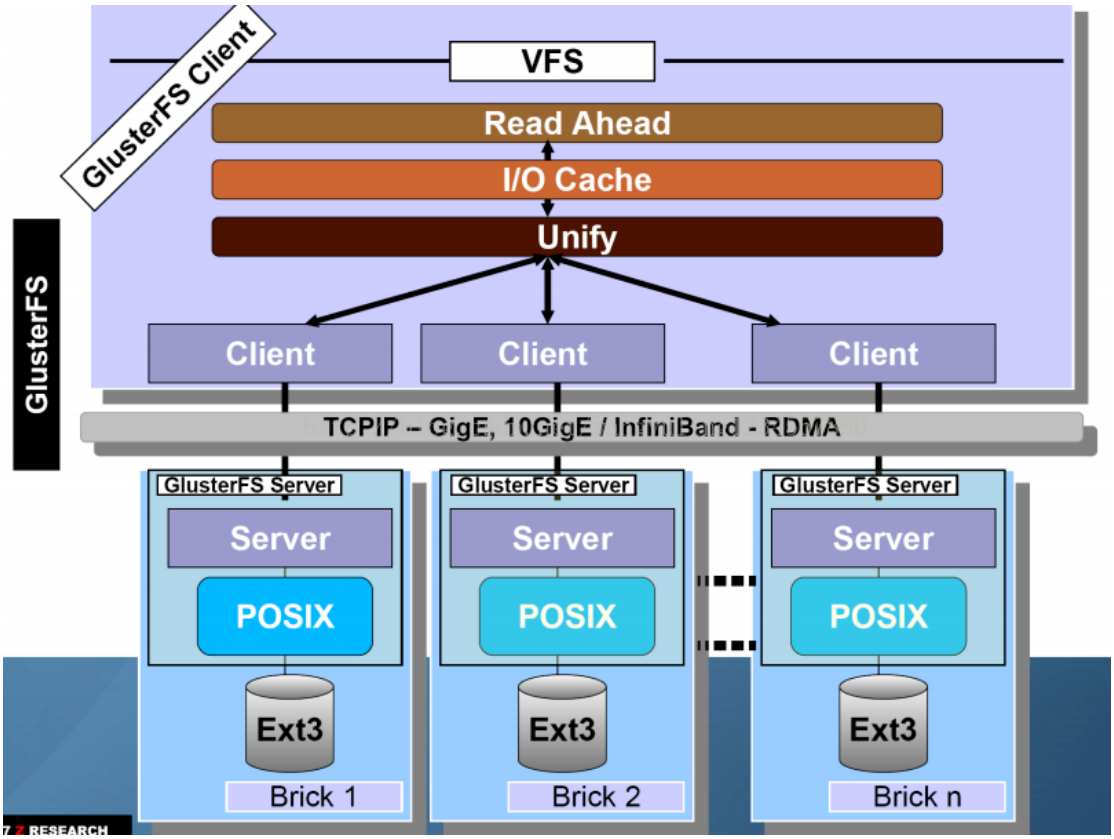
概述：GlusterFS（Google File System）是一个开源的分布式文件系统，Gluster 借助 TCP/IP 网络将存储资源分散存储在网络的不同节点，在通过汇聚为客户端提供统一的资源访问，在存储方面具有很强大的横向扩展能力，通过扩展不同的节点可以支持 PB 级别的存储容量；
Bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、DB、NB

特点：
扩展性与高性能：通过 Scale-out 架构可以增加存储节点的方式来提高容量和性能（磁盘、计算、I/O 资源都可以独立增加），Gluster 弹性哈希（Elastic Hash）解决了 Gluster 服务对元数据服务器的依赖，Gluster 采用弹性哈希算法来确定数据在 chunk 节点中的分布情况，无须元数据服务器，实现了存储的横向扩展，改善了元数据服务器节点的压力以及单点故障；

高可用性：GlusterFS 通过配置不同类型的卷，可以对数据进行自动复制（类似于 RAID1），即使某节点故障，也不影响数据的访问；
通用性：GlusterFS 没有设置独立的私有数据文件系统，而是采用以往的 ext4、ext3 等，数据可以通过传统的磁盘访问方式被客户端所访问；
弹性卷管理：GlusterFS 通过将数据存储在逻辑卷上，逻辑卷从逻辑存储池进行独立逻辑划分，逻辑存储池可以在线进行增加和删除，不会导致业务中断，逻辑卷的数量可以根据实际需求进行自行增加和缩减；

二、GlusterFS 存储架构；





专业术语：

Brick（存储块）：存储池中节点对外提供存储服务的目录；

Volume（逻辑卷）：一个逻辑卷时一组 Brick 的集合，卷是数据存储的逻辑设备，类似 LVM 中的逻辑卷，大部分 GlusterFS 管理操作都是在逻辑卷上进行的；

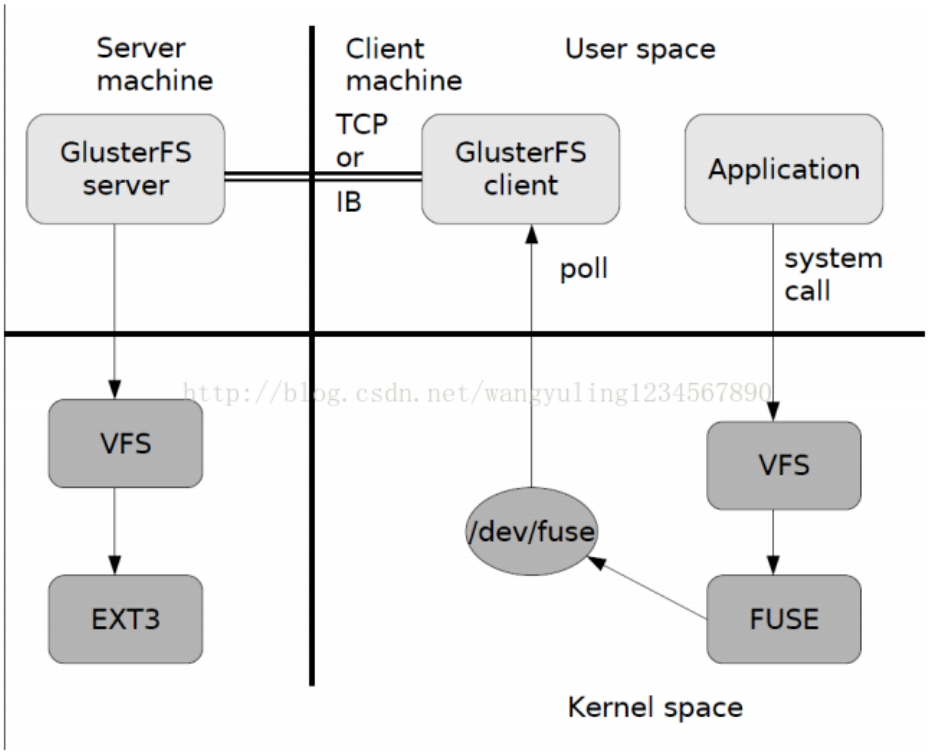
FUSE（用户空间文件系统）：是一个内核模块，用户自行创建挂载的的文件系统；

VFS（接口）：内核空间对用户空间提供的访问磁盘的接口；

Glusterd（后台管理进程）：在存储集群中的每个节点上都要运行；

三、GlusterFS 工作原理；

数据访问流程：



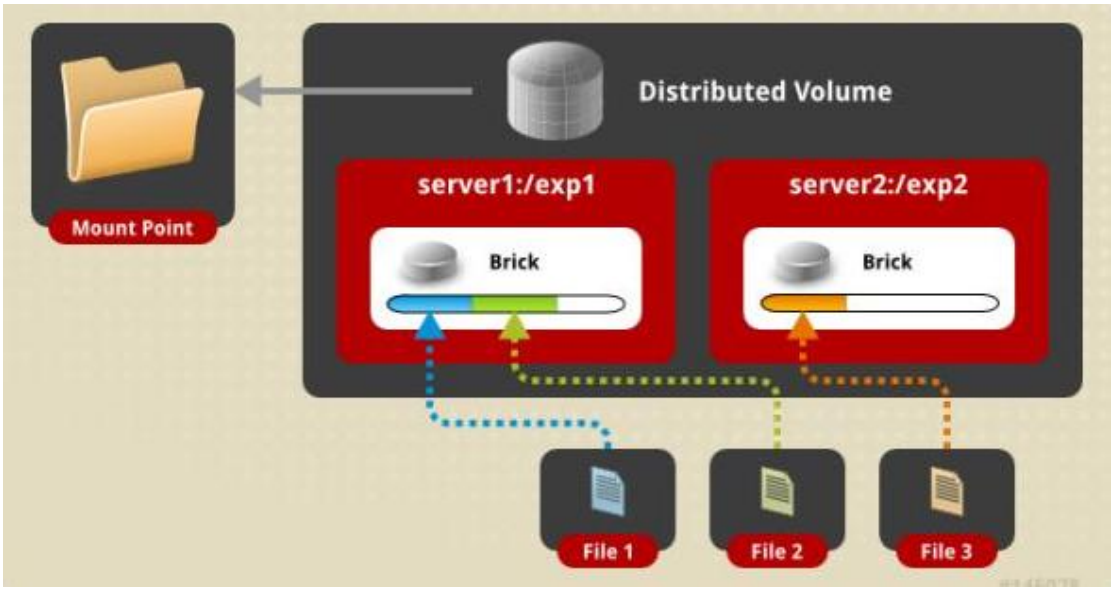
1. 首先是在客户端，用户通过 glusterfs 的 mount point 来读写数据， 对于用户来说， 集群系统的存在对用户是完全透明的， 用户感觉不到是操作本地系统还是远端的集群系统。
2. 用户的这个操作被递交给 本地 linux 系统的 VFS 来处理。
3. VFS 将数据递交给 FUSE 内核文件系统：在启动 glusterfs 客户端以前， 需要想系统注册一个实际的文件系统 FUSE，如上图所示，该文件系统与 ext3 在同一个层次上面，ext3 是对实际的磁盘进行处理，而 fuse 文件系统则是将数据通过/dev/fuse 这个设备文件递交给了 glusterfs client 端。所以我们可以将 fuse 文件系统理解为一个代理。
4. 数据被 fuse 递交给 Glusterfs client 后，client 对数据进行一些指定的处理（所谓的指定，是按照 client 配置文件据来进行的一系列处理， 我们在启动 glusterfs client 时 需 要 指 定 这 个 文 件 ， 其 默 认 位 置 ： /etc/glusterfs/client.vol）。
5. 在 glusterfs client 的处理末端，通过网络将数据递交给 Glusterfs Server，并且将数据写入到服务器所控制的存储设备上。

四、GlusterFS 卷的类型；

分布式卷、条带卷、复制卷、分布式条带卷、分布式复制卷、条带复制卷、分布式条带复制卷；

1. 分布式卷

分布式卷是 GlusterFS 的默认卷，在创建卷时，默认选项是创建分布式卷。在该模式下，并没有对文件进行分块处理，文件直接存储在某个 Server 节点上。由于使用本地文件系统，所以存取效率并没有提高，反而会因为网络通信的原因而有所降低，另外支持超大型文件也会有一定的难度，因为分布式卷不会对文件进行分块处理，一个文件要么在 Server1 上，要么在 Serve2 上，不能分块同时存放在 Sever1 和 Server2 上；



特点：

文件分布在不同的服务器，不具备冗余性；

更容易且廉价地扩展卷的大小；

单点故障会造成数据丢失；

依赖底层的数据保护；

创建方法：

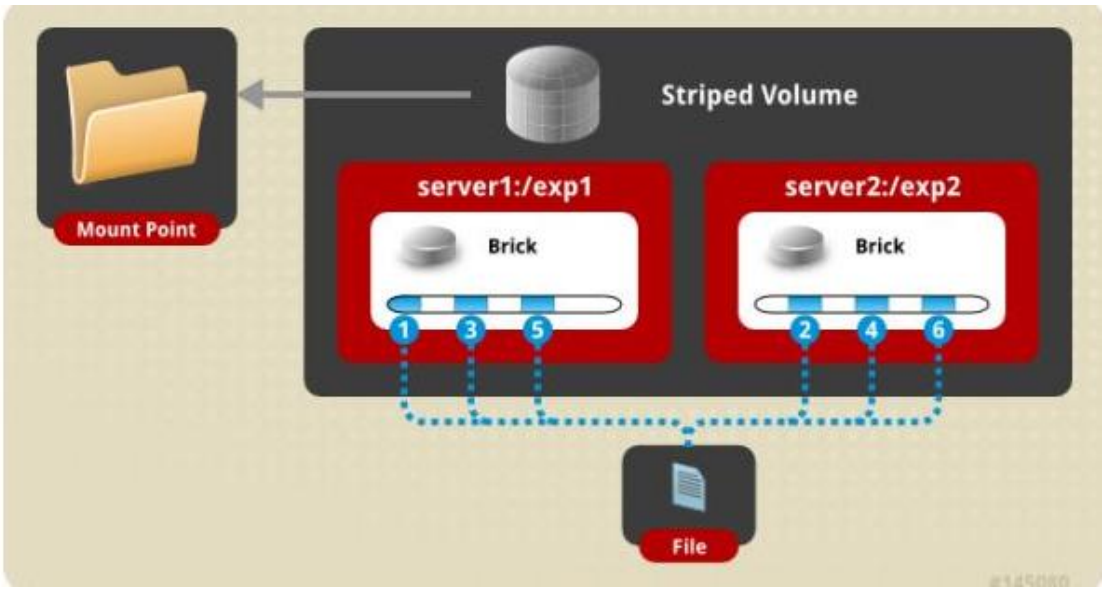
[root@gfs ~]# gluster volume create dis-volume server1:/dir1 server2:/dir2

Creation of dis -volume has been successful

Please start the volume to access data

2.条带卷

Stripe 模式相当于 RAIDO，在该模式下，根据偏移量将文件分成 N 块，轮询地存储在每个 Brick Server 节点。节点把每个数据块都作为普通文件存入本地文件系统中，通过扩展属性记录总块数（Stripe-count） 和每块的序号（Stripe-index），在配置时指定的条带数必须等于卷中 Brick 所包含的存储服务器数，在存储大文件时，性能尤为突出，但是不具备冗余性；



特点：

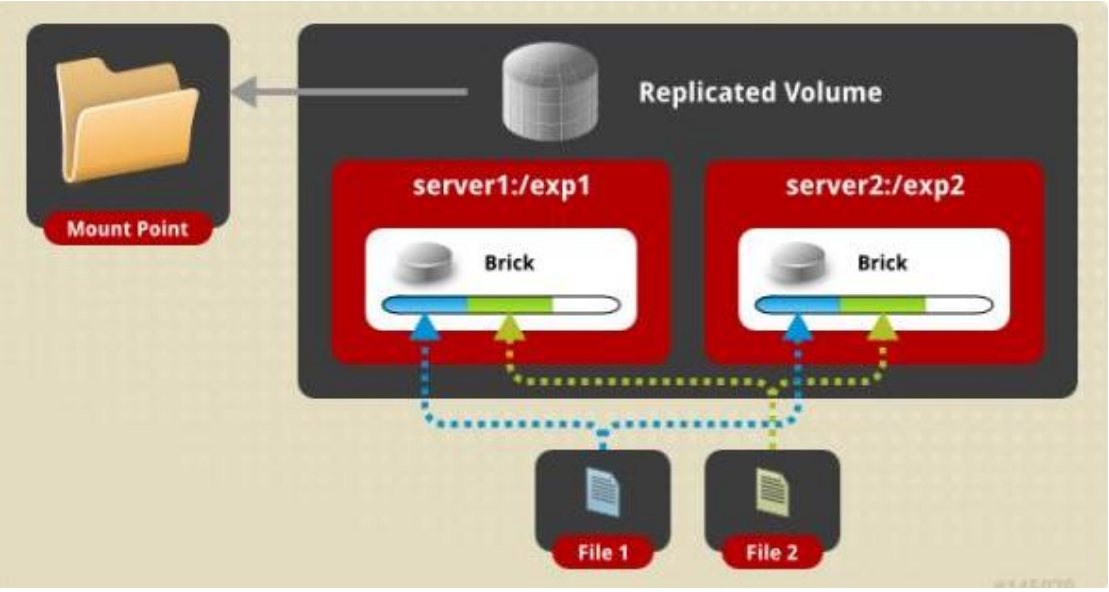
数据被分割成更小块分布到块服务器群中的不同；
分布减少了负载且更小的文件提高了存取速度；
没有数据冗余；

创建方法：

[root@gfs ~]# gluster volume create stripe-volume stripe 2 transport tcp server1:/dir1 server2:/dir2
create of Stripe -volume has been successful
please start the volume to access data

3.复制卷

也称为 AFR（AutGilePepliatio）相当于 RAD1，即同一文件保存一份或多份副本。每个节点上保存相同的内容和目录结构。复制模式因为要保存副本，所以磁盘利用率较低，复制卷时，复制数必须等于卷中 Brick 所包含的存储服务器数，复制卷具备冗余性，即使一个节点损坏，也不影响数据的正常使用；



特点：

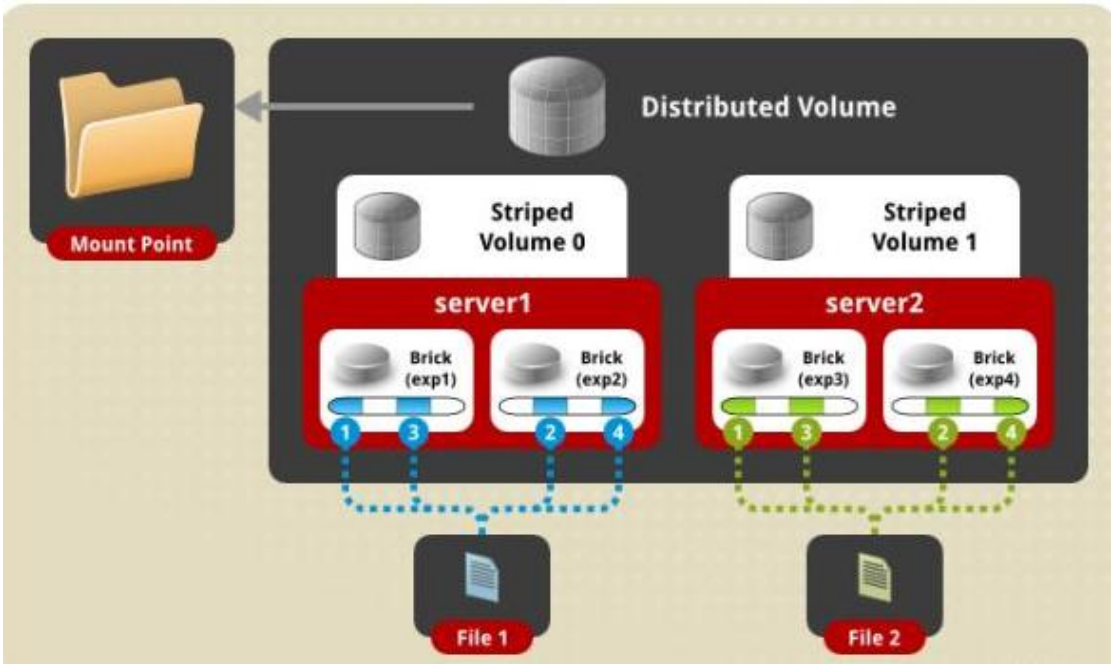
卷中所有的服务器均保存一个完整的副本；
卷的副本数量可由客户创建的时候决定；
最少保证两个块服务器或更多服务器；
具备冗余效果；

创建方法：

[root@gfs ~]# gluster volume create rep-volume replica 2 transport tcp server1:/dir1 server2:/dir2
create of rep -volume has been successful
please start the volume to access data

4.分布式条带卷

分布式条带卷兼顾分布式卷和条带卷的功能，可以理解成为大型的条带卷，主要用于大文件访问处理，创建一个分布式条带，卷最少需要 4 台服务器；



创建方法：

[root@gfs ~]# gluster volume create dis-stripe stripe 2 transport tcp server1:/dir1 server2:/dir2 server3:/dir3 server4:/dir4

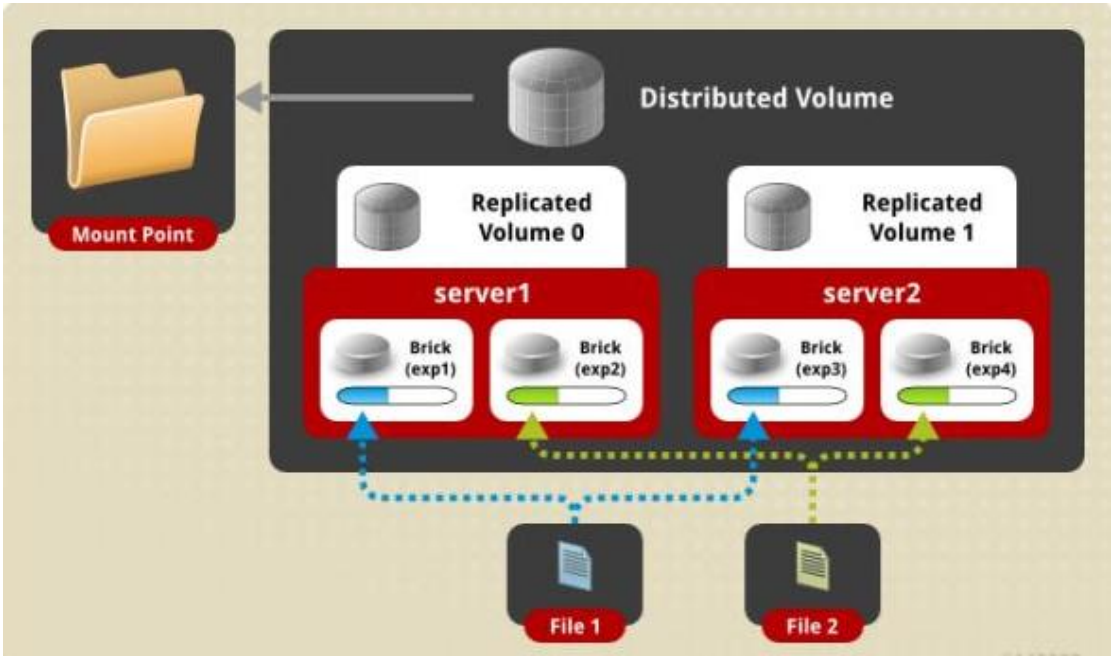
create of dis-stripe has been successful

please start the volume to access data

上述命令创建了一个名为 dis-stripe 的分布式条带卷，配置分布式条带卷时，卷中 Brick 所包含的存储服务器必须是条带数的倍数（大于等于 2 倍），如上述命令，Brick 的数量为 4，条带数为 2；

5.分布式复制卷

分布式复制卷兼顾分布式卷和复制卷的功能，可以理解成为大型的复制卷，主要用于冗余的场景下，创建一个分布式复制卷，最少需要 4 块 brick；



创建方法：

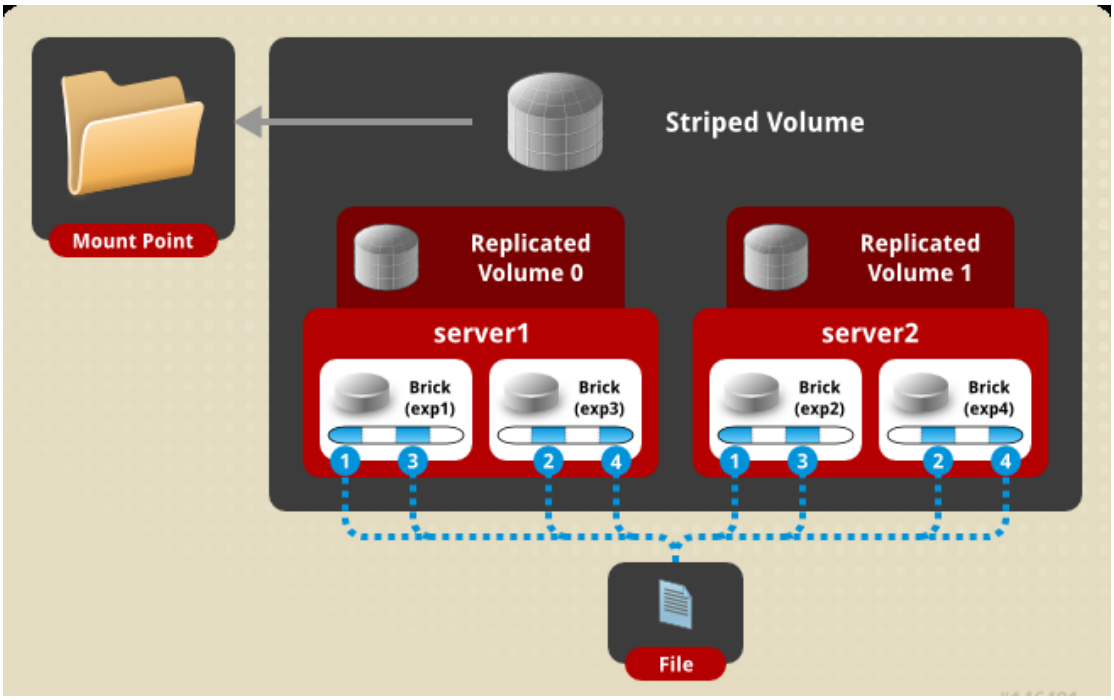
[root@gfs ~]# gluster volume create dis-rep replica 2 transport tcp server1:/dir1 server2:/dir2 server3:/dir3 server4:/dir4

create of dis-rep has been successful

please start the volume to access data

6.条带复制卷

条带复制卷兼顾了条带卷和复制卷两者的优点，相当于 RADI 10，用于存储效率高，备份冗余的场景下，创建条带复制卷，最少需要四个 brick；



创建方法：

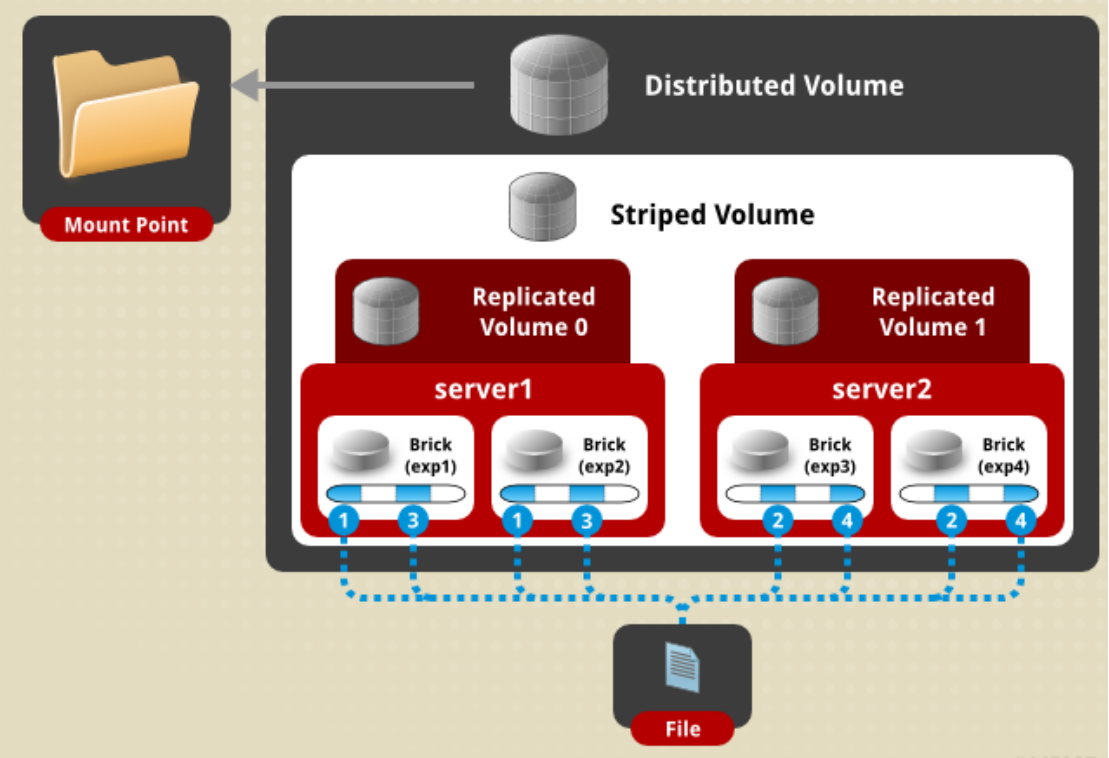
[root@gfs ~]# gluster volume create test-volume stripe 2 replica 2 transport tcp server1:/dir1 server2:/dir2 server3:/dir3 server4:/dir4

create of test-volume has been successful

please start the volume to access data

7.分布式条带复制卷

分布式条带复制卷将分布条带数据在复制卷集群。为了获得最佳效果，可以选择使用分布在高并发的条带复制卷环境下并行访问非常大的文件和性能是至关重要的；



五、案例：搭建 Gluster 分布式文件系统；

案例环境：

系统类型	IP 地址	主机名	所需软件
Centos 7.4 1708 64bit	192.168.100.101	data1.linuxfan.cn	glusterfs glusterfs-server glusterfs-fuse glusterfs-rdma
Centos 7.4 1708 64bit	192.168.100.102	data2.linuxfan.cn	glusterfs glusterfs-server glusterfs-fuse glusterfs-rdma
Centos 7.4 1708 64bit	192.168.100.103	data3.linuxfan.cn	glusterfs glusterfs-server glusterfs-fuse glusterfs-rdma
Centos 7.4 1708 64bit	192.168.100.104	client.linuxfan.cn	glusterfs glusterfs-fuse

案例步骤：

- 配置主机之间的解析（在此所有主机配置相同，在此只列举 data1 节点的配置）；
- 在所有 data 节点上安装 GlusterFS（在此所有主机配置相同，在此只列举 data1 节点的配置）；
- 在 data1 节点上进行创建集群，其他节点会同步配置；
- 在多个 data 节点创建数据存储的位置；

- 在 data1 节点创建数据存储的卷（复制卷），其他节点会同步配置；
- 在 client 客户端节点上安装 gluster 客户端工具并测试挂载；
- client 客户端节点测试存放文件；
- 扩展：Gluster 的管理命令；

- **配置主机之间的解析（在此所有主机配置相同，在此只列举 data1 节点的配置）；**

```
[root@data1 ~]# cat <<END >>/etc/hosts
192.168.100.101 data1.linuxfan.cn
192.168.100.102 data2.linuxfan.cn
192.168.100.103 data3.linuxfan.cn
192.168.100.104 client.linuxafn.cn
END
[root@data1 ~]# ping data1.linuxfan.cn -c 2                    ##ping 命令进行测试
PING data1.linuxfan.cn (192.168.100.101) 56(84) bytes of data.
64 bytes from data1.linuxfan.cn (192.168.100.101): icmp_seq=1 ttl=64 time=0.062 ms
64 bytes from data1.linuxfan.cn (192.168.100.101): icmp_seq=2 ttl=64 time=0.040 ms
```

- **在所有 data 节点上安装 GlusterFS（在此所有主机配置相同，在此只列举 data1 节点的配置）；**

```
[root@data1 ~]# wget -O /etc/yum.repos.d/CentOS-Base.repo http://mirrors.aliyun.com/repo/Centos-7.repo
[root@data1 ~]# yum -y install centos-release-gluster          ##安装 gluster 包的 yum 源
[root@data1 ~]# yum -y install glusterfs glusterfs-server glusterfs-fuse glusterfs-rdma
[root@data1 ~]# systemctl start glusterd
[root@data1 ~]# systemctl enable glusterd
Created symlink from /etc/systemd/system/multi-user.target.wants/glusterd.service to /usr/lib/systemd/system/glusterd.service.
[root@data1 ~]# netstat -utpln |grep glu
tcp        0      0 0.0.0.0:24007           0.0.0.0:*               LISTEN      1313/glusterd
[root@data1 ~]# netstat -utpln |grep rpc
tcp        0      0 0.0.0.0:111             0.0.0.0:*               LISTEN      1311/rpcbind
udp        0      0 0.0.0.0:111             0.0.0.0:*                           1311/rpcbind
udp        0      0 0.0.0.0:634             0.0.0.0:*                           1311/rpcbind
```

- **在 data1 节点上进行创建集群，其他节点会同步；**

```
[root@data1 ~]# gluster peer probe data1.linuxfan.cn          ##添加本机节点
peer probe: success. Probe on localhost not needed
[root@data1 ~]# gluster peer probe data2.linuxfan.cn          ##添加 data2 节点
```

peer probe: success.
[root@data1 ~]# gluster peer probe data3.linuxfan.cn ##添加 data3 节点
peer probe: success.
[root@data1 ~]# gluster peer status ##查看 gluster 集群状态
Number of Peers: 2

Hostname: data2.linuxfan.cn
Uuid: a452f7f4-7604-4d44-8b6a-f5178a41e308
State: Peer in Cluster (Connected)

Hostname: data3.linuxfan.cn
Uuid: b08f1b68-3f2c-4076-8121-1ab17d1517e1
State: Peer in Cluster (Connected)

- **在多个 data 节点创建数据存储的位置；**

[root@data1 ~]# mkdir /data
[root@data1 ~]# gluster volume info
No volumes present

- **在 data1 节点创建数据存储的卷（复制卷），其他节点会同步配置；**

[root@data1 ~]# gluster volume create rep-volume replica 3 transport tcp data1.linuxfan.cn:/data data2.linuxfan.cn:/data data3.linuxfan.cn:/data force ##创建复制卷，名称如上
volume create: rep-volume: success: please start the volume to access data
[root@data1 ~]# gluster volume info
Volume Name: rep-volume
Type: Replicate
Volume ID: ac59612b-e6ce-46ce-85a7-74262fb722b2
Status: Created
Snapshot Count: 0
Number of Bricks: 1 x 3 = 3
Transport-type: tcp
Bricks:
Brick1: data1.linuxfan.cn:/data
Brick2: data2.linuxfan.cn:/data
Brick3: data3.linuxfan.cn:/data
Options Reconfigured:
transport.address-family: inet
nfs.disable: on
performance.client-io-threads: off
[root@data1 ~]# gluster volume start rep-volume ##启动该卷
volume start: rep-volume: success

➢ **在 client 客户端节点上安装 gluster 客户端工具并测试挂载；**

[root@client ~]# yum install -y glusterfs glusterfs-fuse

[root@client ~]# mount -t glusterfs data1.linuxfan.cn:rep-volume /mnt/

[root@client ~]# ls /mnt/

[root@client ~]# df -hT |tail -1

data1.linuxfan.cn:rep-volume fuse.glusterfs 19G 2.0G 17G 11% /mnt

➢ **client 客户端节点测试存放文件；**

[root@client ~]# touch /mnt/{1..10}.file

[root@client ~]# dd if=/dev/zero of=/mnt/1.txt bs=1G count=1

[root@client ~]# ls /mnt/

10.file 1.file 1.txt 2.file 3.file 4.file 5.file 6.file 7.file 8.file 9.file

[root@client ~]# du -sh /mnt/1.txt

1.0G /mnt/1.txt

```
[root@data1 ~]# ls /data/
10.file  1.file  1.txt  2.file  3.file  4.file  5.file  6.file  7.file  8.file  9.file
```

```
[root@data2 ~]# ls /data/
10.file  1.file  1.txt  2.file  3.file  4.file  5.file  6.file  7.file  8.file  9.file
```

```
[root@data3 ~]# ls /data/
10.file  1.file  1.txt  2.file  3.file  4.file  5.file  6.file  7.file  8.file  9.file
```

➢ **扩展：Gluster 的管理命令；**

Gluster peer status	##查看所有的节点信息
Gluster peer probe name	##添加节点
Gluster peer detach name	##删除节点
Gluster volume create xxx	##创建卷
Gluster volume info	##查看卷信息