## glusterfs分布式文件系统

应用场景： 大数据存储、云平台数据存储

glusterfs用于实现分布式存储, 应用于云平台数据的存储、视频流数据存储、集群的共享存储

## glusterfs基本概念

1、brick 集群中节点提供的挂载点目录

2、volume 卷 提供给前端应用服务器的虚拟存储空间

## glusterfs的特性：

1. PB级容量，数千个节点
2. 高可用性
3. 提升读/写性能
4. 基于文件系统级别共享
5. 无metadata(元数据)的存储方式, 使用弹性hash算法实现数据的定位
6. 支持多种挂载方式[FUSE, NFS]

## 部署glusterfs集群

1、项目环境准备

1) 主机名称、IP地址

2) 解析主机名

3) 防火墙、SELinux

4) ntp时间同步

5) ssh密钥远程

6) 配置glusterfs源

2、在集群所有节点上安装glusterfs服务器端软件 ，启动glusterd服务

[root@node01 ~]# for i in 101 102 103 104 105

> do

> ssh 192.168.122.$i yum install -y glusterfs-server glusterfs-fuse glusterfs

> done

[root@node01 ~]# for i in 101 102 103 104 105

> do

> ssh 192.168.122.$i systemctl start glusterd

> ssh 192.168.122.$i systemctl enable glusterd

> done

3、在客户端上安装gluster客户端软件

[root@app\_server ~]# yum install -y glusterfs glusterfs-fuse

4、创建gluster集群【任意节点】

[root@node01 ~]# gluster peer probe node02

peer probe: success.

[root@node01 ~]# gluster peer probe node03

peer probe: success.

[root@node01 ~]# gluster peer probe node04

peer probe: success.

[root@node01 ~]# gluster peer probe node05

peer probe: success.

[root@node01 ~]# gluster peer status

Number of Peers: 4

Hostname: node02

Uuid: 2598d2d9-d387-47cc-8899-d26b905e6758

State: Peer in Cluster (Connected)

Hostname: node03

Uuid: ee6ce928-c65f-4e1e-a994-c8024c878fd1

State: Peer in Cluster (Connected)

Hostname: node04

Uuid: 3a2e41eb-e2fe-4b3f-b9b9-f1cd0a8d71fd

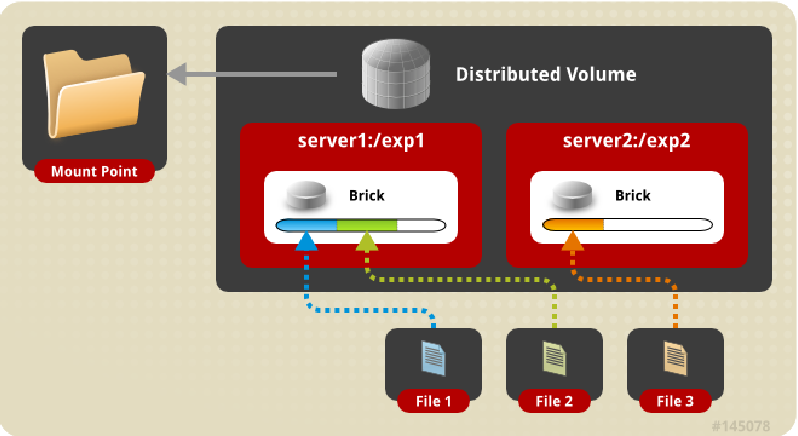
State: Peer in Cluster (Connected)

Hostname: node05

Uuid: 75d1b96d-bb8a-45db-a04a-e29624f91ff3

State: Peer in Cluster (Connected)

## 分布式卷



以整个文件为单位，不同的文件分散存储在不同的brick上，适用于存储大量小文件

无brick数量的限制

默认类型

卷容量 === 所有brick容量之和

提升数据读写速度，无可靠性

[root@node01 ~]# gluster volume create datav1 \

> node01:/data1/br1 \

> node02:/data1/br1

volume create: datav1: success: please start the volume to access data

[root@node01 ~]#

[root@node01 ~]# gluster volume start datav1

[root@node01 ~]# gluster volume info datav1

Volume Name: datav1

Type: Distribute

Volume ID: c9516cde-41ef-4aeb-92ef-11373f64e831

Status: Started

Snapshot Count: 0

Number of Bricks: 2

Transport-type: tcp

Bricks:

Brick1: node01:/data1/br1

Brick2: node02:/data1/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

客户端使用卷

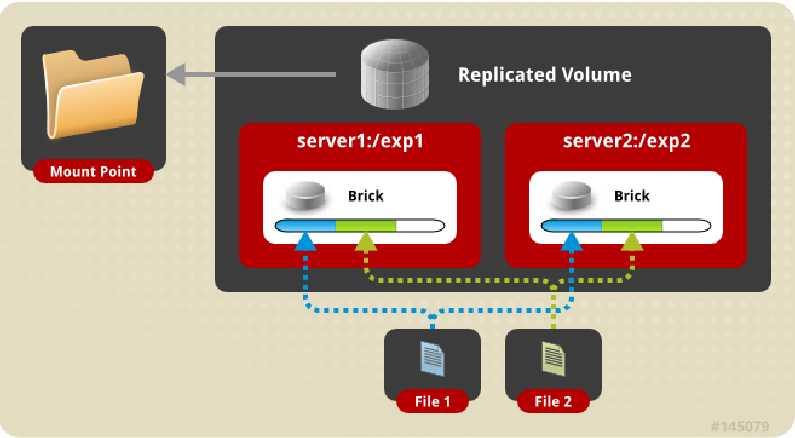
[root@app\_server ~]# mkdir /test

[root@app\_server ~]# mount -t glusterfs node01:/datav1 /test

自动挂载：

node01:/datav1 /test glusterfs defaults,\_netdev 0 0

## 复制卷



每个文件会被复制为brick数量份，分散存储

创建时需要使用参数replica指定文件被复制的份数，该数字要和brick数量一致

提供文件可靠性

[root@node01 ~]# gluster volume create datav2 replica 2 \

> node01:/data2/br1 \

> node02:/data2/br1

volume create: datav2: success: please start the volume to access data

[root@node01 ~]#

[root@node01 ~]# gluster volume start datav2

volume start: datav2: success

[root@node01 ~]# gluster volume info datav2

Volume Name: datav2

Type: Replicate

Volume ID: 79e207c9-ee93-435c-9b11-335999961277

Status: Started

Snapshot Count: 0

Number of Bricks: 1 x 2 = 2

Transport-type: tcp

Bricks:

Brick1: node01:/data2/br1

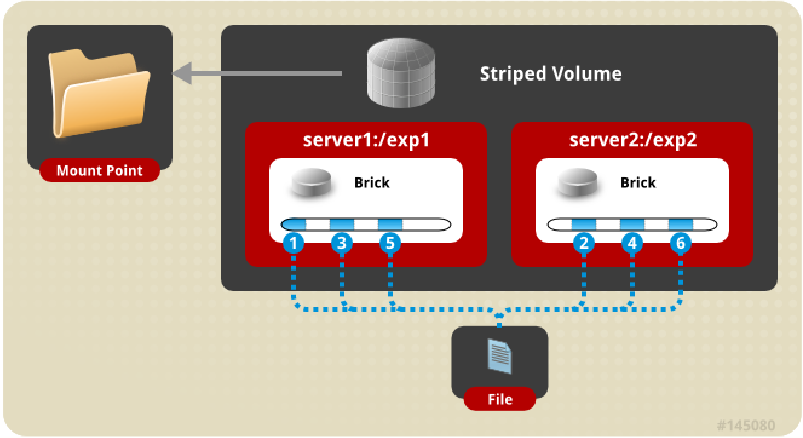
Brick2: node02:/data2/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

## 条带卷



提升读写性能，适用于大文件的存储

创建时通过stripe的参数指定文件被条带的次数, 该数量要和brick数量一致

[root@node01 ~]# gluster volume create datav3 stripe 2 node01:/data3/br1 node02:/data3/br1

volume create: datav3: success: please start the volume to access data

[root@node01 ~]# gluster volume start datav3

volume start: datav3: success

[root@node01 ~]# gluster volume info datav3

Volume Name: datav3

Type: Stripe

Volume ID: 182825bc-8628-4925-a8fd-53f3b102b41f

Status: Started

Snapshot Count: 0

Number of Bricks: 1 x 2 = 2

Transport-type: tcp

Bricks:

Brick1: node01:/data3/br1

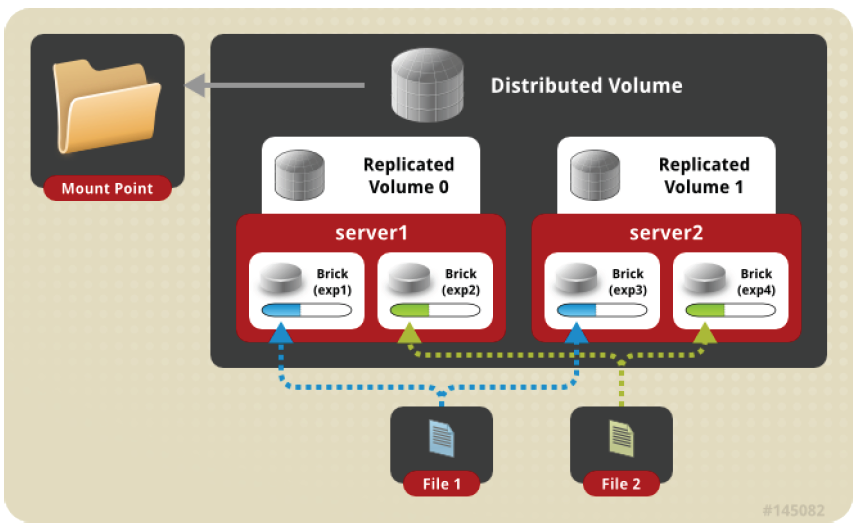
Brick2: node02:/data3/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

## 分布复制卷



存储大量的小文件，并提升文件的可靠性

brick数量是replica参数的复制数的整倍数

[root@node01 ~]# gluster volume create datav4 replica 2 \

> node01:/data4/br1 \

> node02:/data4/br1 \

> node03:/data4/br1 \

> node04:/data4/br1

volume create: datav4: success: please start the volume to access data

[root@node01 ~]#

[root@node01 ~]# gluster volume start datav4

volume start: datav4: success

[root@node01 ~]#

[root@node01 ~]# gluster volume info datav4

Volume Name: datav4

Type: Distributed-Replicate

Volume ID: c405c60a-6be7-4a4f-aabf-771c97b8812d

Status: Started

Snapshot Count: 0

Number of Bricks: 2 x 2 = 4

Transport-type: tcp

Bricks:

Brick1: node01:/data4/br1

Brick2: node02:/data4/br1

Brick3: node03:/data4/br1

Brick4: node04:/data4/br1

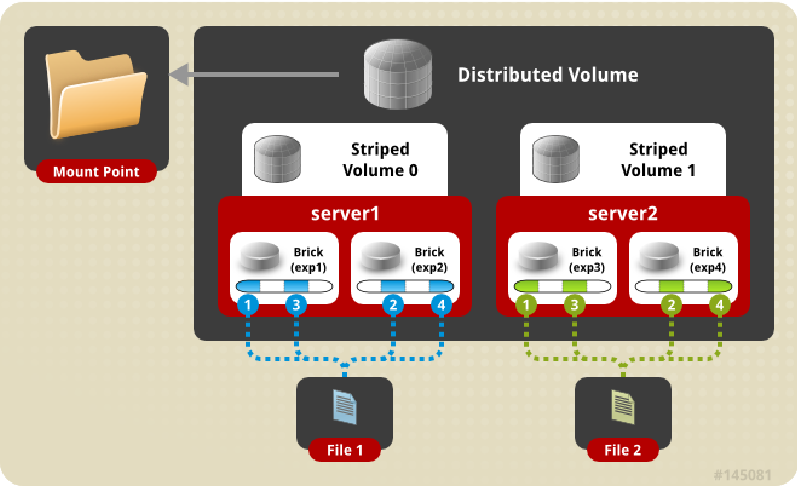
Options Reconfigured:

transport.address-family: inet

nfs.disable: on

[root@node01 ~]#

## 分布条带卷



brick 数是 stripe 的倍数

[root@node01 ~]# gluster volume create datav5 stripe 2 \

> node01:/data5/br1 \

> node02:/data5/br1 \

> node03:/data5/br1 \

> node04:/data5/br1

volume create: datav5: success: please start the volume to access data

[root@node01 ~]#

[root@node01 ~]# gluster volume start datav5

volume start: datav5: success

[root@node01 ~]# gluster volume info datav5

Volume Name: datav5

Type: Distributed-Stripe

Volume ID: 53e5c4d7-0b60-466a-acf7-4e4a8d3b902a

Status: Started

Snapshot Count: 0

Number of Bricks: 2 x 2 = 4

Transport-type: tcp

Bricks:

Brick1: node01:/data5/br1

Brick2: node02:/data5/br1

Brick3: node03:/data5/br1

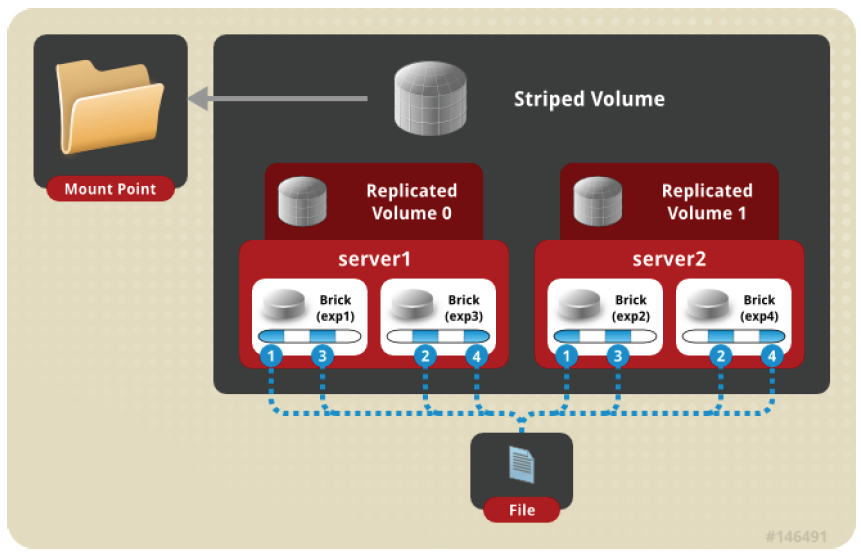
Brick4: node04:/data5/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

## 条带复制卷



适用于大文件的存储，提升文件的读写；提升可靠性

brick数量为stripe条带数、replica复制数的乘积

[root@node01 ~]# gluster volume create datav6 stripe 2 replica 2 \

> node01:/data6/br1 \

> node02:/data6/br1 \

> node03:/data6/br1 \

> node04:/data6/br1 \

>

volume create: datav6: success: please start the volume to access data

[root@node01 ~]# gluster volume start datav6

volume start: datav6: success

[root@node01 ~]#

[root@node01 ~]# gluster volume info datav6

Volume Name: datav6

Type: Striped-Replicate

Volume ID: bcaa5a53-8607-4161-ba9d-eeb692710afe

Status: Started

Snapshot Count: 0

Number of Bricks: 1 x 2 x 2 = 4

Transport-type: tcp

Bricks:

Brick1: node01:/data6/br1

Brick2: node02:/data6/br1

Brick3: node03:/data6/br1

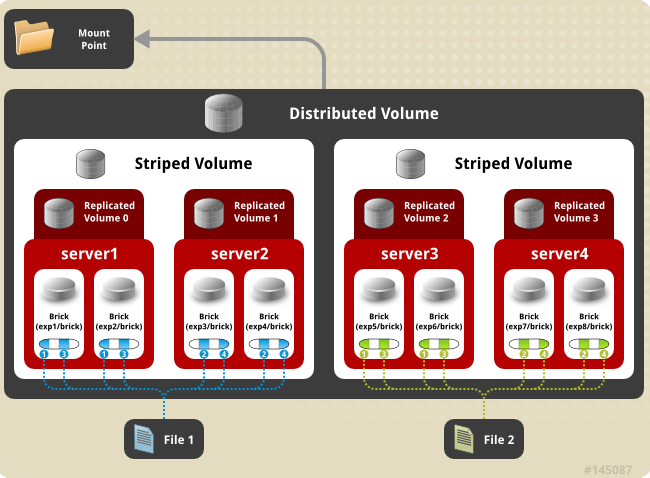
Brick4: node04:/data6/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

## 分布条带复制卷



brick数量为stripe, replica的乘积的整倍数

## 卷管理

### 1、扩容卷

注意：

扩展分布式复制卷、分布式条带卷时，添加的brick数量为replica、stripe参数的整数倍

示例01：扩展分布式卷

1、添加新的brick

[root@node01 ~]# gluster volume add-brick datav1 node05:/data1/br1

volume add-brick: success

2、查看验证

[root@node01 ~]# gluster volume info datav1

Volume Name: datav1

Type: Distribute

Volume ID: c9516cde-41ef-4aeb-92ef-11373f64e831

Status: Started

Snapshot Count: 0

Number of Bricks: 3

Transport-type: tcp

Bricks:

Brick1: node01:/data1/br1

Brick2: node02:/data1/br1

Brick3: node05:/data1/br1

3、重分布卷，保证新的文件能分布到的新的brick上【业务空闲时】

[root@node01 ~]# gluster volume rebalance datav1 start

volume rebalance: datav1: success: Rebalance on datav1 has been started successfully. Use rebalance status command to check status of the rebalance process.

ID: d82e837a-fc93-49ca-9125-2dca298b9c26

[root@node01 ~]#

[root@node01 ~]# gluster volume rebalance datav1 status

Node Rebalanced-files size scanned failures skipped status run time in h:m:s

--------- ----------- ----------- ----------- ----------- ----------- ------------ --------------

localhost 19 0Bytes 55 0 0 completed 0:00:02

node02 33 0Bytes 74 0 0 completed 0:00:04

node05 0 0Bytes 1 0 0 completed 0:00:00

volume rebalance: datav1: success

[root@node01 ~]#

### 2、缩减卷

1、迁移node05 brick上的数据

[root@node01 ~]# gluster volume remove-brick datav1 node05:/data1/br1 start

volume remove-brick start: success

ID: 12139274-2174-403d-b31a-e0fd4bc8d26d

2、查看数据迁移完成

[root@node01 ~]# gluster volume remove-brick datav1 node05:/data1/br1 status

Node Rebalanced-files size scanned failures skipped status run time in h:m:s

--------- ----------- ----------- ----------- ----------- ----------- ------------ --------------

node05 52 0Bytes 52 0 0 completed 0:00:02

3、确认移除brick

[root@node01 ~]# gluster volume remove-brick datav1 node05:/data1/br1 commit

Removing brick(s) can result in data loss. Do you want to Continue? (y/n) y

### 3、替换故障卷

示例01：替换复制卷

[root@node01 ~]# gluster volume replace-brick datav2 node02:/data2/br1 node05:/data2/br1 commit force

[root@node01 ~]# gluster volume info datav2

Volume Name: datav2

Type: Replicate

Volume ID: 79e207c9-ee93-435c-9b11-335999961277

Status: Started

Snapshot Count: 0

Number of Bricks: 1 x 2 = 2

Transport-type: tcp

Bricks:

Brick1: node01:/data2/br1

Brick2: node05:/data2/br1

Options Reconfigured:

transport.address-family: inet

nfs.disable: on

示例02：替换分布式卷

1、添加新brick

2、移除旧brick, 确认数据移动完成后，再commit

### 4、设置卷参数

# gluster volume set 卷名称 参数名称 值

示例01：根据客户端IP地址进行访问控制

针对fuse的方式：

auth.allow

auth.reject

针对nfs方式：

nfs.rpc-auth-reject

nfs.rpc-auth-allow

[root@node01 ~]# gluster volume set datav1 auth.reject 192.168.122.106

volume set: success

auth.allow 仅允许哪些客户端挂载使用

示例02：客户端通过NFS的方式挂载卷

[root@node01 ~]# gluster volume set datav2 nfs.disable off

https://docs.gluster.org/en/latest/Administrator%20Guide/Managing%20Volumes/#tuning-options

其他属性说明:

1、nfs.addr-namelookup

关闭解析客户端主机名，默认值为On, On/Off

2、performance.cache-size

设置读缓存的大小，单位为MB

3、performance.cache-max-file-size

设置缓存的最大文件大小， 2^64 - 1 Bytes

4、performance.cache-min-file-size

设置缓存的最小文件大小 2^64 - 1 Bytes

5、performance.io-thread-count

设置io线程的数量，默认值为16，取值范围 0 --- 65

6、performance.write-behind-window-size

设置每个文件的buffer大小, 单位为MB