GlusterFS是一款非常易于使用的分布式文件存储系统，实现了全部标准POSIX接口，并用fuse实现虚拟化，让用户看起来就像是本地磁盘一样。因此程序想从本地磁盘切换到GlusterFS时是不用修改任何代码的，做到了无缝切换。并且让多台电脑的程序看起来在使用同一块硬盘，简化了很多逻辑。如果您的应用单机磁盘不够放时，不妨考虑下GlusterFS。

**一、    GlusterFS源码安装**

**1. glusterfs 依赖安装**

a. centos下yum安装

|  |  |
| --- | --- |
| 1 | yum install -y flex bison openssl-devel libacl-devel sqlite-devel libxml2-devel libtool automake autoconf gcc attr |

liburcu-bp需源码安装，yum源里面没有

先执行常规命令安装，进入源码目录后

|  |  |
| --- | --- |
| 1  2  3  4 | ./bootstrap  ./configure  make  sudo make install |

执行完常规安装命令后需要执行下面两个命令，才可以让系统找到urcu.

|  |  |
| --- | --- |
| 1  2 | sudo ldconfig  sudo pkg-config --libs --cflags liburcu-bp liburcu |

b. ubuntu下 apt-get安装

|  |  |
| --- | --- |
| 1 | sudo apt-get install flex bison libssl-dev libacl1-dev  libsqlite3-dev libxml2-dev liburcu-dev automake autoconf gcc attr |

d. 可选安装

此外如果要geo 复制功能，需要额外安装，并开启ssh服务：

|  |  |
| --- | --- |
| 1 | yum install -y passwd openssh-client openssh-server |

e. docker下安装额外操作

如果机器只有一台，又想测试集群，可以考虑用docker，然而docker对应用有些功能限制，所以不能直接使用，需要多操作一些。

①需要安装attr

|  |  |
| --- | --- |
| 1 | yum install attr -y |

②没有fuse的时候需要手动建一个

|  |  |
| --- | --- |
| 1 | mknod /dev/fuse c 10 229 |

③运行容器的时候需要提升权限

docker run --privileged=true

例如：

|  |  |
| --- | --- |
| 1  2  3 | sudo docker run --privileged=true -it -h glfs0 -v /dk/d0:/d --name=glfs0  gfs7:2  /bin/bash  或  sudo docker run --privileged=true -it --rm -v /dk/rm0:/d gfs7:2 /bin/bash |

④. 需要加载一个本地卷，将数据文件放在本地卷的目录中，否则磁盘的额外属性不能使用。

**2. glusterfs编译安装**

安装完以上依赖后，我们从官网 <http://www.gluster.org/download/> 下载源码，再编译glusterfs，gluserfs编译命令为常规命令，配置时加上--enable-debug表示编译为带debug信息的调试版本

|  |  |
| --- | --- |
| 1  2  3 | ./configure --prefix=/usr  make  sudo make install |

**二、    GlusterFS服务启停**

glusterfs的大部分命令都需要在root权限下运行，没有root权限会出现各种错误，因此我这里的命令前面都加了sudo，如果您直接用root登录，需免去sudo。

**1. 启动命令**

|  |  |
| --- | --- |
| 1  2  3  4  5 | sudo service glusterd start  或  sudo /etc/init.d/glusterd start  或  sudo glusterd |

**2. 停止命令**

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8 | sudo service glusterd stop  或  sudo /etc/init.d/glusterd stop  或  ps aux|grep glusterd  sudo kill xxxxxx-pid  或ubuntu下  sudo killall glusterd |

直接kill需要先停止各volume才比较安全。

**三、    集群关联**

1. 准备机器（或虚拟机、docker）若干台，我这里启动了4个docker，IP为172.17.0.2 ~ 172.17.0.5

2. 在每台机器上启动glusterFS服务，如上一节。

3. 得到每一台机器的ip或hostname

4. 在第一台机器(172.17.0.2)上执行关联命令，

|  |  |
| --- | --- |
| 1  2  3  4 | sudo gluster peer probe 172.17.0.3  sudo gluster peer probe 172.17.0.4  sudo gluster peer probe 172.17.0.5  ...... |

这样所有机器已经连在一起，注意该命令的意思相当于集群邀请某人加入自己的组织。

**四、    卷/volume操作**

**1．创建volume**

a. 单磁盘，调试环境推荐

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name 172.17.0.2:/d/disk0 |

b. 多磁盘，无raid，试验、测试环境推荐。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 |

c. 多磁盘，有raid1。线上高并发环境推荐。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name replica 2 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 |

注意：以上命令中，磁盘数量必须为复制份数的整数倍。

此外有raid0，raid10，raid5，raid6等方法，但是在线上小文件集群不推荐使用。

**2．启动volume**

刚创建好的volume还没有运行，需要执行运行命令方可使用。

|  |  |
| --- | --- |
| 1 | sudo gluster volume start vol\_name |

**3．挂载volume**

|  |  |
| --- | --- |
| 1  2 | sudo mkdir /local\_mount\_dir  sudo mount -t glusterfs -o acl 172.17.0.2:/vol\_name /local\_mount\_dir |

**4．使用GlusterFS**

a. 挂载了GlusterFS的某个卷后，就可以将其当做本地文件访问，代码中只需使用原生的文件api即可。这种方式使用不一定需要root权限，只要拥有对应目录或文件的权限即可。

b. 直接API方式，这种方式需要root权限才能使用，并且java、python、ruby的api包装目前都不够完整，一般情况不推荐使用。

**5．卸载volume**

卸载与挂载操作是一对。虽然没有卸载也可以停止volume，但是这样做是会出问题，如果集群较大，可能导致后面volume启动失败。

|  |  |
| --- | --- |
| 1 | sudo umount /local\_mount\_dir |

**6．停止volume**

停止与启动操作是一对。停止前最好先卸载所有客户端。

|  |  |
| --- | --- |
| 1 | sudo gluster volume stop vol\_name |

**7．删除volume**

删除与创建操作是一对。删除前需要先停止volume。在生产上一般不会删除volume

|  |  |
| --- | --- |
| 1 | sudo gluster volume delete vol\_name |

**8．在线修复**

当某块磁盘损坏后，需要换一块新的磁盘回去，这时集群中恰好还预留了备用磁盘，因此用备用磁盘替换损坏的磁盘，命令如下两条命令

|  |  |
| --- | --- |
| 1  2 | sudo gluster volume replace-brick vol\_name 172.17.0.3:/d/damaged\_disk 172.17.0.16:/d/new\_disk commit  sudo gluster volume heal vol\_name full |

**9．在线扩容**

随着业务的增长，集群容量不够时，需要添加更多的机器和磁盘到集群中来。

a. 普通情况只需要增加分布的广度就可以，增加的磁盘数量必须为最小扩容单元的整数倍，即replica×stripe，或disperse数的整数倍：

|  |  |
| --- | --- |
| 1 | sudo gluster volume add-brick vol\_name 172.17.0.11:/d/disk0 172.17.0.12:/d/disk0  172.17.0.13:/d/disk0  172.17.0.14:/d/disk0 |

该方法执行完后，需要新增的磁盘可能还没有被实际使用，这时需要平衡数据：

|  |  |
| --- | --- |
| 1 | sudo gluster volume rebalance vol\_name start |

b. 当集群达到一定规模，希望增加备份数时，增加的磁盘数量必须为原分布数量的整数倍。gluster volume info中看到的第一个值，这时需要增加一个参数让系统知道是修改了数据的备份数。假设原先的replica是2，想要改为3，命令如下：

|  |  |
| --- | --- |
| 1 | sudo gluster volume add-brick vol\_name replica 3 172.17.0.11:/d/disk0 172.17.0.12:/d/disk0  172.17.0.13:/d/disk0  172.17.0.14:/d/disk0 |

执行完add-brick命令后，新增的磁盘还没有被实际使用，且系统不会自动复制，这时需要修复数据，让系统达到新指定的备份数

|  |  |
| --- | --- |
| 1 | sudo gluster volume heal vol\_name full |

注意：一次只增加一个备份，如果一次增加多个备份，目前版本可能出错。

**10．在线收缩**

可能原先配置比例不合理，打算将部分存储机器用于其他用途时，跟扩容一样，也分两种情况。

a. 降低分布广度，移除的磁盘必须是一整个或多个存储单元，在volume info的结果列表中是连续的多块磁盘。该命令会自动均衡数据。

|  |  |
| --- | --- |
| 1 | sudo gluster volume remove-brick vol\_name 172.17.0.11:/d/disk0 172.17.0.12:/d/disk0  172.17.0.13:/d/disk0  172.17.0.14:/d/disk0 start |

启动后需要查看删除的状态，实际是自动均衡的状态，直到状态从in progress变为completed。

|  |  |
| --- | --- |
| 1 | sudo gluster volume remove-brick vol\_name 172.17.0.11:/d/disk0 172.17.0.12:/d/disk0  172.17.0.13:/d/disk0  172.17.0.14:/d/disk0 status |

状态显示执行完成后，提交该移除操作。

|  |  |
| --- | --- |
| 1 | sudo gluster volume remove-brick vol\_name commit |

b. 降低备份数，移除磁盘必须是符合要求（好难表达）。在volume info的结果列表中一般是零散的多块磁盘（ip可能是连续的）。该命令不需要均衡数据。

|  |  |
| --- | --- |
| 1 | sudo gluster volume remove-brick vol\_name replica 2 172.17.0.11:/d/disk0 172.17.0.12:/d/disk0  172.17.0.13:/d/disk0  172.17.0.14:/d/disk0 force |

降低备份数时，只是简单删除，而且命令最后用的也是force参数，如果原先系统数据没有复制好，那么也就会出现部分丢失。因此该操作需要极其谨慎。必须先保证数据完整，执行sudo gluster volume heal vol\_name full命令修复，并执行sudo gluster volume heal vol\_name info，和 sudo gluster volume status检查，确保数据正常情况下再进行。

**11．配额设定**

a. 一个volume经常会让多个系统去同时使用，这时为了方便管理，可以为一级或二级目录加上磁盘配额，避免因某个系统的过量使用，而影响其他系统的正常使用。

|  |  |
| --- | --- |
| 1  2  3 | sudo gluster volume quota vol\_name enable  sudo gluster volume quota vol\_name limit-usage /srv\_a 10GB  sudo gluster volume quota vol\_name limit-usage /srv\_b 200MB |

b. 查看当前配额使用量，会以相当直观的列表展示。

|  |  |
| --- | --- |
| 1 | sudo gluster volume quota vol\_name list |

c. 去掉某个目录的配额，

|  |  |
| --- | --- |
| 1 | sudo gluster volume quota vol\_name remove /srv\_a |

d. 停止配额，该方法慎用，否则会全部清除，往往不是自己想要的结果，因为重新enable后，原先设定的配额都已消失。当然，如果打算重新配置所有目录时则比较合适。

|  |  |
| --- | --- |
| 1 | sudo gluster volume quota vol\_name disable |

e. 如果系统不打算将所有磁盘都用于GlusterFS，那么可以在根目录上设置配额。考虑到glusterFS不能充分利用所有的磁盘空间，因此最好将大小设置的比实际空间稍小。

|  |  |
| --- | --- |
| 1 | sudo gluster volume quota vol\_name limit-usage / 100TB |

f. 并且想将这个配额当做磁盘的大小使用，需要执行如下命令，这样df时显示的磁盘大小就是配额了。配额使用的是1024进制的，而非磁盘的1000进制。当配额量超过磁盘量时，df也会显示配额量，因此一定不能这样设置。

|  |  |
| --- | --- |
| 1 | gluster volume set vol\_name quota-deem-statfs on |

以上配额是针对磁盘使用量，另外glusterFS提供文件数量的配额，limit-objects，list-object。可以根据场景使用。

磁盘配额功能gluster volume quota 目录容量达到目标大小时，不是马上生效，而是有一定的时间窗口，（若干秒），在这个时间内，数据还可以写入。这样的特性在配额比较大的时候并不影响，一般不会在短时间内超过太多。

**12．raid选型**

raid1：适合线上中小文件场景，创建命令如前文。

单磁盘，无raid，raid0三种方式只适合于实验环境，允许数据的丢失，一旦数据丢失，基本上需要从头来过。

raid0：适合大文件实验环境。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name stripe 3 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 172.17.0.6:/d/disk0 172.17.0.7:/d/disk0 |

raid10：适合大文件场景。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name replica 2 stripe 3 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 172.17.0.6:/d/disk0 172.17.0.7:/d/disk0 |

raid5，raid6等，一般不适合线上环境，但适合geo备份环境，因为是用软件的方式实现raid5等功能，因此cpu开销较大，而且一旦有磁盘损坏，计算的cpu开销更加大，如果在压力较大的线上环境跑，容易造成较大延迟。如果线上的读写压力很小，也可以考虑使用。

raid5：不很推荐，因为不够平衡，容错性太低，而且开销比较大。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name disperse 6 redundancy 1 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 172.17.0.6:/d/disk0 172.17.0.7:/d/disk0 |

raid6：可以使用，比raid5平衡，容错性比raid5高很多，开销只是稍大。

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name disperse 7 redundancy 2 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 172.17.0.6:/d/disk0 172.17.0.7:/d/disk0 172.17.0.8:/d/disk0 |

更为安全的离线geo备份集群raid推荐：（最大可以允许一半的磁盘损坏，具有极高容错性，数据可用性可达10个9）

|  |  |
| --- | --- |
| 1 | sudo gluster volume create vol\_name disperse 10 redundancy 5 172.17.0.2:/d/disk0 172.17.0.3:/d/disk0 172.17.0.4:/d/disk0 172.17.0.5:/d/disk0 172.17.0.6:/d/disk0 172.17.0.7:/d/disk0 172.17.0.8:/d/disk0 172.17.0.9:/d/disk0 172.17.0.10:/d/disk0 172.17.0.11:/d/disk0 |

**五、    系统特性**

**1．cache、数据一致性**

客户端带有cache，但是该cache并不具备数据一致性，cache的更新是定时更新，默认间隔1秒，也就是说，某个客户端修改或删除一个文件后，需要1秒钟以后，整个集群才能全部感知到，这一秒钟内会存在数据的不一致。因此GlusterFS不适用于数据一致性要求强的数据。对于图片、语音等文件，在应用中做限定，不修改图片，只增加图片，这样数据的一致性问题就不会出现。

客户端的cache会带来性能的提升，因此当集群有一定规模时，合理的规划客户端访问的文件也是有必要的，能够增强cache的利用率。

**2．用户、权限**

GlusterFS所用的用户是linux自身的用户，linux用户有两个属性，一个是用户名，一个是用户号，linux在文件系统总标识一个文件的权限是用用户号的，而这个用户号可以在GlusterFS之间传递。比如某用户名user1，用户号1001，user1用户创建了文件A，权限是0600。

这时另外一台电脑，有用户名user1，用户号1002，这时该用户不能访问A文件。

但是该电脑有一个用户user3，用户号是1001，该用户号与前面的user1用户号相同，可以访问A文件。

为了让用户名和用户号不冲突，在创建系统用户时，指定一个用户号，并且是不容易被系统自动分配到的区间，这样在集群之间能使用一致的用户权限。

**六、    集群规模**

客户端或mount进程会跟所有brick连接，并且端口是小于1024的，因此brick数量一定不能超过1024，

在replica模式下，每个服务器有个glusterfs进程会以客户端形式连接每个brick，如果再在这些服务器上要进行mount，那么最大brick数小于500.

由于端口是使用有限的小于1024的端口，因此要注意保留部分常用的端口，如21,22,80,443端口。避免由于端口被占用导致重要服务无法启动。

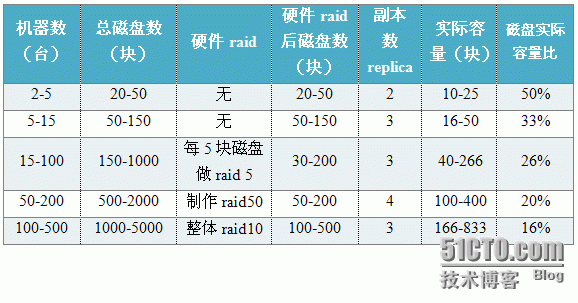
修改 /etc/sysctl.conf 文件，添加一行，具体端口更据需要设定。

|  |  |
| --- | --- |
| 1 | net.ipv4.ip\_local\_reserved\_ports=0-25,80,443 |

然后执行 sysctl 命令使其生效：

|  |  |
| --- | --- |
| 1 | sudo sysctl -p |

线上集群，不同规模下，集群配置如下表。假设表中的每台机器拥有12块磁盘，平均可用于存储的磁盘10块。另外2块用于安装系统、记录日志、备用等功能。

[](http://s5.51cto.com/wyfs02/M02/75/DB/wKioL1ZD_hGwjoY_AAB1BUUccJ0198.png)

系统根据不同时期，不断进行扩容，最大规模可达500台。

系统需要准备一定的备用磁盘或备用机器，以备磁盘或机器损坏时可以及时修复数据。当规模在2-15台时，准备1-3块磁盘作为备用；当规模在15-500台时，准备1-3台电脑作为备用。

系统集群较大时，要保证让数据的多个备份在不同的机器上，这样才可以在系统的某台机器宕机的情况，整个系统还是处于可用状态。

*来自 <*[*http://bangbangba.blog.51cto.com/3180873/1712061*](http://bangbangba.blog.51cto.com/3180873/1712061)*>*

CentOS 7 搭建GlusterFS  <http://www.linuxidc.com/Linux/2015-10/124331.htm>

使用GlusterFS作为KVM后端存储 <http://www.linuxidc.com/Linux/2013-08/89108.htm>

分布式存储系统GlusterFS初体验 <http://www.linuxidc.com/Linux/2013-08/89107.htm>

GlusterFS全局统一命名空间 <http://www.linuxidc.com/Linux/2013-08/89106.htm>

设计新Xlator扩展GlusterFS <http://www.linuxidc.com/Linux/2013-08/89105.htm>

GlusterFS Rebalance简析 <http://www.linuxidc.com/Linux/2013-08/89104.htm>

CentOS 6.0-x86\_64下体验Glusterfs  <http://www.linuxidc.com/Linux/2015-01/111211.htm>

*来自 <*[*http://www.linuxidc.com/Linux/2015-11/125106.htm*](http://www.linuxidc.com/Linux/2015-11/125106.htm)*>*