GlusterFS 系统中文管理手册

作者：黄崇远

时间：2013/11/17

类型	详细
备注	该文档是 GlusterFS 的系统管理指导手册，相应使用过程中的一些经验及一些容易犯错或者理解错误的问题提示。
相关描述	✧ 其他相关文档请参考新浪博客 http://blog.sina.com.cn/huangchongyuan ✧ 有任何其他想法，可以邮件 874450476@qq.com ✧ 文档及相关资料下载请到**个人 360 云盘** http://yunpan.cn/QGf2GDaRFpcDt 及百度文库、新浪爱问搜索。 ✧ 部分文档涉及到源码，有需要的博客留言，关注我的博客。 ✧ 欢迎加入 storm-分布式-IT 技术交流群（**191321336，群中有详细的资料**），一起讨论技术，一起分享代码，一起分享设计。

目录

1 文档说明

该文档主要内容出自 www.gluster.org 官方提供的英文系统管理手册《Gluster File System 3.3.0 Administration Guide》。可以看成是管理手册中文版吧。

区别在于：
(1)它是英文的，个人整理的是中文的。所以不想看英文文档的人可以参考参考~~！
(2)它是 3.3.0 的管理手册，个人系统管理使用实践是 3.4.1(不过 3.3.1 也实践过了)。
(3)它包含了很多内容，本文档只摘取其中系统管理部分(系统管理命令使用)，并且进行了归类整理，方便理解。
(4)另外附加了很多系统管理中很多需要注意的地方，我相信这个是比较重要的(血泪经验谈)。

写这个文档的目的一是给自己做个备份，毕竟自己看英文文档感觉也是很累的，需要的时候可以翻翻，另外就是使用实践的过程中却是遇到很多需要注意的地方，自己当时也吃亏了很多次才搞明白，所以希望这些能够对那些也是使用研究 GlusterFS 的人有些许的帮助。

看完这个文档，有时间还是去看官方文档《Gluster File System 3.3.0 Administration Guide》吧，呵呵，毕竟那才是正版，而且也比较全面。

2 系统维护管理

2.1 管理说明

在解释系统管理时会提供实例，首先给大家提供一个环境说明。所有管理实践都是在 3.4.1 的版本上做的，所以只针对该版本进行说明。

系统节点：

IP	别名	Brick		
192.168.2.100	server0	/mnt/sdb1	/mnt/sdc1	/mnt/sdd1
192.168.2.101	server1	/mnt/sdb1	/mnt/sdc1	/mnt/sdd1
192.168.2.102	server2	/mnt/sdb1	/mnt/sdc1	/mnt/sdd1

实践时个人使用自己的笔记本创建了三个节点，并每台虚拟机 mount 三块磁盘作为 Brick 使用，每个 brick 分配了 30G 的虚拟容量。

实例约定：

AFR 卷名： afr_vol
DHT 卷名： dht_vol
Stripe 卷名： str_vol
客户端挂载点： /mnt/gluster

2.2 系统部署

//先从如何部署一个完整的系统说起吧。

(1)在每个节点上启动 glusterd 服务

#service glusterd start

(2)添加节点到存储池，在其中一个节点上操作，如 server0

#gluster peer probe server1

#gluster peer probe server2

//可以使用 gluster peer status 查看当前有多少个节点，显示不包括该节点

(3)创建系统卷,部署最常见的分布式卷，在 server0 上操作

#gluster volume create dht_vol 192.168.2.{100,101,102}:/mnt/sdb1

//分别使用 server0/1/2 的磁盘挂载目录/mnt/sdb1 作为 brick

(4)启动系统卷，在 server0 上操作

#gluster volume start dht_vol

(5)挂载客户端，例如在 server2 上

#mount.glusterfs server0:/dht_vol /mnt/gluster

//将系统卷挂载到 server2 上的/mnt/gluster 目录下就可以正常使用了。该目录聚合了三个不同主机上的三块磁盘。

//从启动服务到提供全局名字空间，整个部署流程如上。

2.3 基本系统管理

2.3.1 节点管理

 # gluster peer command

2.3.1.1 节点状态

#gluster peer status　　//在 serser0 上操作，只能看到其他节点与 server0 的连接状态

Number of Peers: 2

Hostname: server1

Uuid: 5e987bda-16dd-43c2-835b-08b7d55e94e5

State: Peer in Cluster (Connected)

Hostname: server2

Uuid: 1e0ca3aa-9ef7-4f66-8f15-cbc348f29ff7

State: Peer in Cluster (Connected)

2.3.1.2 添加节点

 # gluster peer probe HOSTNAME

#gluster peer probe server2　　将 server2 添加到存储池中

2.3.1.3 删除节点

 # gluster peer detach HOSTNAME

#gluster peer detach server2　　将 server2 从存储池中移除

//移除节点时，需要确保该节点上没有 brick，需要提前将 brick 移除

2.3.2 卷管理

2.3.2.1 创建卷

 # gluster volume create NEW-VOLNAME [transport [tcp | rdma | tcp,rdma]] NEW-BRICK...

创建分布式卷(DHT)

#gluster volume create dht_vol 192.168.2.{100,101,102}:/mnt/sdb1

//DHT 卷将数据以哈希计算方式分布到各个 brick 上，数据是以文件为单位存取，基本达到分布均衡，提供的容量和为各个 brick 的总和。

创建副本卷(AFR)

#gluster volume create afr_vol replica 3 192.168.2.{100,101,102}:/mnt/sdb1

//AFR 卷提供数据副本，副本数为 replica，即每个文件存储 replica 份数，文件不分割，以文件为单位存储；副本数需要等于 brick 数；当 brick 数是副本的倍数时，则自动变化为 Replicated-Distributed 卷。

#gluster volume create afr_vol replica 2 192.168.2.{100,101,102}:/mnt/sdb1 192.168.2.{100,101,102}:/mnt/sdc1

//每两个 brick 组成一组，每组两个副本，文件又以 DHT 分布在三个组上，是副本卷与分布式卷的组合。

创建条带化卷(Stripe)

#gluster volume create str_vol stripe 3 192.168.2.{100,101,102}:/mnt/sdb1

//Stripe 卷类似 RAID0，将数据条带化，分布在不同的 brick，该方式将文件分块，将文件分成 stripe 块，分别进行存储，在大文件读取时有优势；stripe 需要等于 brick 数；当 brick 数等于 stripe 数的倍数时，则自动变化为 Stripe-Distributed 卷。

#gluster volume create str_vol stripe 3 192.168.2.{100,101,102}:/mnt/sdb1 192.168.2.{100,101,102}:/mnt/sdc1

//没三个 brick 组成一个组，每组三个 brick，文件以 DHT 分布在两个组中，每个组中将文件条带化成 3 块。

创建 Replicated-Stripe-Distributed 卷

#gluster volume create str_afr_dht_vol stripe 2 replica 2 192.168.2.{100,101,102}:/mnt/sdb1 192.168.2.{100,101,102}:/mnt/sdc1 192.168.2.{100,101}:/mnt/sdd1

//使用 8 个 brick 创建一个组合卷，即 brick 数是 stripe*replica 的倍数，则创建三种基本卷的组合卷，若刚好等于 stripe*replica 则为 stripe-Distrbuted 卷。

2.3.2.2 卷信息

#gluster volume info

//该命令能够查看存储池中的当前卷的信息，包括卷方式、包涵的 brick、卷的当前状态、卷名及 UUID 等。

2.3.2.3 卷状态

#gluster volume status

//该命令能够查看当前卷的状态，包括其中各个 brick 的状态，NFS 的服务状态及当前 task 执行情况，和一些系统设置状态等。

2.3.2.4 启动/停止卷

 # gluster volume start/stop VOLNAME

//将创建的卷启动，才能进行客户端挂载；stop 能够将系统卷停止，无法使用；此外 gluster 未提供 restart 的重启命令

2.3.2.5 删除卷

 # gluster volume delete VOLNAME

//删除卷操作能够将整个卷删除，操作前提是需要将卷先停止

2.3.3 Brick 管理

2.3.3.1 添加 Brick

若是副本卷，则一次添加的 Bricks 数是 replica 的整数倍；stripe 具有同样的要求。

 # gluster volume add-brick VOLNAME NEW-BRICK

#gluster volume add-brick dht_vol server3:/mnt/sdc1

//添加 server3 上的/mnt/sdc1 到卷 dht_vol 上。

2.3.3.2 移除 Brick

若是副本卷，则移除的 Bricks 数是 replica 的整数倍；stripe 具有同样的要求。

 # gluster volume remove-brick VOLNAME BRICK start/status/commit

#gluster volume remove-brick dht_vol start

//GlusterFS_3.4.1 版本在执行移除 Brick 的时候会将数据迁移到其他可用的 Brick 上，当数据迁移结束之后才将 Brick 移除。执行 start 命令，开始迁移数据，正常移除 Brick。

#gluster volume remove-brick dht_vol status

//在执行开始移除 task 之后，可以使用 status 命令进行 task 状态查看。

#gluster volume remove-brick dht_vol commit

//使用 commit 命令执行 Brick 移除，则不会进行数据迁移而直接删除 Brick，符合不需要数据迁移的用户需求。

PS：系统的扩容及缩容可以通过如上节点管理、Brick 管理组合达到目的。

(1)扩容时，可以先增加系统节点，然后添加新增节点上的 Brick 即可。

(2)缩容时，先移除 Brick，然后再进行节点删除则达到缩容的目的，且可以保证数据不丢失。

2.3.3.3 替换 Brick

 # gluster volume replace-brick VOLNAME BRICKNEW-BRICK start/pause/abort/status/commit

#gluster volume replace-brick dht_vol server0:/mnt/sdb1 server0:/mnt/sdc1 start

//如上，执行 replcace-brick 卷替换启动命令，使用 start 启动命令后，开始将原始 Brick 的数据迁移到即将需要替换的 Brick 上。

#gluster volume replace-brick dht_vol server0:/mnt/sdb1 server0:/mnt/sdc1 status

//在数据迁移的过程中，可以查看替换任务是否完成。

#gluster volume replace-brick dht_vol server0:/mnt/sdb1 server0:/mnt/sdc1 abort

//在数据迁移的过程中，可以执行 abort 命令终止 Brick 替换。

#gluster volume replace-brick dht_vol server0:/mnt/sdb1 server0:/mnt/sdc1 commit

//在数据迁移结束之后，执行 commit 命令结束任务，则进行 Brick 替换。使用 volume info 命令可以查看到 Brick 已经被替换。

2.4 系统扩展维护

2.4.1 系统配额

2.4.1.1 开启/关闭系统配额

 # gluster volume quota VOLNAME enable/disable

//在使用系统配额功能时，需要使用 enable 将其开启；disable 为关闭配额功能命令。

2.4.1.2 设置(重置)目录配额

 # gluster volume quota VOLNAME limit-usage /directory limit-value

#gluster volume quota dht_vol limit-usage /quota 10GB

//如上，设置 dht_vol 卷下的 quota 子目录的限额为 10GB。

PS：这个目录是以系统挂载目录为根目录"/",所以/quota 即客户端挂载目录下的子目录 quota

2.4.1.3 配额查看

 # gluster volume quota VOLNAME list

 # gluster volume quota VOLNAME list /directory name

//可以使用如上两个命令进行系统卷的配额查看，第一个命令查看目的卷的所有配额设置，第二个命令则是执行目录进行查看。

//可以显示配额大小及当前使用容量，若无使用容量(最小 0KB)则说明设置的目录可能是错误的(不存在)。

2.4.2 地域复制(geo-replication)

 # gluster volume geo-replication MASTER SLAVE start/status/stop

地域复制是系统提供的灾备功能，能够将系统的全部数据进行异步的增量备份到另外的磁盘中。

#gluster volume geo-replication dht_vol 192.168.2.104:/mnt/sdb1 start

//如上，开始执行将 dht_vol 卷的所有内容备份到 2.104 下的/mnt/sdb1 中的 task，需要注意的是，这个备份目标不能是系统中的 Brick。

2.4.3 I/O 信息查看

Profile Command 提供接口查看一个卷中的每一个 brick 的 IO 信息。

#gluster volume profile VOLNAME start

//启动 profiling，之后则可以进行 IO 信息查看

#gluster volume profile VOLNAME info

//查看 IO 信息，可以查看到每一个 Brick 的 IO 信息

 # gluster volume profile VOLNAME stop

//查看结束之后关闭 profiling 功能

2.4.5 Top 监控

Top command 允许你查看 bricks 的性能例如：read, write, file open calls, file read calls, file write calls, directory open calls, and directory real calls

所有的查看都可以设置 top 数，默认 100

 # gluster volume top VOLNAME open [brick BRICK-NAME] [list-cnt cnt]

//查看打开的 fd

#gluster volume top VOLNAME read [brick BRICK-NAME] [list-cnt cnt]

//查看调用次数最多的读调用

#gluster volume top VOLNAME write [brick BRICK-NAME] [list-cnt cnt]

//查看调用次数最多的写调用

 # gluster volume top VOLNAME opendir [brick BRICK-NAME] [list-cnt cnt]

 # gluster volume top VOLNAME readdir [brick BRICK-NAME] [list-cnt cnt]

//查看次数最多的目录调用

 # gluster volume top VOLNAME read-perf [bs blk-size count count] [brick BRICK-NAME] [list-cnt cnt]

//查看每个 Brick 的读性能

 # gluster volume top VOLNAME write-perf [bs blk-size count count] [brick BRICK-NAME] [list-cnt cnt]

//查看每个 Brick 的写性能