GlusterFS

Подкопаев Антон, podkoav239@gmail.com
Алексеев Антон, anton.m.alexeyev@gmail.com

Computer Science Center

14 марта 2013

Распределенные файловые системы

- Доступ с многих хостов
- Инкапсуляция расположения файлов
- Реплики и отказоустойчивость
- Параллельный доступ
- Масштабируемость

Распределенные файловые системы. Основные компоненты

- Клиент
- Сервер данных
- Сервер метаданных

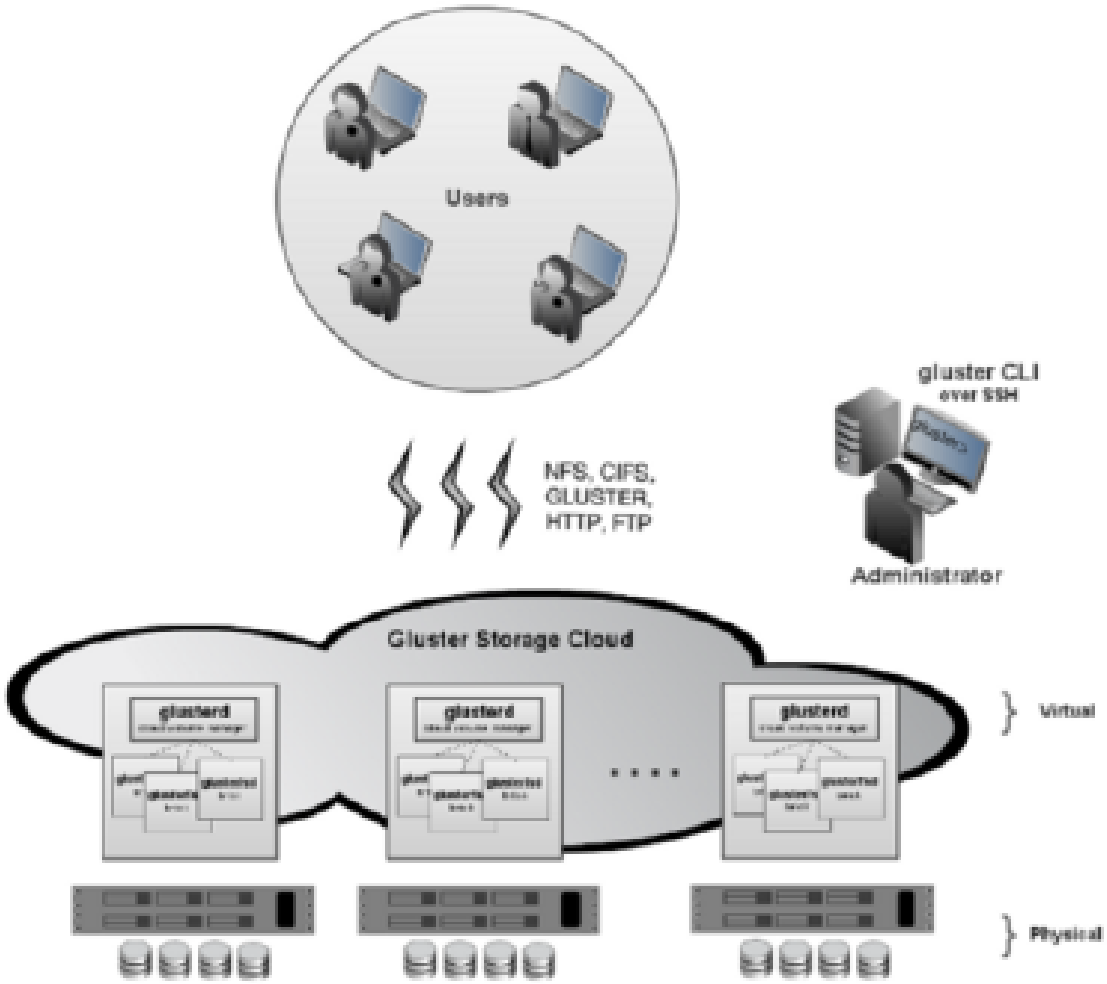
Распределенные файловые системы. Основные компоненты

- Клиент
- Сервер данных
- Сервер метаданных

GlusterFS:
Сервер данных в том числе выполняет и функции сервера метаданных

GlusterFS

- Сетевая фаловая система, работающая в user space
- До нескольких петабайт данных в одной точке монтирования



Термины GlusterFS

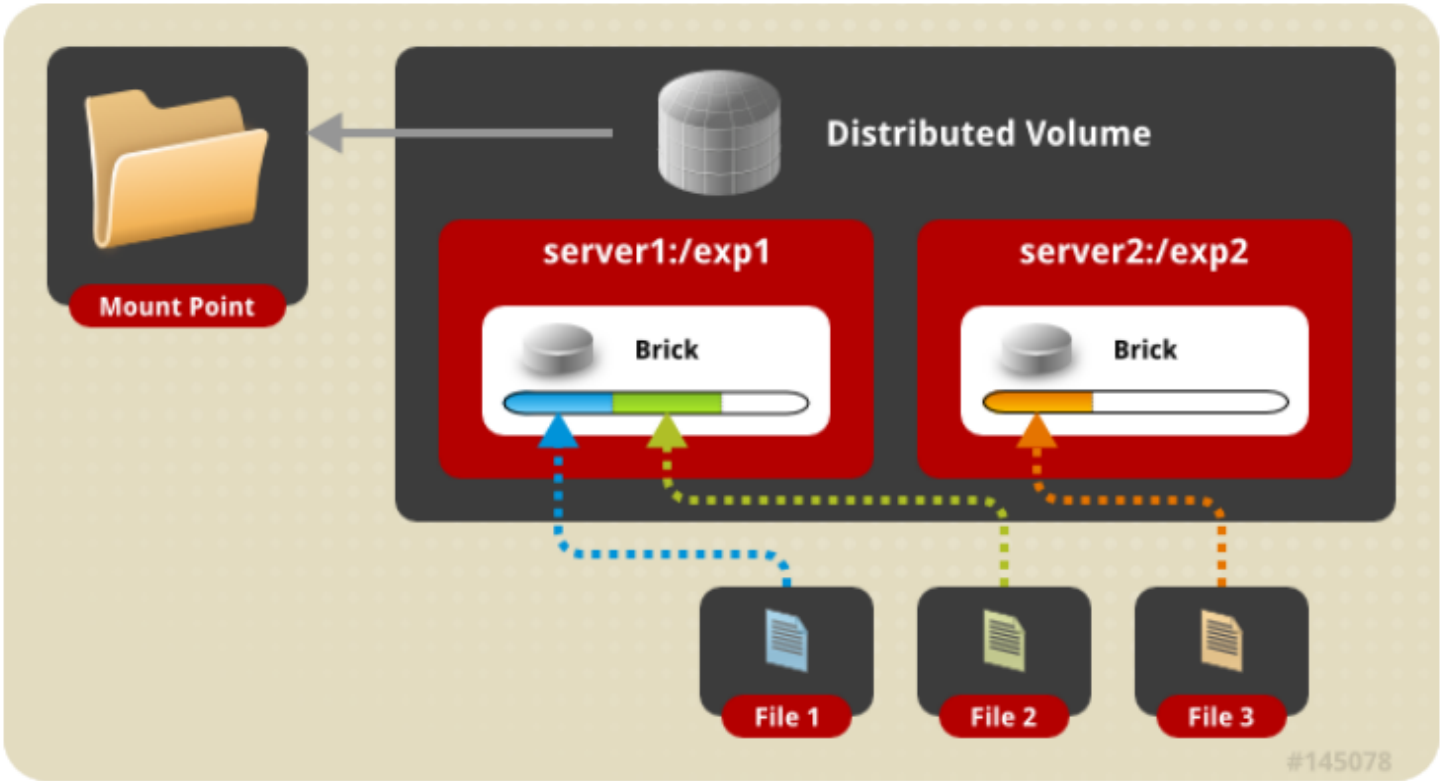
- Brick
- Логический диск
- Доверенные хранилища

Типы логических дисков

- Распределеные
- Реплицируемые
- Разделяющие

- Распределенные разделяющие
- Распределенные реплицируемые
- Разделяющие реплицируемые

Распределенные логические диски (1)



Распределенные логические диски (2)

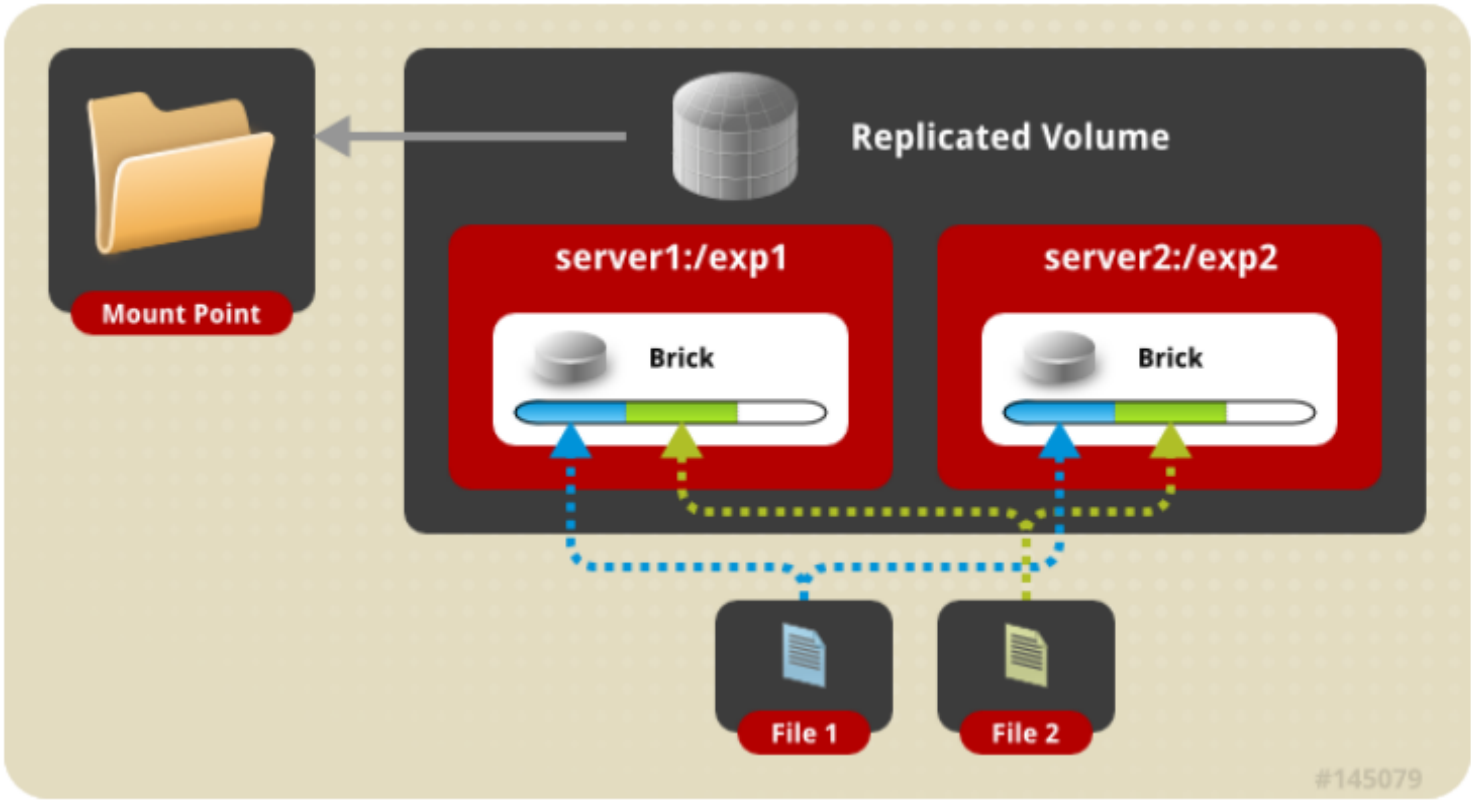
Плюсы

- Больше серверов => выше производительность при параллельном доступе
- Увеличение диска = добавление сервера (можно во время работы)

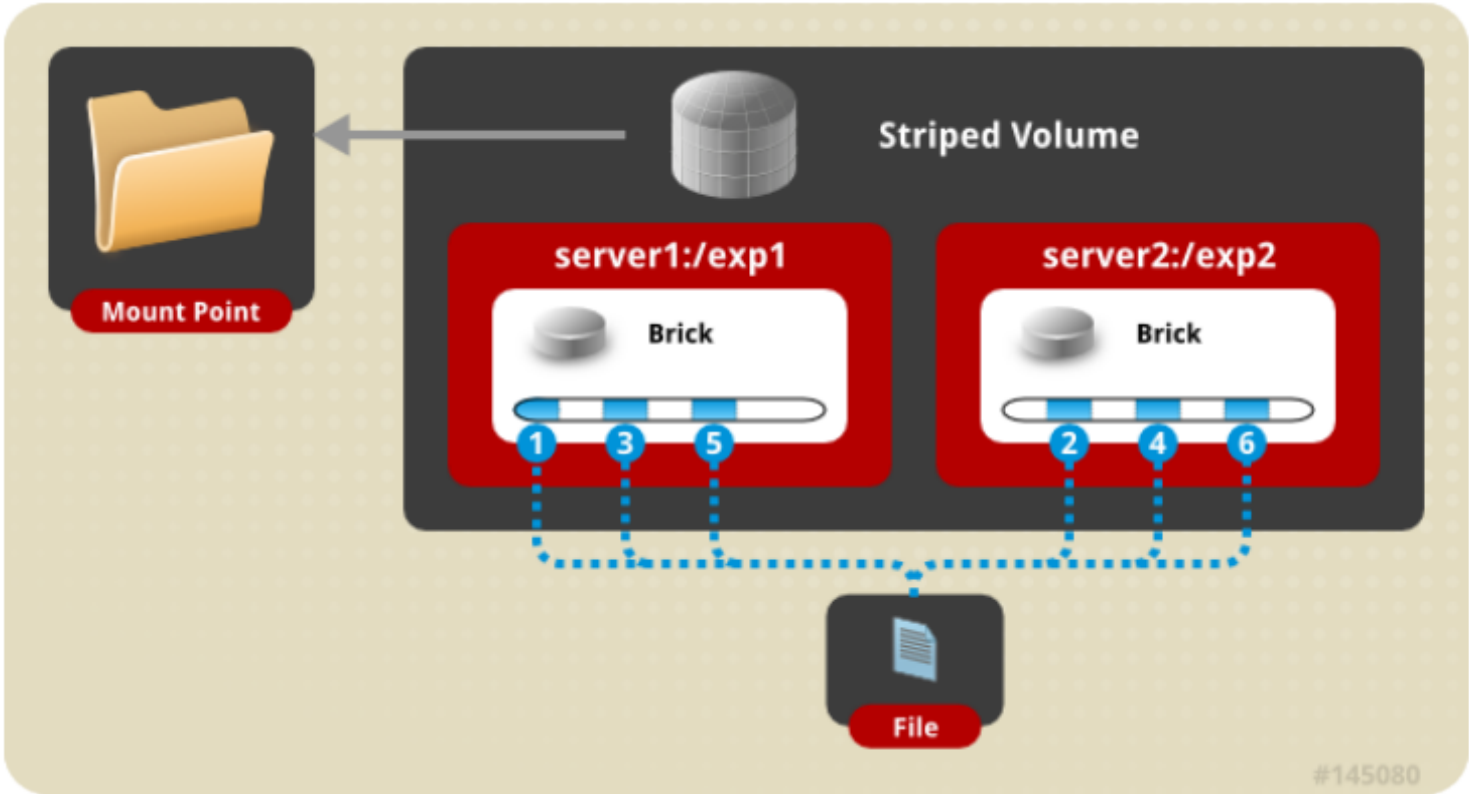
Минусы

- Потеря сервера = потеря данных на нем
- Файл не может быть больше размера узла
- Смена имени файла => дополнительное время на lookup

Реплицируемые логические диски



Разделяющие логические диски



Запуск GlusterFS (1)

```
apt-get install glusterfs-server
service glusterfs-server start
service glusterfs-server status
```

Запуск GlusterFS (1)

```
apt-get install glusterfs-server
service glusterfs-server start
service glusterfs-server status
```

```
mkfs.xfs disk-image
mount disk-image gluster_disk
```

Запуск GlusterFS (1)

```
apt-get install glusterfs-server
service glusterfs-server start
service glusterfs-server status
```

```
mkfs.xfs disk-image
mount disk-image gluster_disk
```

```
gluster peer probe 192.168.1.105

gluster volume create gv0 replica 2
192.168.1.114:/home/us1/gluster_disk
192.168.1.105:/home/us2/gluster_disk

gluster volume start gv0
gluster volume info
mount -t glusterfs 192.168.1.105:/gv1 /mnt
```

Запуск GlusterFS (2)

Распределенные

gluster volume create test-volume server1:/exp1 server2:/exp2

Реплицируемые

gluster volume create test-volume replica 2 server1:/exp1 server2:/exp2

Разделяющие

gluster volume create test-volume stripe 2 server1:/exp1 server2:/exp2

Добавление серверов

```
gluster peer probe new_server
```

Добавление серверов

```
gluster peer probe new_server
gluster volume add-brick vol_name new_brick
```

Добавление серверов

```
gluster peer probe new_server
gluster volume add-brick vol_name new_brick
gluster volume info
```

Добавление серверов

```
gluster peer probe new_server
gluster volume add-brick vol_name new_brick
gluster volume info
```

Для расширения реплицируемых (разделяющих) дисков надо добавлять количество серверов, кратное фактору реплики (разбиения)

Удаление серверов

```
gluster volume remove-brick vol_name new_brick start
```

Удаление серверов

```
gluster volume remove-brick vol_name new_brick start
gluster volume remove-brick vol_name new_brick status
```

Удаление серверов

```
gluster volume remove-brick vol_name new_brick start
gluster volume remove-brick vol_name new_brick status
gluster volume remove-brick vol_name new_brick commit
```

Рычажки

Все настройки осуществляются так

gluster volume set vol_name option param

Пример

gluster volume set myvolume performance.cache-size 256MB

Рычажки

auth.allow	IP-адреса клиентов, с джокерами: 192.168.1.
auth.reject	IP-адреса заблокированных клиентов
client.grace-timeout	время жизни замков, поставленных клиентом, после разрыва соединения, в секундах (10–1800)
cluster.self-heal-window-size	максимальное число блоков в файле, для которых self-heal происходит одновременно (0-1025)

Рычажки

cluster.data-self-heal-algorithm	тип self-heal: full (копирование всего файла), diff (копирование только несинхронизированных блоков), reset (если файла нет — полное копирование, если размер файла близок к размеру страницы диска — чтение несколькими операциями)

Рычажки

cluster.min-free-disk	сколько процентов диска должны оставаться незанятыми
cluster.stripe-block-size	размер страйпа (в байтах)
cluster.self-heal-daemon	включение / выключение фонового self-heal на репликах (on/off)
diagnostics.brick-log-level	уровень логов bricks: INFO, DEBUG, WARNING, ERROR, CRITICAL, NONE, TRACE
diagnostics.client-log-level	уровень логов клиентов
diagnostic.dump-fd-stats	статистика по операциям над файлами (on/off)

Рычажки

feature.read-only	монтирование диска в режиме read-only для всех клиентов (on/off)
features.lock-heal	self-healing замков при разрыве соединения (on/off)
features.quota-timeout	объём памяти для каждой директории кэшируется; можно указать макс. допустимое время пребывания любой директории в кэше (0-3600 сек.)

Рычажки

geo-replication.indexing	автоматическая синхронизация изменений от master к slave
network.frame-timeout	время, через которое операция объявляется «мёртвой», если сервер не отвечает
network.ping-timeout	время ожидания клиентом ответа от сервера (default: 42 сек.!); после разрыва все данные, связанные с клиентом, уничтожаются; при восстановлении соединения — восстановление всех ресурсов и замков. Очень, очень дорого. Не надо так делать.

Рычажки

nfs.volume-access	тип доступа для sub-volume (read-write/read-only)
nfs.trusted-write	запрещать клиенту коммитить, если произошёл UNSTABLE write; STABLE writes — синхронизируются (on/off)
nfs.trusted-sync	все записи и запросы на коммит считаются асинхронными (на момент ответа не гарантируется, что запись произведена на все диски)
nfs.rpc-auth-unix	включение аутентификации AUTH_UNIX (default: On)
nfs.rpc-auth-null	включение аутентификации AUTH_NULL (default: On)

Рычажки

nfs.rpc-auth-allow	IP-адреса и hostnames, которым разрешается устанавливать соединение с сервером
nfs.rpc-auth-reject	те, кому нельзя
nfs.ports-insecure	разрешать соединения непривилегированным хостам (default: off)
nfs.port	для связывания Gluster NFS с недефолтным номером порта (38465–38467)
nfs.disable	запретить экспорт диска NFS (default: off)

Рычажки

performance.io-thread-count	число потоков транслятора управляющего IO
performance.cache-max-file-size	max размер кэша транслятора (до $2^{64} - 1$ байт)
performance.cache-min-file-size	min размер кэша транслятора (от 0 байт)
performance.cache-refresh-timeout	интервал, с которым кэш будет обновляться (0-61 сек.)
performance.cache-size	размер кэша чтения (default: 32 MB)

Рычажки

server.allow-insecure	разрешать соединения непривилегированным хостам (default: on)
server.grace-timeout	время жизни замка после разрыва связи (10–1800 сек.)
server.statedump-path	путь к дампу состояния системы (default: /tmp brick'a)

Ссылки

- http://www.gluster.org/