Распределенные ФС

Подкопаев Антон, podkoav239@gmail.com

СП, СПбГУ

14 октября 2013

Файловая система

- Модель данных, программные компоненты, персистентные структуры данных и API
- Абстракция для доступа к данным, находящимся на физическом носителе
- Традиционная модель
  - Файл — объект с именем и некоторым содержанием
  - Каталог — список файлов и подкаталогов
  - Каталоги и файлы — пространство имен
  - Файл уникально идентифицируется путем

Метаинформация

Для ФС важна метаинформация — название файла, список блоков, время модификации, права доступа

Локальные ФС

- Некоторая интеграция с ядром ОС
- Данные на локальном HDD
- Блоки размером в несколько килобайт
- Кеширование страниц

Распределенная ФС

- Компоненты распределены по разным машинам
- Распределенность существенно влияет на принимаемые решения

Ваш К.О.

Компоненты РФС

- Клиент
  - API прикладных приложений и код для коммуникации с сервером
- Сервер данных
  - Содержимое файлов
- Сервер метаданных
  - Информация о местоположении файла и еще кое-что

Аспекты функционирования

- Прозрачность размещения файлов
- Совместный доступ
- Кеширование
- Репликация
- Масштабируемость

Прозрачность размещения файлов

- Прикладному ПО известен только путь
- Чем меньше информации о физическом местоположении, тем лучше

Совместный доступ и кеширование

- Централизованная ФС
  - Атомарные чтения и запись, блокировки, журналирование
- Распределенная ФС
  - Сетевые задержки, репликация усложняет жизнь

Варианты управления СД

- Синхронные чтение/запись
- Write-through cache
- Файлы неизменяемые после создания
- Append-only
- Уведомление о изменениях для клиентов, открывших файл
- Полноценные транзакции

Подкопаев Антон

Репликация

- Синхронная и асинхронная
- Политика согласованности реплик
- Запись в реплики

Масштабируемость

- Стремимся к линейной
  - было N дисков и K машин
  - стало на 2N данных — добавили N дисков, сохранили пропускную способность
  - добавили K машин — получили в два раза большую пропускную способность
- На практике есть препятствия
  - пропускная способность сети, сетевых интерфейсов серверов, производительсность сервера метаданных, блокировки

Подкопаев Антон

NFS у всех на слуху...

NFS

Подкопаев Антон

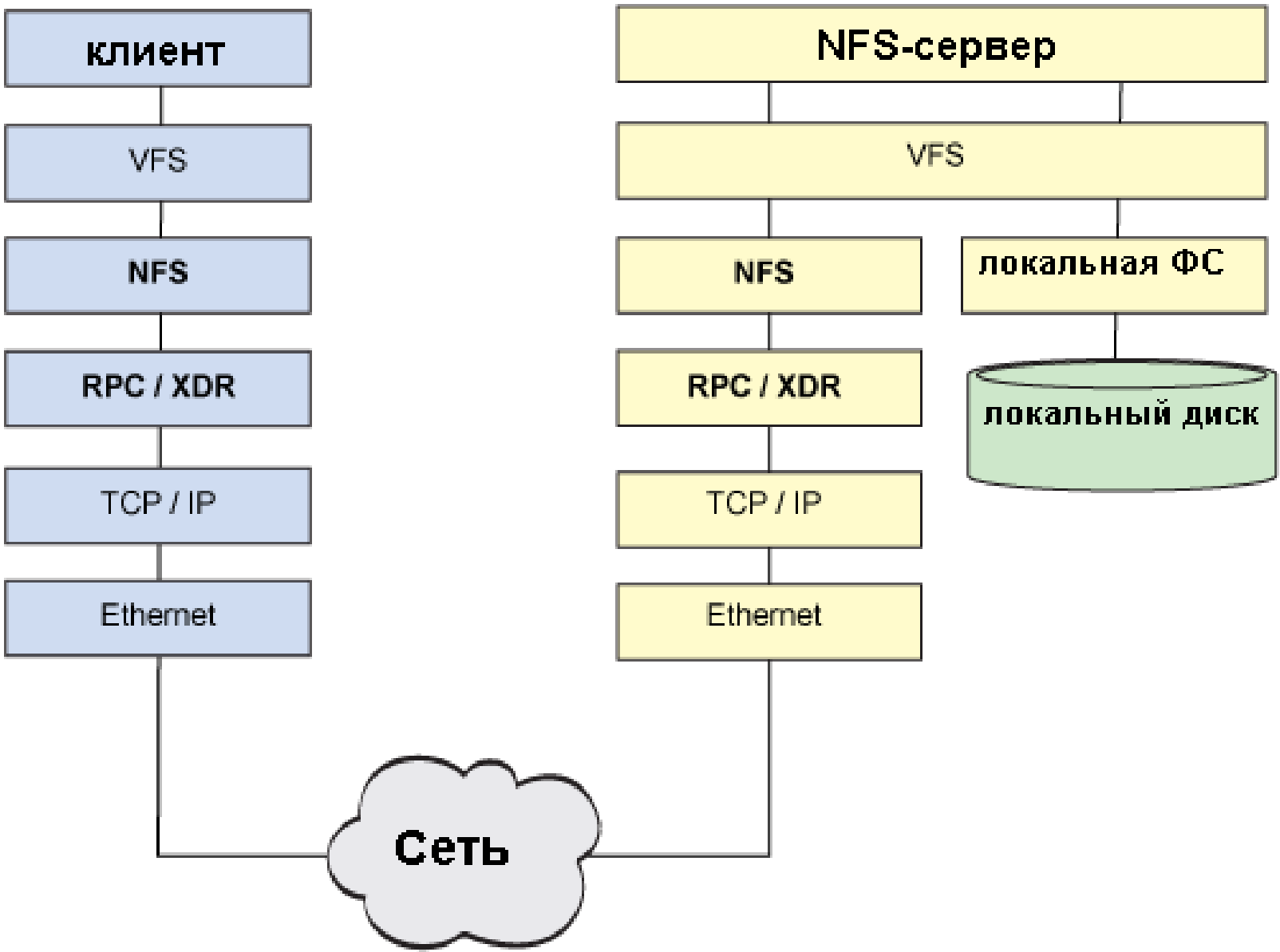
NFS

Подкопаев Антон

- Network File System
- Sun, 1984
- POSIX API
- На сервере NFS работает с интерфейсом файловой системы
- Поддержка блокировок и сессионного кеширования

NFS. Устройство

Подкопаев Антон



AFS

Подкопаев Антон

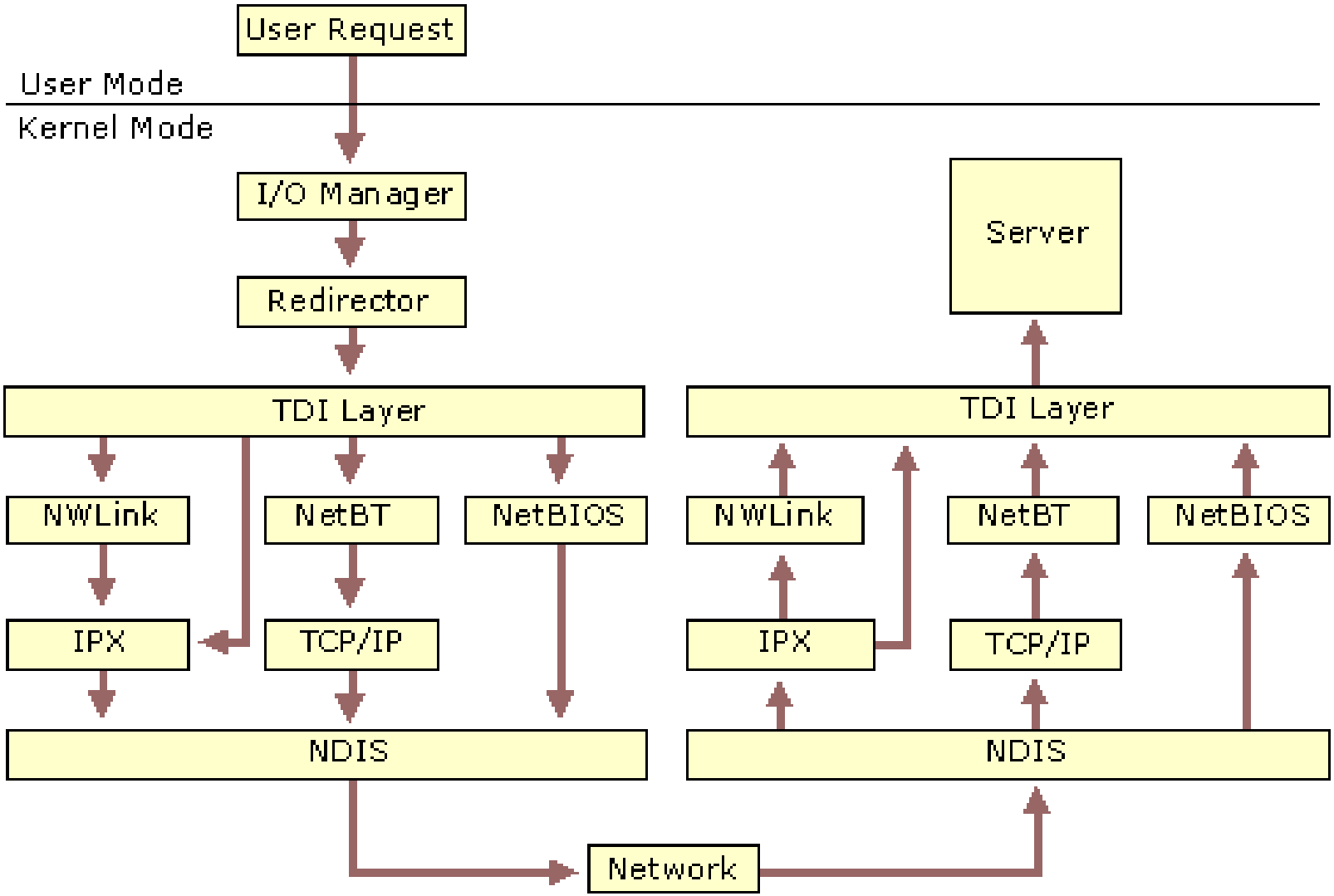
- Andrew File System
- Carnegie Mellon, 1980-е
- Сессионое кеширование, нотификации об изменениях, блокировки файлов
- Моментальные read-only снимки томов
- Не POSIX API

CIFS

- Common Internet File System
- ...aka SMB
  - Server Message Block
  - Samba
- ...aka Windows Shared Folders
- IBM, 1983
- Microsoft, 1996
- Уступающие блокировки, блокировки файлов
- Аутентификация коллективного доступа
- Уведомление об изменении каталога

CIFS. Устройство

Подкопаев Антон



GFS

- Google File System
- Начало 2000-ых

Предпосылки

- Большие файлы (N Gb) записываются и читаются пакетными процессами (creawler, indexer)
- Пропускная способность важнее случайного доступа
- Ширпотребные компьютеры

Концепции архитектуры

- Много файловых серверов, один активный сервер метаданных (мастер)
- Файлы хранятся фрагментами по 64 Mb
- Три реплики каждого фрагмента на различных файловых серверах
- Приоритетные операции с файлом
  - Большое последовательное чтение
  - Конкурентное наращивание
- Кеширование на клиенте не производится
- Не POSIX API

Развертывание GFS

- Ячейка — единица развертывания
- В ячейке один мастер и много файловых серверов
- Ячейка GFS соответствует физическому датацентру

Взаимодействие клиена и мастера при чтении

- Приложение собирается прочитать фрагмент
- GFS библиотека звонит мастеру, тот возвращает адреса реплик — файловых серверов, хранящих фрагмент
- GFS библиотека напрямую звонит одному из файловых серверов с просьбой вернуть нужный диапазон внутри данного фрагмента
- Дальше прямое общение клиента и файлового сервера

Требования меняются

- Google 2010-х годов — интерактивные приложения
- Файлы меньше в размерах и больше в количестве
- Требование во времени произвольного доступа жестче

Требования меняются

- Google 2010-х годов — интерактивные приложения
- Файлы меньше в размерах и больше в количестве
- Требование во времени произвольного доступа жестче

GFS в Google больше не используется, на смену пришел Colossus

Реализации, похожие на GFS

- Реализация Google File System закрыта
- Открытые проекты с аналогичной архитектурой
  - Apache HDFS: реализация на Java из проекта Hadoop
  - QFS: реализация на C++

GlusterFS

Подкопаев Антон

- OpenSource
- Сервер данных в том числе выполняет и функции сервера метаданных

Термины

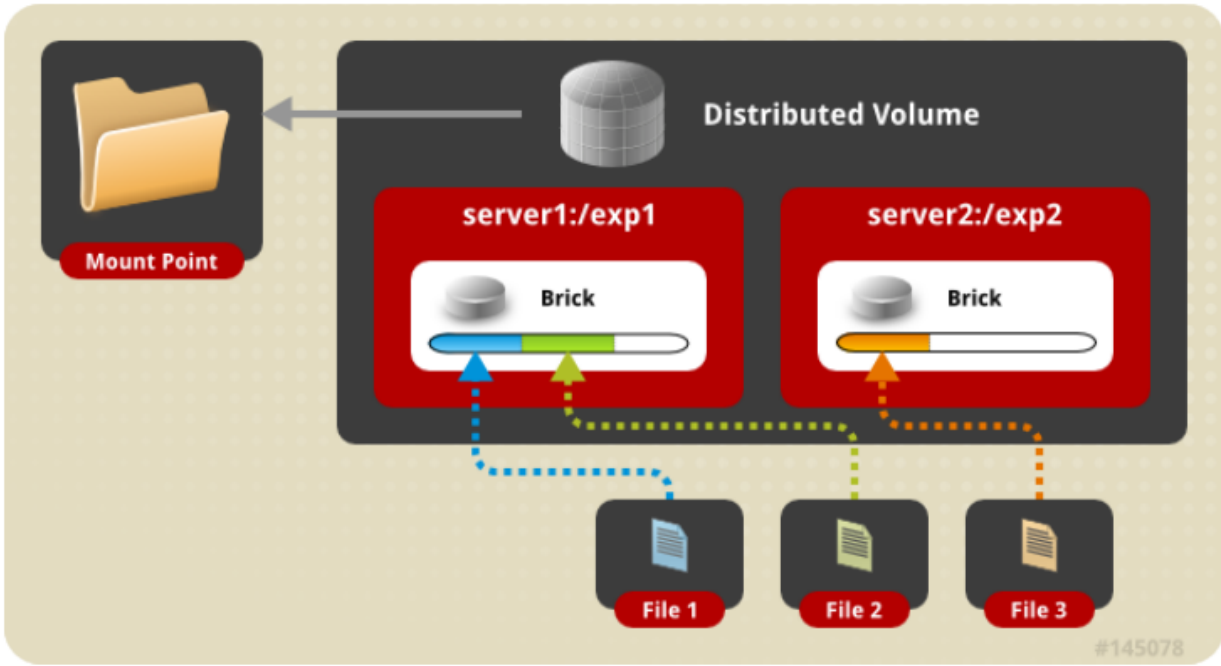
- Brick
- Логический диск
- Доверенные хранилища

Типы логических дисков

- Распределеные
- Реплицируемые
- Разделяющие

- Распределенные разделяющие
- Распределенные реплицируемые
- Разделяющие реплицируемые

Распределенные логические диски (1)



Распределенные логические диски (2)

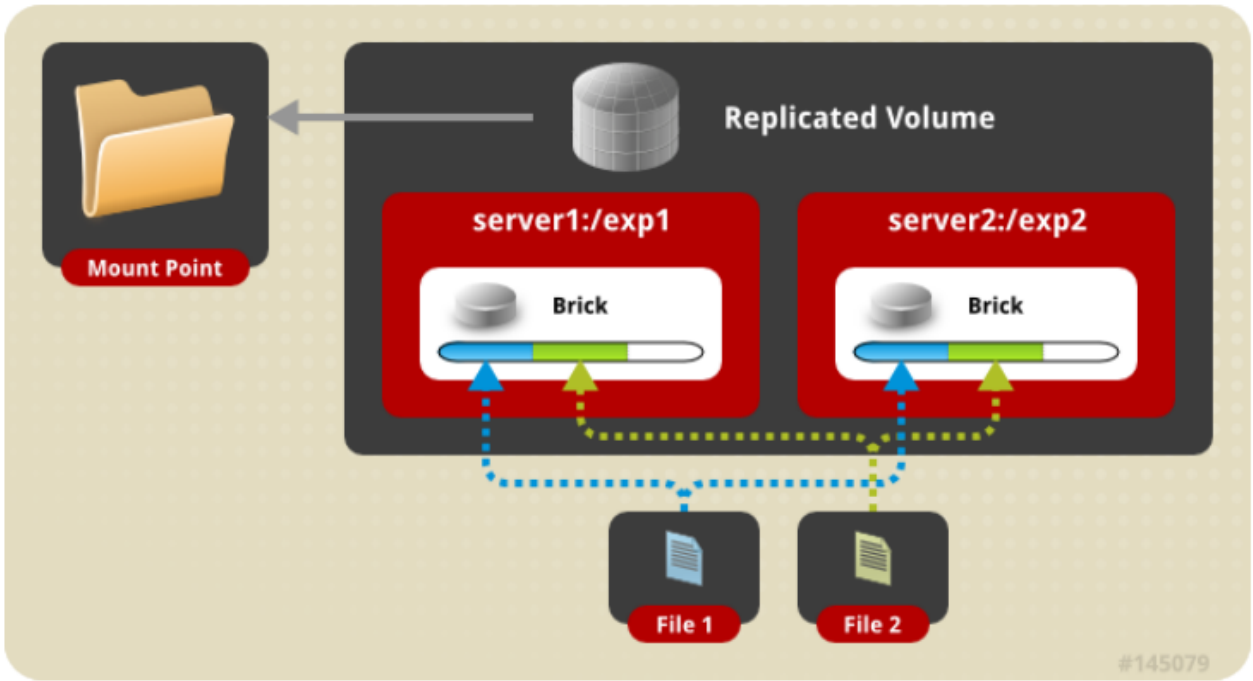
Плюсы

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

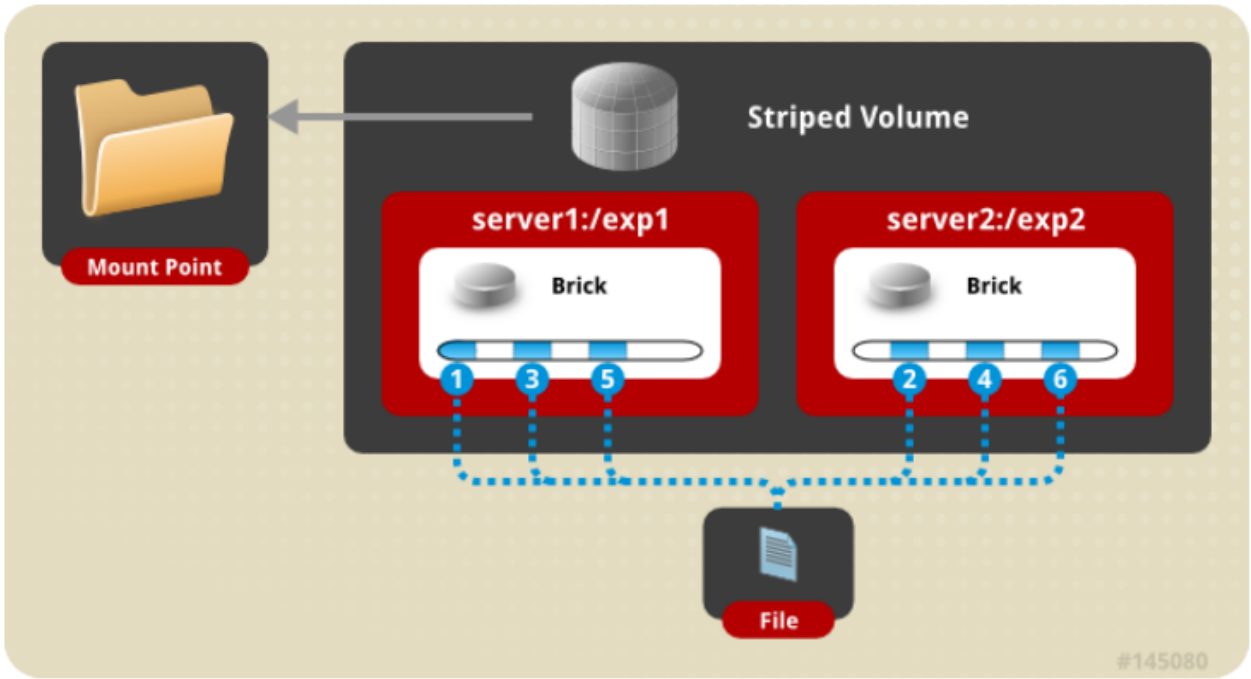
Минусы

- 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)

Подкопаев Антон


```
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