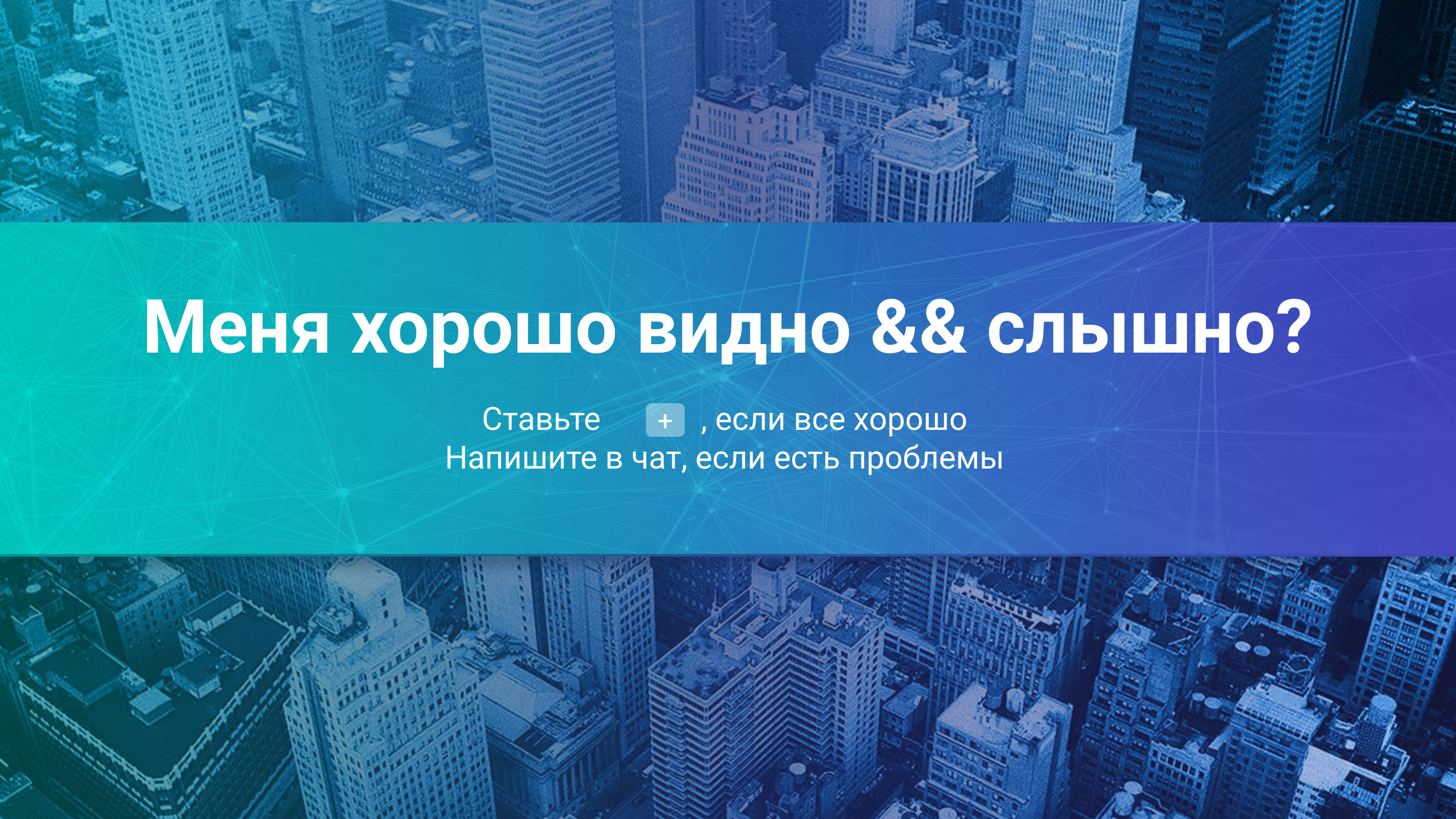

OTUS
ОНЛАЙН-ОБРАЗОВАНИЕ
Онлайн-образование

# Проверить, идет ли запись!



# Меня хорошо видно && слышно?

Ставьте + , если все хорошо
Напишите в чат, если есть проблемы

# Распределенные файловые системы: GlusterFS

Лапа Викентий Анатольевич

Test Automation Engineer

whamcloud.com

nop@tut.by

# Преподаватель

## Лапа Викентий

- более 10  лет опыта работы в среде Linux
  - Cluster File Systems
- последние 3 года преподаватель курса “Основы администрирования Linux”
  - выпущено 26 групп
- интересно наблюдать как человек развивается

# Правила вебинара

Активно участвуем: выполняем задания, отвечаем на вопросы

Если возникли сложности задаем вопрос в чат

На вопросы постараюсь отвечать сразу, но возможны паузы

# Маршрут вебинара

Знакомство и настройка на обучение

Разбираем примеры и определения

Проверяем чему научились

# **Цели вебинара** | После занятия вы сможете

1. Установить кластер с файловой системой GlusterFS
2. Выбрать конфигурацию и тип хранилища
3. Проверять состояние файловой системы

# Зачем нам распределенная файловая система.

- С какими распределенными файловыми системами знакомы?
- Что это такое распределенная файловая система?
- В чем отличие от локальной файловой системы?
- Для каких задач?

# **Смысл** | Зачем вам это уметь

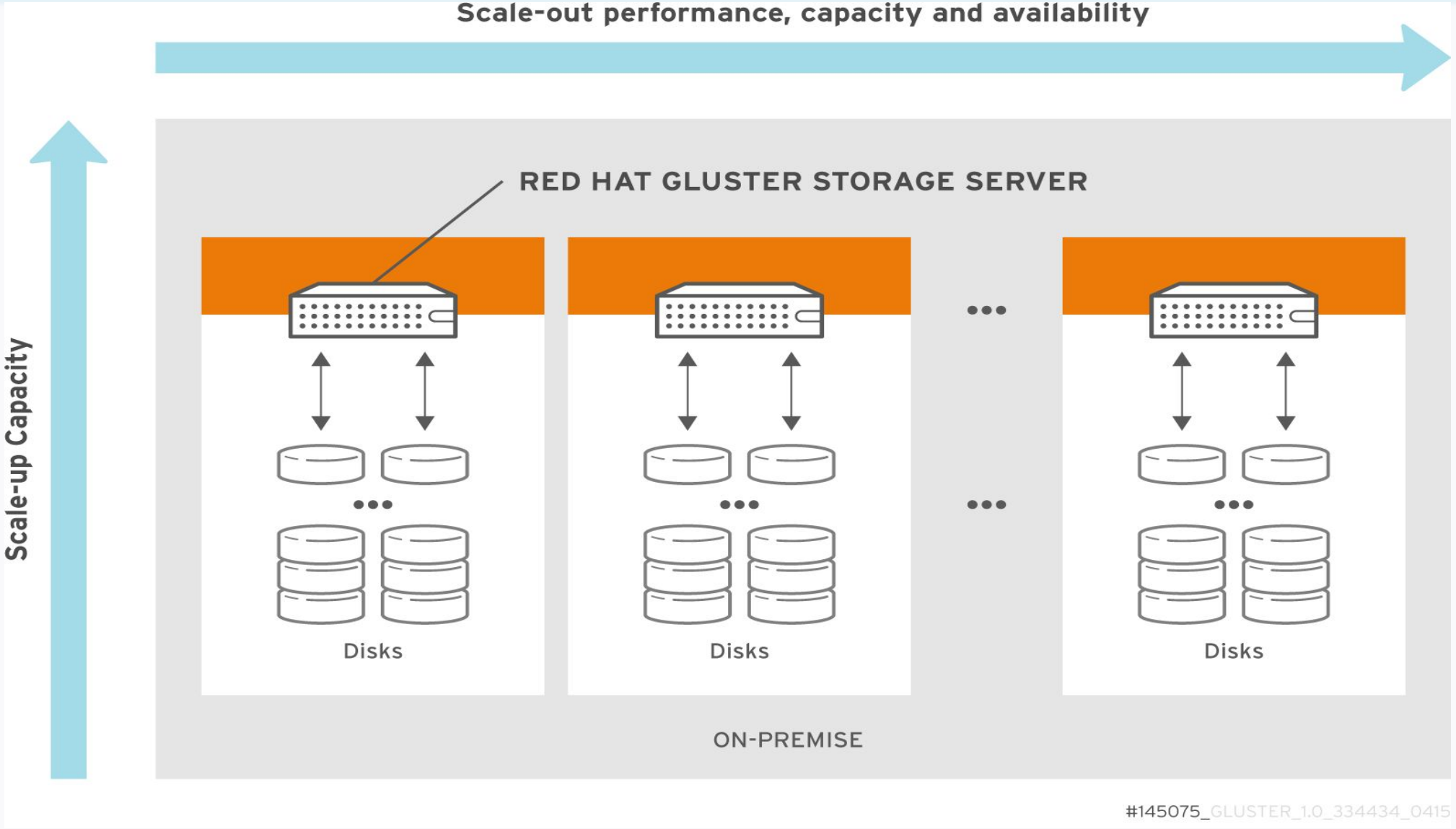
1. представление сервиса распределенной файловой системы
2. обеспечить отказоустойчивость
3. построить хранилище больших размеров

# Термины и определения

# Проверим себя. Что знаем о GlusterFS?

- Проходим тест. Ссылку пришлю в чате
- Прошли тест - оценку пишем в чат
- Время на выполнение 5 минут

# Пример масштабирования файловой системы



# Проверим работу тестового окружения

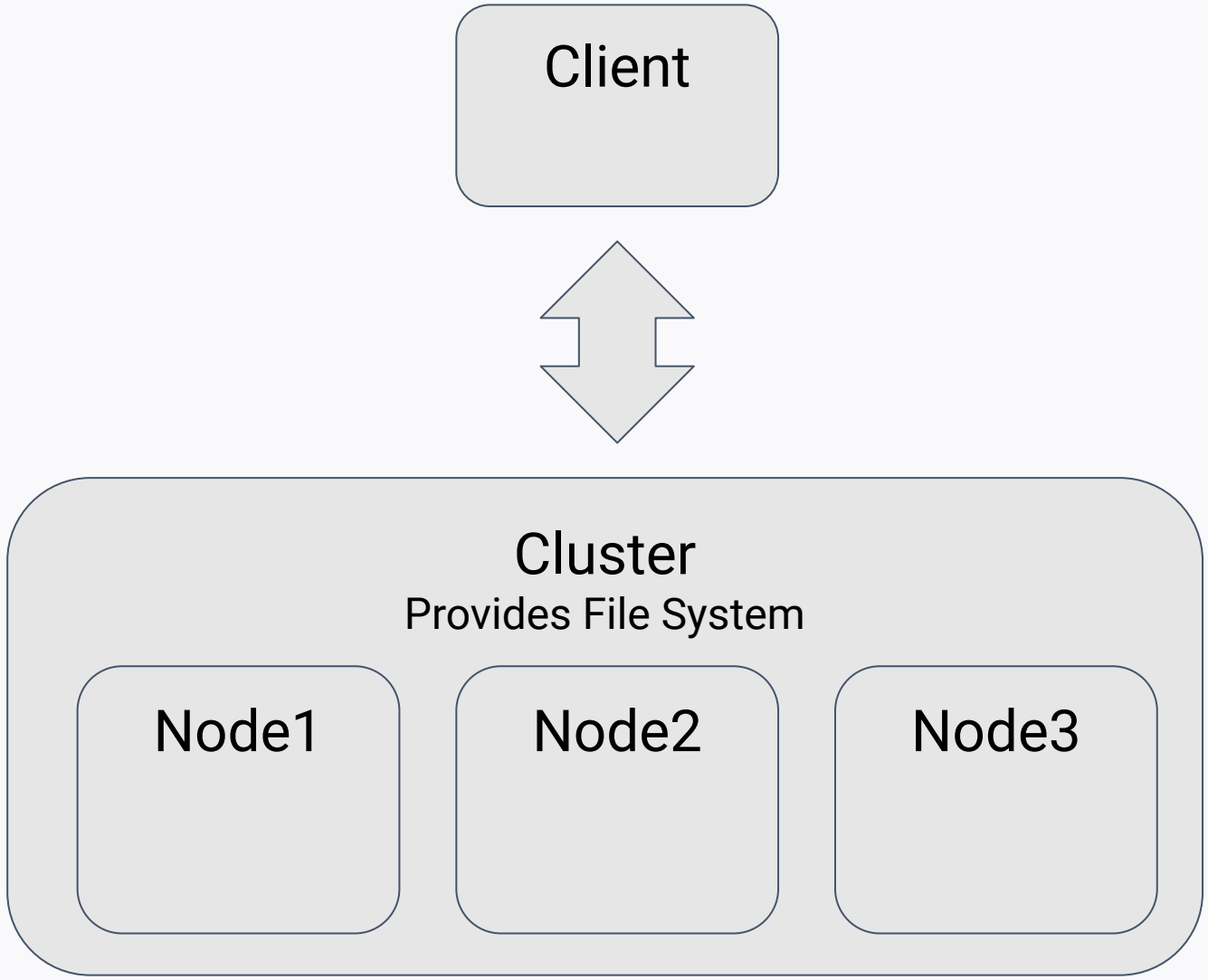
```
git clone git@github.com:nixuser/otus-gluster.git
cd otus-gluster/
cat README.md
```

Дальше выполняем команды из README.md

# Кластер

**Кластер** - набор серверов работающих вместе, предоставляющих единый ресурс.

В документации называется **Trusted Storage Pool (TSP)**



# **Блок** (brick)

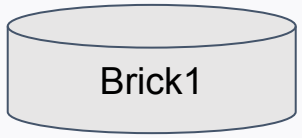
**Блок (brick)** - единица хранения glusterFS, директория экспортируемая на сервере из пула серверов (truster server pool).

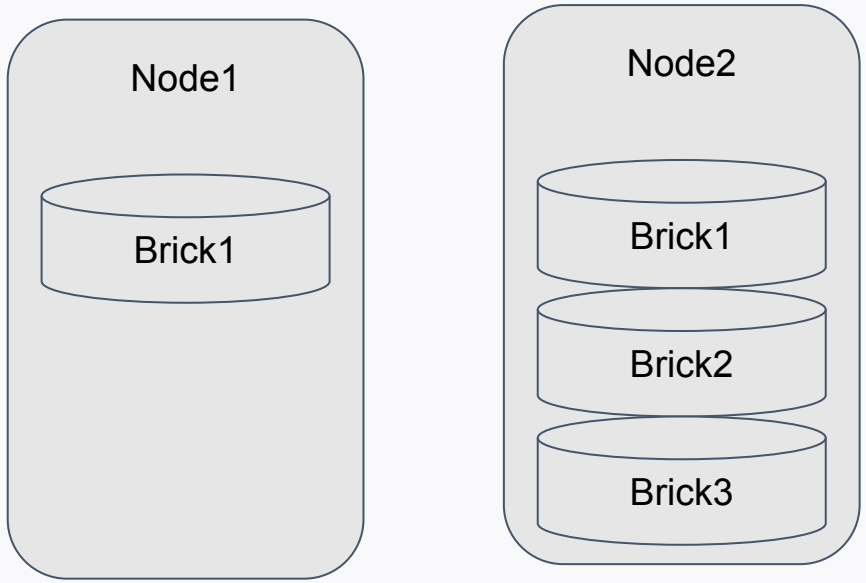
Формат записи:

```
SERVER:EXPORT
```

Например:

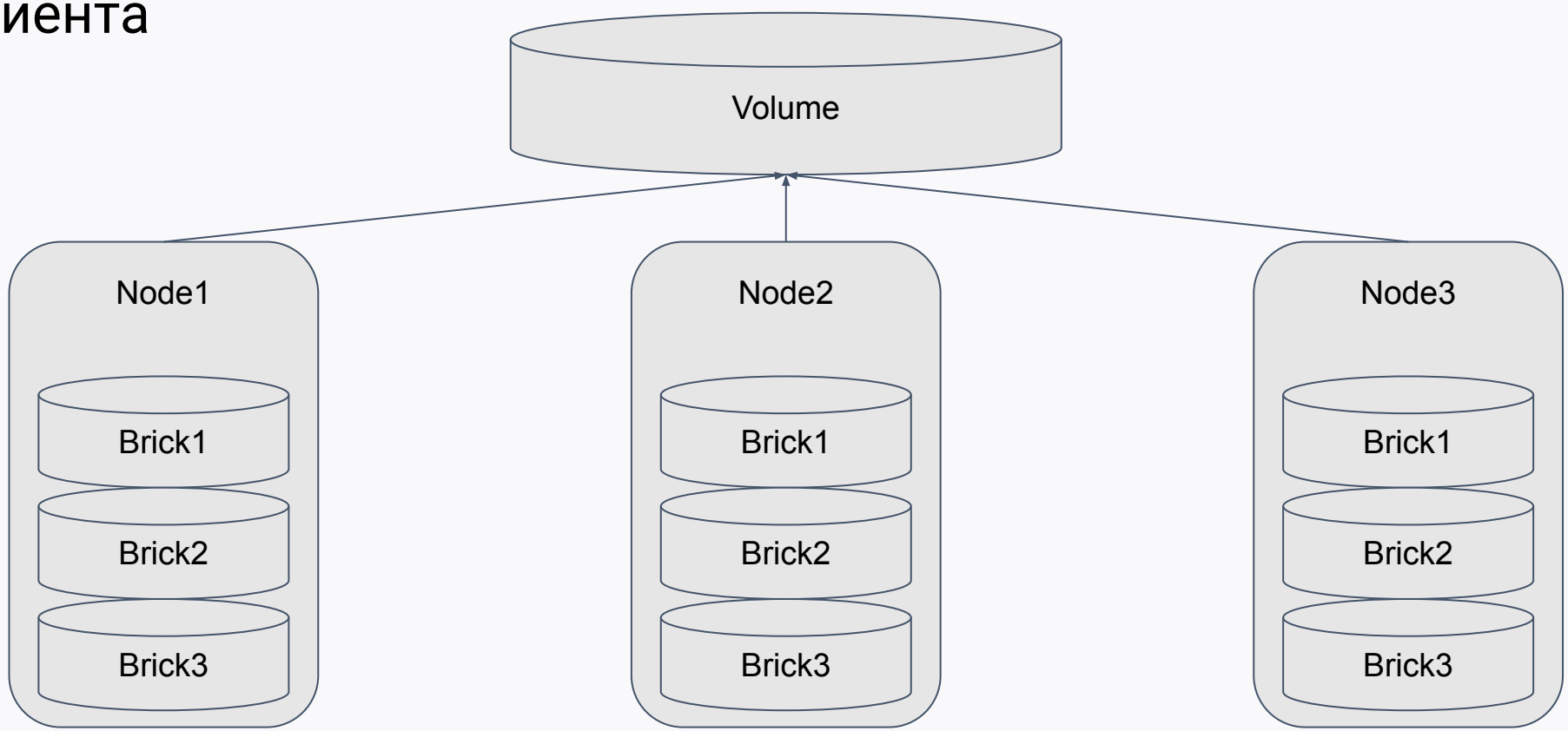
```
myhostname:/exports/myexportdir/
192.168.7.150:/srv/gluster/brick
```



# **Том** (volume)

**Том (volume)** это логически объединенные блоки (bricks). Имя volume используется в команде монтирования клиента



# Изучаем кластер.

**Выполнить команды и узнать:**

- Сколько узлов в кластере?
- Какое имя у volume? Какой тип volume?
- Сколько bricks в volume и сколько bricks приходиться на один сервер?
- Какая точка монтирования для брика?
- Какая файловая система для brick?
- Сколько места в кластере?
- Точка монтирования клиента GlusterFS?

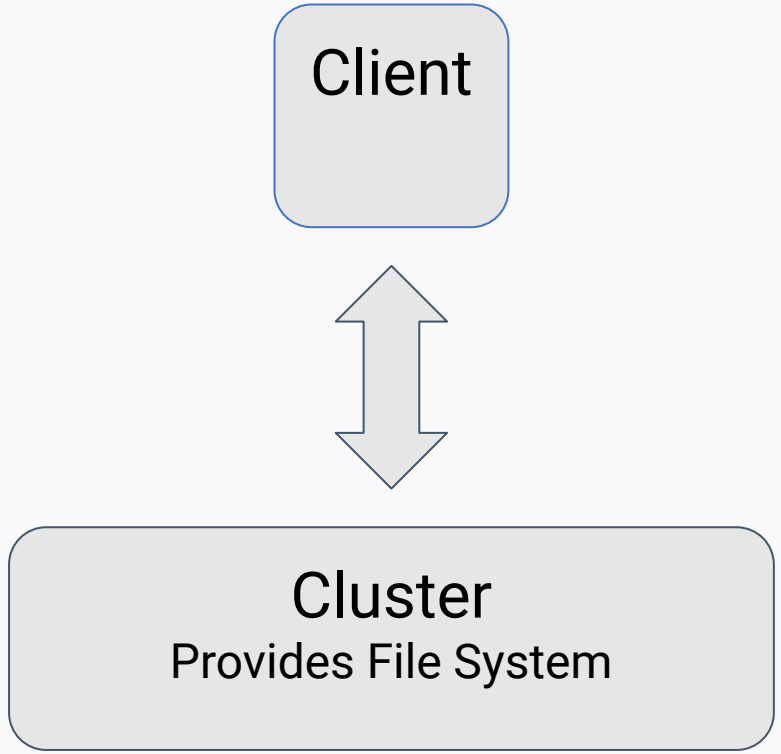
Ответы пишите в чат

```
gluster peer status
gluster pool list
gluster volume list
gluster volume status
gluster volume info
```

# **Клиенты** GlusterFS

Интерфейсы клиента

- **libgfapi** (клиент работает напрямую с Volumes минуя FUSE. Скорость.)
  - QEMU  storage layer
  - Samba VFS plugin
  - NFS-Ganesha
- **Native Client (FUSE)**



# Пример работы с клиентом GlusterFS


- монтируем Native Client

```
sudo mount -t glusterfs gluster1:/shara
/mnt/gluster_one_server
```

- запускаем I/O с клиента

```
for i in {1..20}
do dd if=/dev/urandom of=/mnt/gluster/test_fs/file_$i bs=1M
count=1
done
```

- изучаем расположение файлов на bricks

```
clush --hostfile=nodes ls /srv/gluster/brick/test_fs
```

# **Разбираемся что такое** Translator

Транслятор – модуль,
конвертирующий  запросы

- от пользователей к хранилищу
- от запроса к запросу
- реализация возможностей
- построение стека

Например:

**Storage** транслятор самого низкого уровня, хранит и получает доступ к данным из локальной файловой системы.

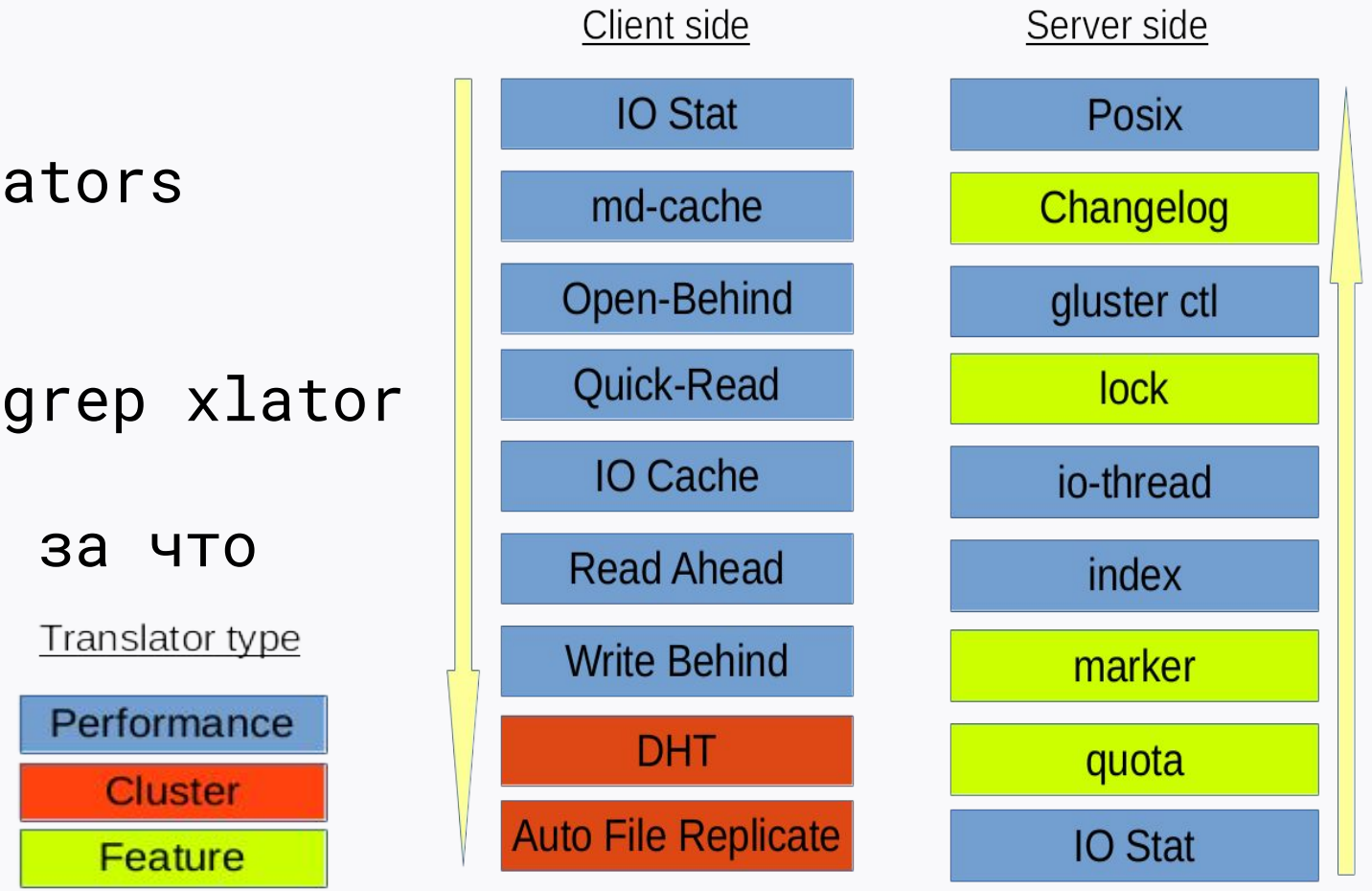
**Cluster** транслятор управляет распределением и репликацией данных.

# Пример. Найдем трансляторы

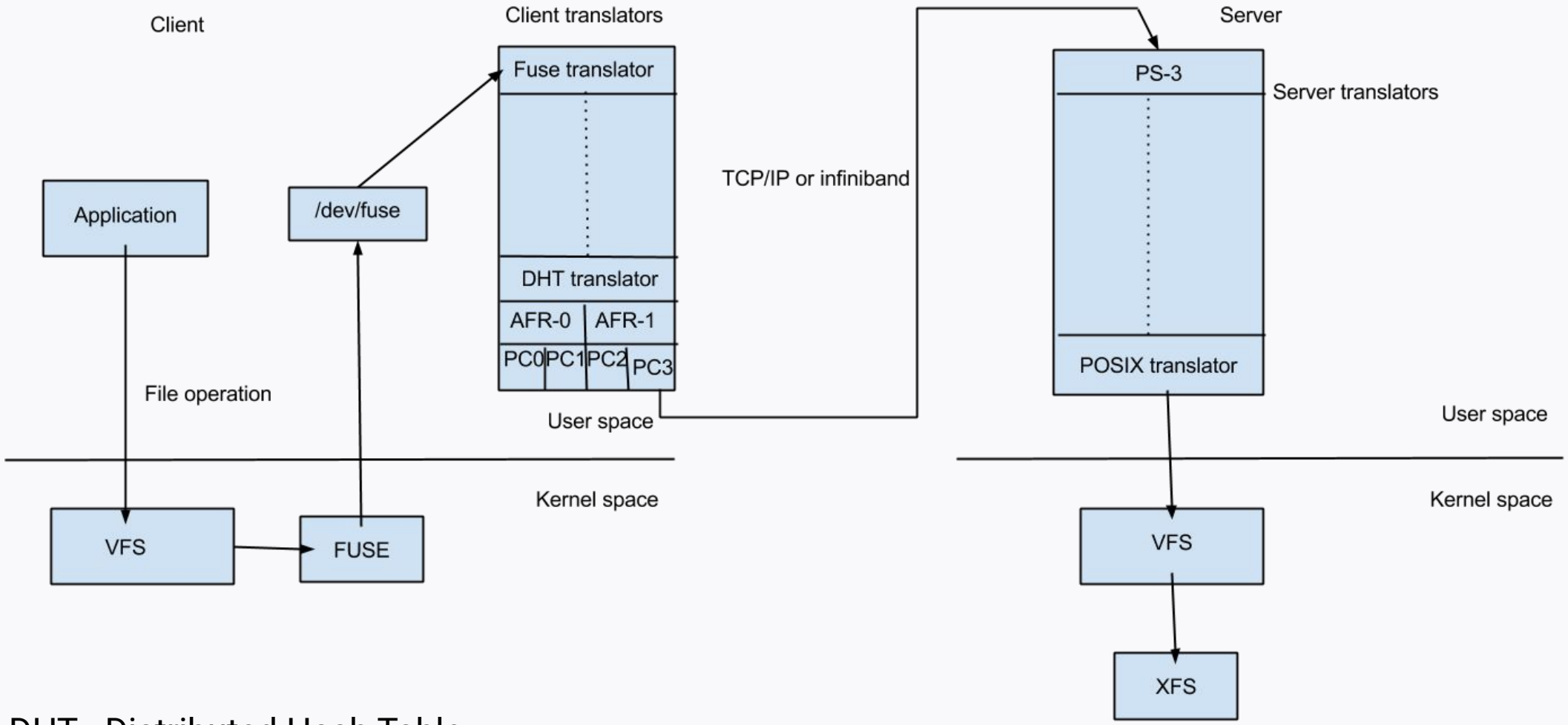
Выведем трансляторы клиента
```
rpm -ql glusterfs-client-xlators
```

Выведем трансляторы сервера
```
rpm -ql glusterfs-server | grep xlator
```

По названиям определить кто за что отвечает?



# Пример работы транслятора DHT




DHT - Distributed Hash Table

# Демоны для работы трансляторов

**glusterd** = management daemon

управляет вольюмом, остальными демонами

**glusterfsd** = per-brick daemon

**glustershd** = self-heal daemon

ребилд данных у реплицированных вольюмов в случаях отказа нод кластера.

**glusterfs** = client deamon

на клиенте, но может быть и на NFS серверах

.

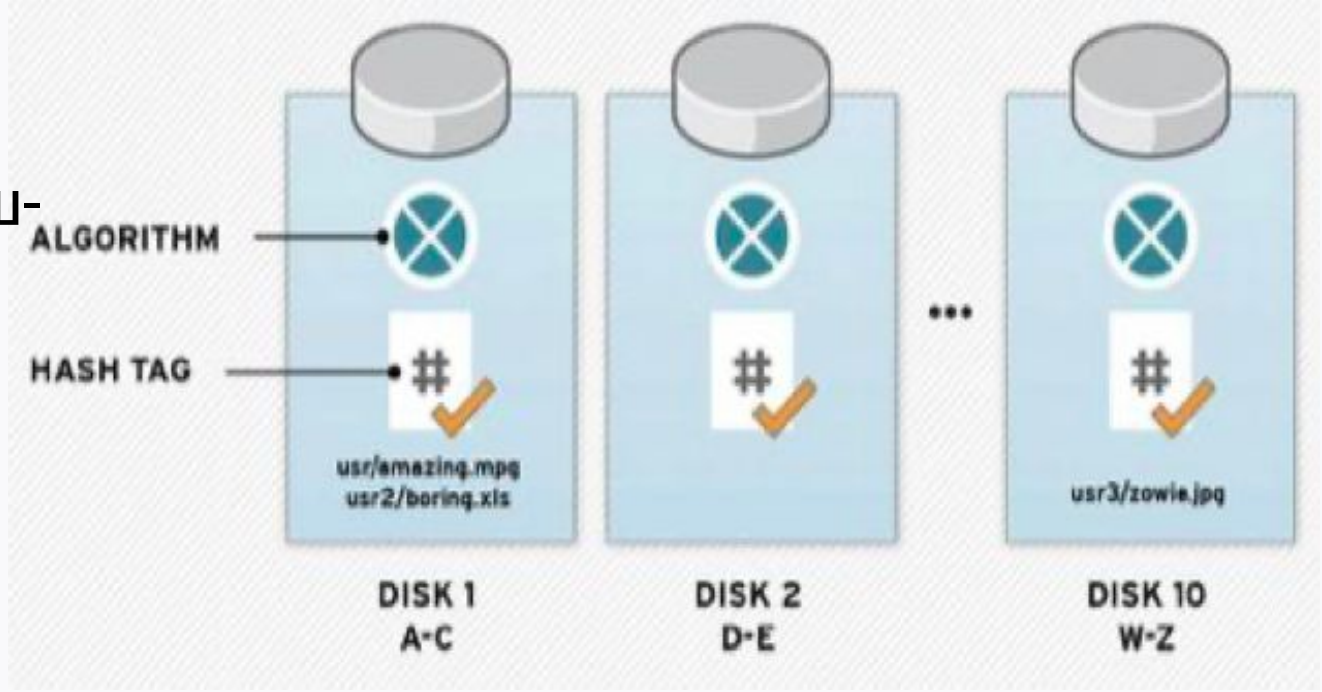
# Демоны для работы трансляторов

```
ps axf | grep gluster
ps axf | grep glusterd

systemctl status glusterd
systemctl cat glusterd

systemctl status glusterfsd
systemctl cat glusterfsd

ls /var/log/glusterfs/

ps axf | grep glusterfs
```

# Алгоритм работы

Нет сервера метаданных

- Месторасположение файла определяется уникальным хэш-тэгом **GFID** (Global file ID)
- Тэги хранятся на той же файловой системе
- Файлы распределяются на основе расчета
- Операции создания тома, расширения/сжатия выполняются без прерывания доступа к данным



# Найдем GFID файла

```
ls -i /srv/gluster/brick/test
  100704281 file_11  100672337 file_18  100704279 file_8

find /srv/gluster/brick/test -inum 100704279
   /srv/gluster/brick_b/shara/.glusterfs/b0/55/b0559aa3-a
   523-4602-b69d-9515cb09711d
```

# Найдем GFID файла по кластеру

```
clush --hostfile=nodes sudo ls
/srv/gluster/brick/.glusterfs/f6/65/f665aecd-400d-4202-96
15-cf4632d4977b

gluster2:
/srv/gluster/brick/.glusterfs/f6/65/f665aecd-400d-4202-96
15-cf4632d4977b
gluster1:
/srv/gluster/brick/.glusterfs/f6/65/f665aecd-400d-4202-96
15-cf4632d4977b
```

# **Удалим** volume

Отмонтируем всех клиентов

```
findmnt -t fuse.glusterfs
clush --hostfile=nodes sudo umount /mnt/gluster
clush --hostfile=nodes sudo findmnt -t fuse.glusterfs
```

Удалим volume

```
sudo gluster volume list
sudo gluster volume delete gluster
```

# **Типы томов** (volumes)

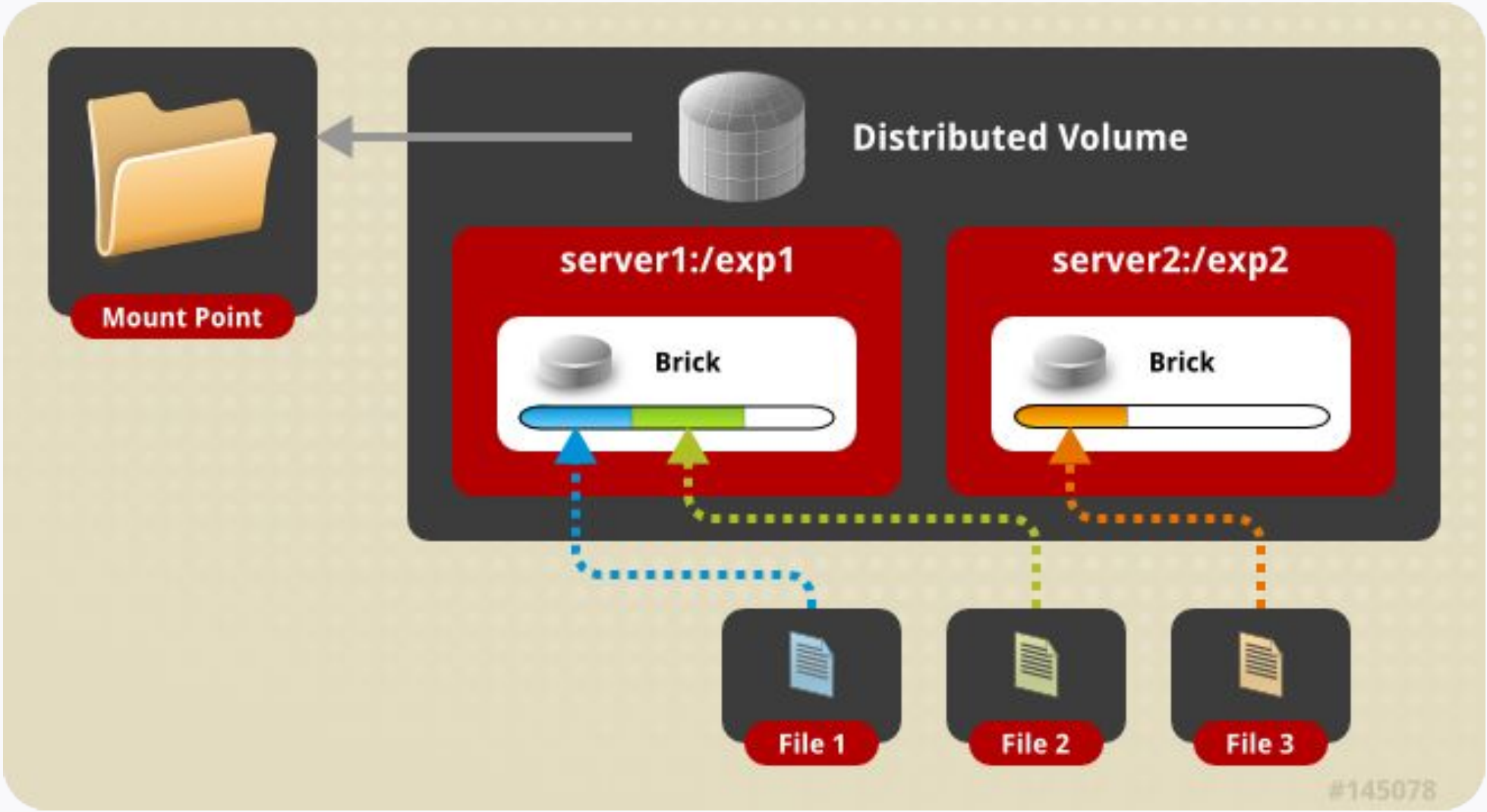
- **distributed** - (распределенный)
- **replicated** - (реплицированный)
- **distributed** replicated - (распределенный реплицированный)
- **dispersed** - (рассредоточенный)

Тип тома указывается при создании командой

```
gluster volume create [stripe | replica | disperse]
[transport tcp | rdma | tcp,rdma ]
```

- если тип не указан то по умолчанию создается distributed

# **Распределенный том** (distributed)



# **Распределенный том** (distributed)

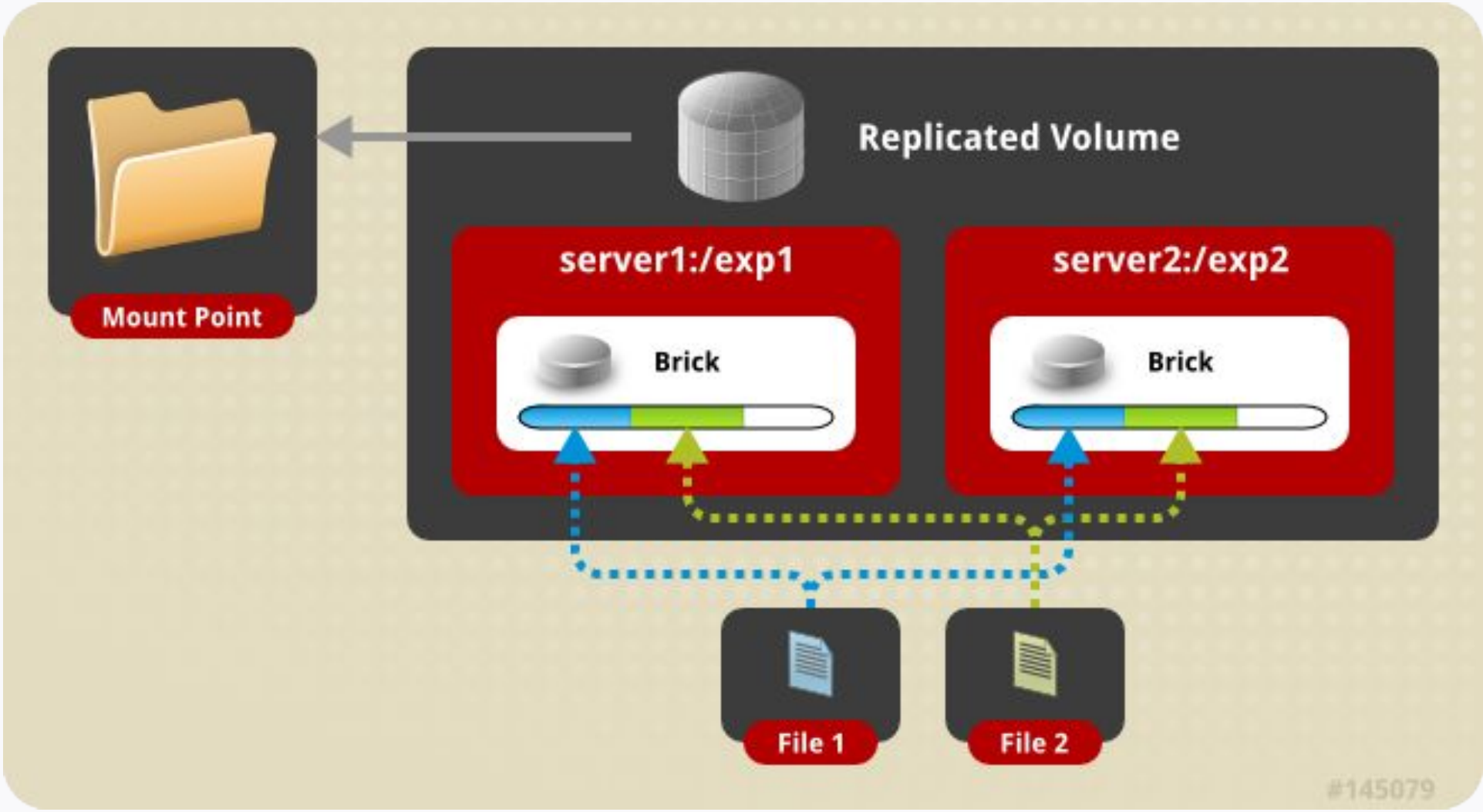
Создаем распределенный том на 3х серверах (1 brick на сервер)

```
gluster volume create  test-volume  host1:brick1
host2:brick2 host3:brick3
```

Создаем распределенный том на 4х серверах (2 brick на сервер)

```
gluster volume create dist-volume
server1:/exp1 server2:/exp2 server3:/exp3 server4:/exp4
```

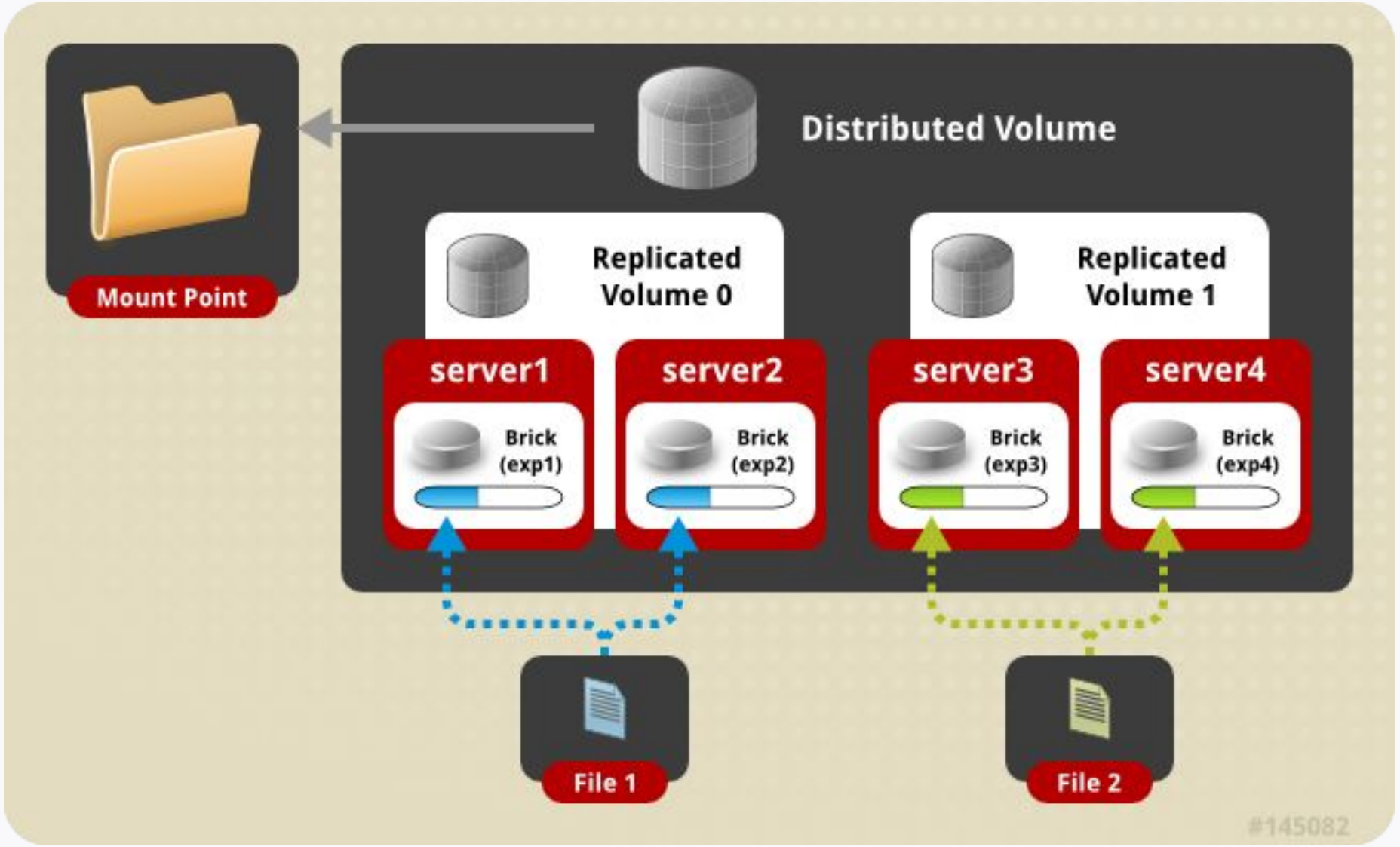
# **Реплицированный том** (replicated)



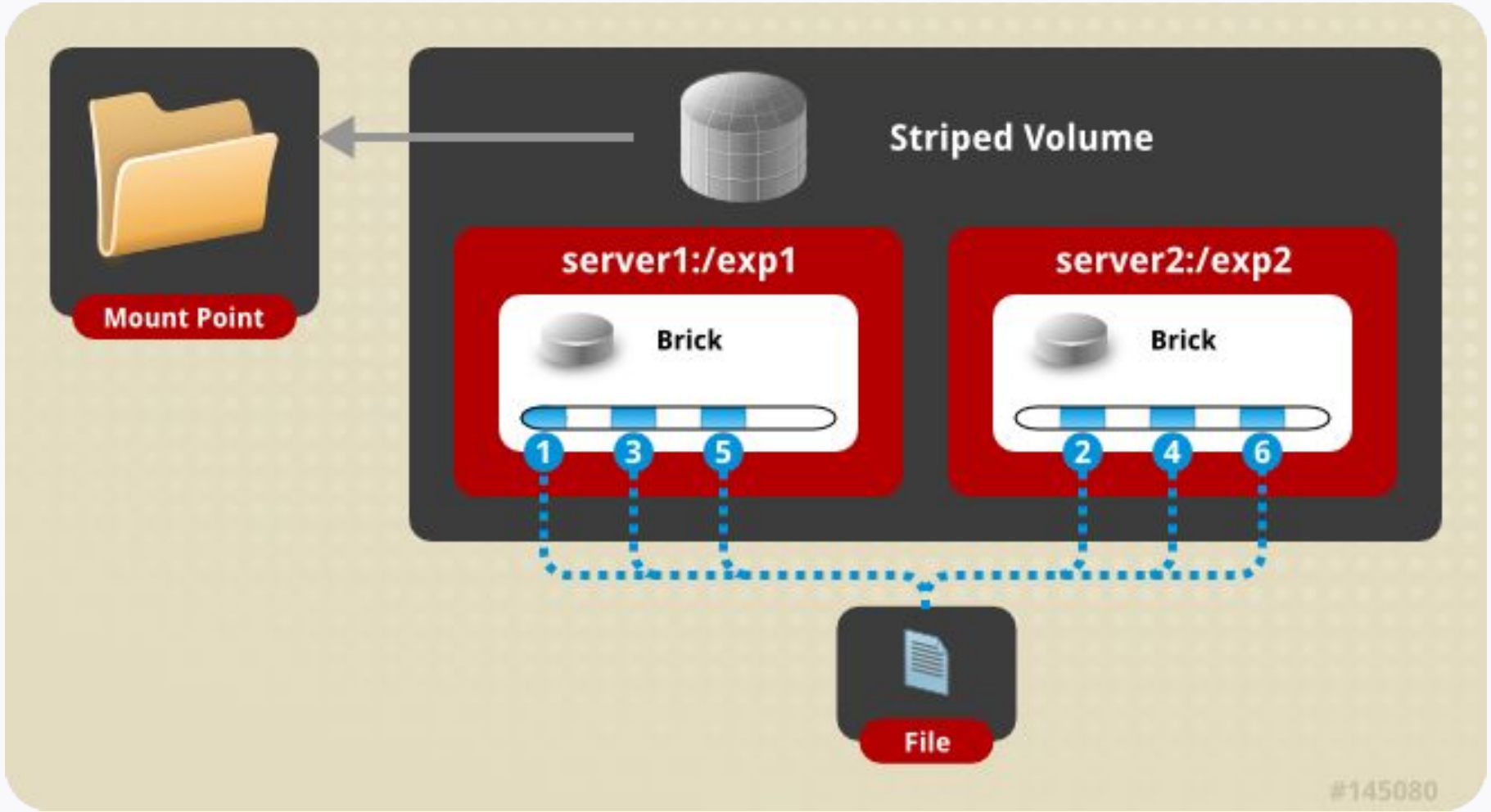
# **Реплицированный том** (replicated)

```
gluster volume create shara2 replica 3 arbiter 1
transport tcp gluster1:/srv/gluster/brick_b/shara
gluster2:/srv/gluster/brick_b/shara
gluster3:/srv/gluster/brick_b/shara
```

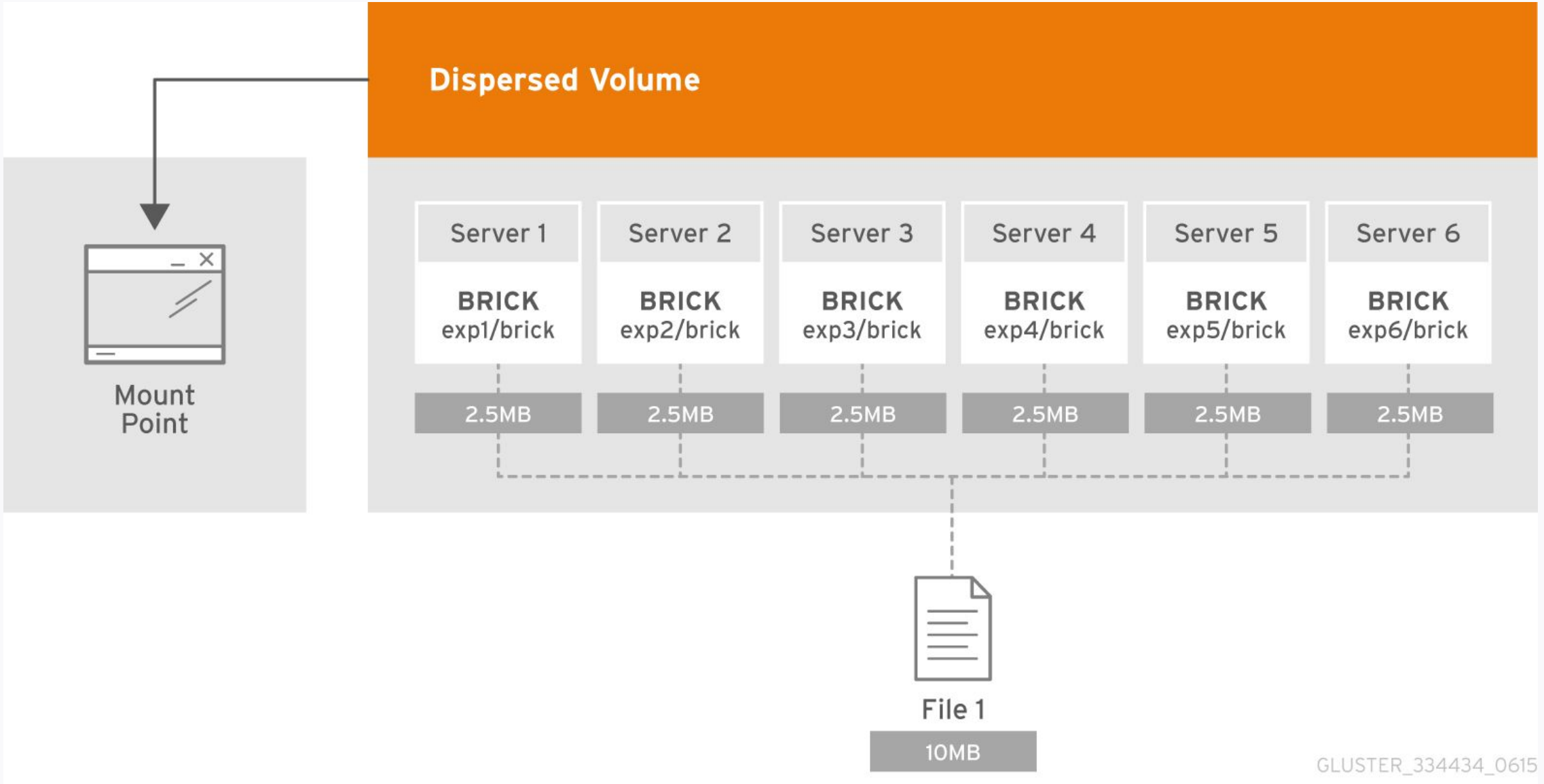
# Распределенный реплицированный том



# Страйпы


Striped Volume
Mount Point
server1:/exp1
server2:/exp2
Brick
Brick
1
3
5
2
4
6
File
#145080

# **Рассредоточенный** (Dispersed Volume)



# **Рассредоточенный** (Dispersed Volume)

```
gluster volume create shara disperse 3 redundancy 1
gluster{1..3}:/srv/gluster/brick_b/shara
sudo gluster volume start shara
```

# Distributed-replicated vs Distributed-dispersed

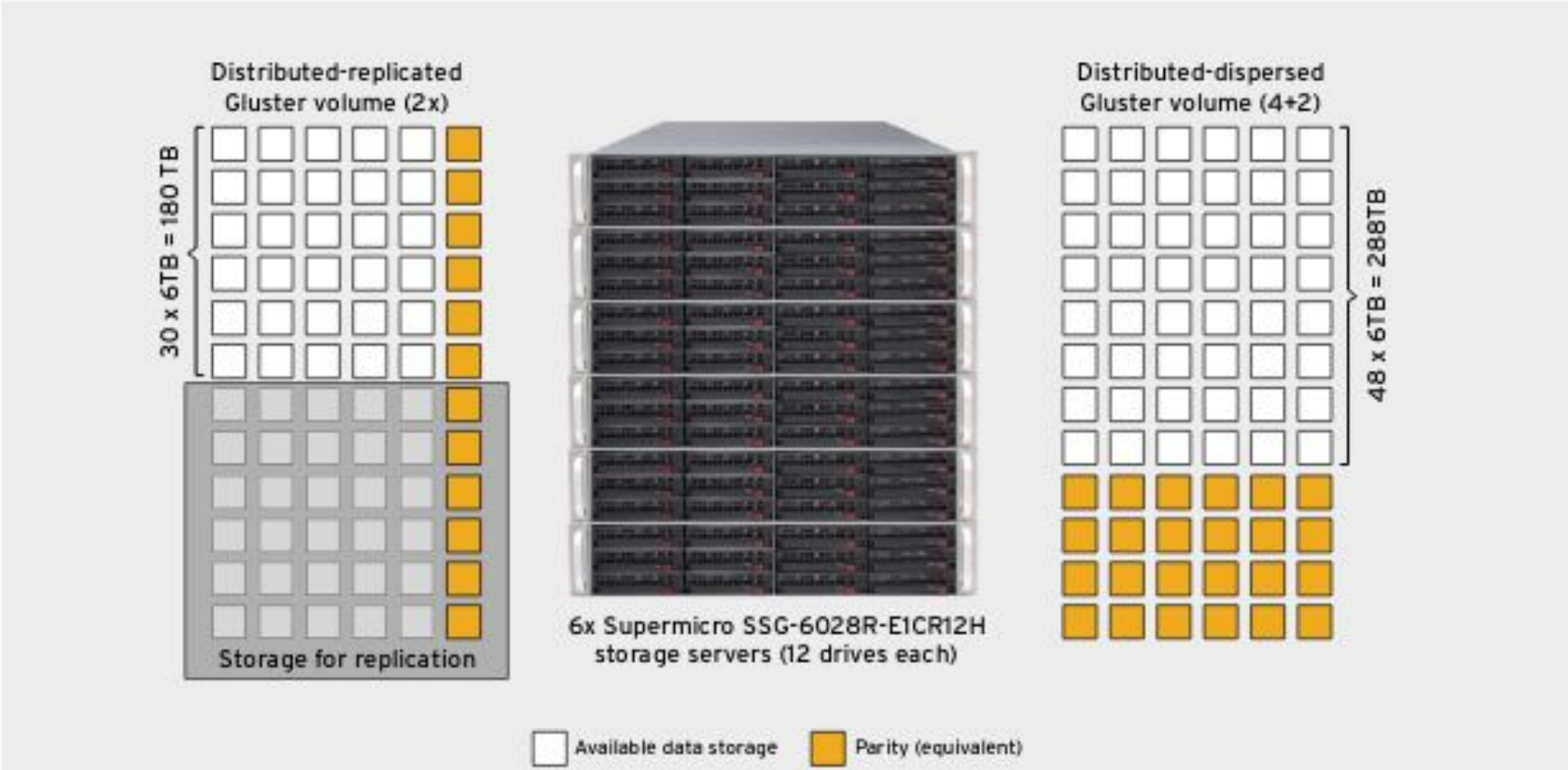



**Figure 3.** *Figure 3. Dispersed Gluster volumes can yield significantly more usable storage than replicated volumes (space for parity depicted logically).*

# Практика

# Выполняем инструкции из файла

- Создаем блоки (bricks)
- Добавляем узлы в кластер
- Собираем вольюм, disperse 3+1
- Собираем вольюм, replicated 2 + 1
- Дополнительные настройки
- Проверяем отказоустойчивость

# Дополнительные настройки

Смотрим настройку ping-timeout:

```
gluster volume get shara network.ping-timeout
```

Меняем настройку

```
gluster volume set shara network.ping-timeout "5"
```

Проверяем значение применилось

```
gluster volume get shara network.ping-timeout
```

Все настройки

```
gluster volume get gluster all
```

# Подводим итоги

# Вопросы

- У кого получилось настроить кластер?
- Какой тип volume выбрали?

# Рефлексия

Проходим тест

Как изменились результаты?

# Следующий вебинар

## Тема:

Ссылка на вебинар будет в ЛК за 15 минут

Материалы к занятию в ЛК — можно изучать

Обязательный материал обозначен красной лентой

# Следующий вебинар

## Тема:

Дату сообщу дополнительно в чате группы

Ссылка на вебинар будет в ЛК за 15 минут

Домашнее задание и краткое содержание занятия здесь

# Список материалов для изучения

- https://github.com/gluster/community/tree/master/meetings
- сравнение CephFS и GlusterFS
  https://m.habr.com/ru/company/croccloudteam/blog/430474/
- тестирование Dispersed Volume
  https://m.habr.com/ru/company/0/blog/353666/
- тестирование GlusterFS
  https://m.habr.com/ru/company/croccloudteam/blog/417475/
- подробнее про работу DHT

  https://glusterdocs-beta.readthedocs.io/en/latest/overview-concepts/translators.html#

# Спасибо за внимание!
# Приходите на следующие вебинары

Лапа Викентий Анатольевич

Test Automation Engineer
whamcloud.com
nop@tut.by