Основы построения файловых систем

Сегодня мы поговорим про NFS

NFS (Network File System) – механизм для того, чтобы локальную ФС одного компьютера сделать доступной по сети.

Основные цели

- Независимость от аппаратной платформы и операционной системы,
- Восстановление после сбоев и падений,
- Прозрачность: доступ к удалённой ФС должен быть неотличим от доступа к локальной,
- Поддержка POSIX-семантики,
- Достаточная производительность.

Протокол NFS

- Null -> ()
- Lookup(dirfh, name) -> (fh, attr)
- Create(dirfh, name, attr) -> (newfh, attr)
- Remove(dirfh, name) -> (status)
- Getattr(fh) -> (attr)
- Setattr(fh, attr) -> attr
- Read(fh, offset, count) -> (attr, data)
- Write(fh, offset, count, data) -> (attr)
- Rename(dirfh, name, tofh, toname) -> (status)
- Link(dirfh, name, tofh, toname) -> (status)
- Symlink(dirfh, name, string) -> (status)
- Readlink(fh) -> (string)
- Mkdir(dirfh, name, attr) -> (fh, newattr)
- Rmdir(dirfh, name) -> (status)
- Readdir(dirfh, cookie, count) -> (direntries)
- Statfs(fh) -> (fsstats)

Протокол NFS

Нет открытия файла по имени – протокол подстроен под то, как ядро делает проход по пути.

Протокол не имеет состояния – все данные, нужные запросу, передаются в аргументах.

Протокол NFS

Нет открытия файла по имени – протокол подстроен под то, как ядро делает проход по пути.

Протокол не имеет состояния – все данные, нужные запросу, передаются в аргументах.

- Упрощает восстановление после сбоев – его просто нет.

Протокол NFS

Нет открытия файла по имени – протокол подстроен под то, как ядро делает проход по пути.

Протокол не имеет состояния – все данные, нужные запросу, передаются в аргументах.

- Упрощает восстановление после сбоев – его просто нет.
- Записи должны быть синхронными.

Протокол NFS

Нет открытия файла по имени – протокол подстроен под то, как ядро делает проход по пути.

Протокол не имеет состояния – все данные, нужные запросу, передаются в аргументах.

- Упрощает восстановление после сбоев – его просто нет.
- Записи должны быть синхронными.
- NFS file handles существенно зависят от того, что в Unix File System к содержимому файла можно получить не зная его пути, а только зная номер иноды.

Протокол NFS

Нет открытия файла по имени – протокол подстроен под то, как ядро делает проход по пути.

Протокол не имеет состояния – все данные, нужные запросу, передаются в аргументах.

- Упрощает восстановление после сбоев – его просто нет.
- Записи должны быть синхронными.
- NFS file handles существенно зависят от того, что в Unix File System к содержимому файла можно получить не зная его пути, а только зная номер иноды.
- В локальных файловых системах тоже просочилось знание об NFS:

```
__u32   i_block[EXT2_N_BLOCKS];/* Pointers to blocks */
__u32   i_generation;   /* File version (for NFS) */
__u32   i_file_acl; /* File ACL */
__u32   i_size_high;    /* Formerly i_dir_acl, directory ACL */
__u32   i_faddr;    /* Fragment address */
```

Трудности с NFS (очевидные)

- Открытые удалённые файлы (silly rename)
- Разница во времени на клиенте и сервере
- Права доступа и аутентификация
- Блокироки файлов
- Неидемпотентность операций delete и rename

Трудности с NFS (неочевидные)

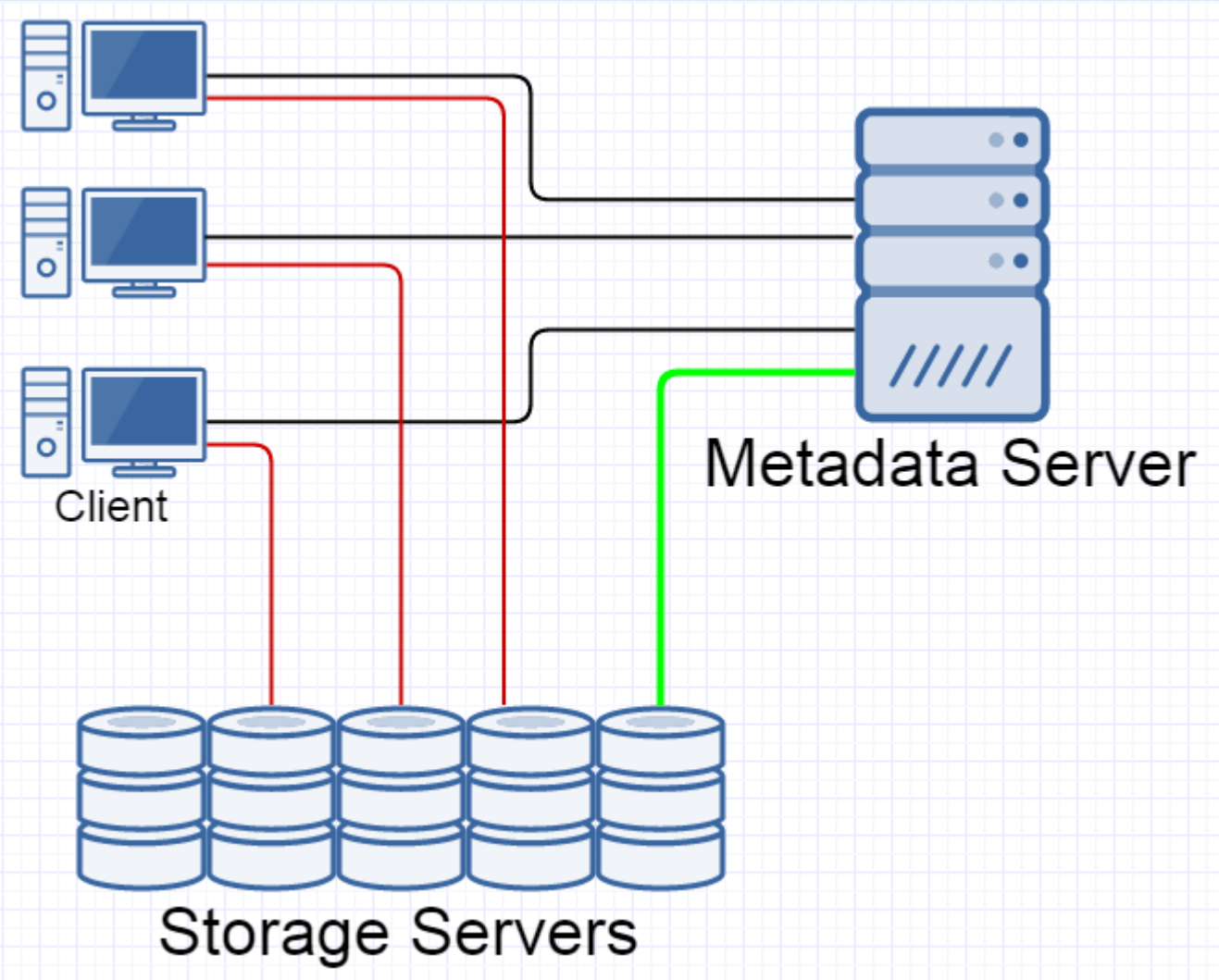
- UDP fragmentation:
  чтение и запись ядро будет делать блоками по 4096 байт (или больше),
  что больше MTU, поэтому пакеты с запросами будут разбиваться на несколько;
  разрезанным пакетам присваивается 16-битный идентификатор, который используется для их сборки в правильном порядке; reassembly timeout для UDP составляет 30 секунд, а 65536 пакетов на 1Gbps интерфейсе можно отправить меньше, чем за 1 секунду

Трудности с NFS (неочевидные)

- UDP fragmentation:
  чтение и запись ядро будет делать блоками по 4096 байт (или больше),
  что больше MTU, поэтому пакеты с запросами будут разбиваться на несколько;
  разрезанным пакетам присваивается 16-битный идентификатор, который используется для их сборки в правильном порядке; reassembly timeout для UDP составляет 30 секунд, а 65536 пакетов на 1Gbps интерфейсе можно отправить меньше, чем за 1 секунду

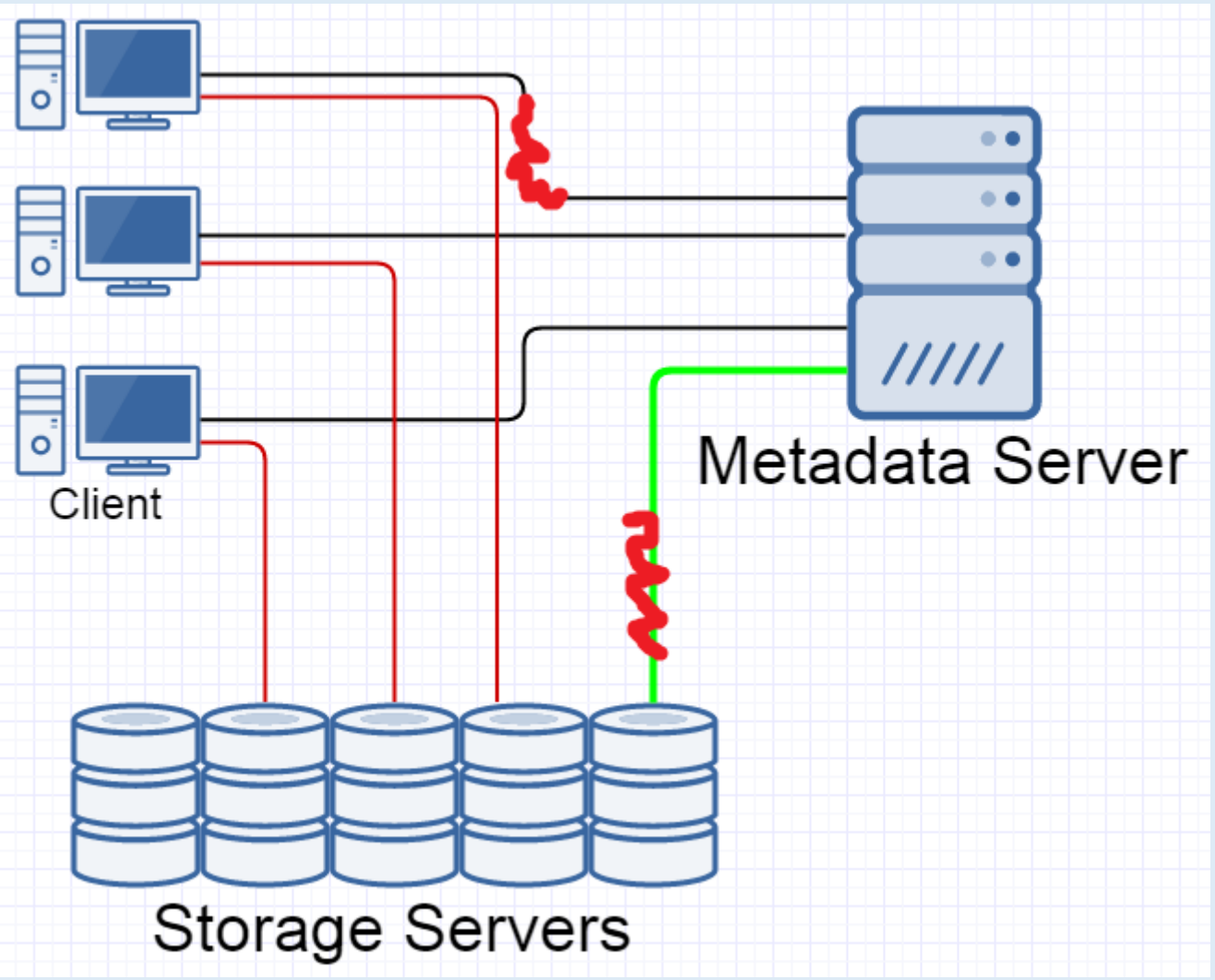
- Ответ от Sun RPC portmapper может быть 20 раз длиннее запроса
  при использовании UDP ответ будет послен не туда, откуда пришёл пакет, а по адресу, указанному в пакете
  Это даёт возможность проводить Bandwidth amplification attacks.

NFSv4.1 и Parallel NFS



- NFS layouts:
  file
  block
  object storage
- File & directory delegations

NFSv4.1 и Parallel NFS: network partitions



Network partition – ситуация, когда теряется связь между частью узлов в сети.

- Layout recall
- Stateids,
- Client fencing