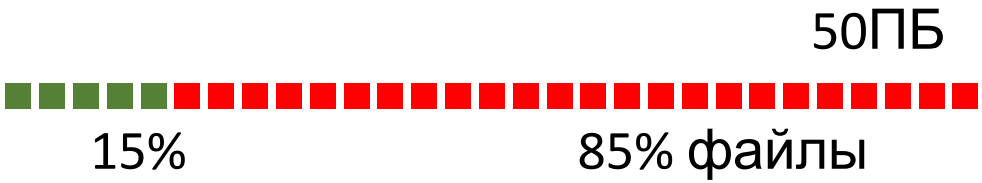
PGDAY'17 RUSSIA

Как мигрировать 50Пб в 32ПБ

Андрей Сумин, CTO of Mail Services at Mail.ru

Из чего состоит почта



Структура базы

FileDB

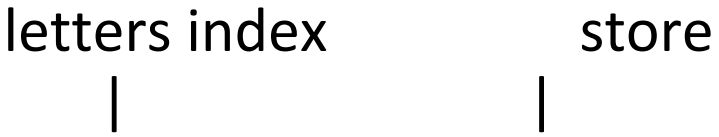
FileDB

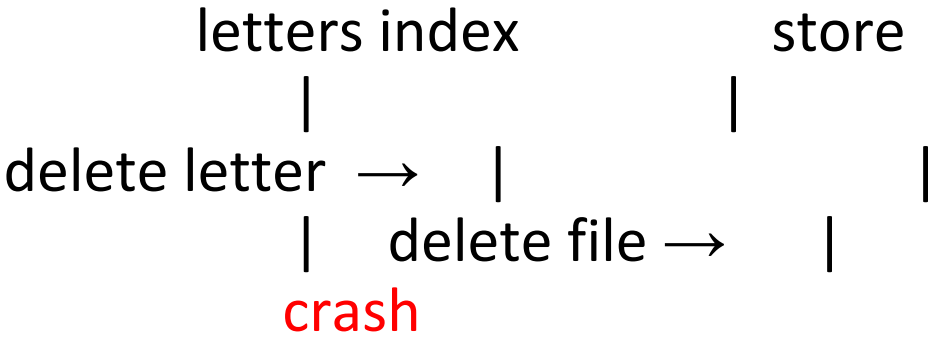
sha1 (от содержимого файла)

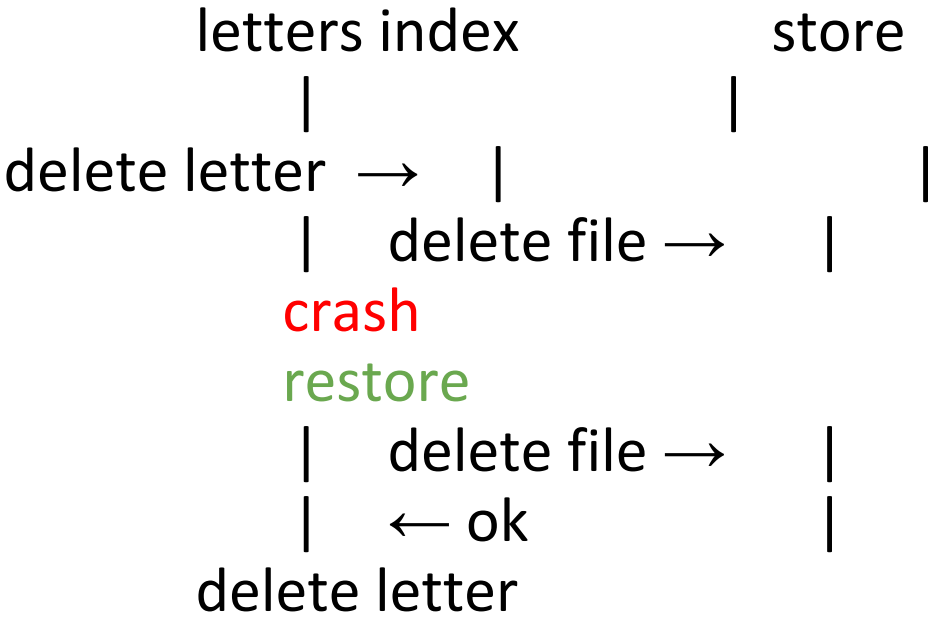
FileDB

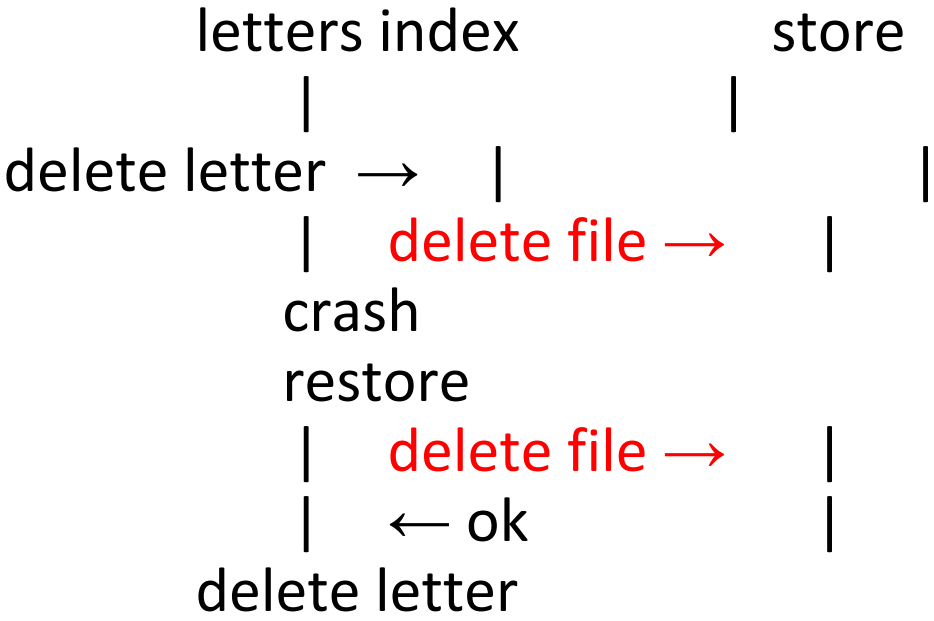
sha1

counter


letters index
store







FileDB

sha1

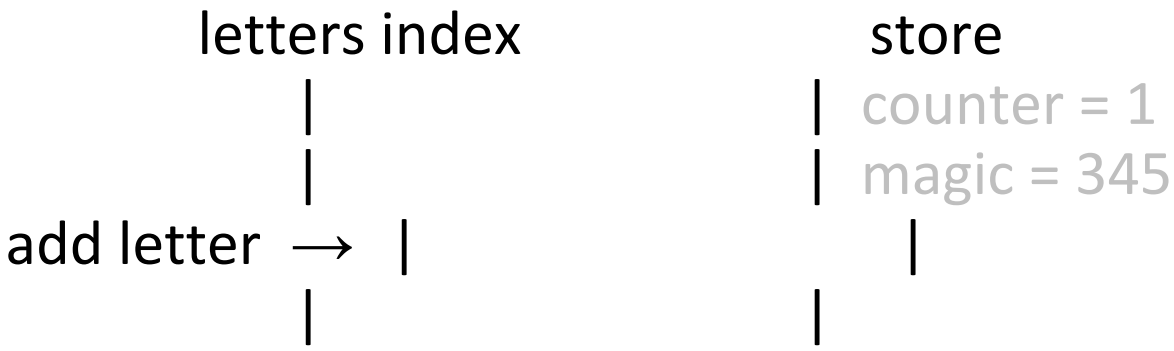
counter

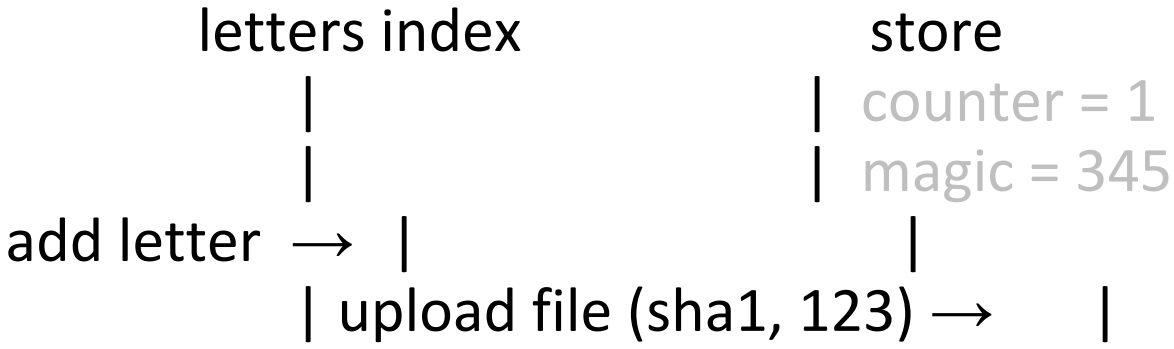
magic

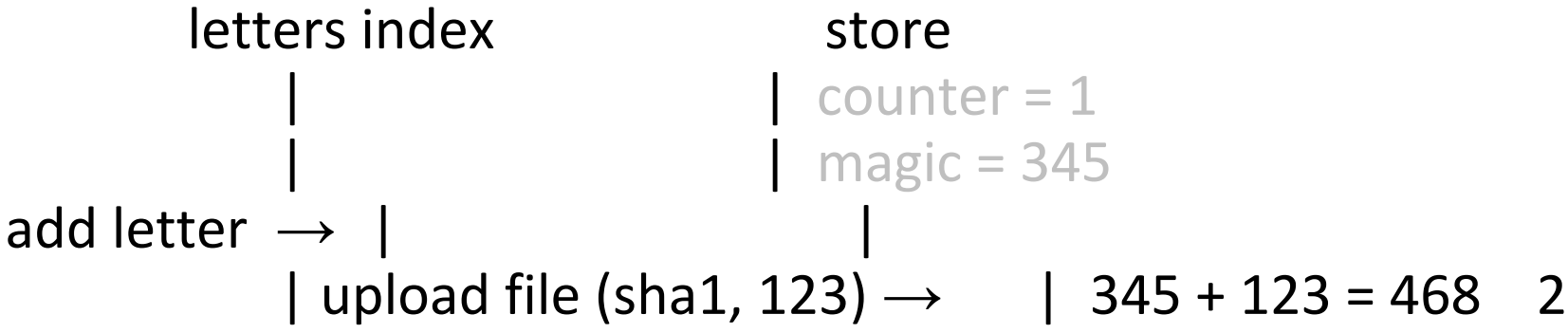
в индексах писем лежит sha1 и magic

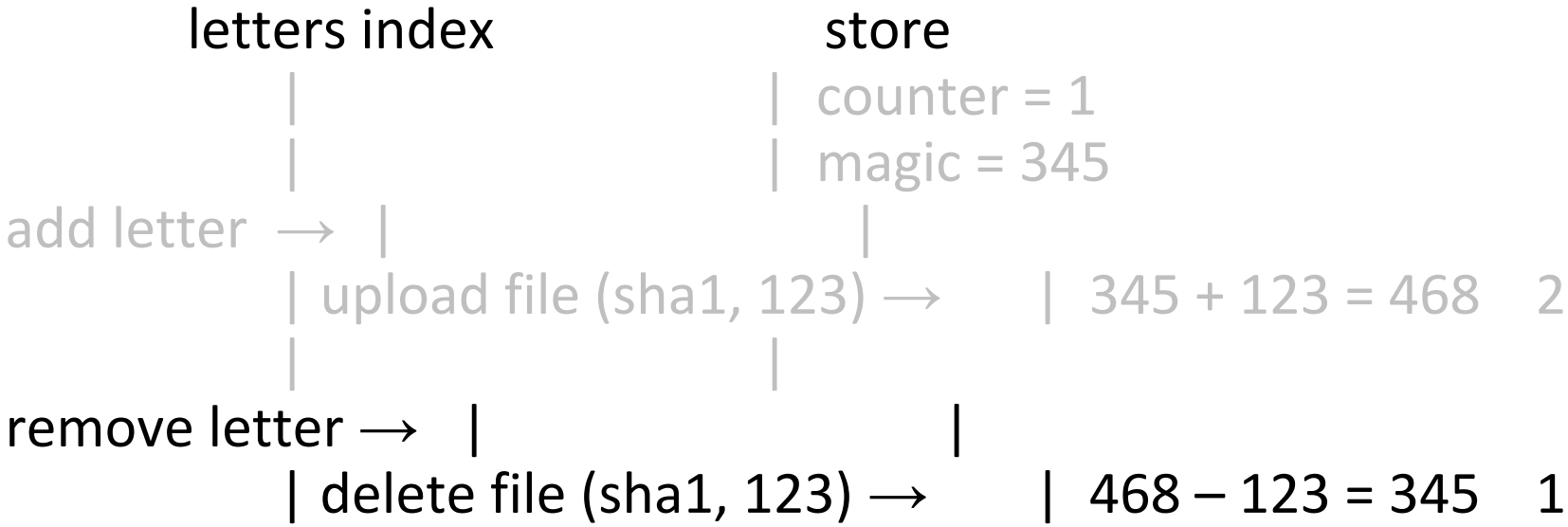
counter = 1
magic = 345

counter = 1
magic = 345


letters index
store
counter = 1
magic = 345
add letter

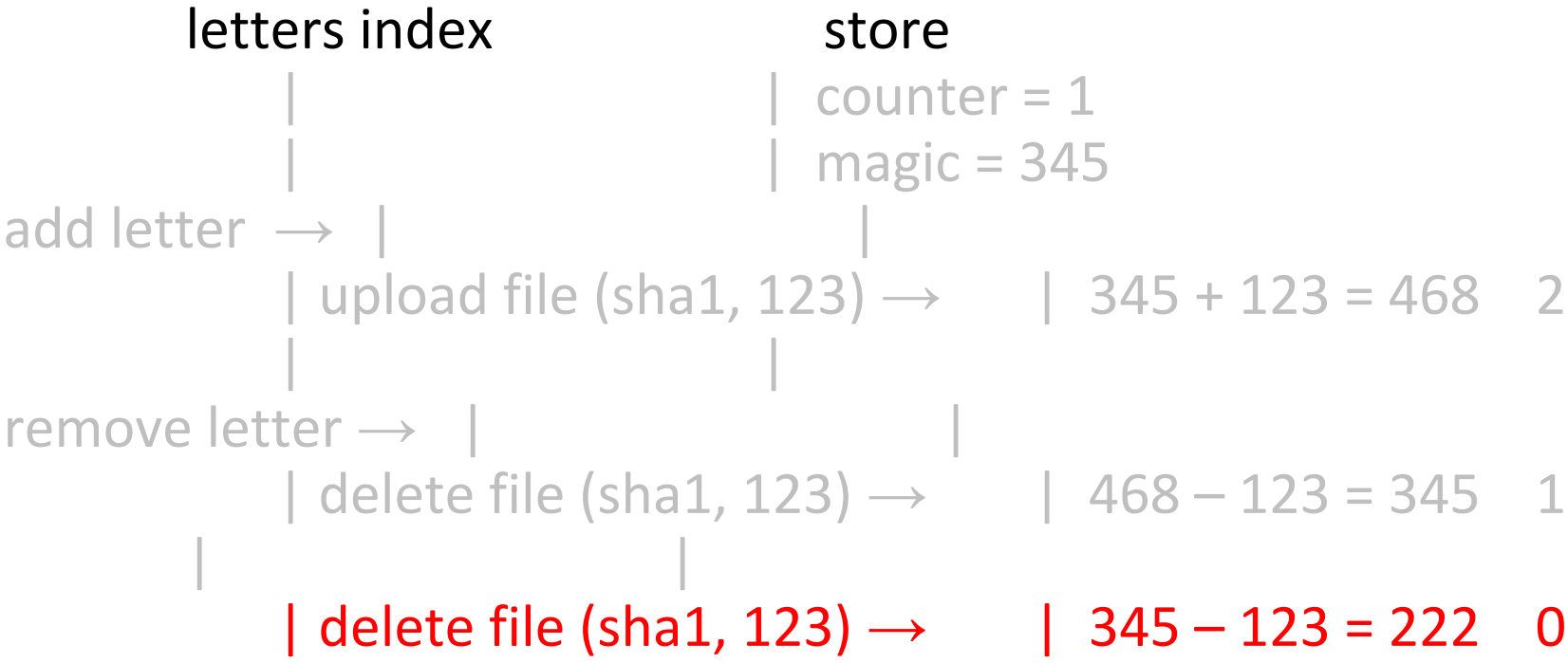


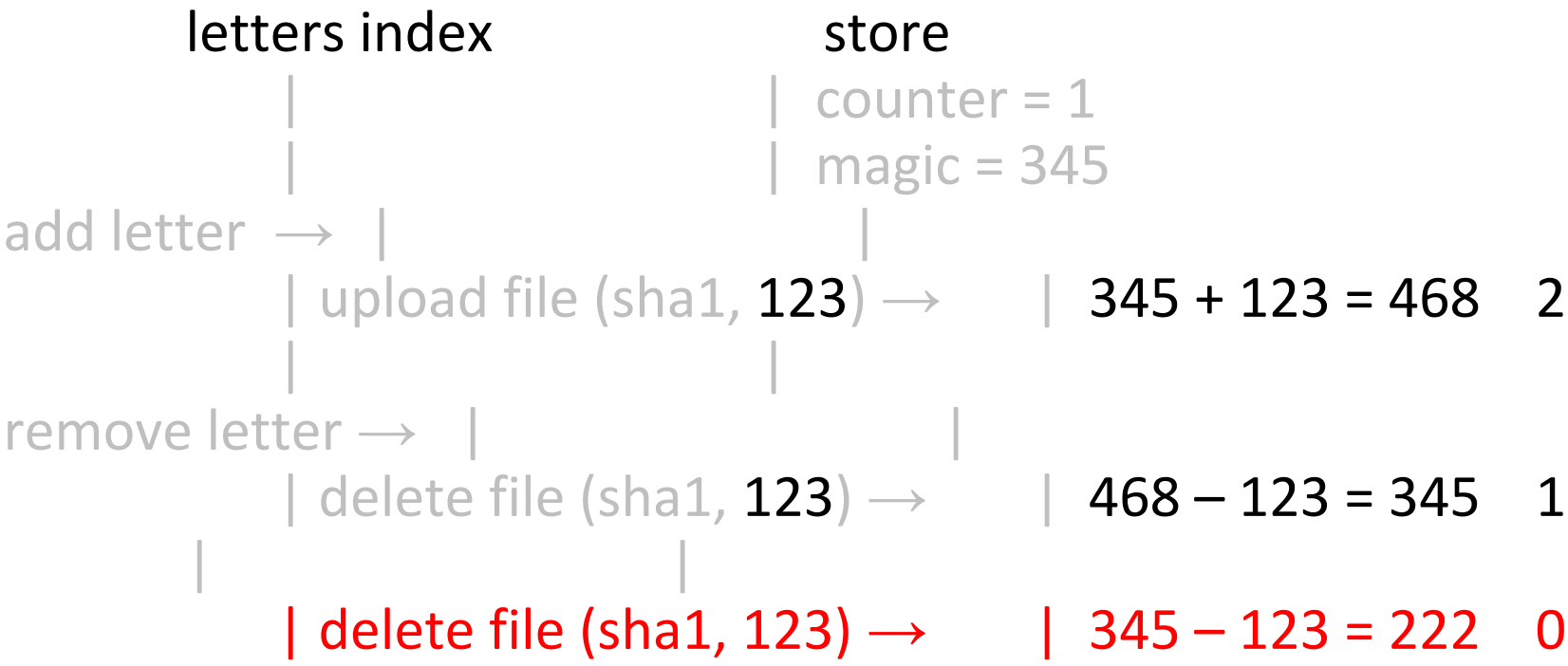




```
          letters index                    store
               |                          | counter = 1
               |                          | magic = 345
add letter  →  |                          |
               | upload file (sha1, 123) →  | 345 + 123 = 468   2
               |                          |
remove letter →  |                        |
               | delete file (sha1, 123) →  | 468 – 123 = 345   1

               | delete file (sha1, 345) →  | 345 – 345 = 0     0
```




letters index
store
counter = 1
magic = 345
add letter
upload file (sha1, 123)
345 + 123 = 468 2
remove letter
delete file (sha1, 123)
468 – 123 = 345 1
delete file (sha1, 123)
345 – 123 = 222 0


letters index
store
counter = 1
magic = 345
add letter
upload file (sha1, 123)
345 + 123 = 468 2
remove letter
delete file (sha1, 123)
468 – 123 = 345 1
delete file (sha1, 123)
345 – 123 = 222 0
delete file (sha1, 345)
222 – 345 = -123 0

FileDB

sha1

counter

magic

flags

FileDB

sha1
counter
magic
flags
IP0
disk0
IP1
disk1

FileDB

sha1

counter

magic

flags

pair_id → IP0

disk0

IP1

disk1

PairDB

id

IP0

disk0

IP1

disk1

PairDB

id

IP0

disk0

IP1

disk1

flags

PairDB

id
IP0
disk0
IP1
disk1
flags
free0
free1

FileDB		
sha1	20	
counter		4
magic		4
flags	4	
pair_id		4

36 (данные)

36 + 5 (длина полей)

36 + 5 + 16 (заголовки)

36 + 5 + 16 = 57 байт

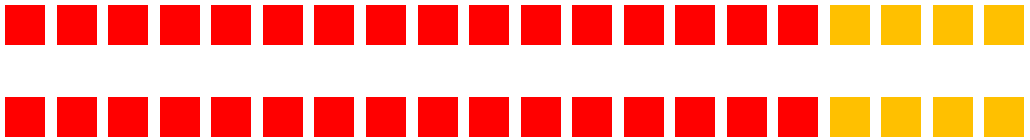
57 * 12 * 10^9 = 637 GB

12 * 12 * 10^9 = 179GB

637 GB 179 GB

637 GB

179 GB



1600 GB

PairDB содержит 14000 записей

Loader API

inc (sha1, magic)

inc (sha1, magic)

upload (sha1, magic)

inc (sha1, magic)

upload (sha1, magic)

dec (sha1, magic)

inc (sha1, magic)

upload (sha1, magic)

dec (sha1, magic)

GET /sha1

iproto

|sync cmd len|

iproto

|sync cmd len|flags|origin-len|sha1|magic|

Выбираем пару для заливки файла

|---1---|----------2----------|-----3-----|--------4--------|

Проблема чистой пары

|--1--|----------2----------|--3--|--4--|

Проблема чистой пары

|--√1--|----------√2----------|--√3--|--√4--|

Пара выбрана

|--√1--|----------√2----------|--√3--|--√4--|

Пара выбрана

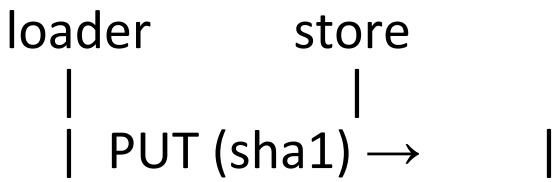
|--√1--|----------√2----------|--√3--|--√4--|

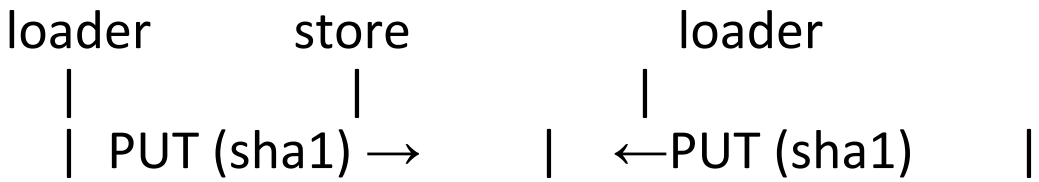
Отправляем небольшой файл на эти диски.

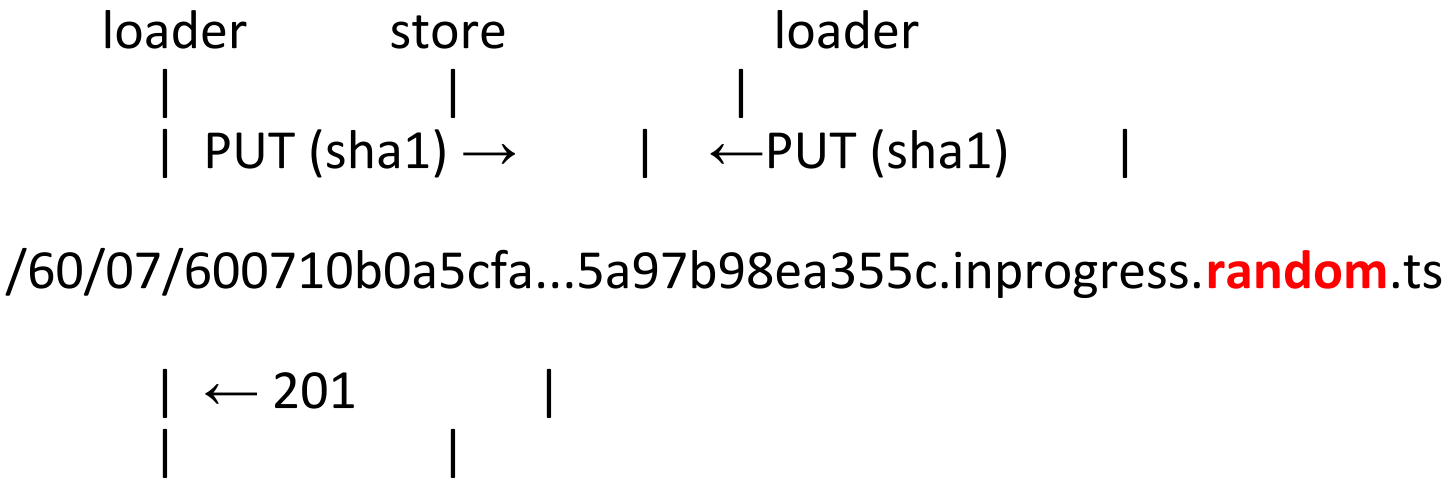
|sync cmd len|flags|origin-len|sha1|magic|filecontent.....|

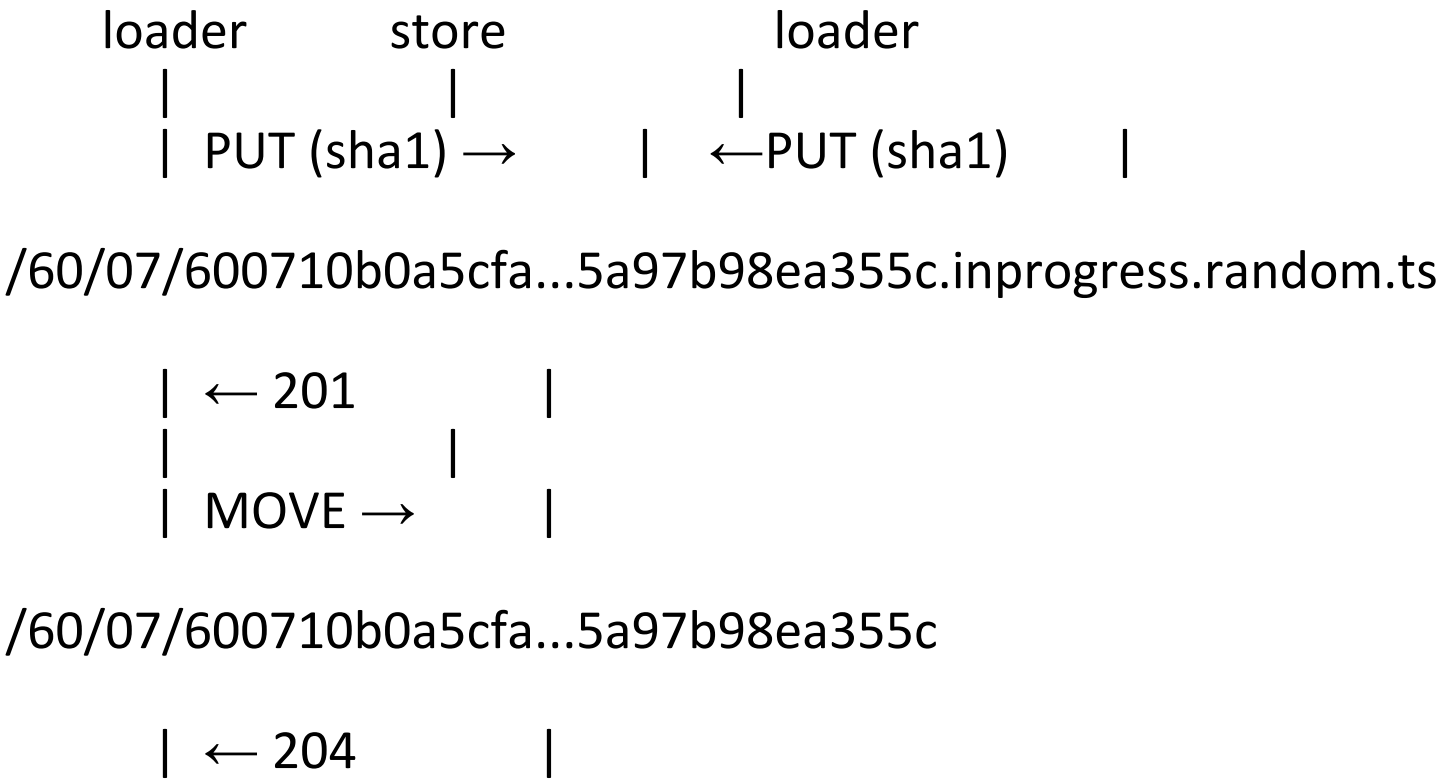
sha1 считаем на лету

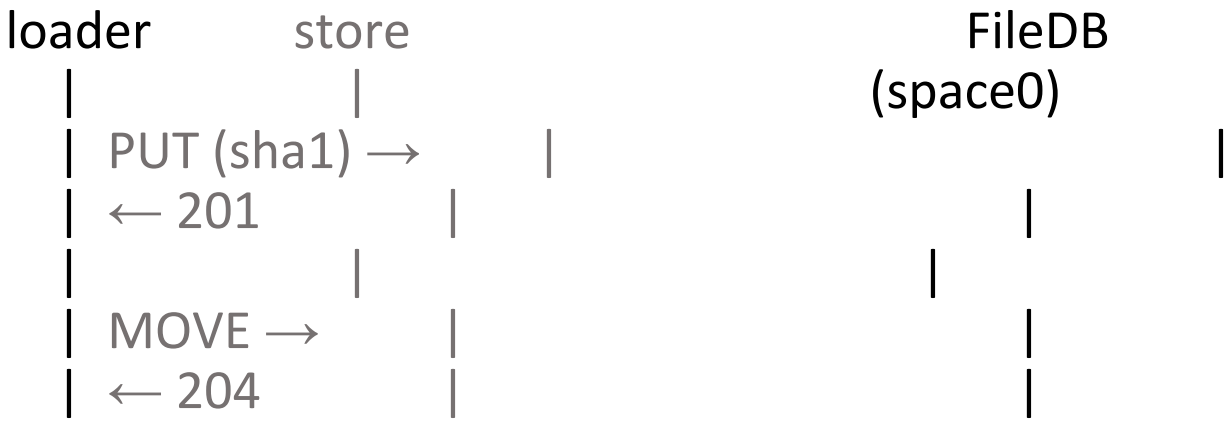
nginx + webdav

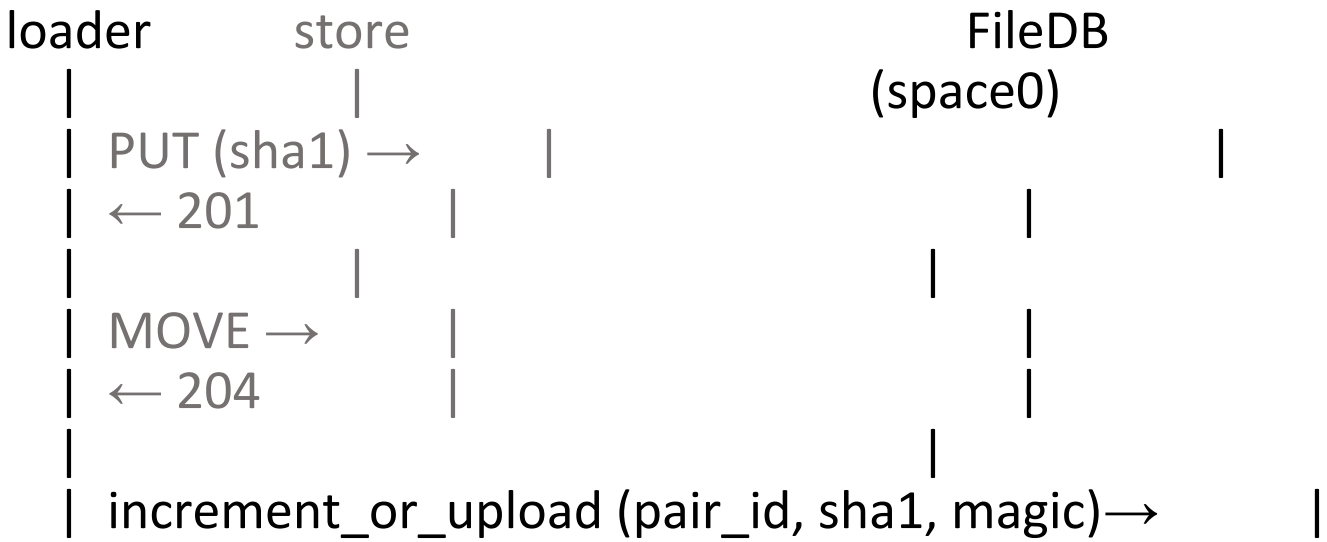


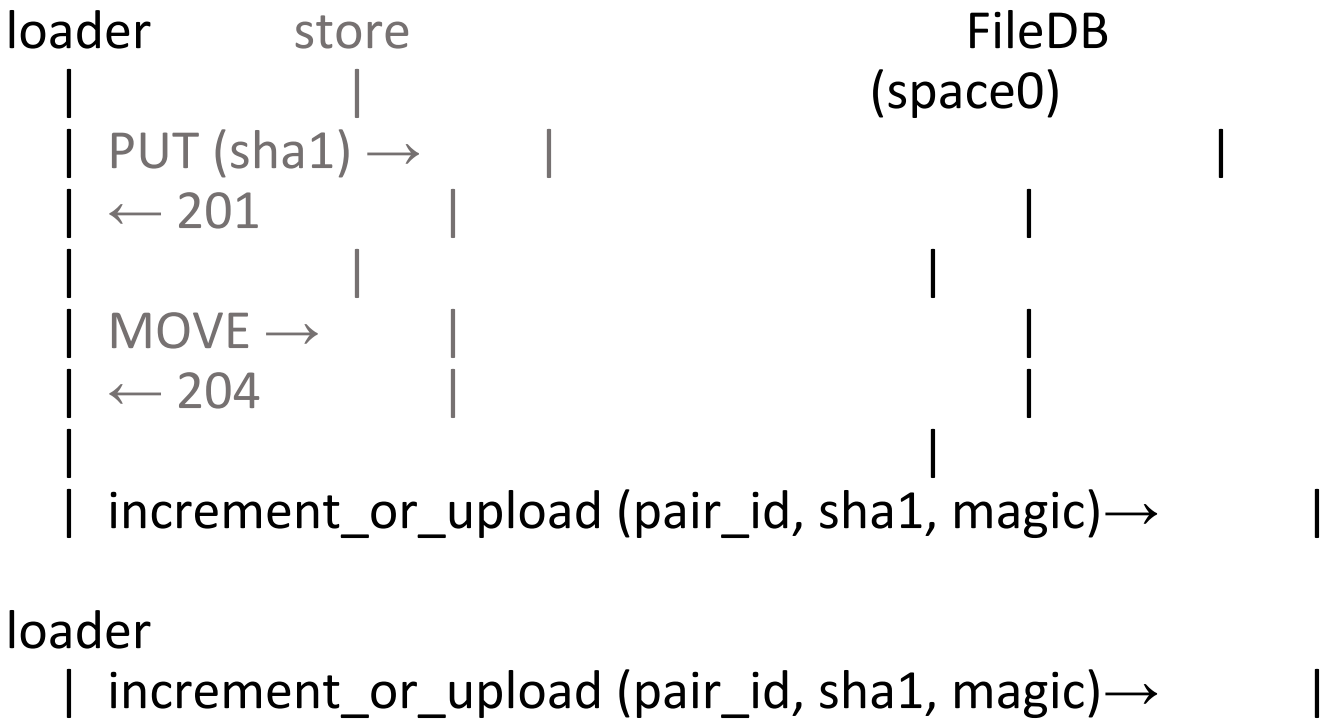



loader
store
loader
PUT (sha1) →
←PUT (sha1)
/60/07/600710b0a5cfa...5a97b98ea355c.inprogress.random.ts
← 201




loader
store
FileDB
(space0)
PUT (sha1) →
← 201
MOVE →
← 204





Удаление файла

- Гарантированно записать файл
- Быстро отдать файл
- Держать метаданные консистентными

Удаление файла

- Гарантированно записать файл
- Быстро отдать файл
- Держать метаданные консистентными

Удалять можно в offline

Удаление файла

- Гарантированно записать файл
- Быстро отдать файл
- Держать метаданные консистентными

Удалять можно в offline

При удалении только уменьшаем счетчик

decrement (sha1, magic)

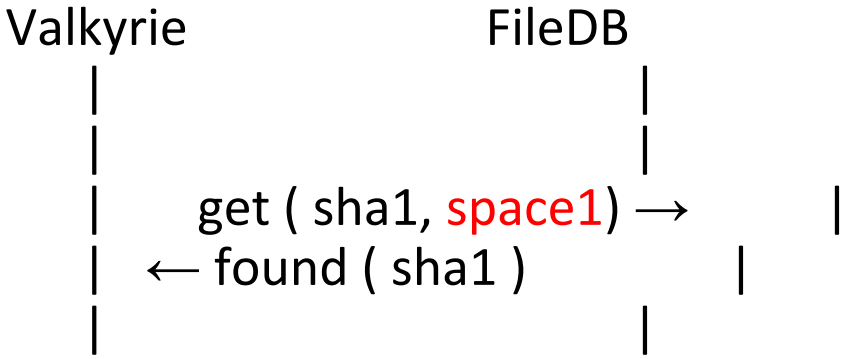
counter--
current_magic –= magic

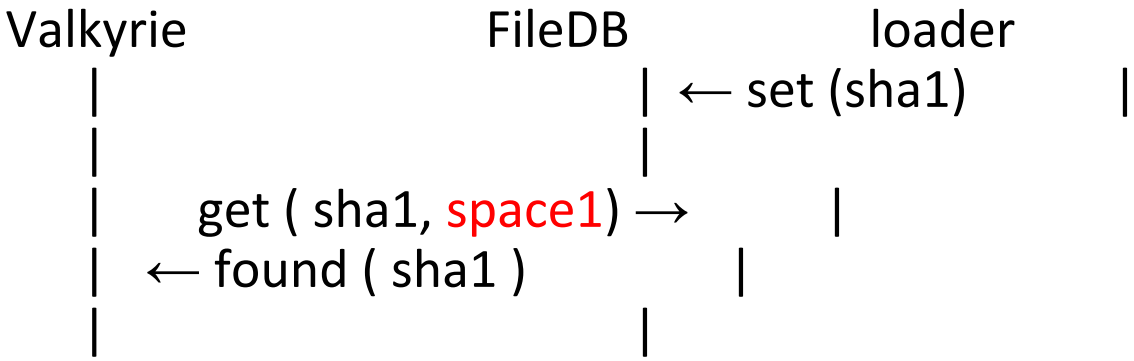
decrement (sha1, magic)

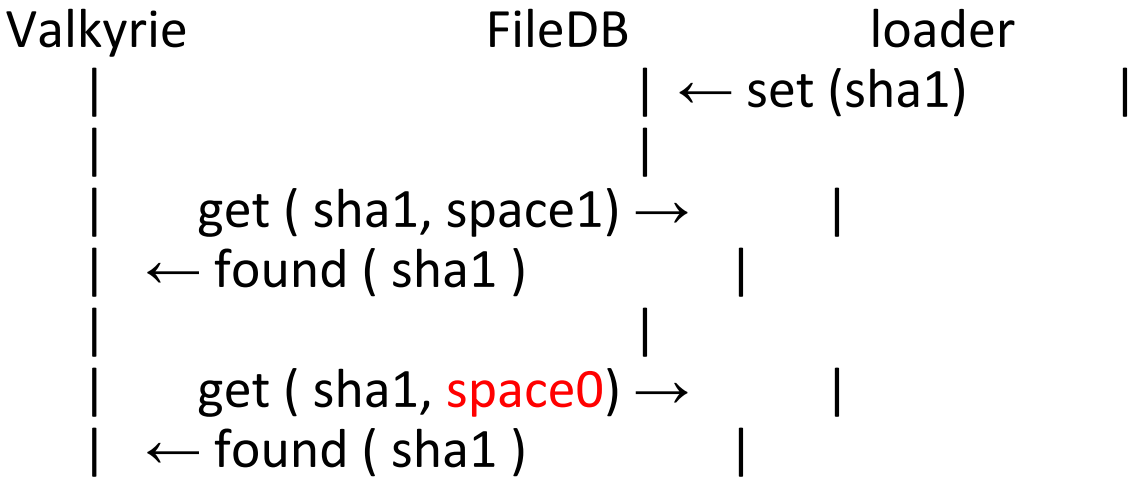
counter--
current_magic –= magic

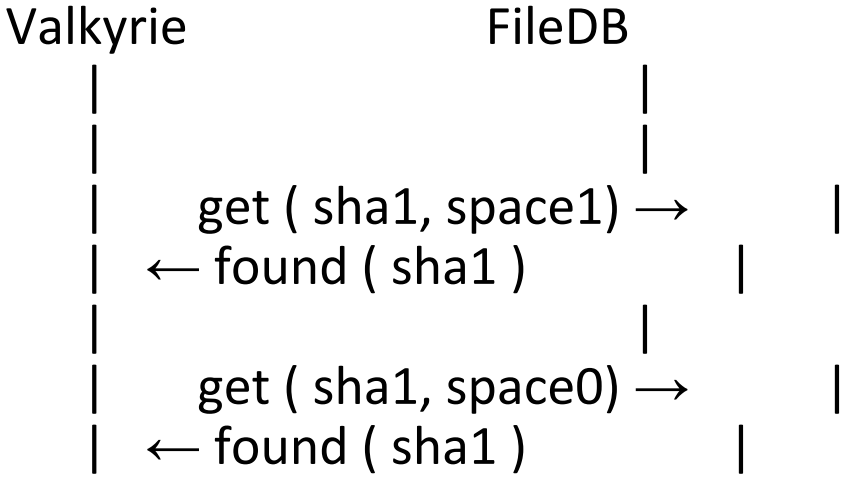
If (counter == 0 && current_magic== 0){
    move(sha1, space1)
}

Valkyrie

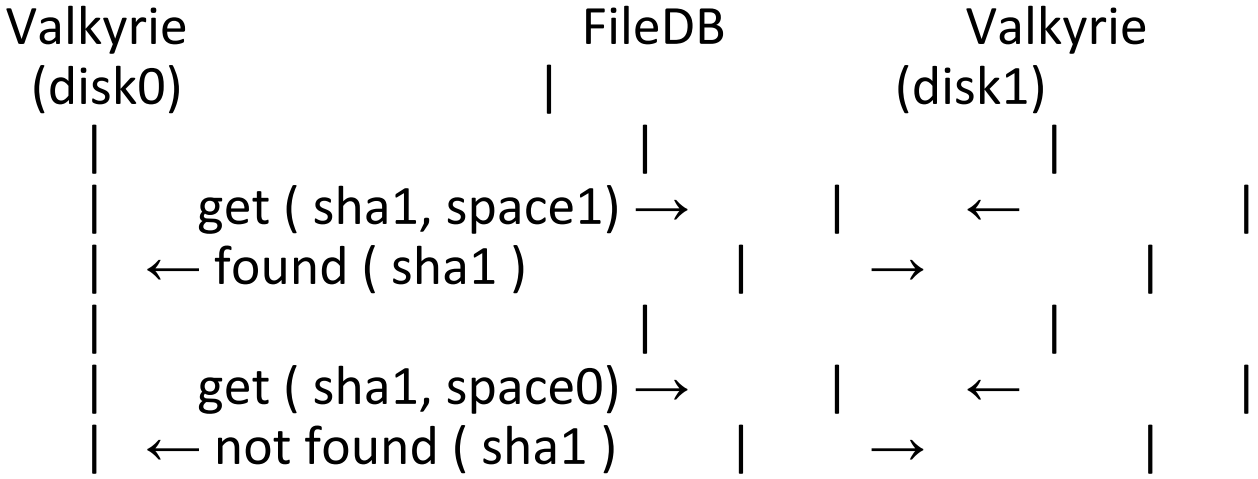








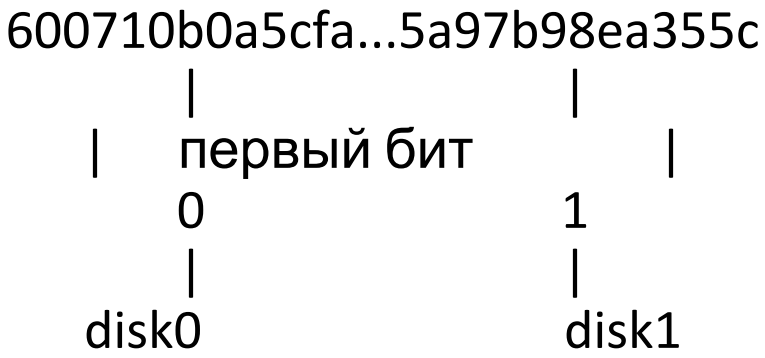
rename sha1 → sha1.deleted.ts (карантин)



Когда удалять запись из space1?

Решим кто мастер

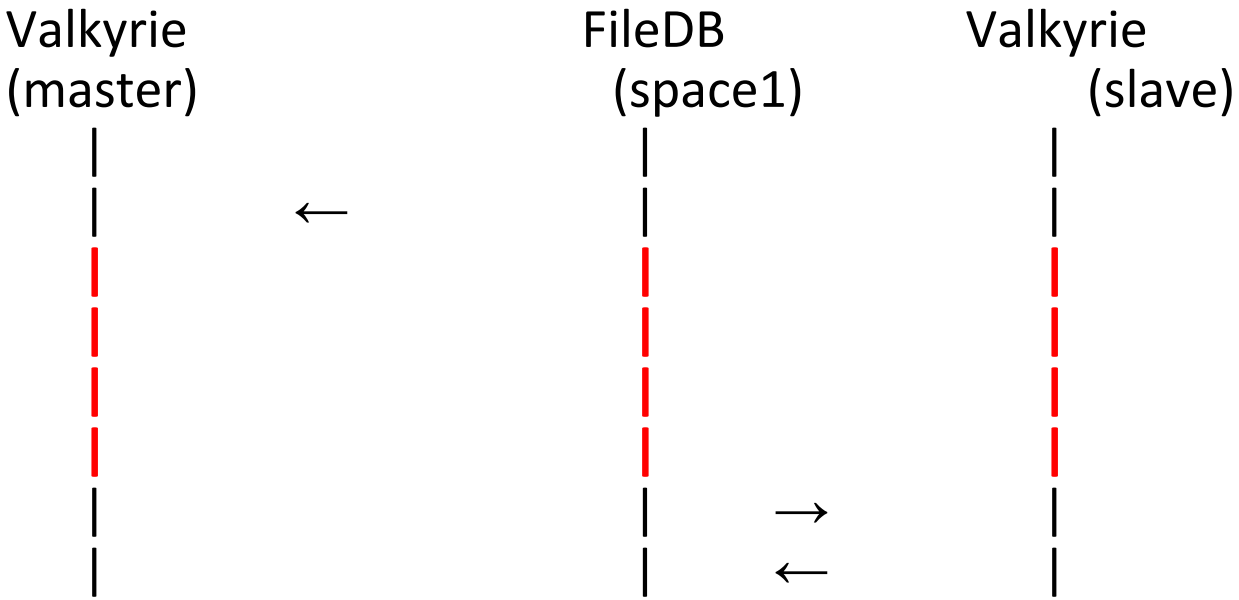
Решим кто мастер



Разнесем по времени

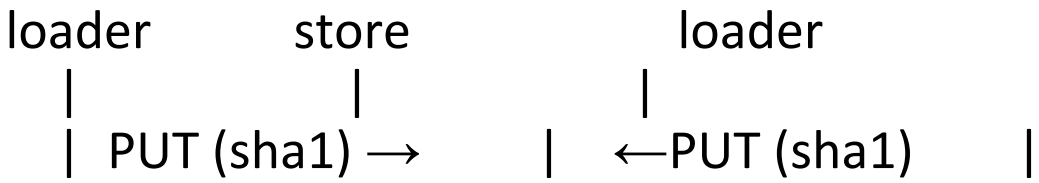
magic (space0) → проверка консистентности

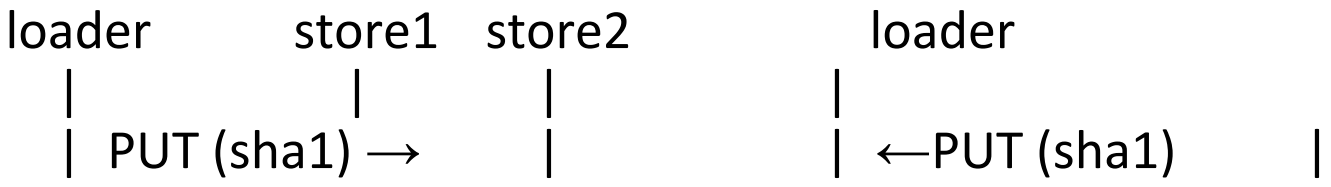
magic (space1) → timestamp удаления


Valkyrie
(master)
FileDB
(space1)
Valkyrie
(slave)

Valkyrie находит на диске файл sha1:

1. Записи о нем нет в FileDB — на карантин через sha1.deleted.ts






loader
store1
FileDB
(space0)
PUT (sha1) →
MOVE →
increment_or_upload (pair_id1, sha1, magic)→
loader
store2
PUT (sha1) →
MOVE →
increment_or_upload (pair_id2, sha1, magic)→

Valkyrie находит на диске файл sha1:

1. Записи о нем нет в FileDB — на карантин через sha1.deleted.ts
2. Запись есть, но указывает на другую пару — проверить на другой паре и удалить

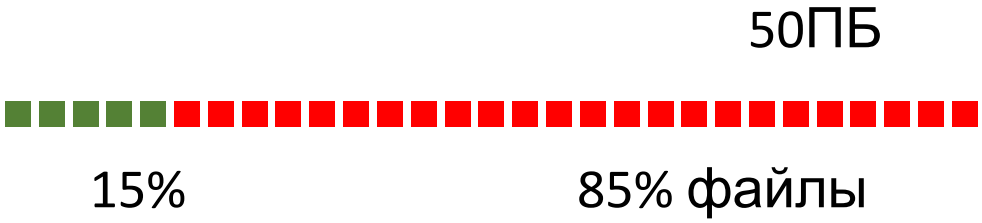
Valkyrie находит на диске файл sha1:

1. Записи о нем нет в FileDB — на карантин через sha1.deleted.ts
2. Запись есть, но указывает на другую пару — проверить на другой паре и удалить
3. В FileDB запись указывает на текущую пару — сходить HEAD на второй диск, на текущем диске проверить целостность.

Disk1 — оказался проблемным

Disk2 — readonly + размув

Из чего состоит почта



Из чего состоит почта