facebook

GFProxy: Scaling the FUSE Client

Shreyas Siravara

Production Engineer

October 7th, 2016

Our Team

Lachlan Mulcahy

David Hasson

Richard Wareing

Kevin Vigor

Max Rijevski

Drake Diedrich

Shreyas Siravara

Agenda

1	Gluster Native FUSE Client
2	GFProxy Server & Client
3	Failover
4	Usage & Performance
5	Questions?

The FUSE Client

Advantages

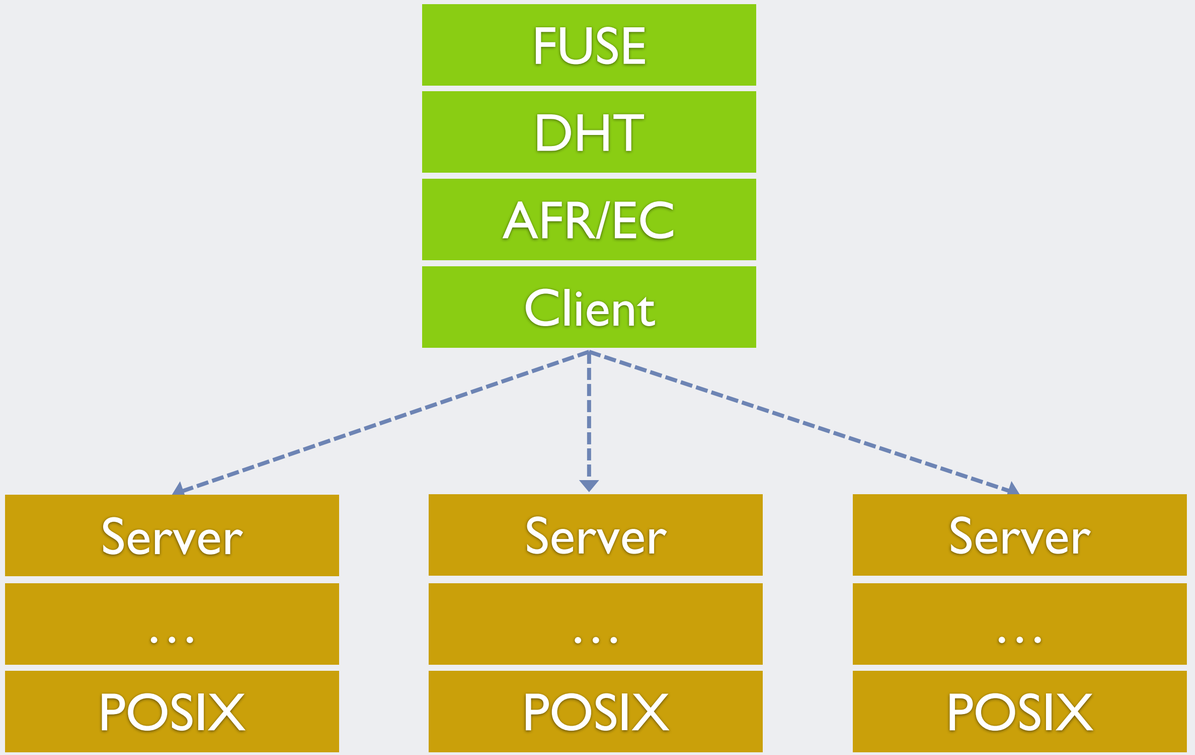
- FUSE-based driver
  - Less vulnerable to stuck mounts (e.g., NFS kernel mounts)
  - Userspace, easier to patch & update
- Better support for file-locking
- More efficient for write-heavy workloads
  - Fewer syscalls @ the brick than NFS.

The FUSE Client

Disadvantages

- Connections to all the bricks in the cluster
  - ~10k connections per brick in some cases
- Operational challenges
  - Difficult to track down and upgrade 1000s of clients
- Client-side network magnification when using replication

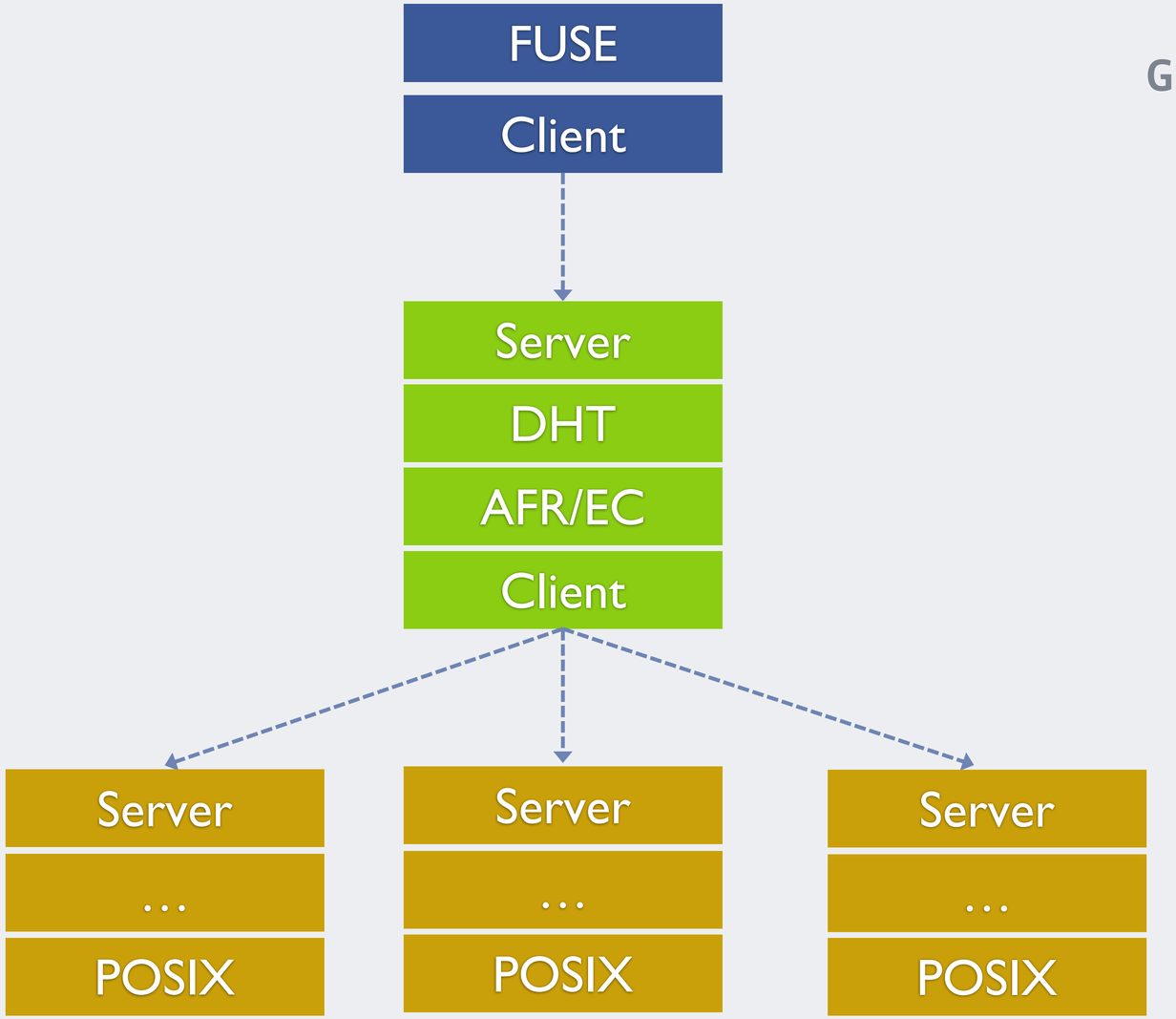
Native FUSE Translator Stack



Native FUSE Client

Brick Servers

GFProxy Translator Stack

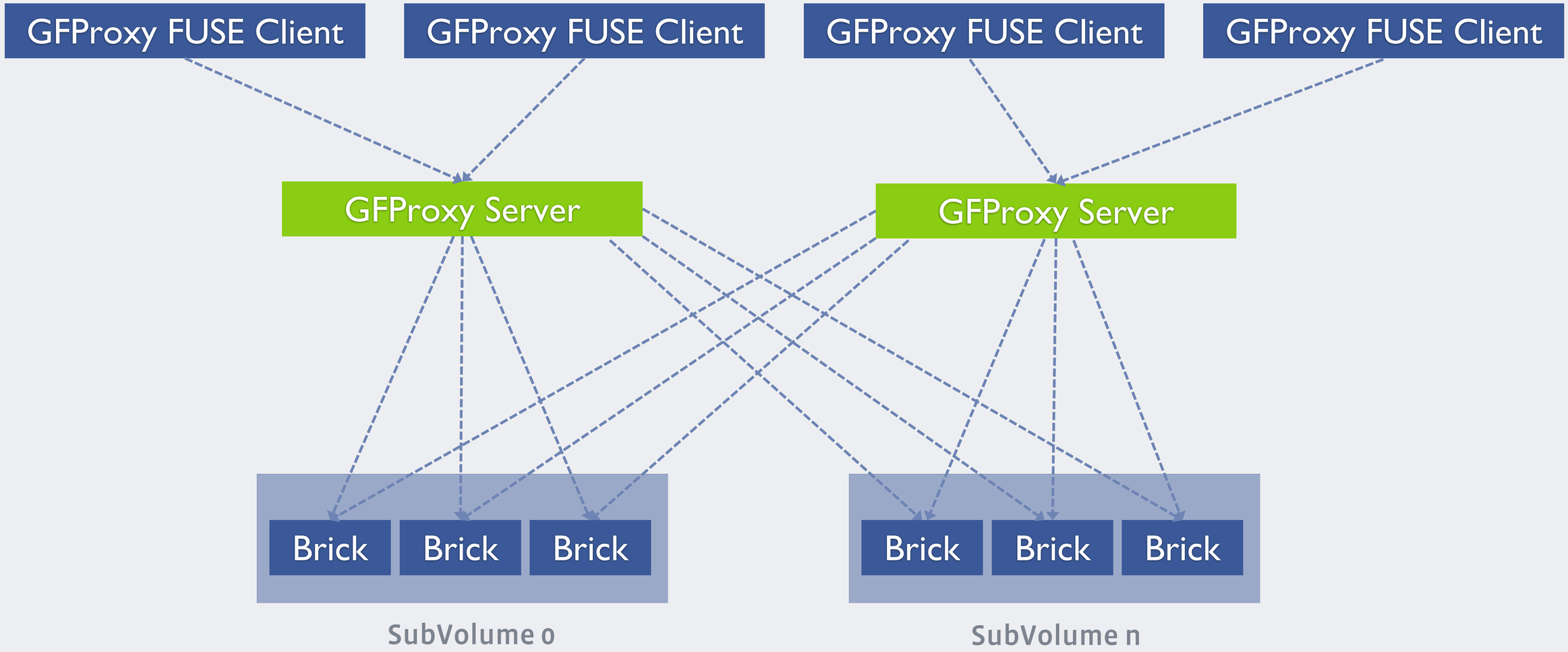


GFProxy FUSE Client

GFProxy Server

Brick Servers

GFProxy Servers & Clients



The FUSE Client

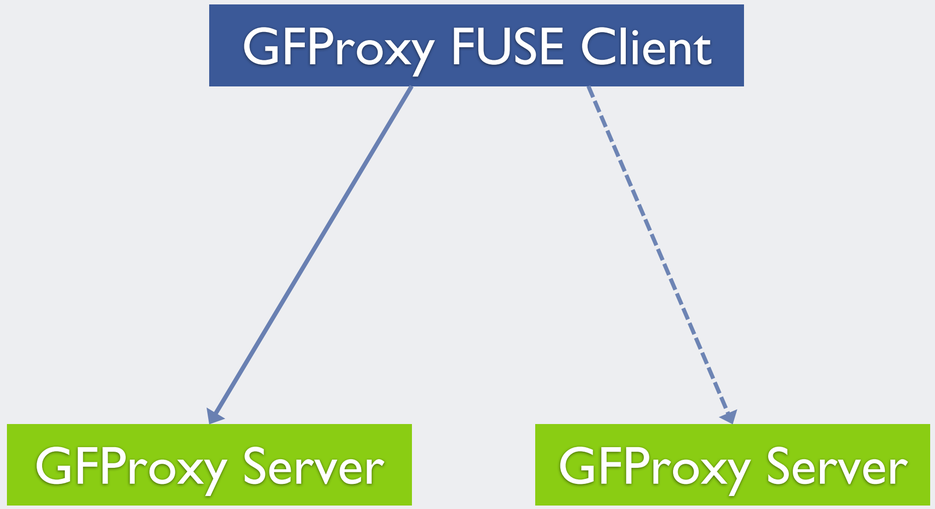
Improvements

- Single connection to a GFProxy Server
- Upgrades can happen on server-side for core changes
- No client-side network magnification

Failover

What happens when a GFProxy endpoint dies?

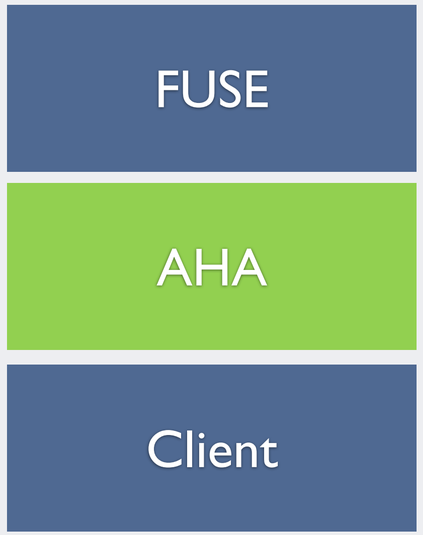
- Client translator gets a disconnect event
  - Unwinds the FOPs with `ENOTCONN`
- Applications using the FUSE mount receive `ENOTCONN`
- Interrupts workloads during unexpected failure or planned maintenance (e.g., upgrades)

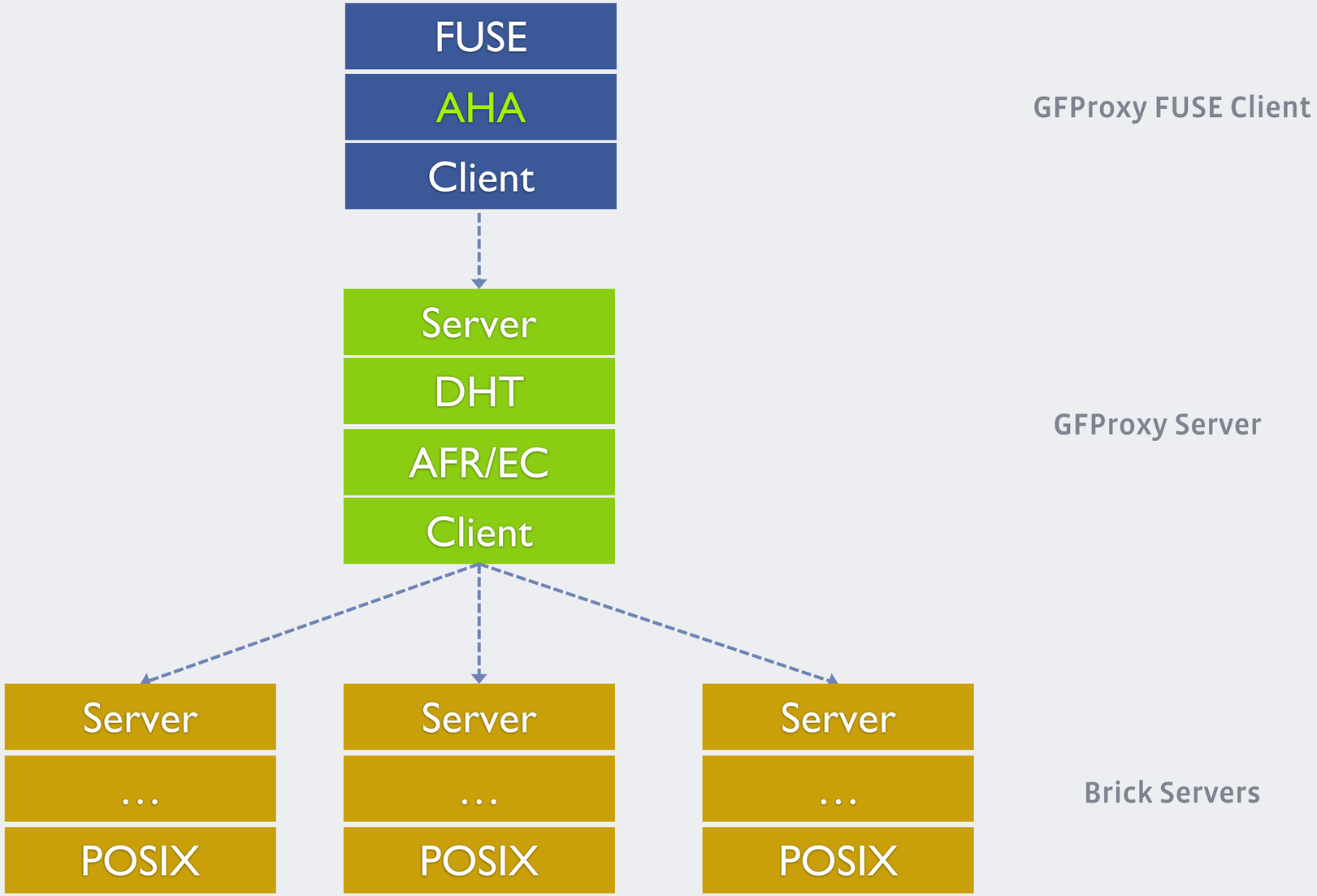


Failover

AHA (Advanced High Availability) Translator

- Sits in the FUSE client
- If a FOP returns with `ENOTCONN`:
  - Queue the FOP to be retried later
- Upon receiving `GF_EVENT_CHILD_UP`:
  - Retry all the FOPs in the queue




FUSE
AHA
Client
Server
DHT
AFR/EC
Client
Server
…
POSIX
Server
…
POSIX
Server
…
POSIX
GFProxy FUSE Client
GFProxy Server
Brick Servers

Usage

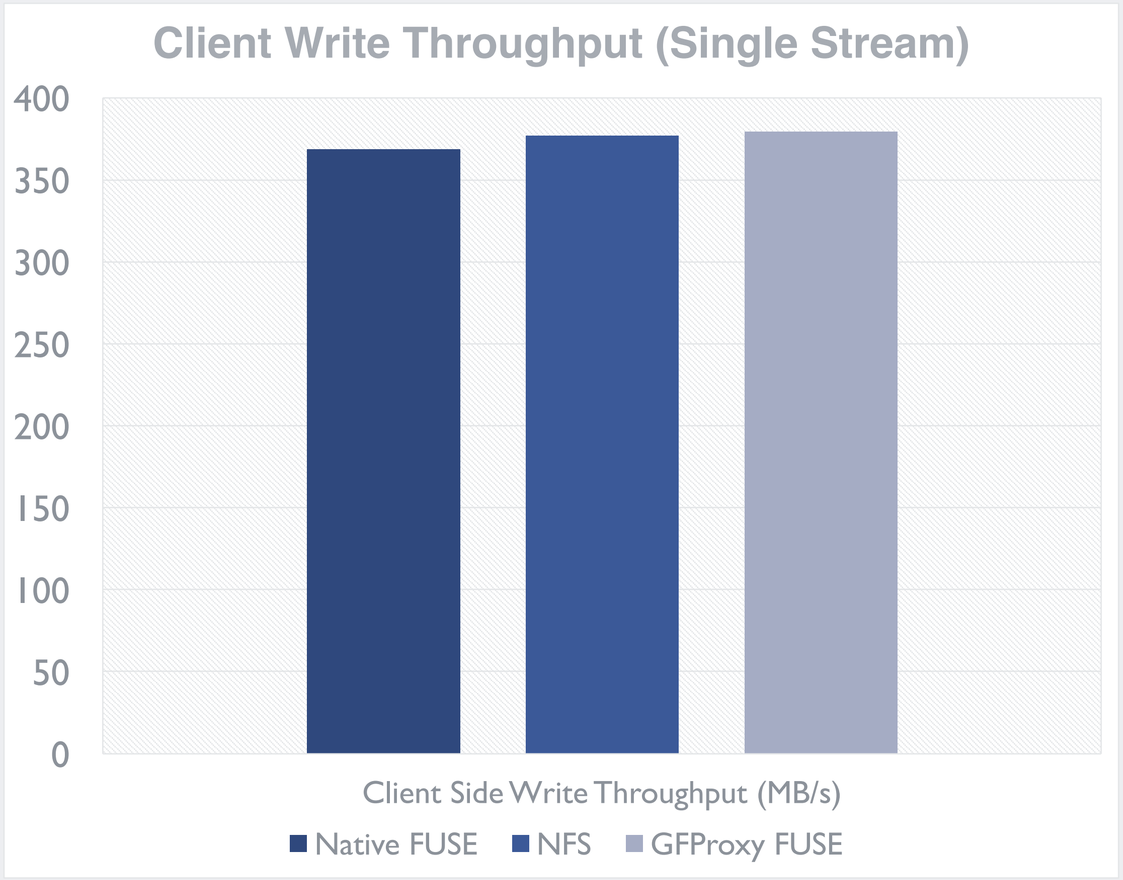
- GFProxy server starts with glusterd
- Volume files for client & server generated when creating the volume

```
mount -t glusterfs -o gfproxy host:/volume /mnt
```

```
glusterfs --volfile-id=gfproxy-client/volume host /mnt
```

Performance

Compared with NFS & Native FUSE Client



Single Client, writing a 4GB file

6 clients, writing 4GB each

Future Work

- Currently supports single volume only
  - Needs better integration with glusterd portmapper
- Open Source

Thank You!

facebook