GlusterFS
Challenges and Futures

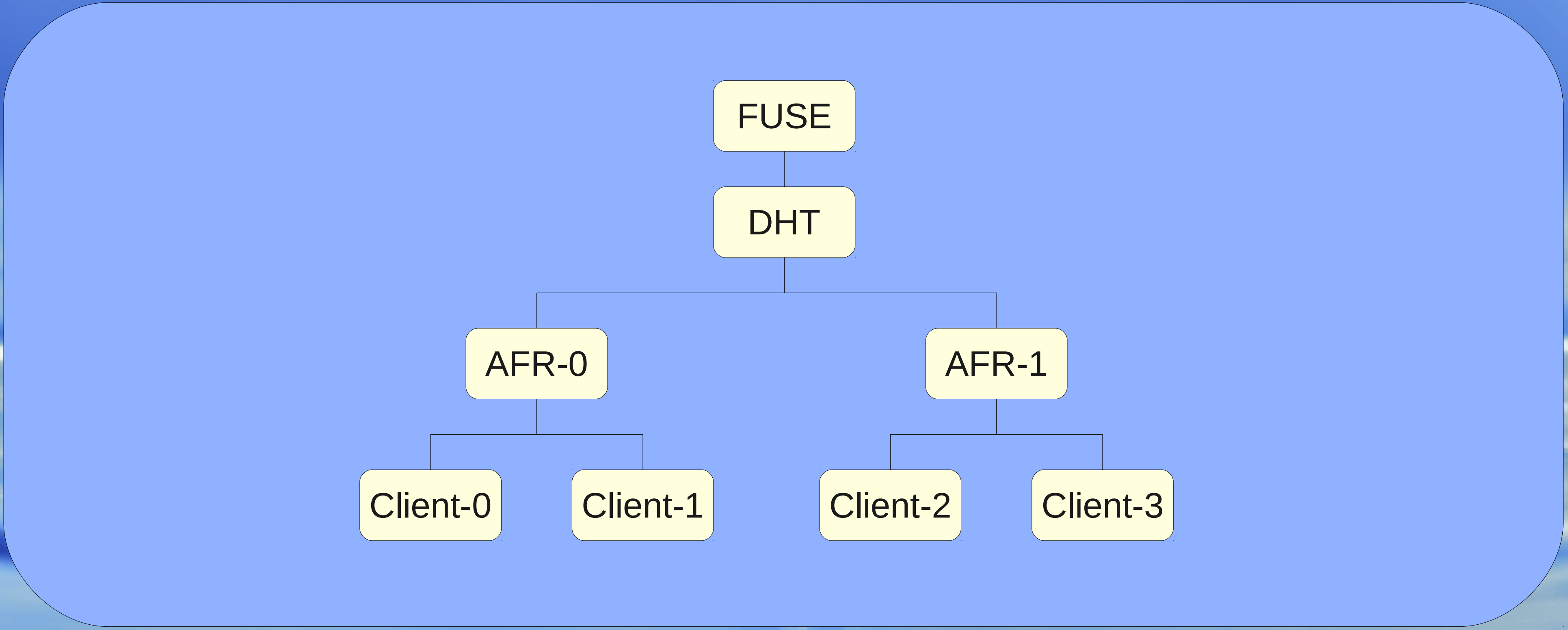
Jeff Darcy
Storage Developer Conference
September 17, 2012

What Is GlusterFS?

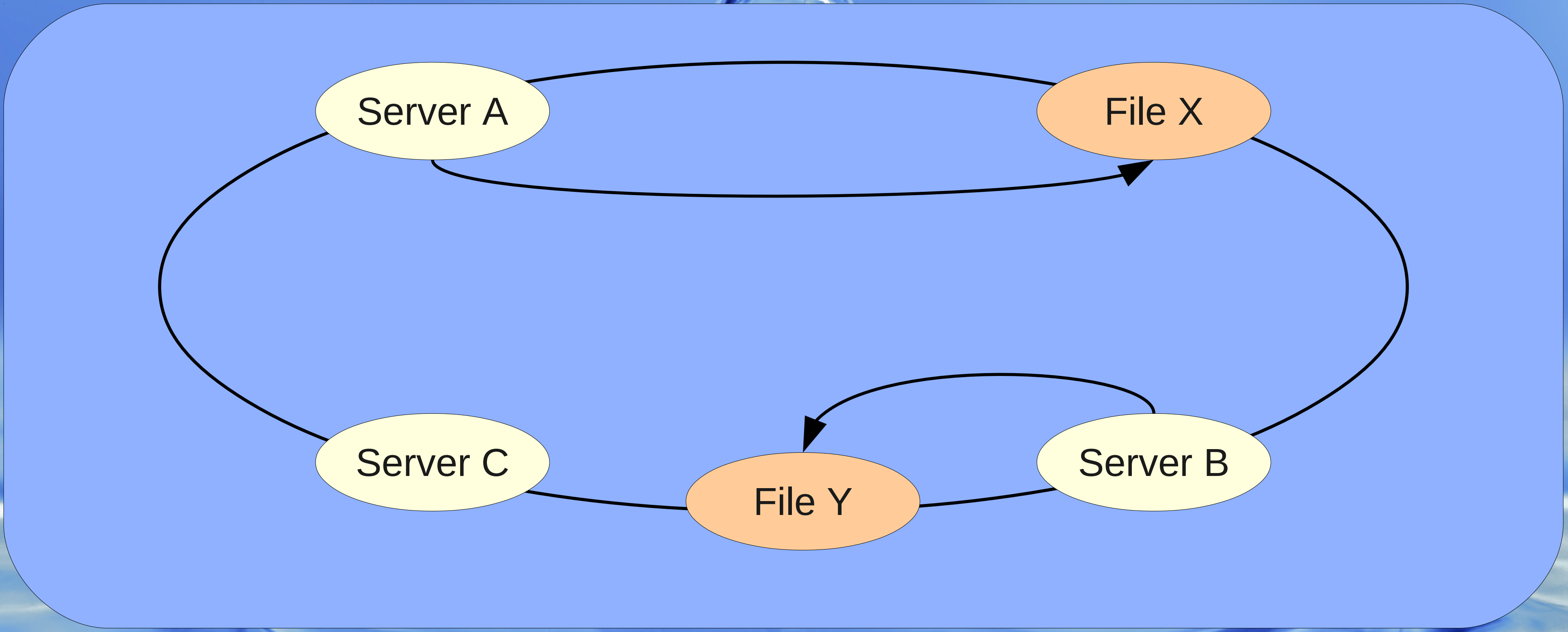
- Just your basic distributed file system
  - sharding, replication, striping
- Decentralized
  - no central metadata server
  - core functionality on clients
- Modular
  - “translators”

Translator Stacking

- One to one, one to many, one to zero (?)
- Rearrange, move from client to server, ...



Distribution (now)



Replication (now)

- Based on changelog (“dirty flags”)
  - set flags, do operation, clear flags
  - use flags to determine repair (“self-heal”) after failure
- Latency sensitive
  - 3+ network round trips per user request
  - implementation heavily optimized

Challenge: Multitenancy

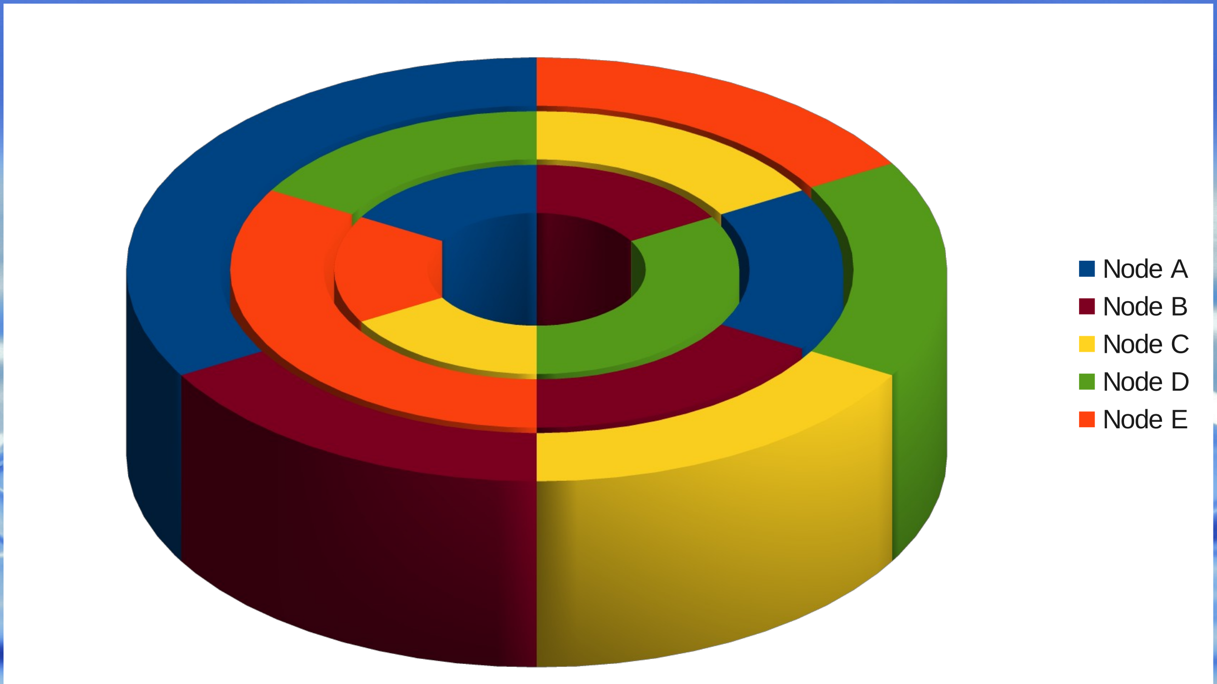
- Focus of CloudFSHekaFS
  - to be merged with GlusterFS soon-ish
- Isolate name and ID spaces
- Encrypt in flight and at rest
- Auth*
- Quota (next slide)
- Performance isolation
  - cgroups

Challenge: Distributed Quota

- Can't trust clients to enforce quota
- Can't just divide equally among servers
  - unequal usage (e.g. due to explicit placement)
  - EDQUOT on one while still space on another
- “Quota rebalancing daemon”
  - monitor/adjust continuously
  - interesting problem at high scale

Challenge: Better Rebalancing

- Optimal placement vs. minimal data movement
- Different kinds of weighting



Challenge: Replication Latency

- Reaching limits of current approach
- Have to go async?
  - but still ordered
  - exploit compute/data locality (e.g. Hadoop)
  - journaling, conflict resolution

Challenge: Directory Traversal

- Piggyback attrs (and xattrs) on readdir
- Even better: cursor approach
  - read everything in opendir
  - zero network activity for readdir
  - less current, but more consistent

Challenge: Many Small Files

- Prefetch whole directories
    - if marked, below size threshold, ...
- Exploit async journal
    - only works if compute/data colocated
- Weaken consistency?
    - allow create/write/close to be buffered
    - directory-level fsync (magic xattr)

Conclusions

- Most of these challenges are not unique to GlusterFS
- Modularity and incremental progress are preferable to monolithic “solve all problems at once”
- GlusterFS provides a good environment in which to experiment with solutions