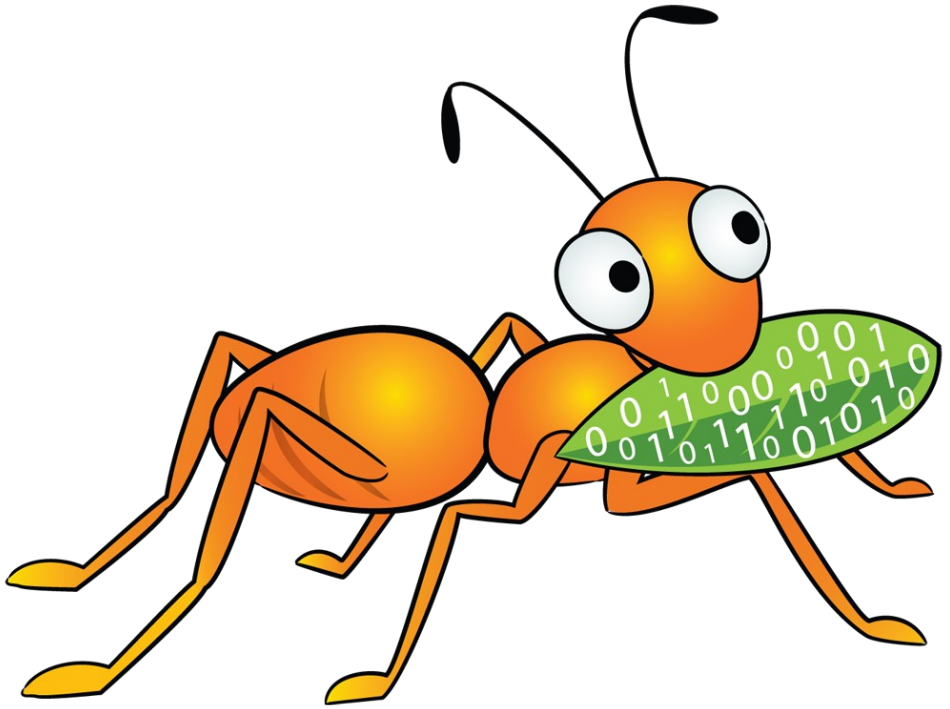
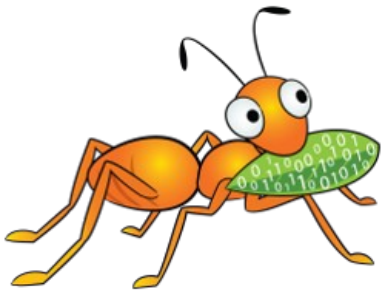
The State of the Gluster Community



John Mark Walker
Gluster Community Guy
November 8, 2012

Topics

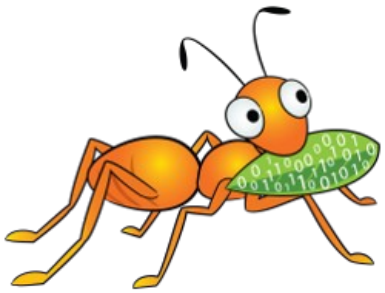
- What is GlusterFS
- GlusterFS 3.3
- Ways to Access GlusterFS Volumes
  - GlusterFS client (FUSE)
  - Libgfapi – QEMU block device driver
  - Translators
  - Swift API – Unified File and Object
- The Roadmap



Simple Economics

- **Simplicity, scalability, less cost**

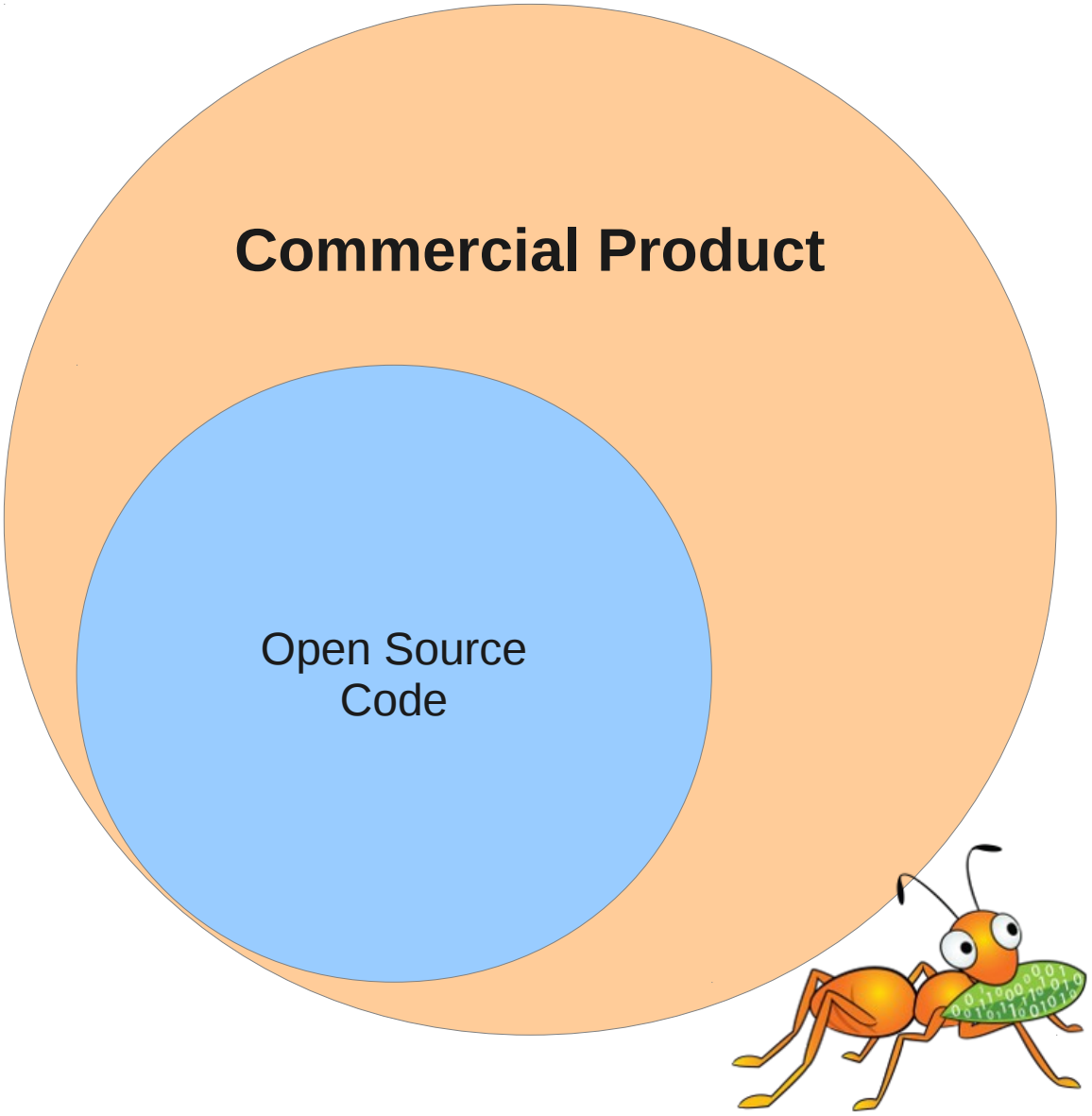
Virtualized	Multi-Tenant	Automated	Commoditized
Scale on Demand	In the Cloud	Scale Out	Open Source



What is Open Source

“Open Core”

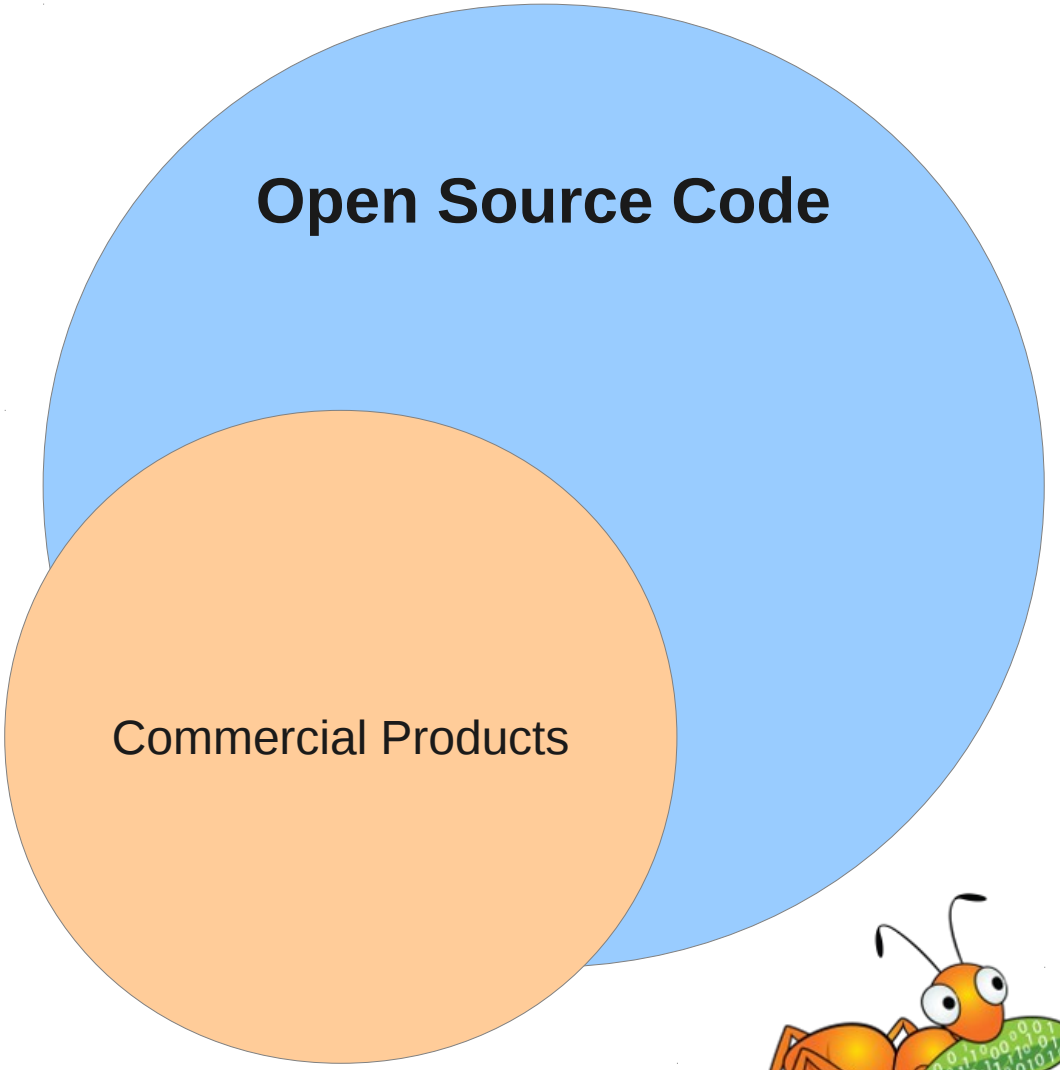
- All engineering controlled by project/product sponsor
- No innovation outside of core engineering team
- All open source features also in commercial product
- Many features in Commercial product not in open source code



What is Open Source

“Real” Open Source

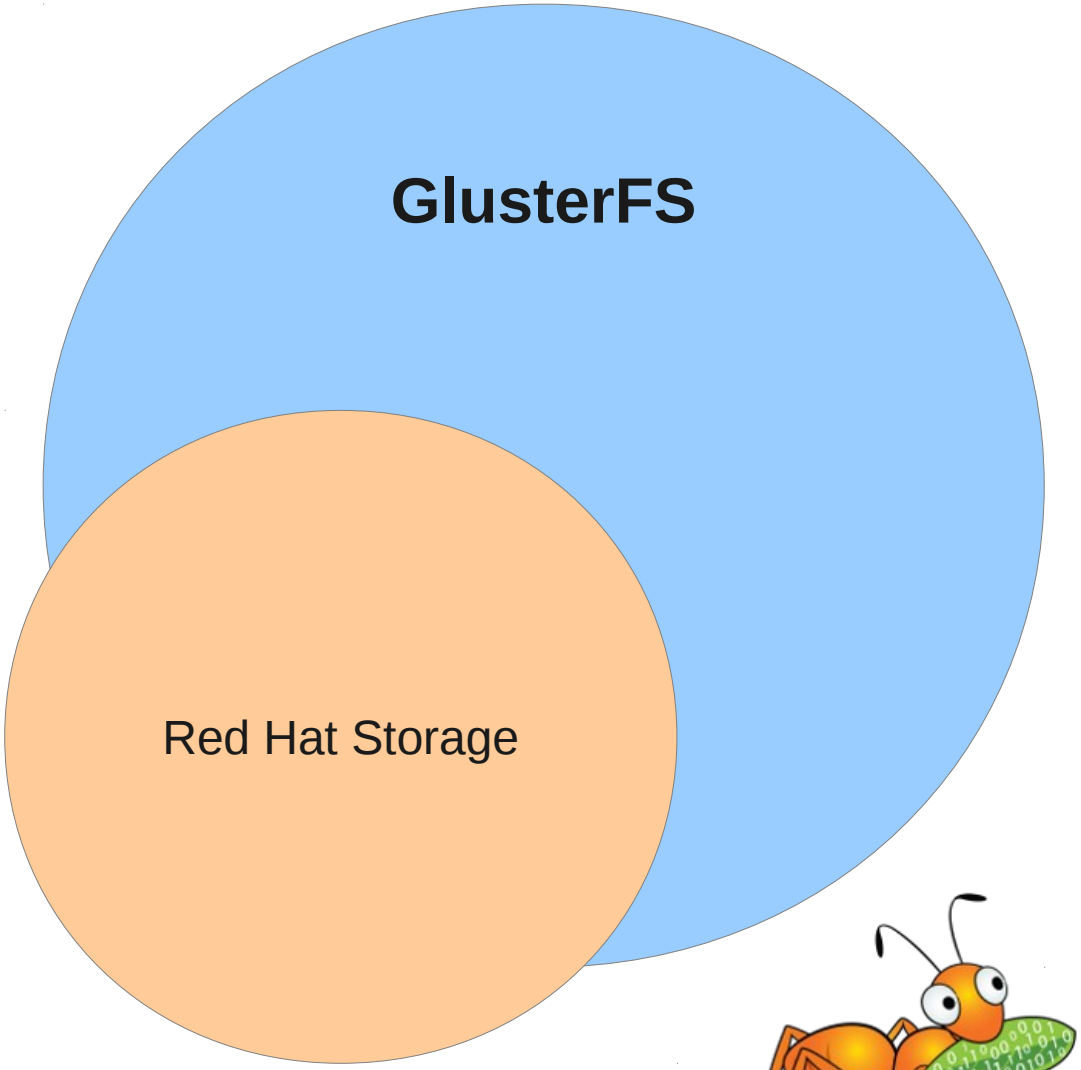
- Many points of collaboration and innovation in open source project
- Engineering team from multiple sources
- Project and product do not completely overlap
- Commercial products are hardened, more secure and thoroughly tested



What is Open Source

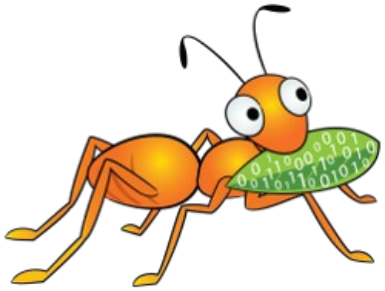
“Real” Open Source

- Enables more innovation on the fringes
- Engineering team from multiple sources
- Open source project is “upstream” from commercial product
- “Downstream” products are hardened, more secure and thoroughly tested



Simplicity Bias

- FC, FCoE, iSCSI → HTTP, Sockets
- Modified BSD OS → Linux / User Space / C, Python & Java
- Appliance based → Application based



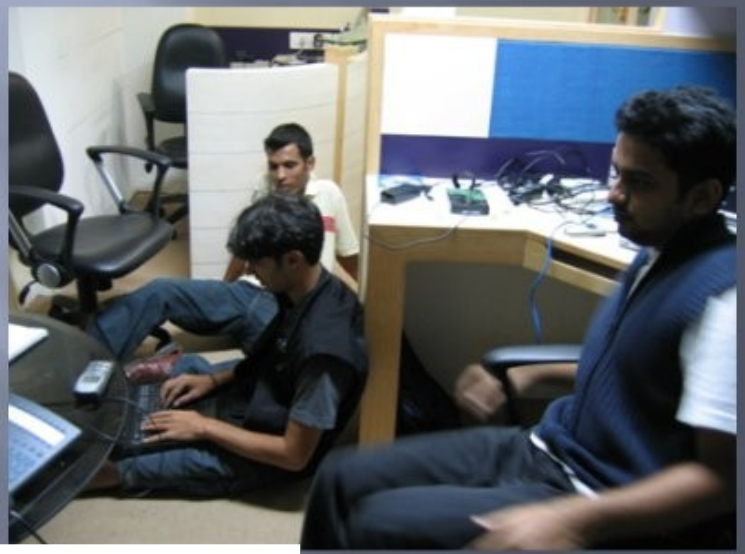
Scale-out Open Source is the winner

Conference Room

All work and no play?

US Head Office

Meeting Round 1

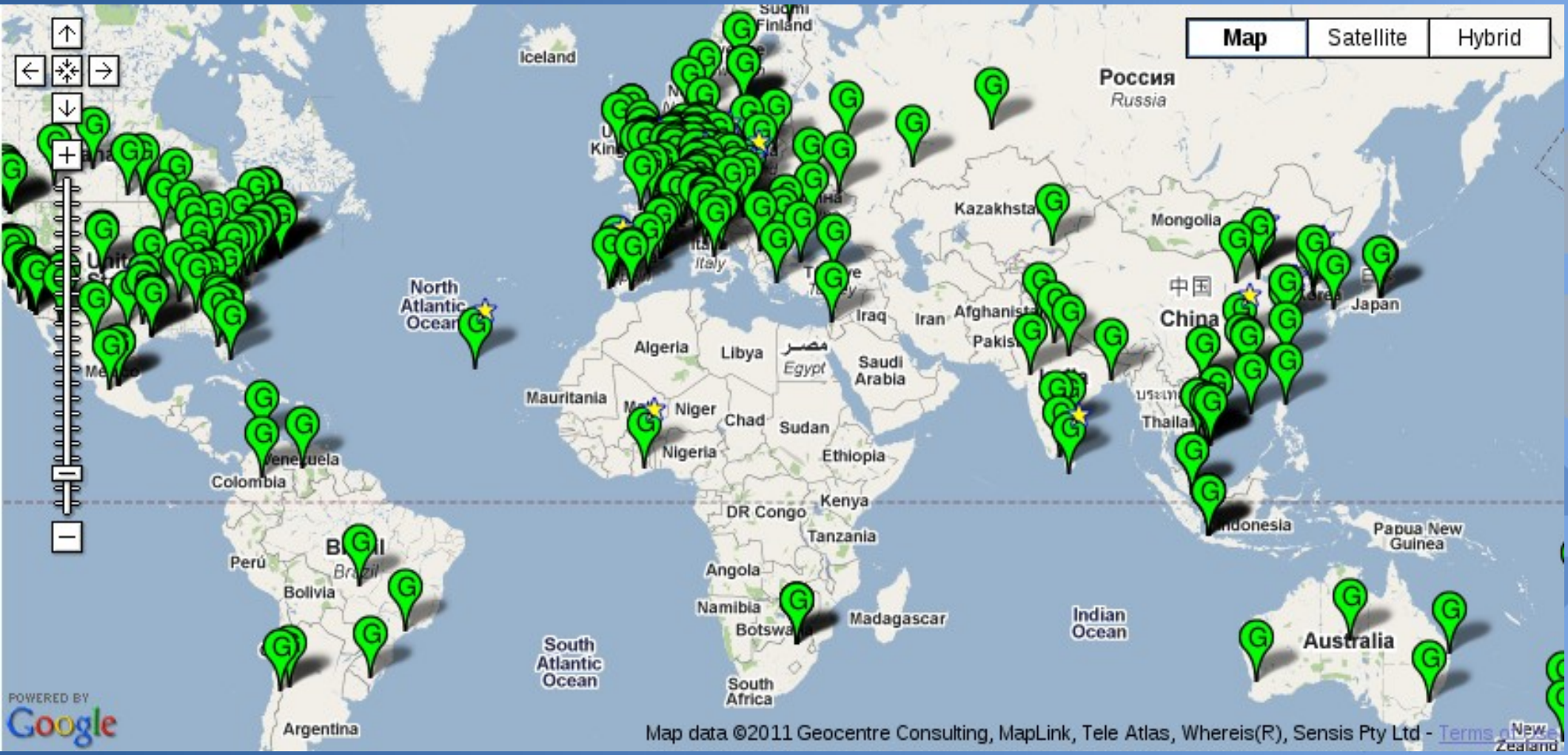


Bengaluru Office

Bedroom

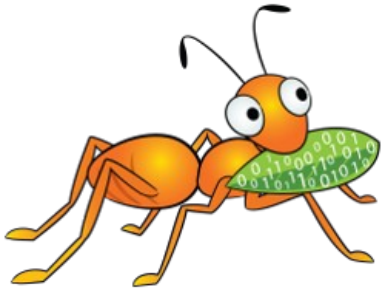
Bengaluru Office

Community Deployments



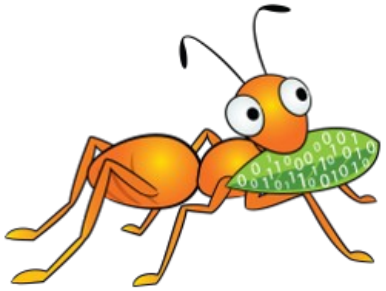
Not a Storage Company

- At first a cluster-building company
- Engineering team excelled at building open source HPC systems



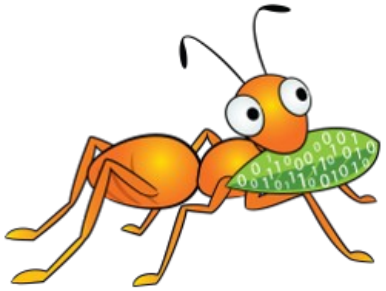
Necessity:

The Mother of Invention



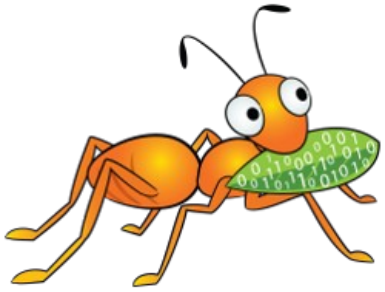
The big idea:

Storage should be simple



What is Simple Storage?

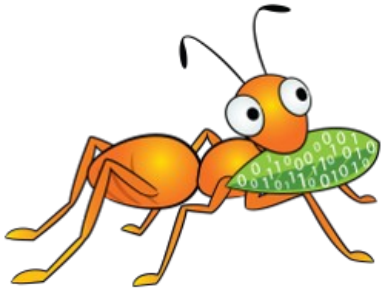
- Low-risk, easy to deploy and administer, data consistency, open source, software-only, user space



What is GlusterFS, Really?

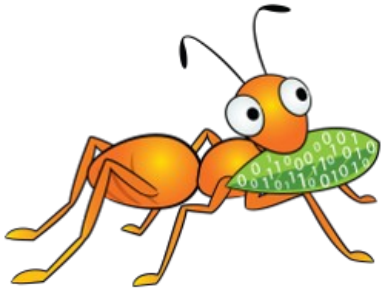
Gluster is a unified, distributed storage system

- User space, global namespace, stackable, POSIX-y, scale-out NAS platform, inspired by GNU Hurd



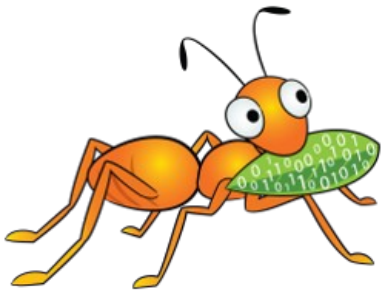
Some Features

- No single point of failure
  - DHT
- Synchronous and asynchronous replication
- Proactive self-healing



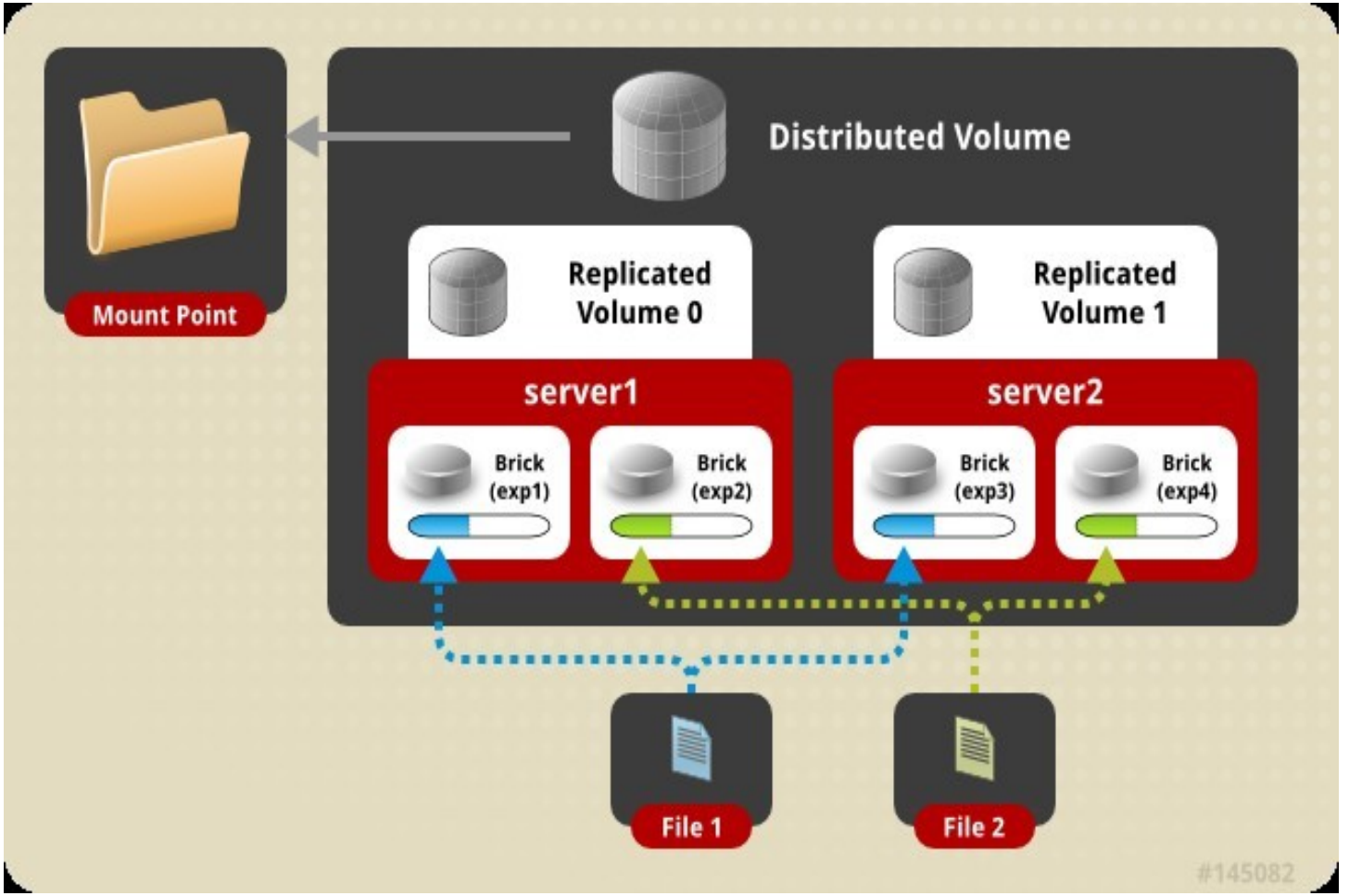
What Can You Do With It?

- **Media –** Docs, Photos, Video
- **Shared storage** – multi-tenant environments
- **Big Data** – Log Files, RFID Data
- **Objects** – Long Tail Data



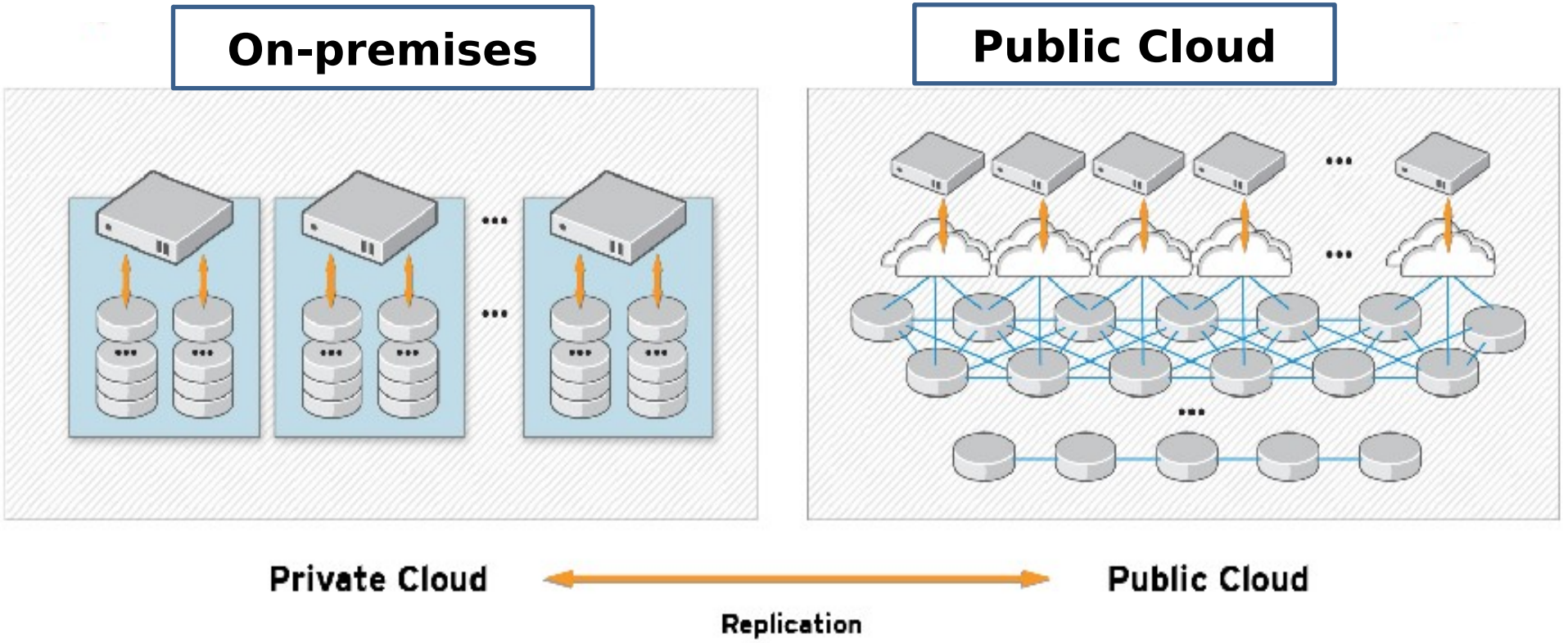
Standard Deployment

- Distributed over multiple servers
- Replicate volumes
- On top of disk FS (XFS, Ext4, ie. Xattrs)
- Multi-protocol access



Storage for Any Environment

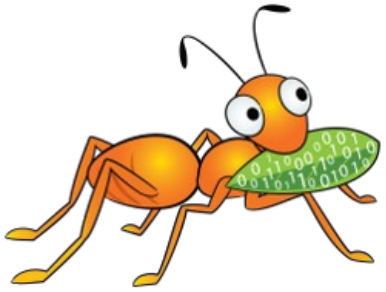
Scale-out NAS for On-premises and Public Clouds



- Standardized NAS infrastructure
- On-premise and public cloud
- POSIX-ish
- Apps move easily between environments
- Replicate between both

First Versions

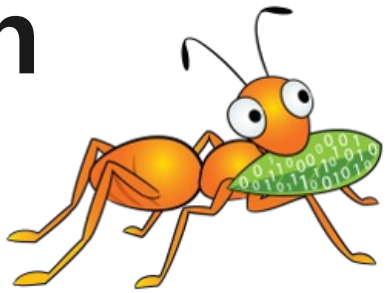
- Toolkit for building storage systems
- Very hacker-friendly
- Community integral part of development
  - Drove feature development
  - Repeatable use cases



Mid-2011 Snapshot

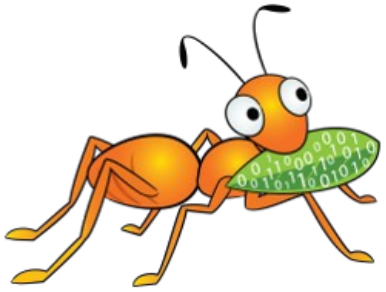
- Scale-out NAS
- Distributed and replicated
- NFS, CIFS and native GlusterFS
- User-space, stackable architecture
- Lots of users, not many devs

→ **A good platform to build on**



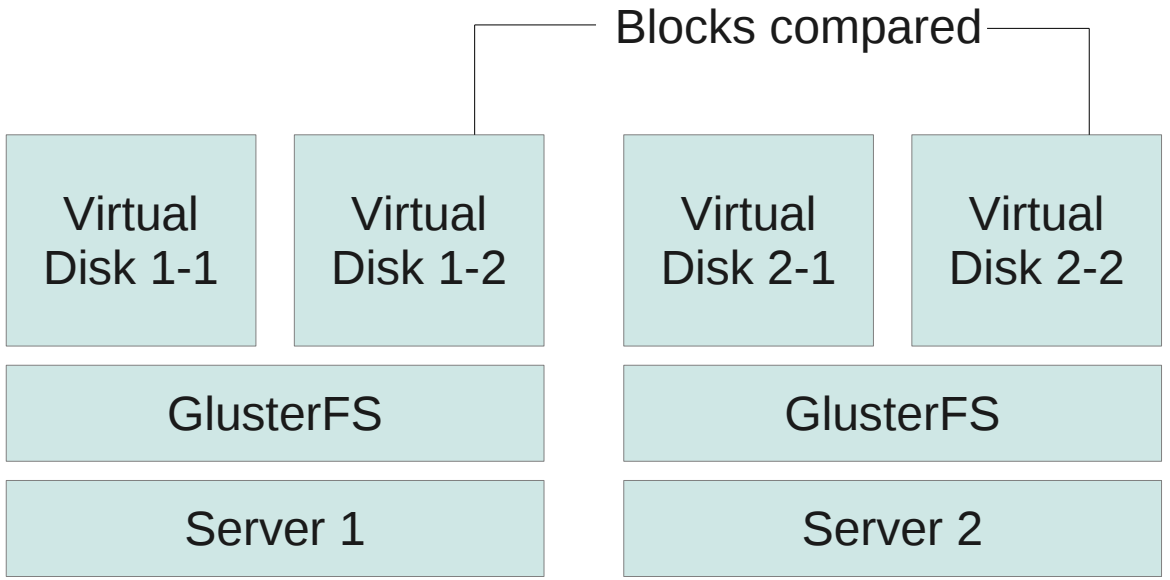
GlusterFS 3.3: Building on the Foundation

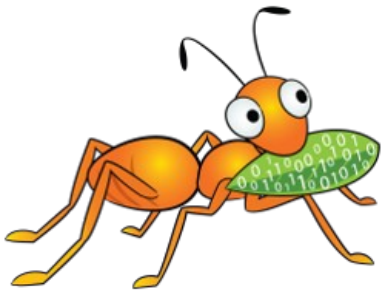
- Granular locking
- Proactive self-healing
- Improved rebalancing
- More access methods



Granular Locking

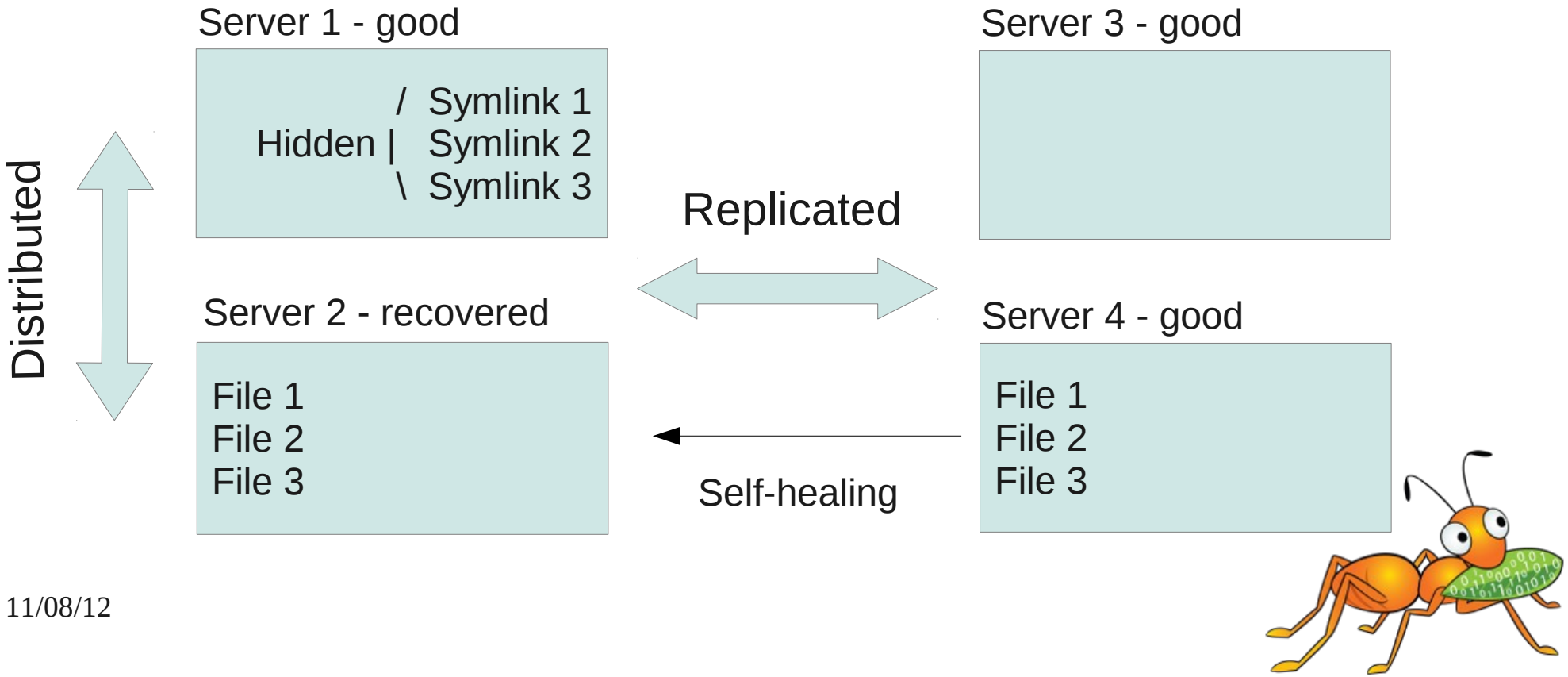
- Server fails, comes back
- Files evaluated
- Block-by-block until healed





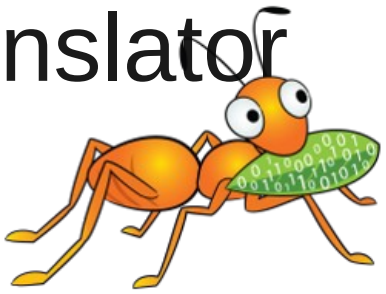
Proactive Self-healing

- Performed server-to-server
- Recovered node queries peers



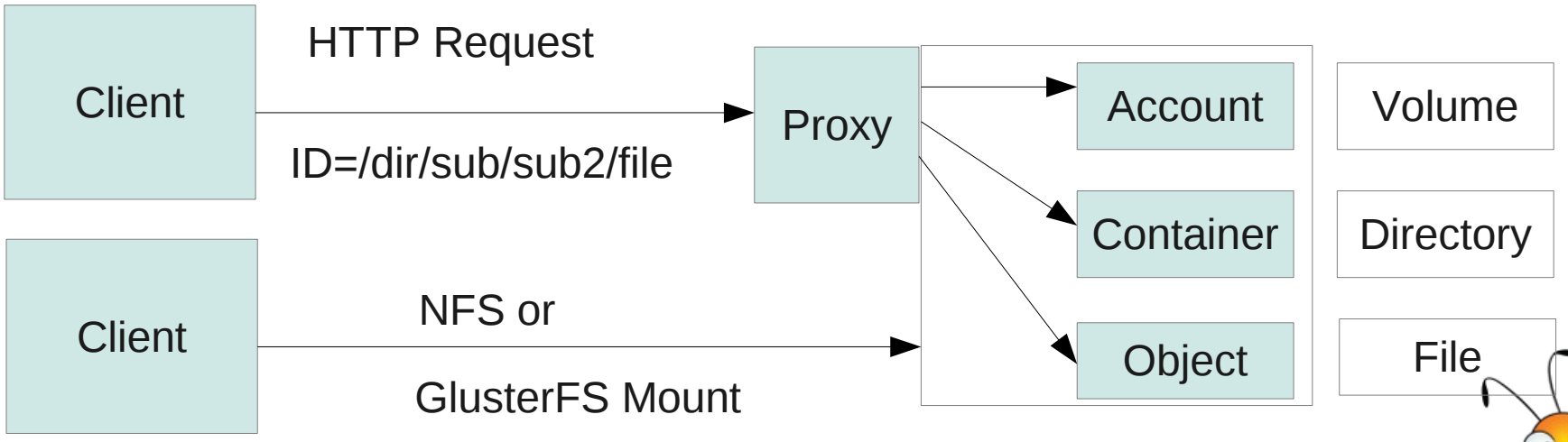
Easier Rebalancing

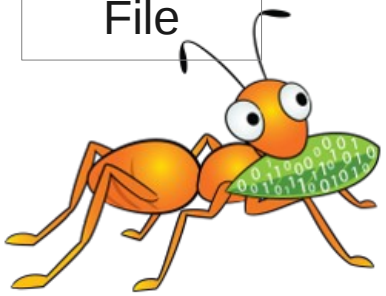
- Now faster
  - Previously, created entire new hash set, moving data unnecessarily
  - Now recreates hash map and compares to old
- Easier to decommission server nodes
- Proof point for synchronous translator API



Unified File and Object (UFO)

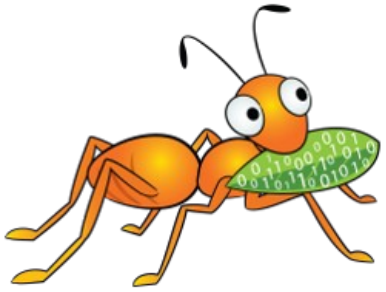
- S3, Swift-style object storage
- Access via UFO or Gluster mount





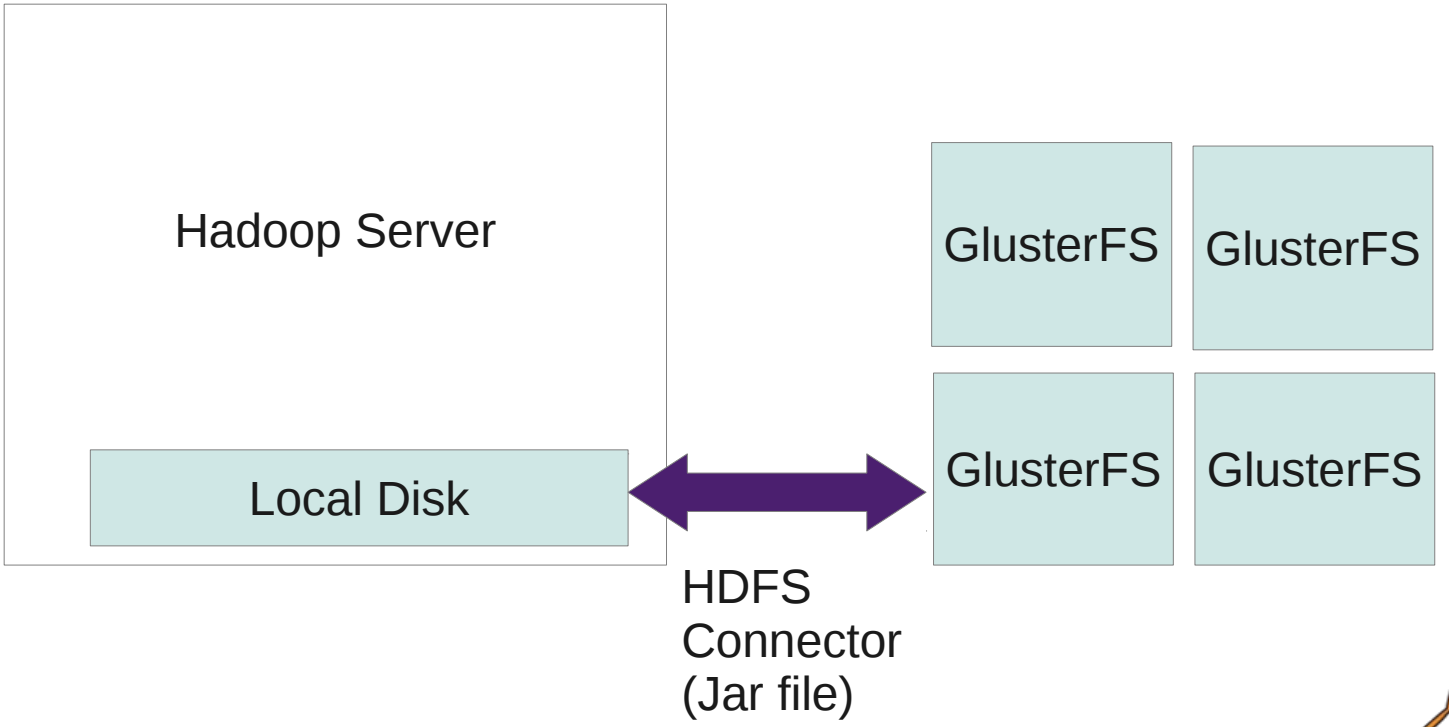
Unified File and Object (UFO)

- Your gateway to the cloud
- Your data, accessed your way

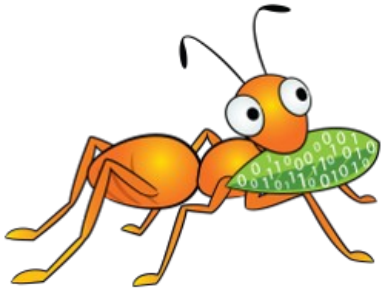


HDFS Compatibility

- Run MapReduce jobs on GlusterFS
- Add unstructured data to Hadoop

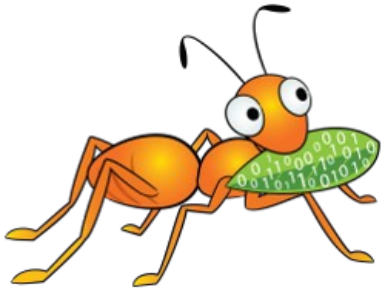


4. Coming Attractions



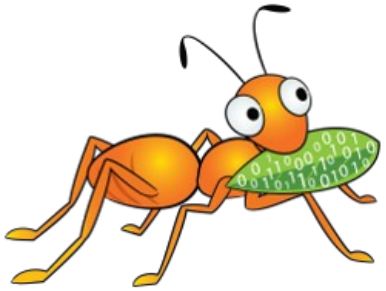
API Check

- Ways to interface with GlusterFS
  - Translators
    - Stackable, async and sync
  - FUSE mount
    - GlusterFS client
  - Libgfapi
    - FUSE bypass



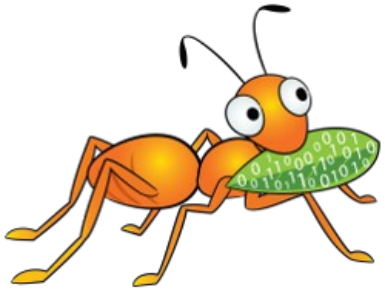
API Check

- Ways to interface with GlusterFS
  - Marker framework
    - Geo-replication, quickly ID changes
  - UFO RESTful API
  - HDFS library
  - Management API
    - oVirt 3.1



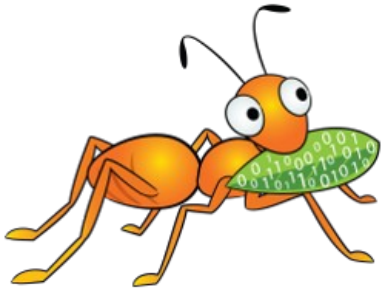
Better VM Image Handling

- Better responsiveness for random I/o use cases
- Contribution: Block Device Translator

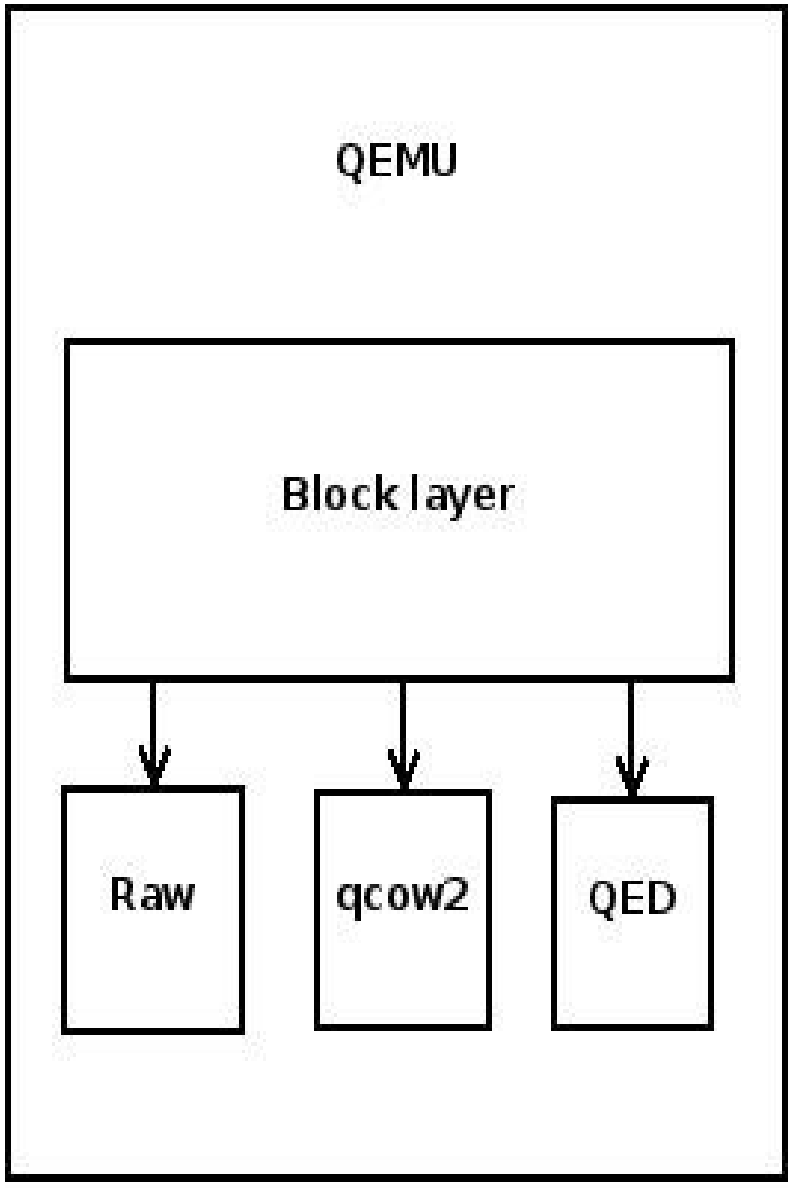


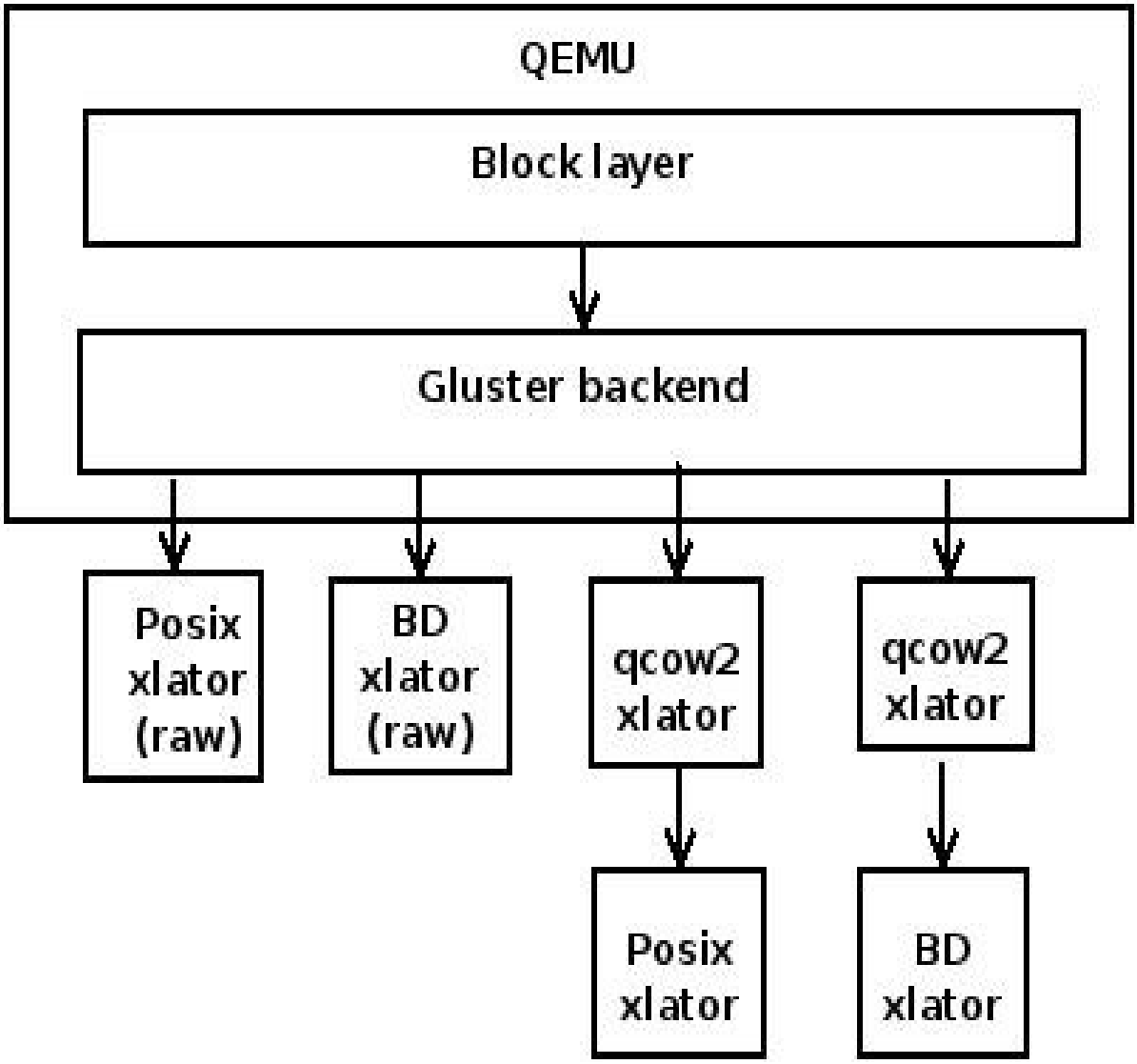
Enabling GlusterFS for Virtualization use

- QEMU-GlusterFS integration
  - Native integration, no FUSE mount
  - Gluster as QEMU block back end
  - QEMU talks to gluster and gluster hides different image formats and storage types underneath
  - Block device support in GlusterFS via Block Device translator
  - Logical volumes as VM images



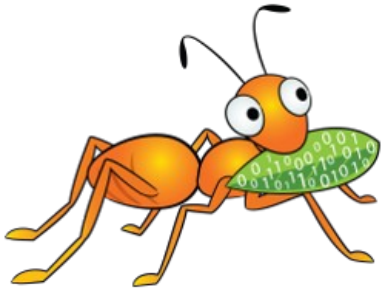
GlusterFS & QEMU





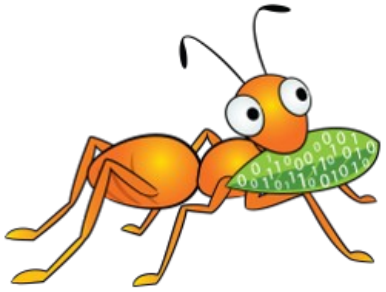
Libglusterfs Client API

- Previously abandoned
- Brought back to life
  - In part because of QEMU Fuse bypass contributions



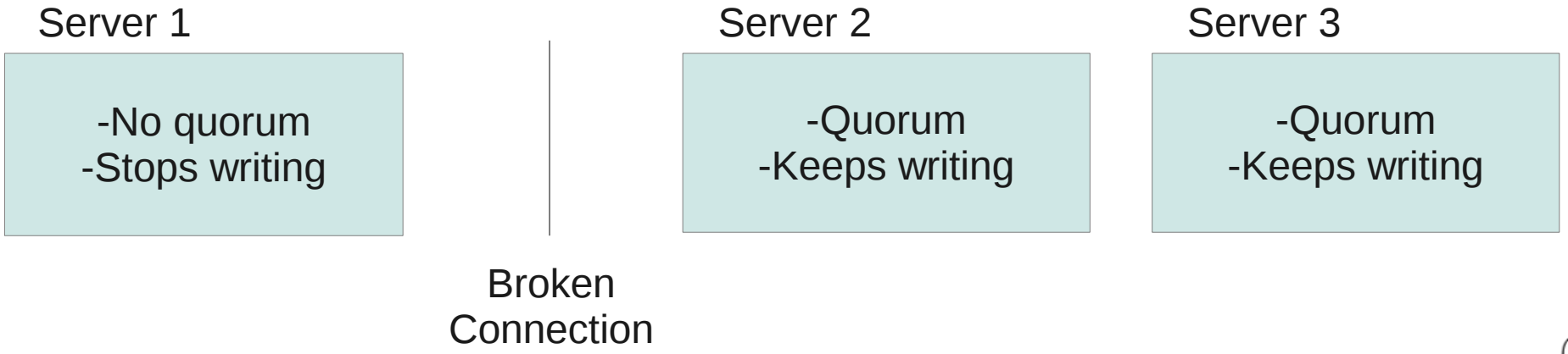
Split Brain

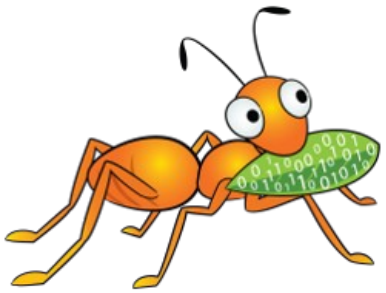
- Nodes cannot see each other, but can all still write
- Often due to network outages
- Sometimes results in conflicts
- Up to 3.2, GlusterFS had no concept of “quorum”



Quorum Enforcement

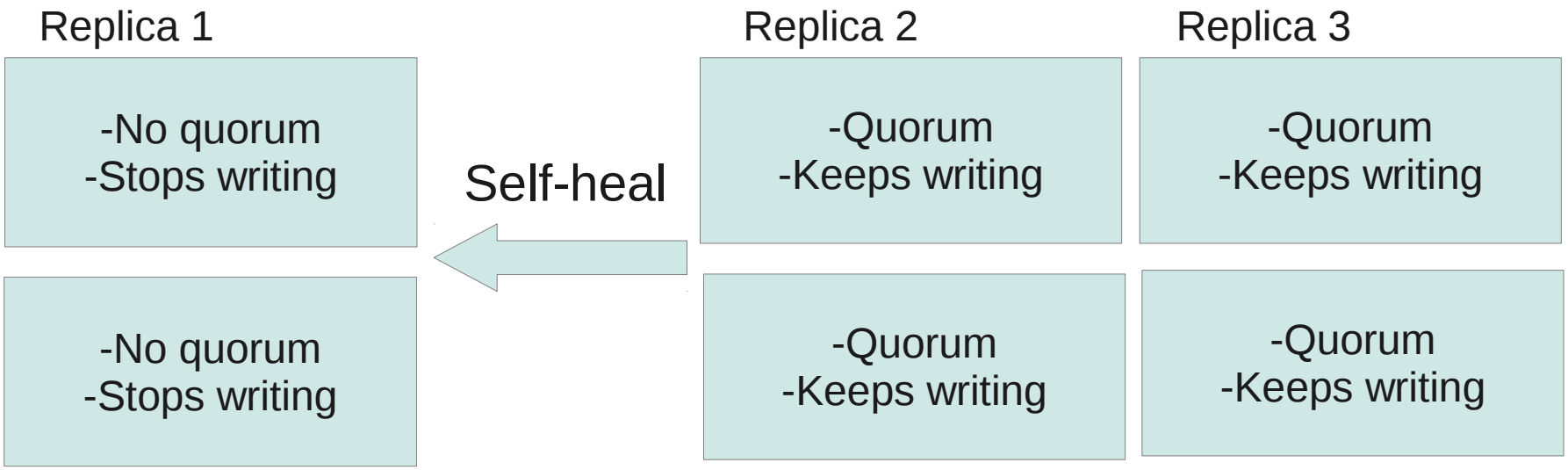
- Which node has valid data?
- If quorum, keep writing, else stop
  - Configurable option

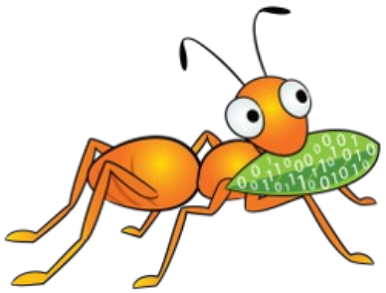




Quorum Enforcement

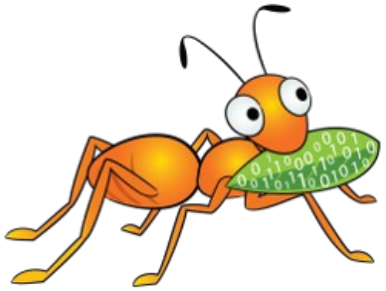
- After connection restored, self-heal kicks off





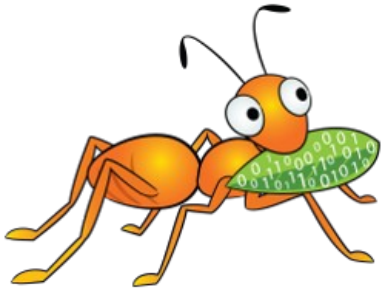
Enhanced Quorum

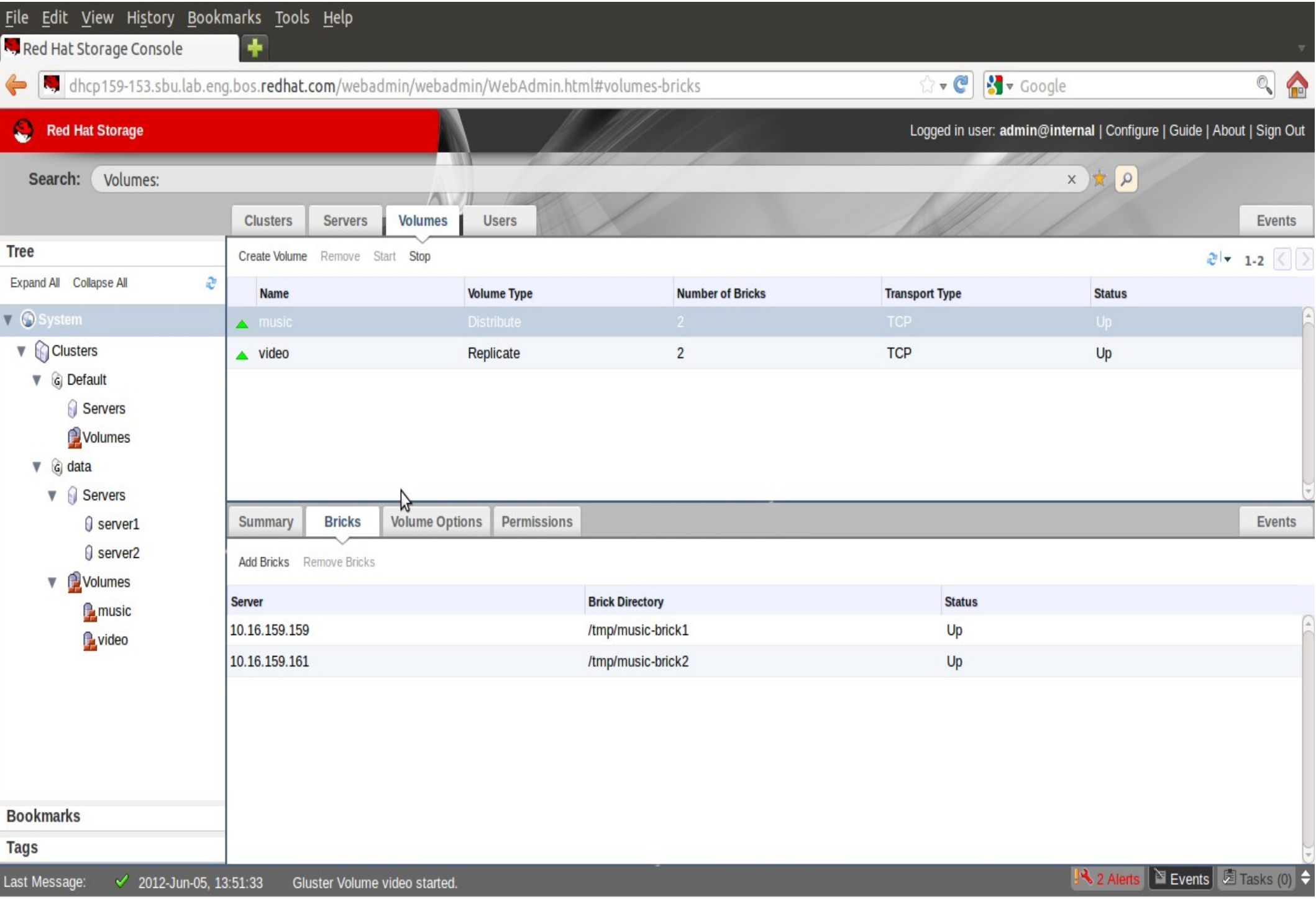
- Quorum tracking on the servers
- Need quorum for any management changes
- 3rd party arbiters / observers so never N=2



Management UI & REST API

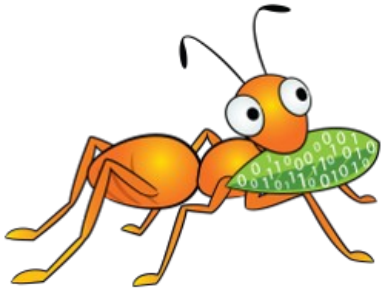
- Collaboration with oVirt project
- Management GUI for admins
- RESTful gateway for devs
- First community release... ?




File Edit View History Bookmarks Tools Help
Red Hat Storage Console
dhcp159-153.sbu.lab.eng.bos.redhat.com/webadmin/webadmin/WebAdmin.html#volumes-bricks
Google
Red Hat Storage
Logged in user: admin@internal | Configure | Guide | About | Sign Out
Search: Volumes:
Clusters
Servers
Volumes
Users
Events
Tree
Expand All
Collapse All
System
Clusters
Default
Servers
Volumes
data
Servers
server1
server2
Volumes
music
video
Create Volume
Remove
Start
Stop
1-2
Name
Volume Type
Number of Bricks
Transport Type
Status
music
Distribute
2
TCP
Up
video
Replicate
2
TCP
Up
Summary
Bricks
Volume Options
Permissions
Events
Add Bricks
Remove Bricks
Server
Brick Directory
Status
10.16.159.159
/tmp/music-brick1
Up
10.16.159.161
/tmp/music-brick2
Up
Bookmarks
Tags
Last Message:
2012-Jun-05, 13:51:33
Gluster Volume video started.
2 Alerts
Events
Tasks (0)

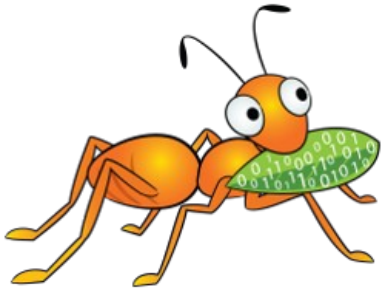
Multi-tenancy & Encryption

- HekaFS created this for cloud deployments
- In-flight data encryption



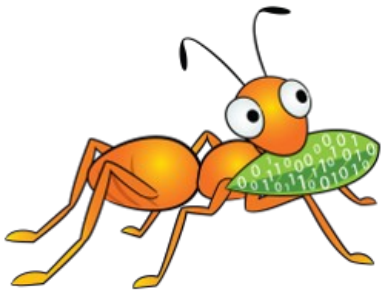
Down the Road

- Multi-master Geo-rep
- Snapshots
- Versioning
- GeoRep Sparse Replicas
- File compression & de-dupe



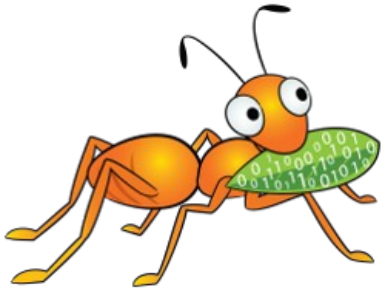
Server-side Processing

- Implementing gfind, glocate
- Fast traversal of metadata in xattrs
  - Find and locate responsive
- Inotify-esque behavior: triggers based on i/o activity, ie. file close
  - Why rely on Hadoop batch-processing?

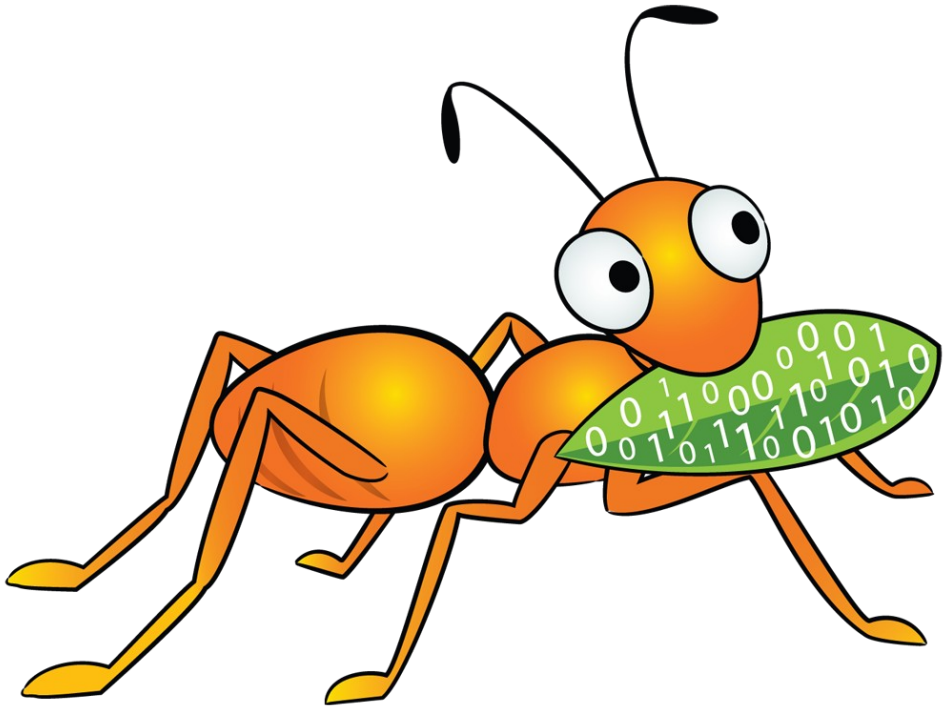


Goal: Intelligent Storage

- Just storing and retrieving data is not enough
- Should be able to store, analyze, transform, mutilate, and retrieve
- Intelligent storage gives sysadmins and developers the ultimate data swiss army knife



Thank you!



John Mark Walker
Gluster Community Guy
johnmark@redhat.com