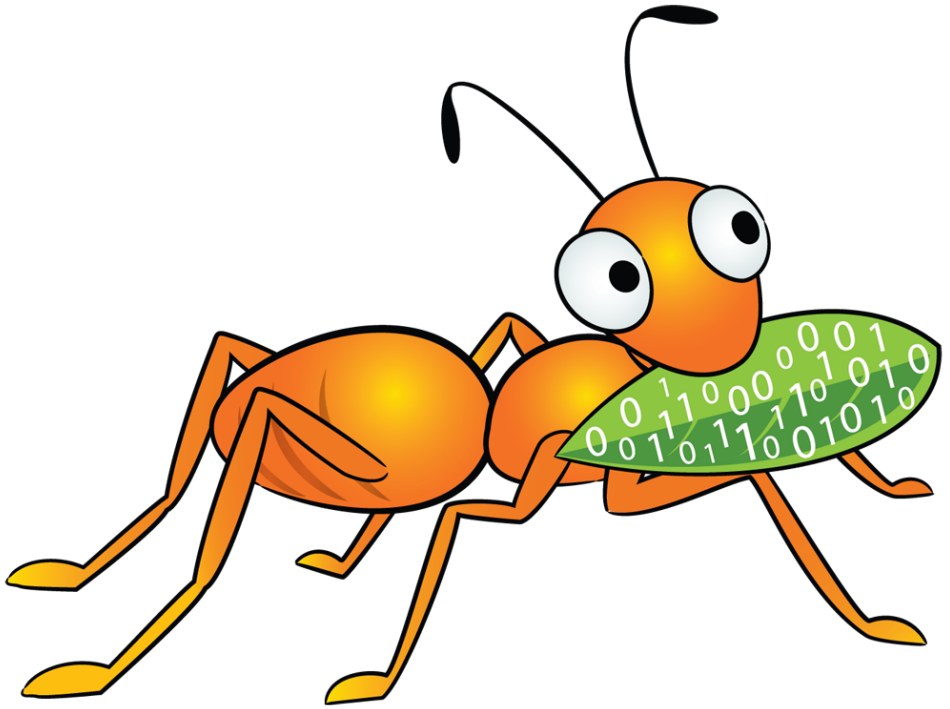
High-availability data storage and access on Gluster



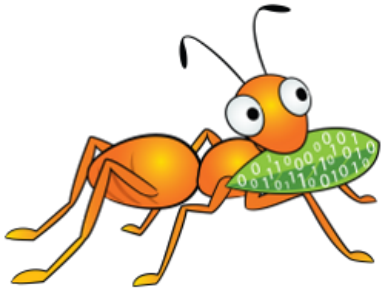
Niels de Vos
Red Hat Storage Engineer
GlusterFS co-maintainer
ndevos@redhat.com

T-DOSE
The place where experts meet

November 12, 2016
Eindhoven

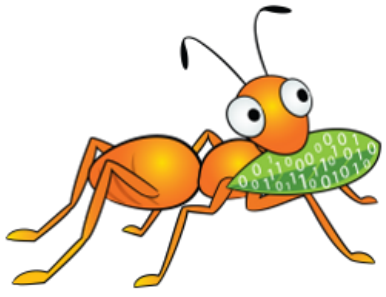
Agenda

- What is Gluster?
  - Access Protocols/Methods
  - Basic High-Availability
- High-Availability projects
- Failure Scenarios
  - backend, frontend and applications
- Deployment Examples



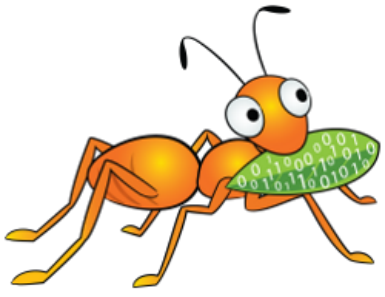
What is Gluster ?

Gluster is a distributed scale out filesystem that allows rapid provisioning of additional storage based on your storage consumption needs. It incorporates automatic failover as a primary feature. All of this is accomplished without a centralized metadata server.



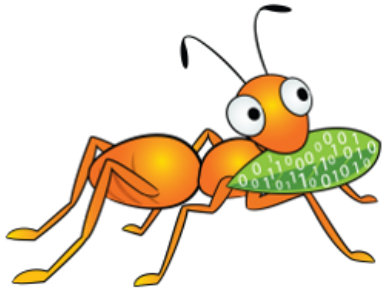
Gluster in Keywords

- Scalable, general-purpose storage platform
  - POSIX-y Distributed File System
  - Object storage (swift)
  - Flexible storage (libgfapi)
- No Metadata Server
- Heterogeneous Commodity Hardware
- Flexible and Agile Scaling
  - Capacity – Petabytes and beyond
  - Performance – Thousands of Clients



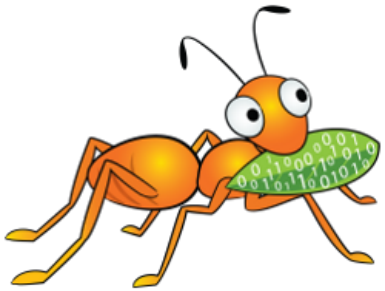
Gluster Use-Cases

- Content Delivery Networks
  - Media Streaming
  - Download Servers
- Archival
  - Backup services
  - Long term media archives
- Virtual Machine images
- High Performance / Distributed Computing
  - Rendering Farms
  - Big Data

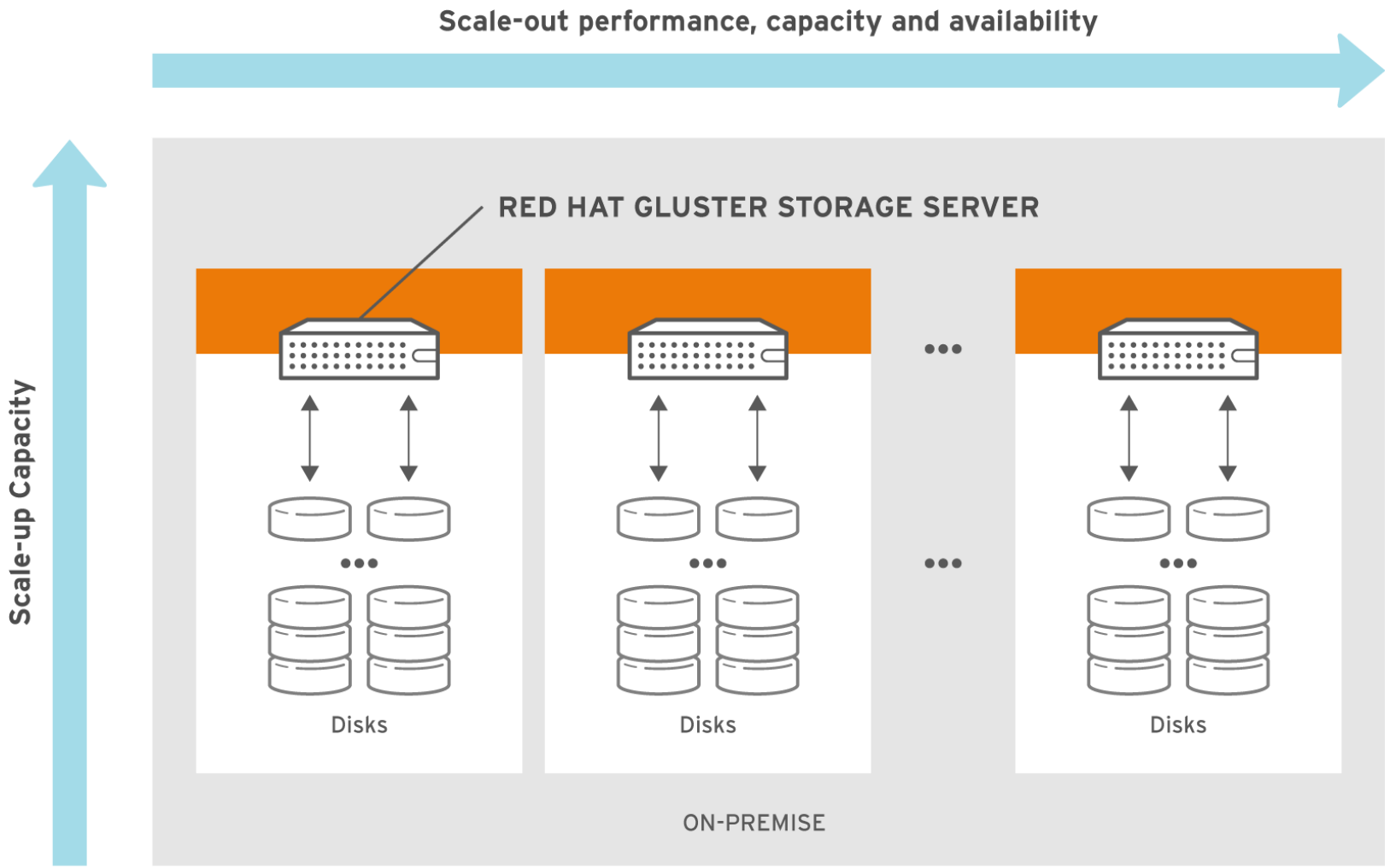


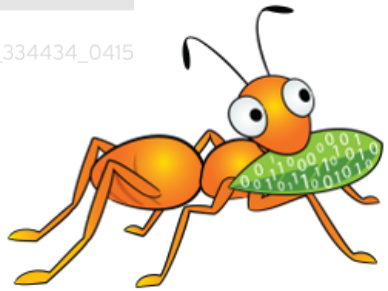
Data Access Overview

- GlusterFS Native Client
  - Filesystem in Userspace (FUSE)
- NFS
  - Built-in Service, NFS-Ganesha with libgfapi
- SMB/CIFS
  - Samba server required (libgfapi based module)
- Gluster For OpenStack Swift (Glusterswift)
- libgfapi flexible abstracted storage
  - Integrated with QEMU, Bareos and others



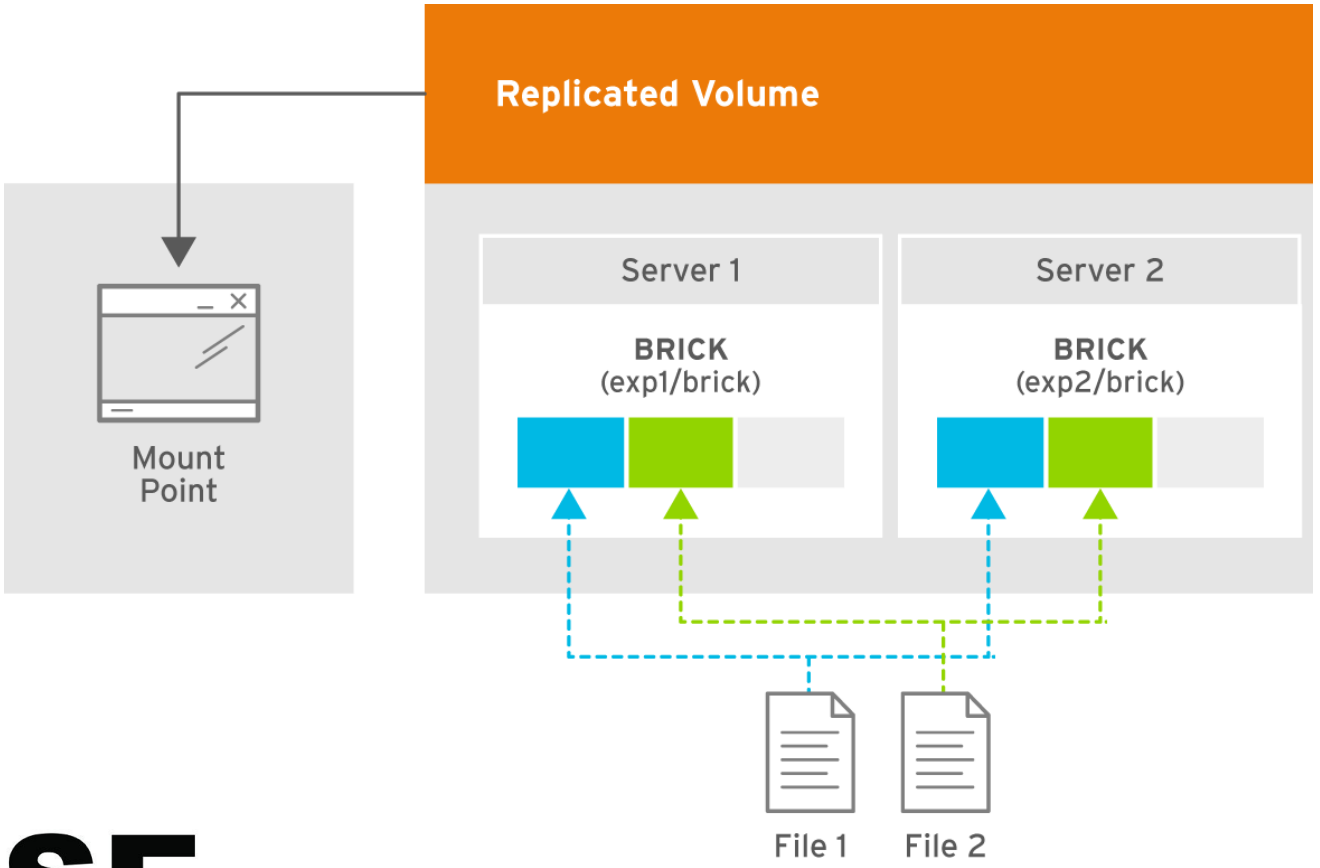
Scale-out and Scale-up





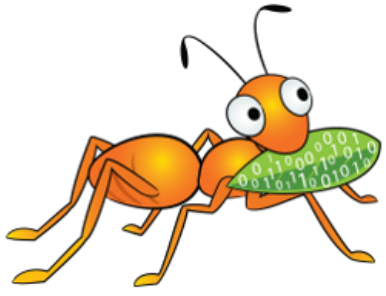
Replicated Volume

- Copies files to multiple bricks
- *Similar* to file-level RAID 1



High-Availability projects

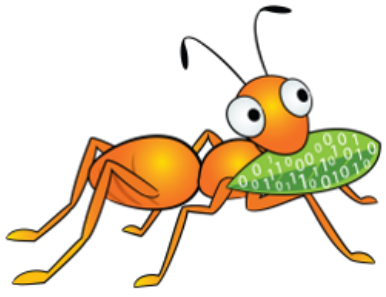
- CTDB from the Samba project
  - Used for Samba and Gluster/NFS
- Pacemaker from Cluster Labs
  - Used for NFS-Ganesha
  - Storhaug for common Samba + NFS-Ganesha
- keepalived
  - For high-availability + loadbalancing (webservers etc)



Failure Scenarios: backend

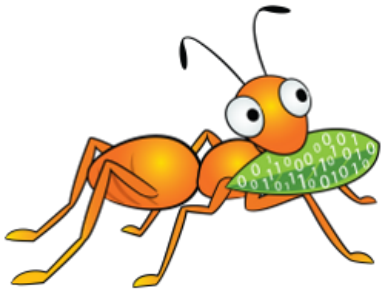
- Network failures within the storage communication
- Gluster clients can detect brick failures
- Gluster servers can detect storage-server failures

- Quorum is used for preventing split-brain scenarios
- Ping-timeout (client-side) for keeping connections active



Failure Scenarios: frontend

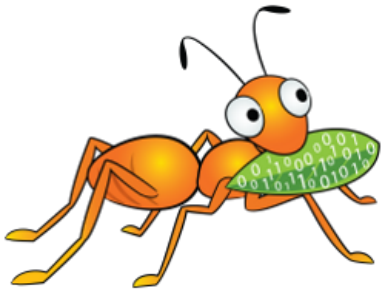
- NFS-Ganesha or Samba fail-over
- Standard HA-projects for IP-address relocation
  - Gluster integrated pacemaker with 10 sec. interval
- ‘tickle’ TCP-connections
- Potentially needs to notify clients for lock-reclaim
  - Gluster servers release locks after connection loss



Failure Scenarios: applications

- Transparent for most applications
- Unless clients use file- or byte-range locks
  - When should a lock of a failing client be released?

- NFSv4 has a RENEW operation
- NFS-clients need to RECLAIM locks after failover



Deployment Examples: NFS failover

- NFS-Ganesha with NFSv4 clients
- Assume a NFS-Ganesha failure
- pacemaker detects the failure
  - By default a 10 second interval
- IP-address relocation
- NFS-Ganesha invokes GRACE on all servers
  - special state for NFS-server, by default 90 seconds
  - new lock requests (and I/O) get blocked/stalled
- NFS-clients need to reclaim their obtained locks

Deployment Examples: Hypervisor and VMs

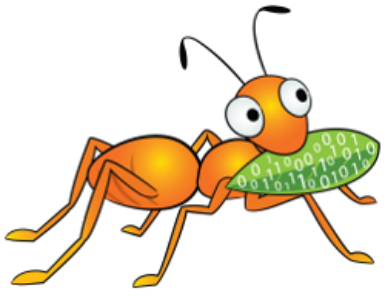
- oVirt management with QEMU/gfapi access to Gluster
- Assume a storage-server failure
- QEMU/gfapi detects the failure after ping-timeout
  - By default the ping-timeout is 42 seconds
- Filesystem in the VM may detect I/O delay sooner
  - By default SCSI emulation times-out after 30 seconds

- Problem: filesystems in VMs may become read-only

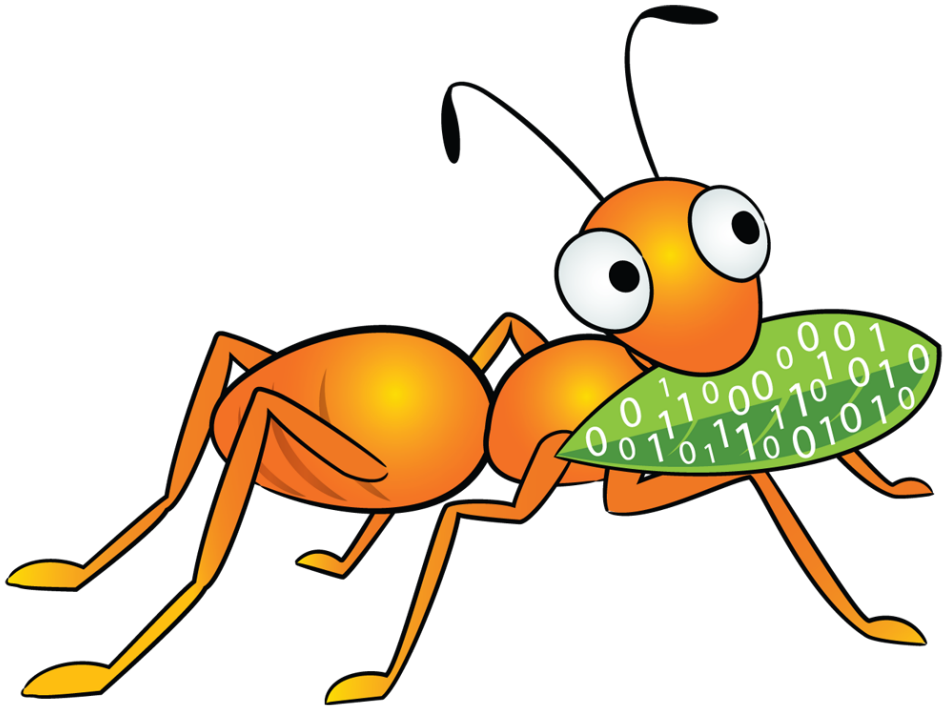
Deployment Examples: failover planning

- Assume 3 frontend servers (each has 33% of the load)
- One server fails, 33% of the load needs to be relocated
- One of the two remaining servers needs to handle 66%
  - Servers need double the resources for failover

- A problem when each server has only one virtual-IP
- Can be reduced when each server has two virtual-IPs
  - Servers need 1.5x the resources for failover



Thanks!



T-DOSE

The place where experts meet