RED HAT SUMMIT
10 YEARS and counting
SAN FRANCISCO | APRIL 14-17, 2014

Red Hat Storage Server Administration Deep Dive

Dustin L. Black, RHCA
Sr. Technical Account Manager
Red Hat Global Support Services

***** This session will include a live demo from 6-7pm *****

Dustin L. Black, RHCA

Sr. Technical Account Manager

Red Hat, Inc.

dustin@redhat.com

@dustinlblack

redhat.
CERTIFIED
ARCHITECT

Gluster

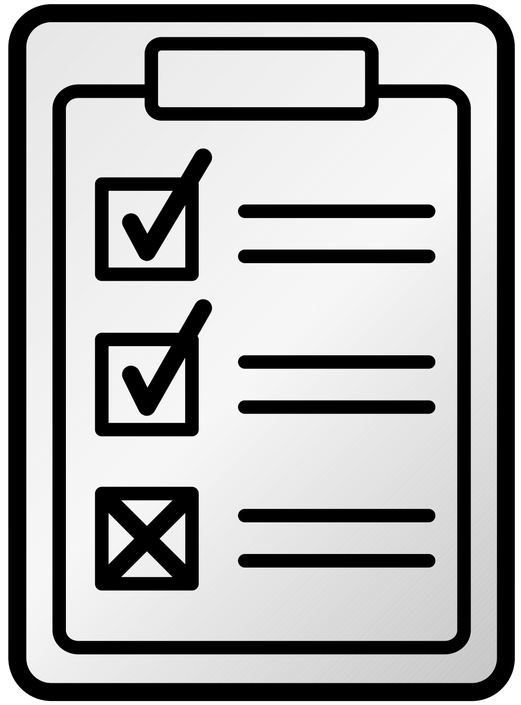
redhat.

WTH is a TAM?

- Premium named-resource support
- Proactive and early access
- Regular calls and on-site engagements
- Customer advocate within Red Hat and upstream
- Multi-vendor support coordinator
- High-touch access to engineering
- Influence for software enhancements
- NOT Hands-on or consulting

Agenda

- Technology Overview & Use Cases
- Technology Stack
- Under the Hood
- Volumes and Layered Functionality
- Asynchronous Replication
- Data Access
- SWAG Intermission
- Demo Time!



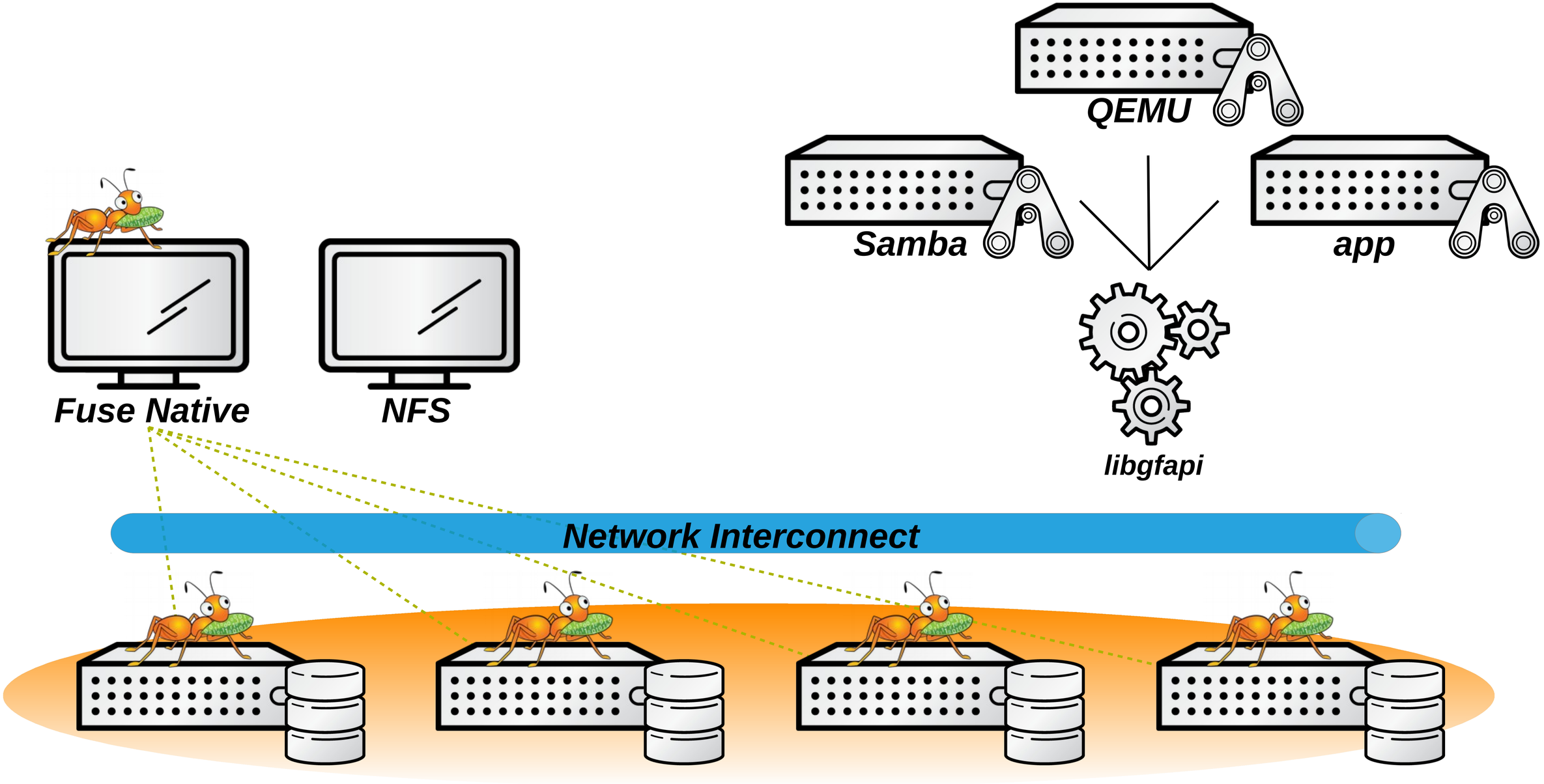
RED HAT SUMMIT
10 YEARS and counting
SAN FRANCISCO | APRIL 14-17, 2014

Technology Overview

Red Hat Storage Server Administration Deep Dive

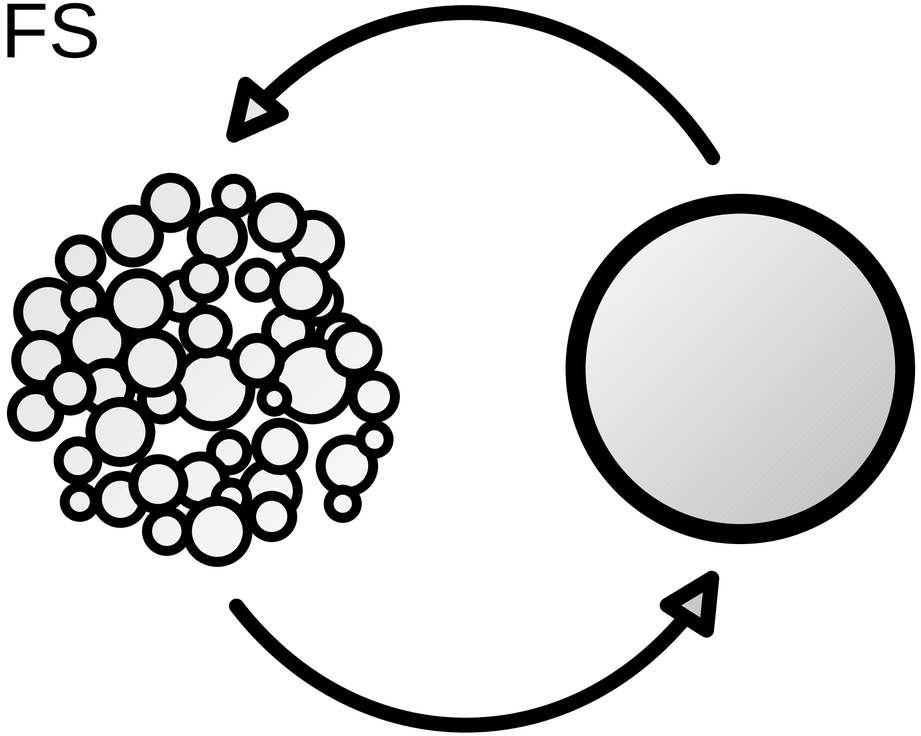
What is GlusterFS?

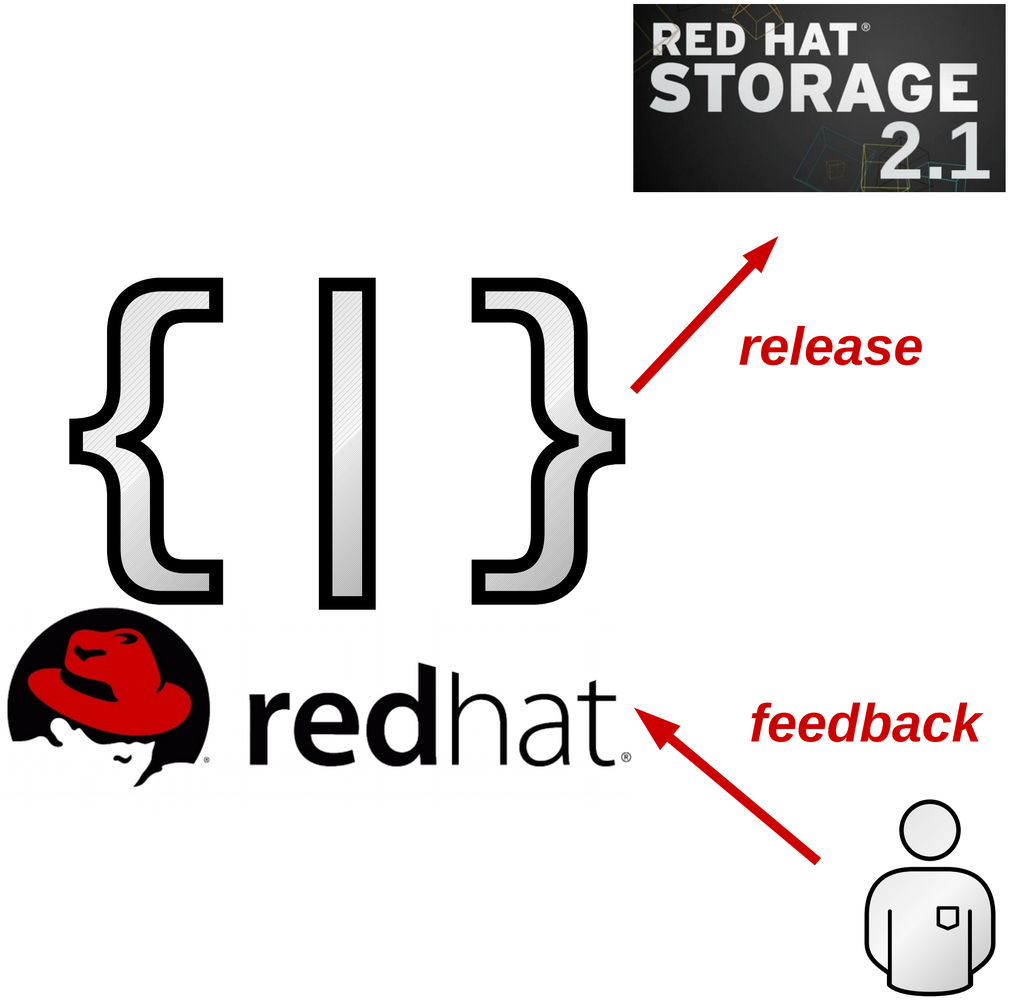
- Clustered Scale-out **General Purpose** Storage Platform
  - POSIX-y Distributed File System
  - ...and so much more
- Built on Commodity systems
  - x86_64 Linux ++
  - POSIX filesystems underneath (XFS, EXT4)
- No Metadata Server
- Standards-Based – Clients, Applications, Networks
- Modular Architecture for Scale and Functionality

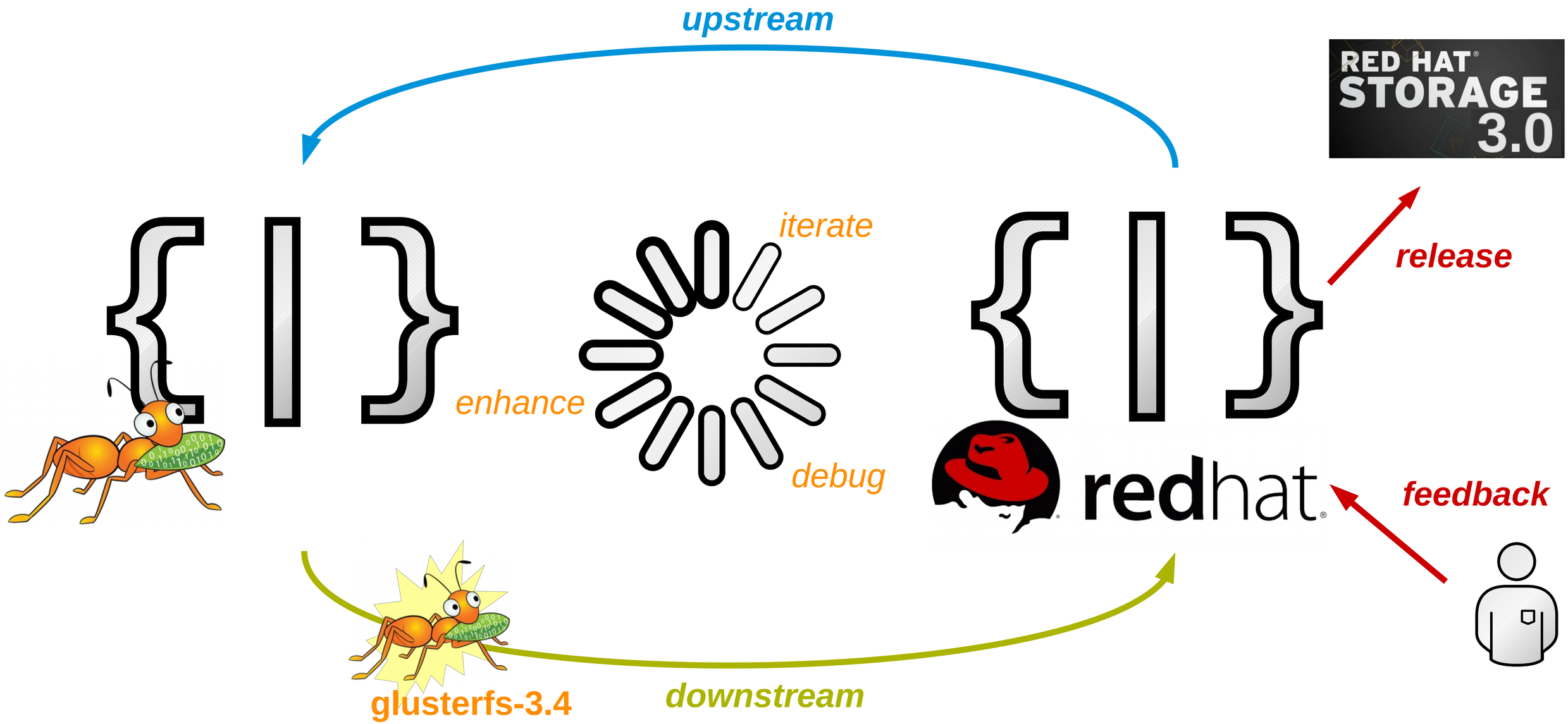

QEMU
Samba
app
libgfapi
Fuse Native
NFS
Network Interconnect

What is Red Hat Storage?

- Enterprise Implementation of GlusterFS
- Integrated Software Appliance
  - RHEL + XFS + GlusterFS
- Certified Hardware Compatibility
- Subscription Model
- 24x7 Premium Support




RED HAT®
STORAGE
2.1
release
redhat.
feedback


upstream
RED HAT® STORAGE 3.0
iterate
enhance
debug
release
redhat.
feedback
glusterfs-3.4
downstream

GlusterFS vs. Traditional Solutions

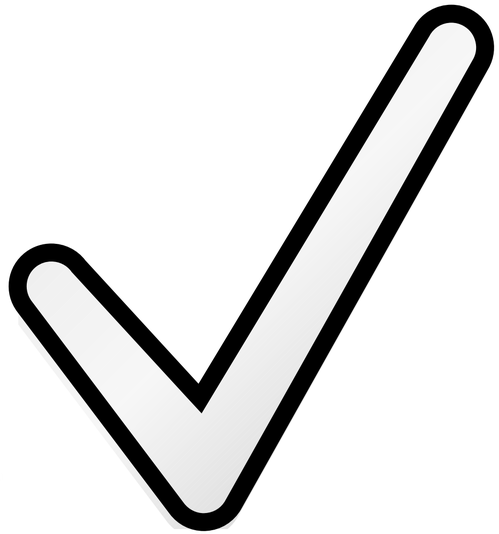
- A basic NAS has limited scalability and redundancy
- Other distributed filesystems are limited by metadata service
- SAN is costly & complicated, but high performance & scalable
- *GlusterFS is...*
  - *Linear Scaling*
  - *Minimal Overhead*
  - *High Redundancy*
  - *Simple and Inexpensive Deployment*

Use Cases

Red Hat Storage Server Administration Deep Dive

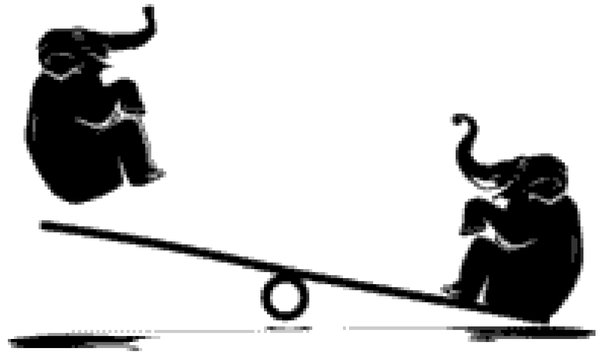
Common Solutions

- Large Scale File Server
- Media / Content Distribution Network (CDN)
- Backup / Archive / Disaster Recovery (DR)
- High Performance Computing (HPC)
- Infrastructure as a Service (IaaS) storage layer
- Database offload (blobs)
- Unified Object Store + File Access

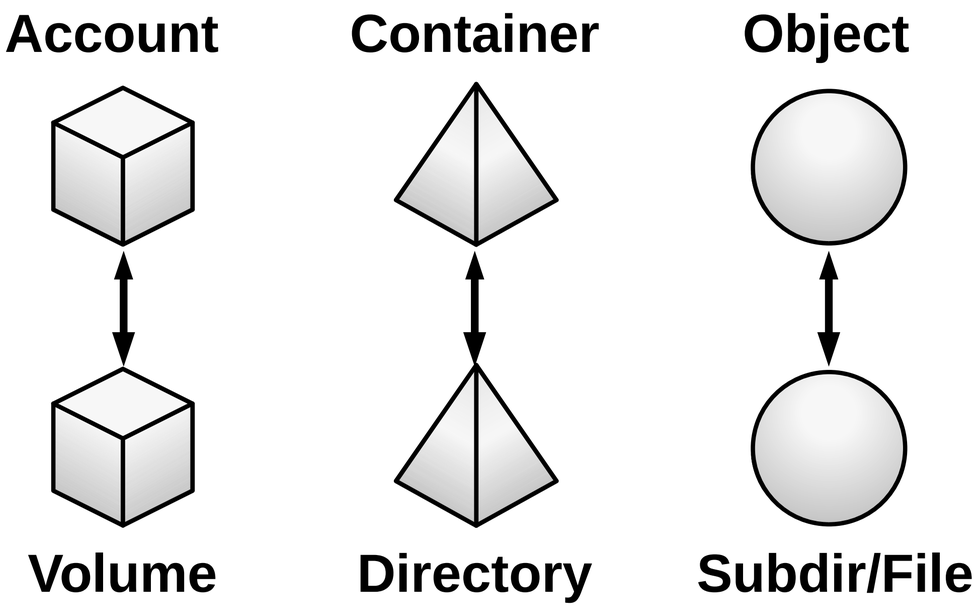


Hadoop – Map Reduce

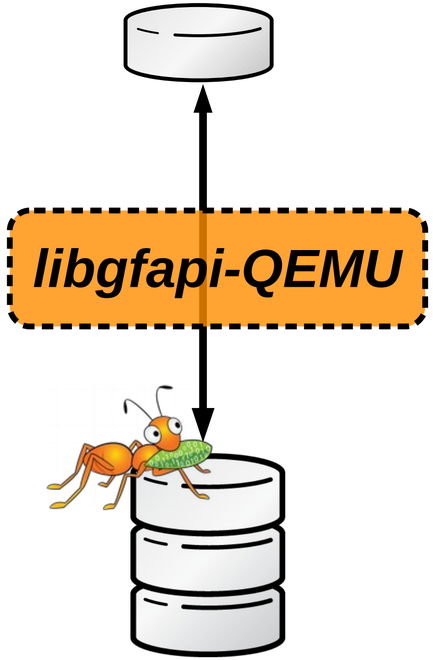
- Access data within and outside of Hadoop
- No HDFS name node single point of failure / bottleneck
- Seamless replacement for HDFS
- Scales with the massive growth of big data



Swift



Cinder / Glance



Gluster

redhat.

RED HAT SUMMIT
10 YEARS and counting
SAN FRANCISCO | APRIL 14-17, 2014

Technology Stack

Red Hat Storage Server Administration Deep Dive

Terminology

- Brick
  - Fundamentally, a filesystem mountpoint
  - A unit of storage used as a ***capacity*** building block
- Translator
  - Logic between the file bits and the Global Namespace
  - Layered to provide GlusterFS ***functionality***

Everything is Modular

Terminology

- Volume
  - Bricks combined and passed through translators
  - Ultimately, what's presented to the end user
- Peer / Node
  - Server hosting the brick filesystems
  - Runs the gluster daemons and participates in volumes

Disk, LVM, and Filesystems

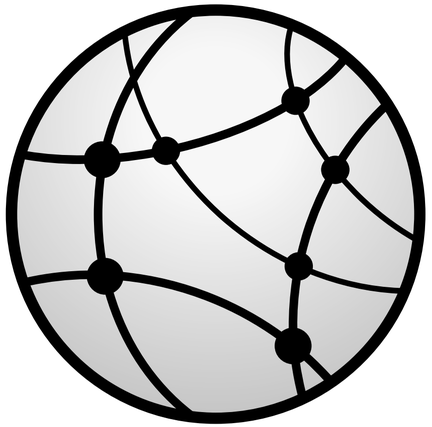
- Direct-Attached Storage (DAS)

  -or-

- Just a Bunch Of Disks (JBOD)
- Hardware RAID
  - RHS: RAID 6 required
- Logical Volume Management (LVM)
- POSIX filesystem w/ Extended Attributes (EXT4, XFS, BTRFS, ...)
  - RHS: XFS required

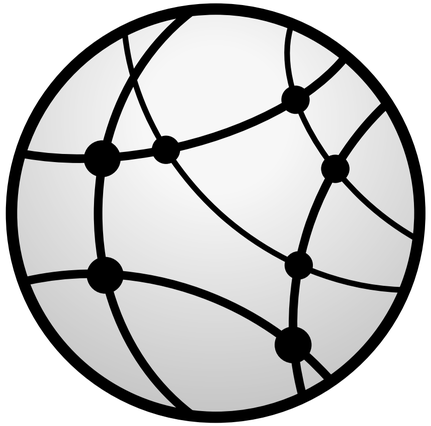
Data Access Overview

- GlusterFS Native Client
  - Filesystem in Userspace (FUSE)
- NFS
  - Built-in Service
- SMB/CIFS
  - Samba server required; NOW libgfapi-integrated!



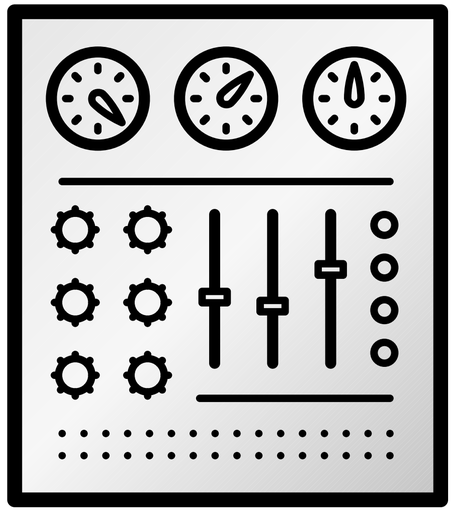
Data Access Overview

- Gluster For OpenStack (G4O; aka UFO)
  - Simultaneous object-based access via OpenStack Swift
- NEW! libgfapi flexible abstracted storage
  - Integrated with upstream Samba and Ganesha-NFS



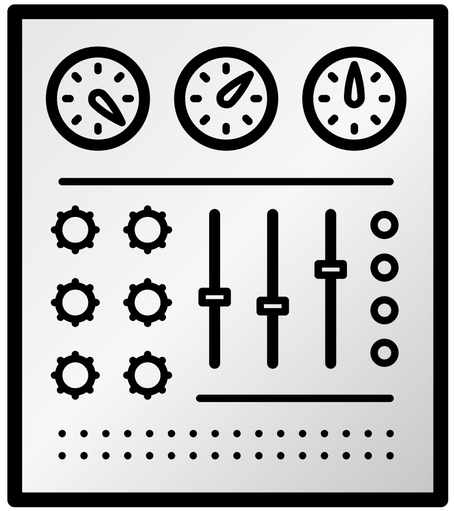
Gluster Components

- `glusterd`
  - Management daemon
  - One instance on each GlusterFS server
  - Interfaced through `gluster` CLI
- `glusterfsd`
  - GlusterFS brick daemon
  - One process for each brick on each server
  - Managed by `glusterd`

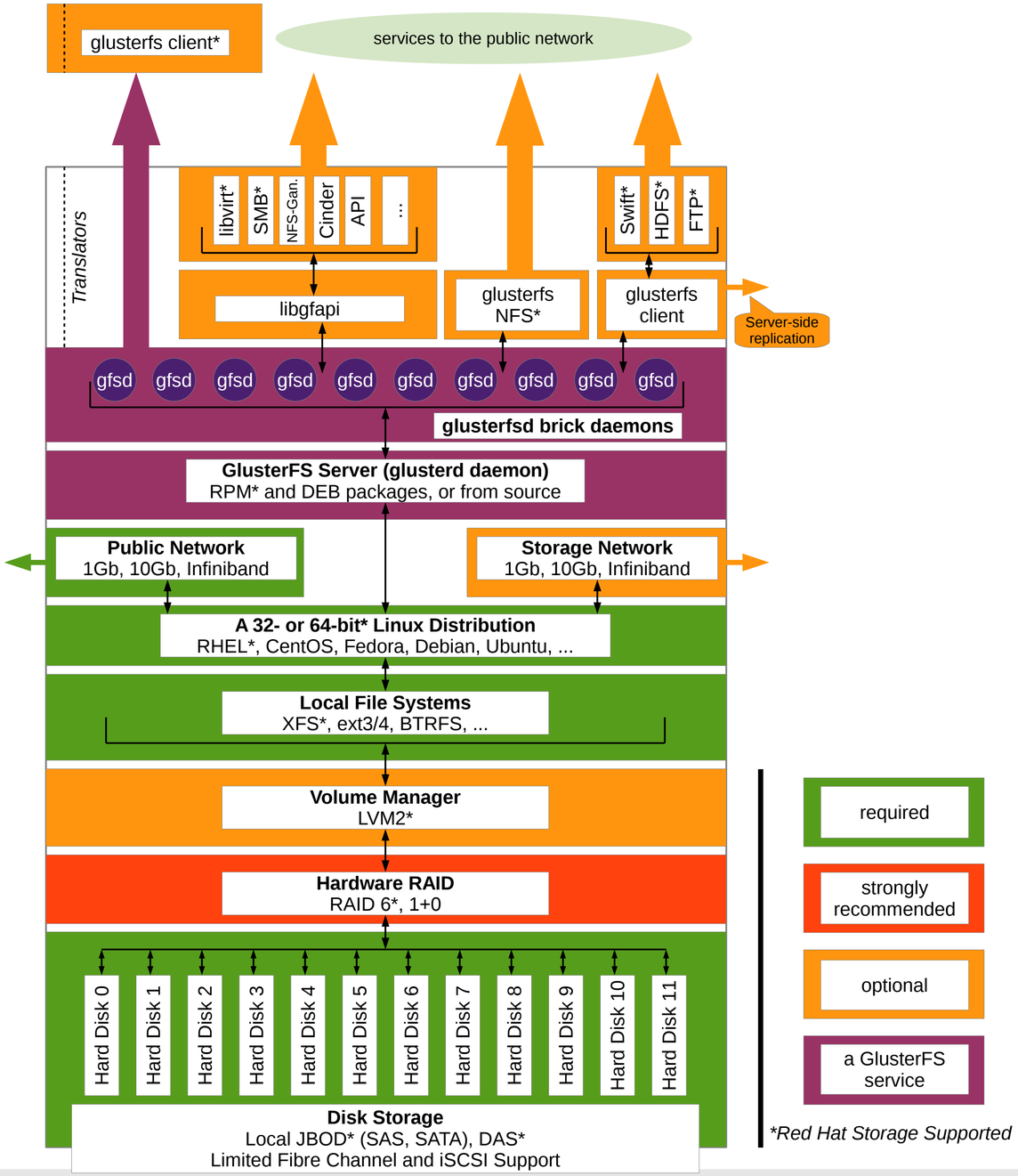


Gluster Components

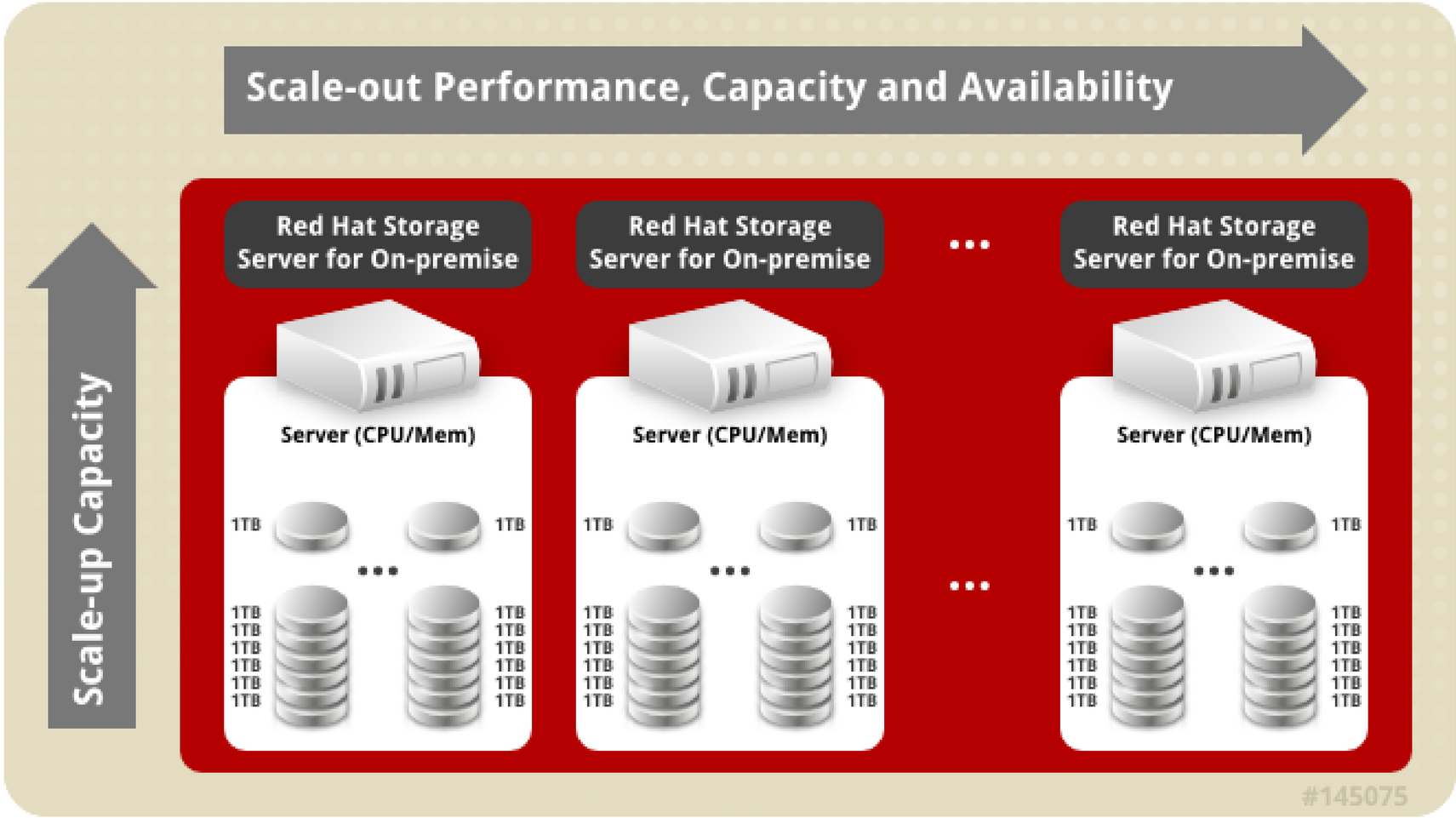
- `glusterfs`
  - Volume service daemon
  - One process for each volume service
    - NFS server, FUSE client, Self-Heal, Quota, ...
- `mount.glusterfs`
  - FUSE native client mount extension
- `gluster`
  - Gluster Console Manager (CLI)



Putting it Together



Up and Out!


Scale-out Performance, Capacity and Availability
Scale-up Capacity
Red Hat Storage Server for On-premise
Red Hat Storage Server for On-premise
Red Hat Storage Server for On-premise
Server (CPU/Mem)
Server (CPU/Mem)
Server (CPU/Mem)
1TB
#145075

10 YEARS *and counting*

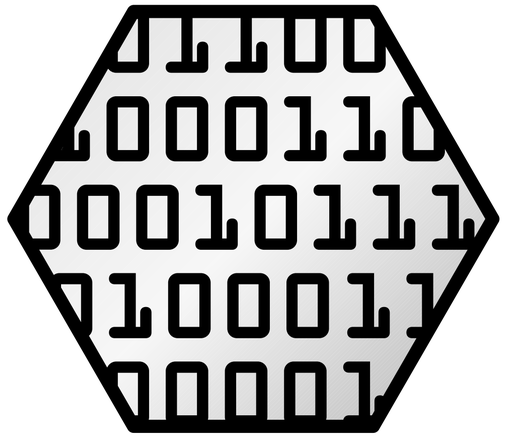
SAN FRANCISCO | APRIL 14-17, 2014

Under the Hood

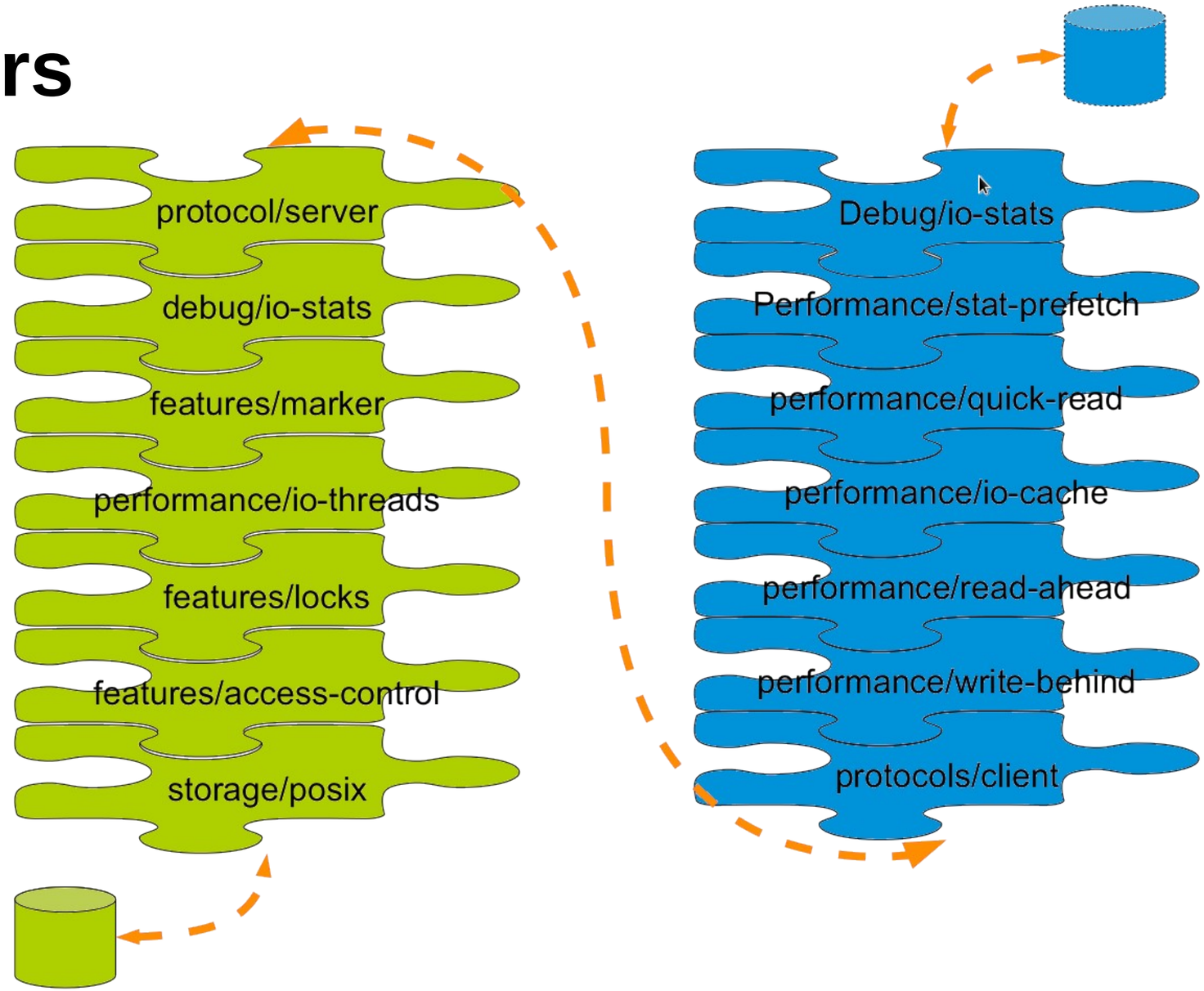
Red Hat Storage Server Administration Deep Dive

Elastic Hash Algorithm

- No central metadata
  - No Performance Bottleneck
  - Eliminates risk scenarios
- Location hashed intelligently on filename
  - Unique identifiers, similar to md5sum
- The “Elastic” Part
  - Files assigned to virtual volumes
  - Virtual volumes assigned to multiple bricks
  - Volumes easily reassigned on the fly



Translators



Your Storage Servers are Sacred!

- Don't touch the brick filesystems directly!
- They're Linux servers, but treat them like storage appliances
  - Separate security protocols
  - Separate access standards
- Don't let your Jr. Linux admins in!
  - A well-meaning sysadmin can quickly break your system or destroy your data

RED HAT SUMMIT
10 YEARS and counting
SAN FRANCISCO | APRIL 14-17, 2014

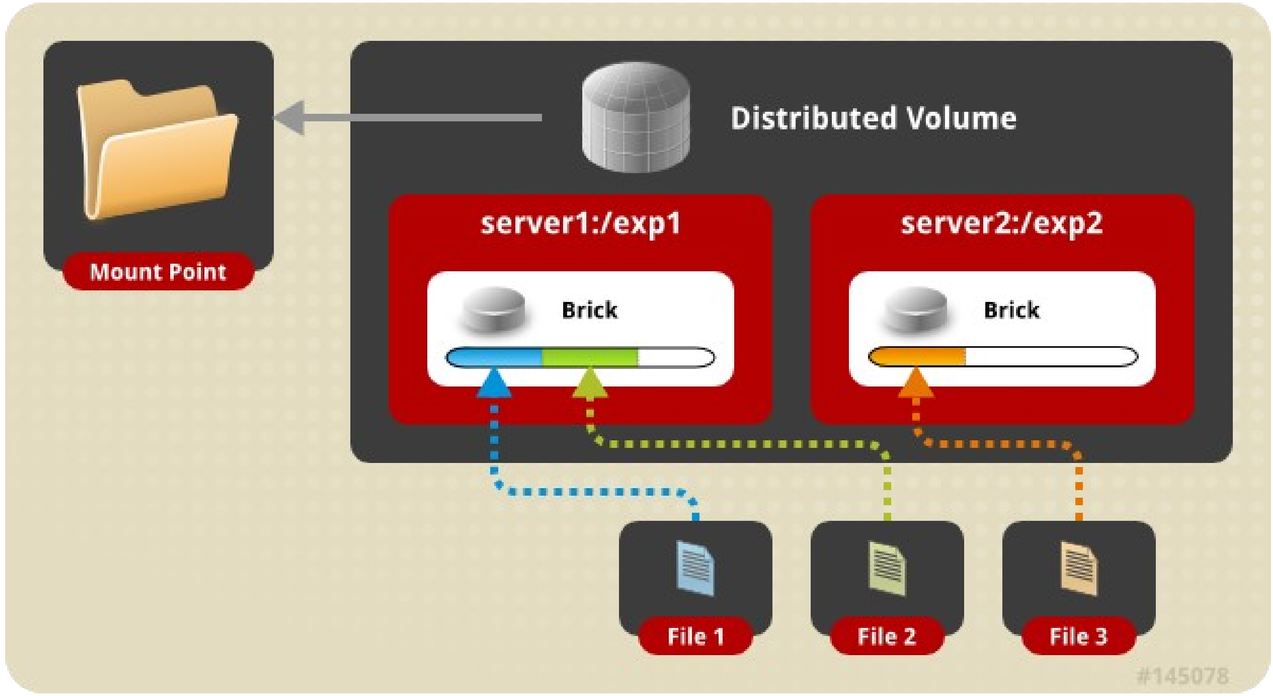
Basic Volumes

Red Hat Storage Server Administration Deep Dive

redhat.

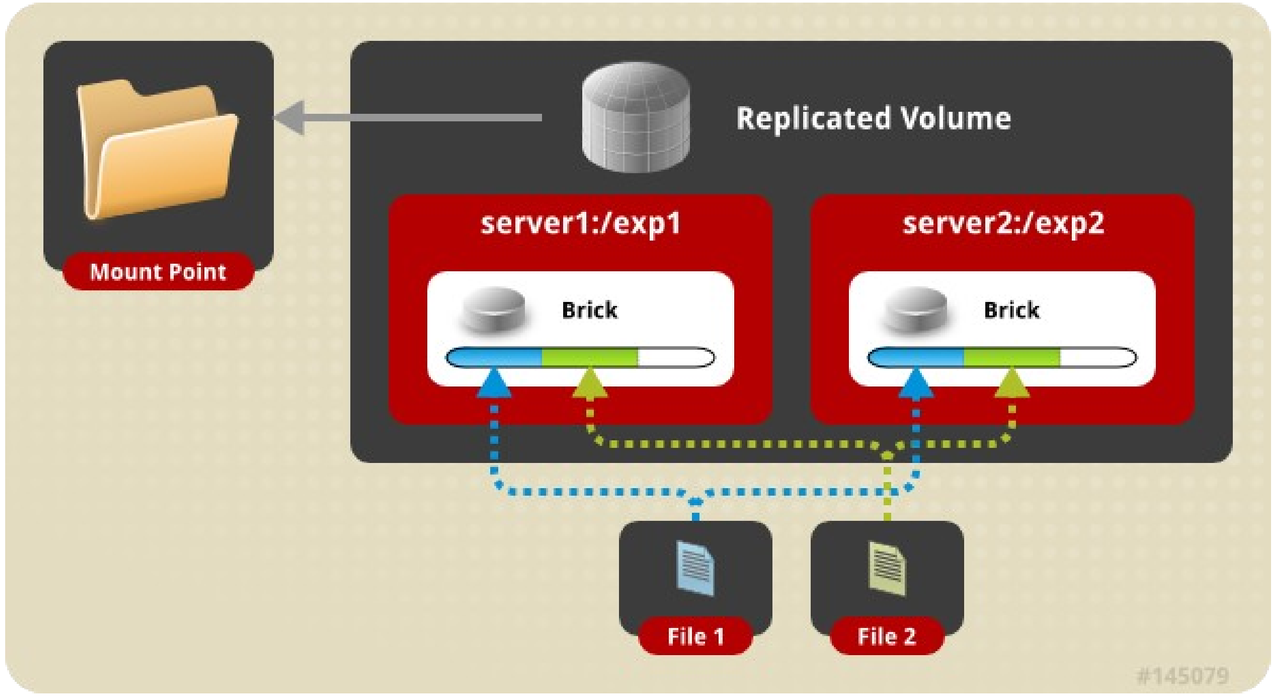
Distributed Volume

- The default configuration
- Files “evenly” spread across bricks
- Similar to file-level RAID 0
- Server/Disk failure could be catastrophic



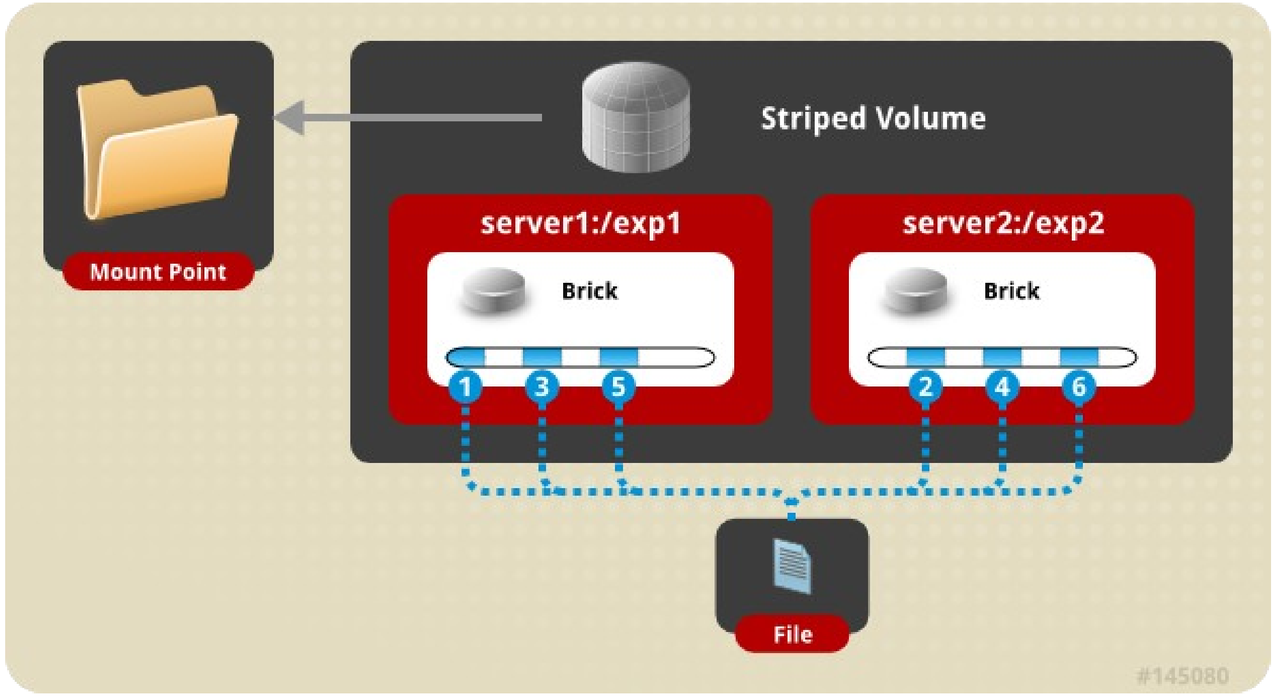
Replicated Volume

- Files written synchronously to replica peers
- Files read synchronously, but ultimately serviced by the first responder
- Similar to file-level RAID 1



Striped Volumes

- Individual files split among bricks (sparse files)
- Similar to block-level RAID 0
- Limited Use Cases
  - HPC Pre/Post Processing
  - File size exceeds brick size



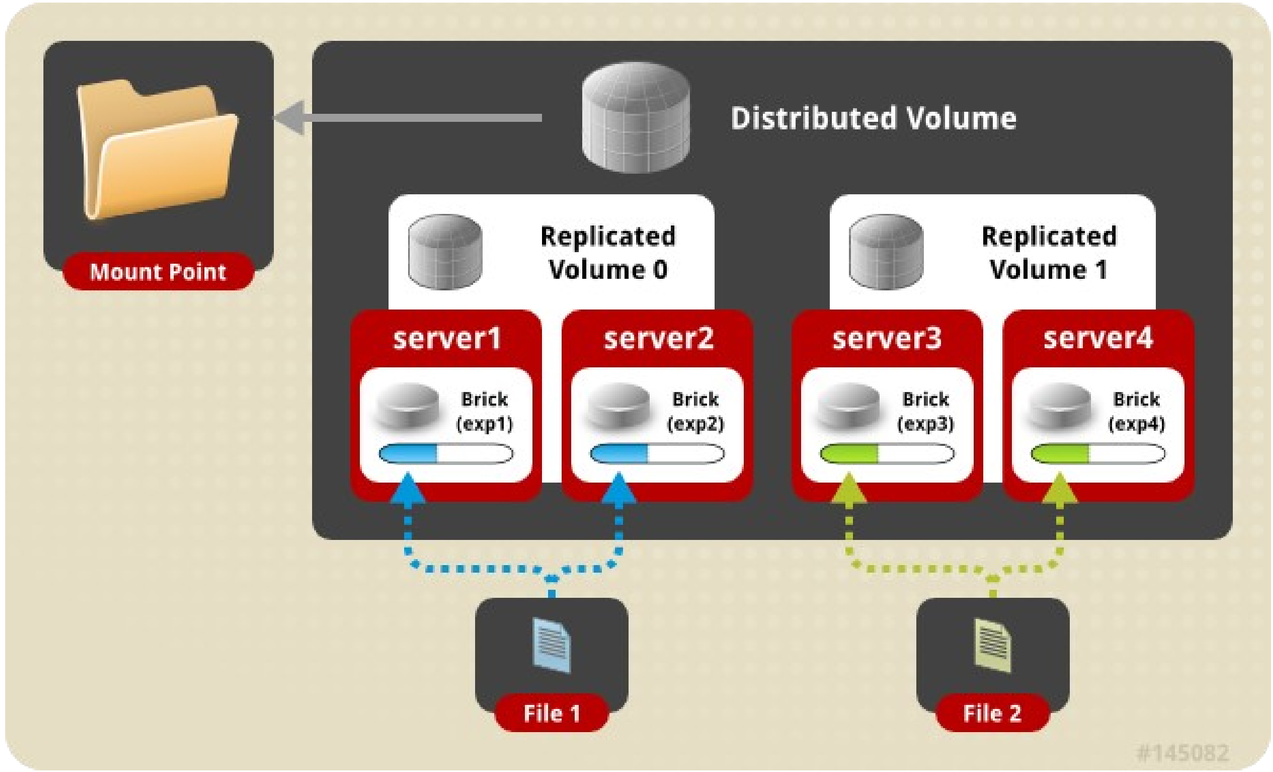
RED HAT SUMMIT
10 YEARS and counting
SAN FRANCISCO | APRIL 14-17, 2014

Layered Functionality

Red Hat Storage Server Administration Deep Dive

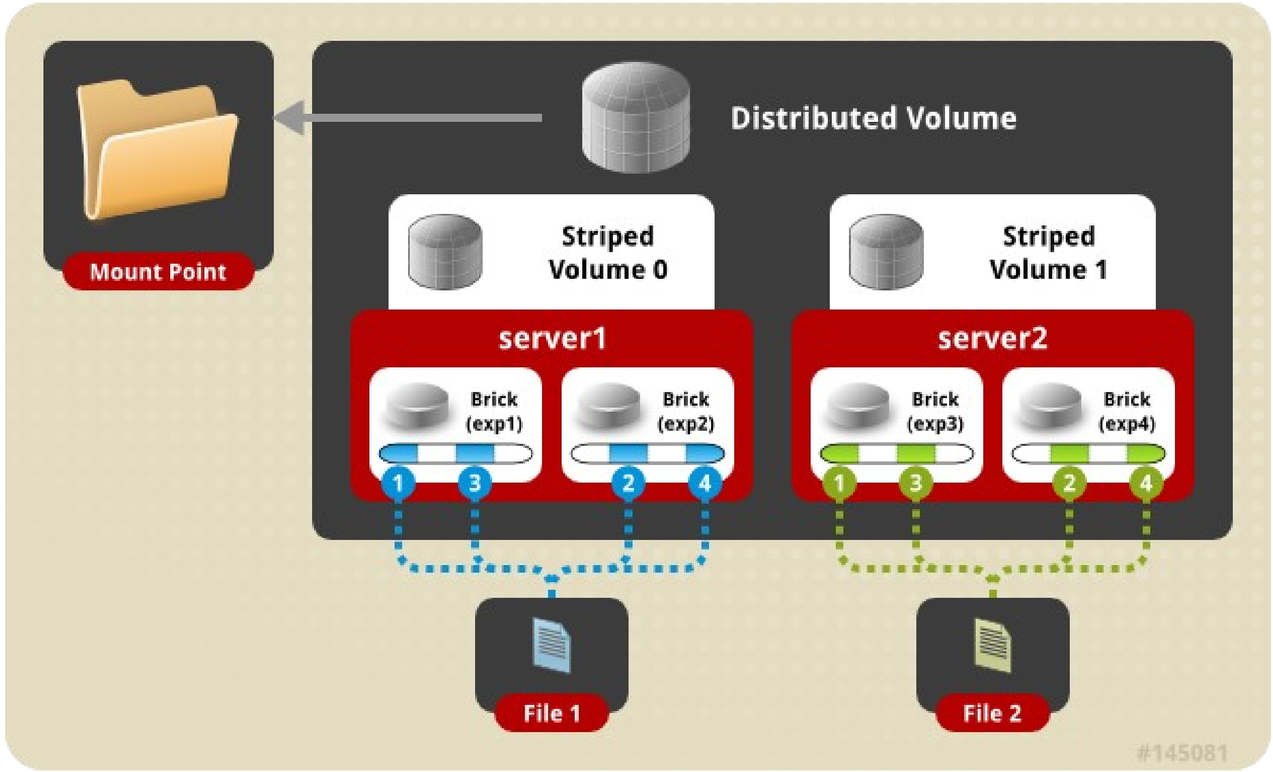
Distributed Replicated Volume

- Distributes files across multiple replica sets



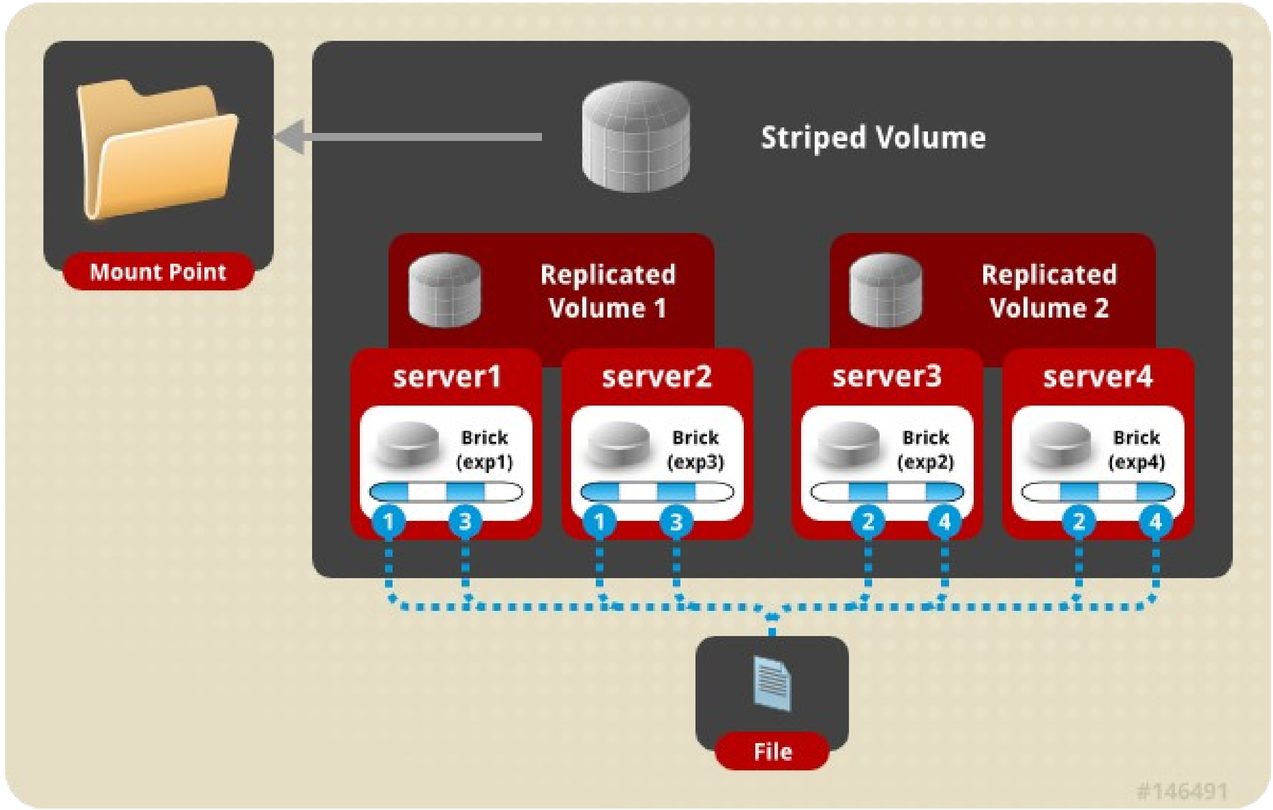
Distributed Striped Volume

- Distributes files across multiple stripe sets
- Striping plus scalability



Striped Replicated Volume

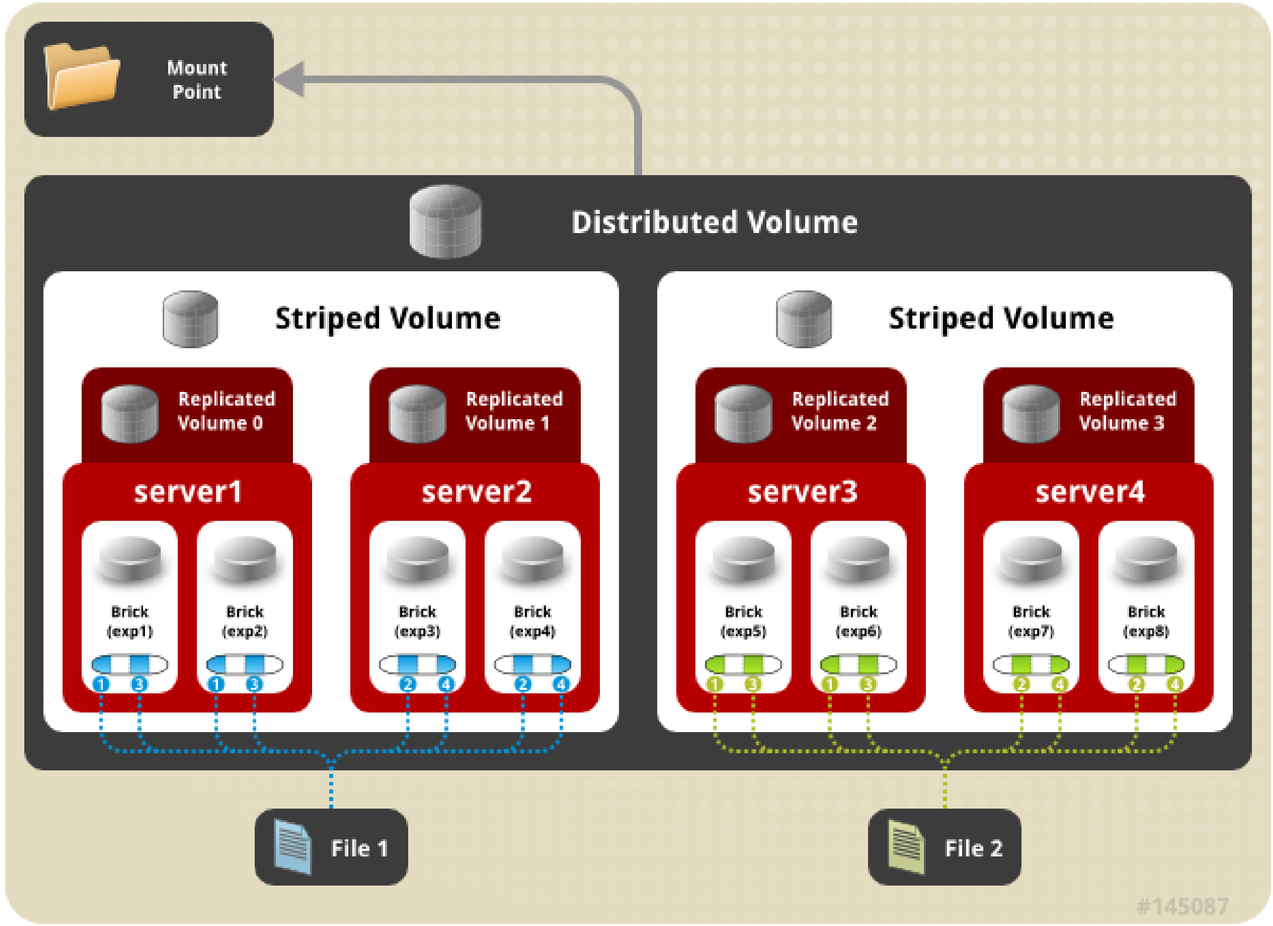
- Replicated sets of stripe sets
- Similar to RAID 10 (1+0)



Distributed Striped Replicated Volume

- Limited Use Cases – Map Reduce

Don't do it like this - ->

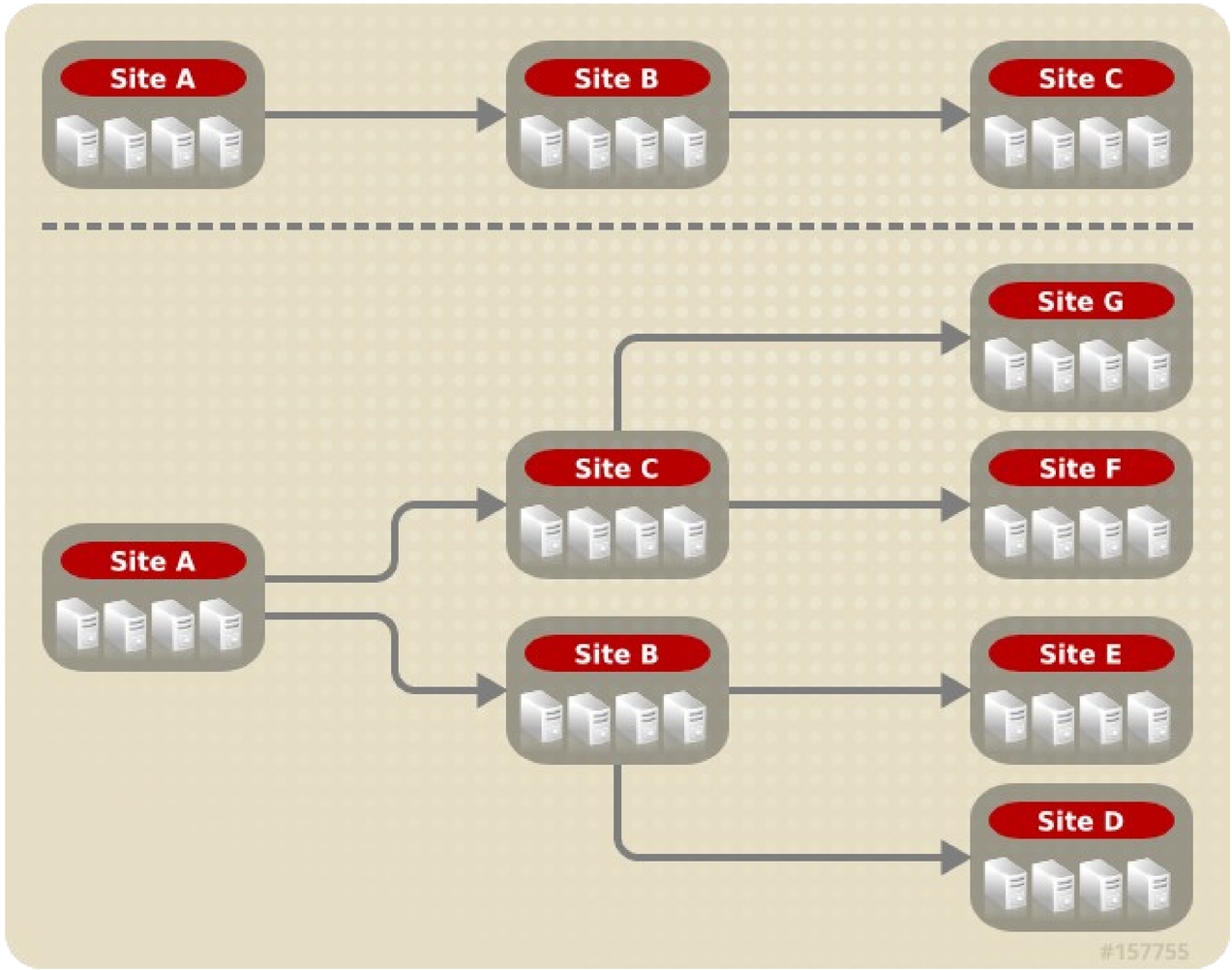


Asynchronous Replication

Red Hat Storage Server Administration Deep Dive

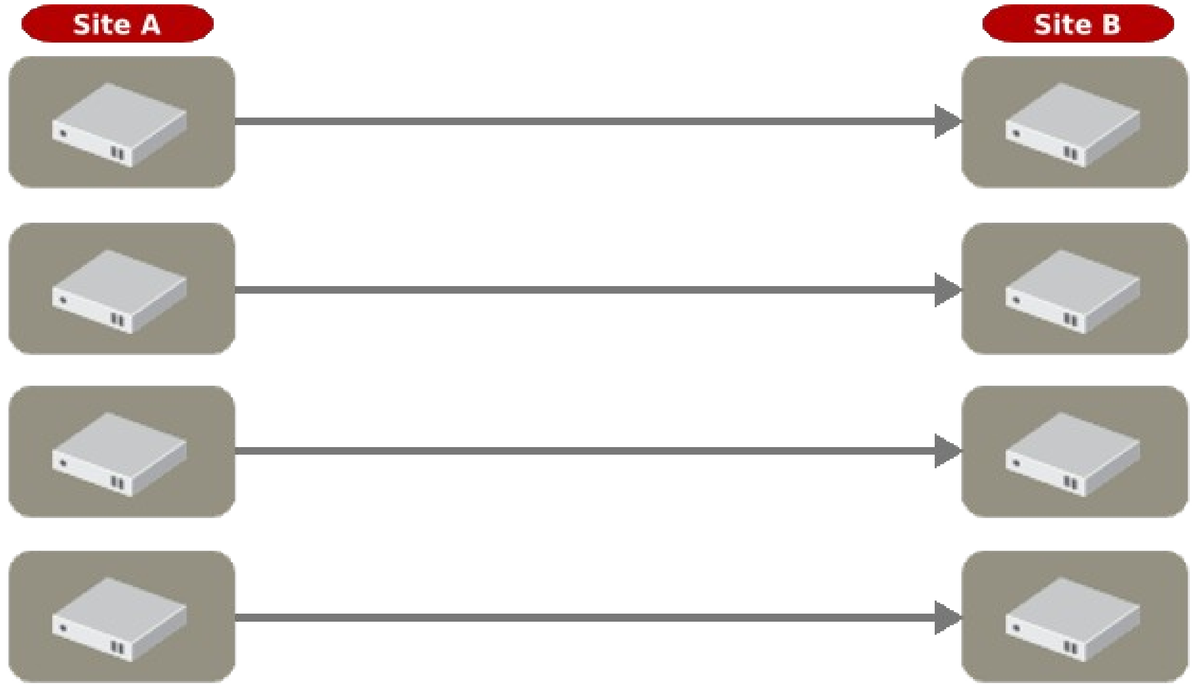
Geo Replication

- Asynchronous across LAN, WAN, or Internet
- Master-Slave model
  - Cascading possible
- Continuous and incremental
- One Way



NEW! Distributed Geo-Replication

- Drastic performance improvements
  - Parallel transfers
  - Efficient source scanning
  - Pipelined and batched
  - File type/layout agnostic
- Available now in RHS 2.1
- Planned for GlusterFS 3.5



Distributed Geo-Replication

- Drastic performance improvements
  - Parallel transfers
  - Efficient source scanning
  - Pipelined and batched
  - File type/layout agnostic
- Perhaps it's not just for DR anymore...

http://www.redhat.com/resourcelibrary/case-studies/intuit-leverages-red-hat-storage-for-always-available-massively-scalable-storage

Data Access

Red Hat Storage Server Administration Deep Dive

GlusterFS Native Client (FUSE)

- FUSE kernel module allows the filesystem to be built and operated entirely in userspace
- Specify mount to any GlusterFS server
- Native Client fetches volfile from mount server, then communicates directly with all nodes to access data
- Recommended for high concurrency and high write performance
- Load is inherently balanced across distributed volumes

NFS

- Standard NFS v3 clients
- Standard automounter is supported
- Mount to any server, or use a load balancer
- GlusterFS NFS server includes Network Lock Manager (NLM) to synchronize locks across clients
- Better performance for reading many small files from a single client
- Load balancing must be managed externally

NEW! libgfapi

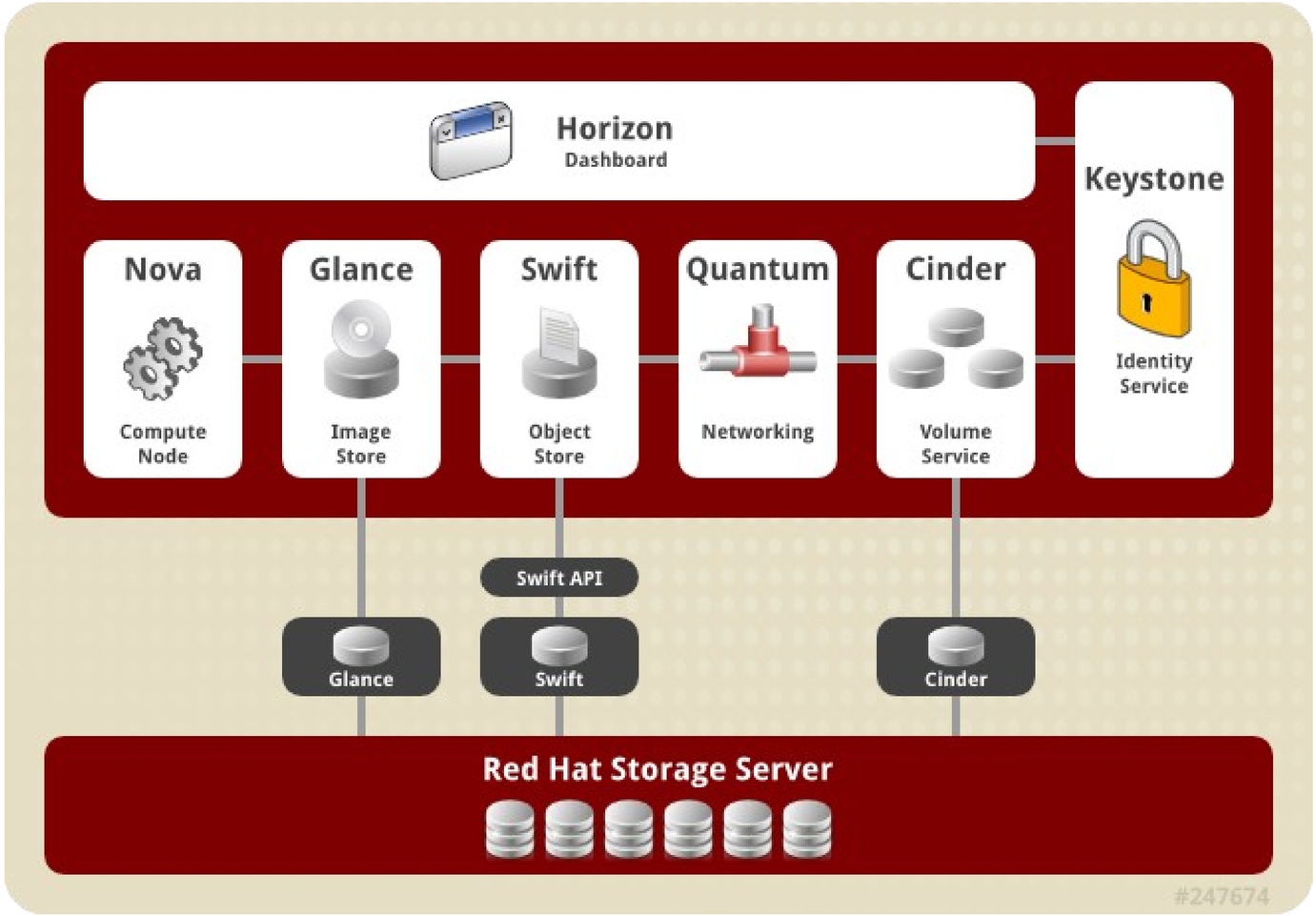
- Introduced with GlusterFS 3.4
- User-space library for accessing data in GlusterFS
- Filesystem-like API
- Runs in application process
- no FUSE, no copies, no context switches
- ...but same volfiles, translators, etc.

SMB/CIFS

- NEW! In GlusterFS 3.4 – Samba + libgfapi
  - No need for local native client mount & re-export
  - Significant performance improvements with FUSE removed from the equation
- Must be setup on each server you wish to connect to via CIFS
- CTDB is required for Samba clustering

HDFS Compatibility

Gluster 4 OpenStack (G4O)☥



☥The feature formerly known as UFO

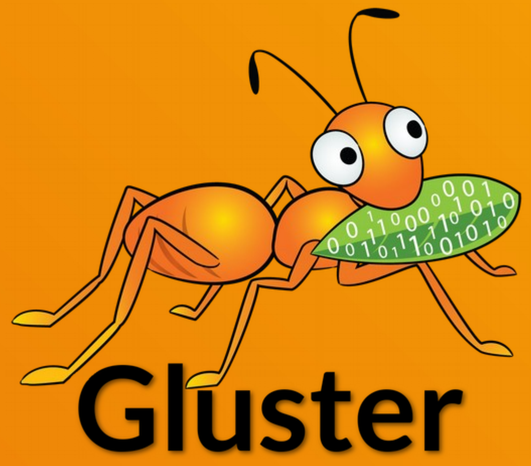
Gluster

redhat.

RED HAT SUMMIT

10 YEARS *and counting*

SAN FRANCISCO | APRIL 14-17, 2014



SWAG Intermission

Red Hat Storage Server Administration Deep Dive

Gluster

redhat.

Demo Time!

Red Hat Storage Server Administration Deep Dive

Do it!

Red Hat Storage Server Administration Deep Dive

Do it!

- Build a test environment in VMs in just minutes!
- Get the bits:
  - Fedora 20 has GlusterFS packages natively: fedoraproject.org
  - RHS 2.1 ISO available on the Red Hat Portal: access.redhat.com
  - Go upstream: gluster.org
  - Amazon Web Services (AWS)
    - Amazon Linux AMI includes GlusterFS packages
    - RHS AMI is available

Check Out Other Red Hat Storage Activities at The Summit

- Enter the raffle to win tickets for a $500 gift card or trip to LegoLand!
  - Entry cards available in all storage sessions - the more you attend, the more chances you have to win!
- Talk to Storage Experts:
  - Red Hat Booth (# 211)
    - Infrastructure
    - Infrastructure-as-a-Service
- Storage Partner Solutions Booth (# 605)
- Upstream Gluster projects
  - Developer Lounge

Follow us on Twitter, Facebook: @RedHatStorage

RED HAT SUMMIT

10 YEARS *and counting*

SAN FRANCISCO | APRIL 14-17, 2014

Thank You!

***** Please Leave Your Feedback in the Summit Mobile App Session Survey *****

- Contact
  - dustin@redhat.com
  - storage-sales@redhat.com
- Resources
  - www.gluster.org
  - www.redhat.com/storage/
  - access.redhat.com/support/offerings/tam/
- Twitter
  - @dustinlblack
  - @gluster
  - @RedHatStorage

Red Hat Storage Server Administration Deep Dive

Slides Available at: people.redhat.com/dblack

redhat.