GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

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

<venue> -- <date>

Dustin L. Black, RHCA
Sr. Technical Account Manager
Red Hat, Inc.

dustin@redhat.com
@dustinlblack

redhat.
CERTIFIED
ARCHITECT

#whatis 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
- Scaling Up and Out
- A Peek at GlusterFS Logic
- Redundancy and Fault Tolerance
- Data Access
- General Administration
- Use Cases
- Common Pitfalls

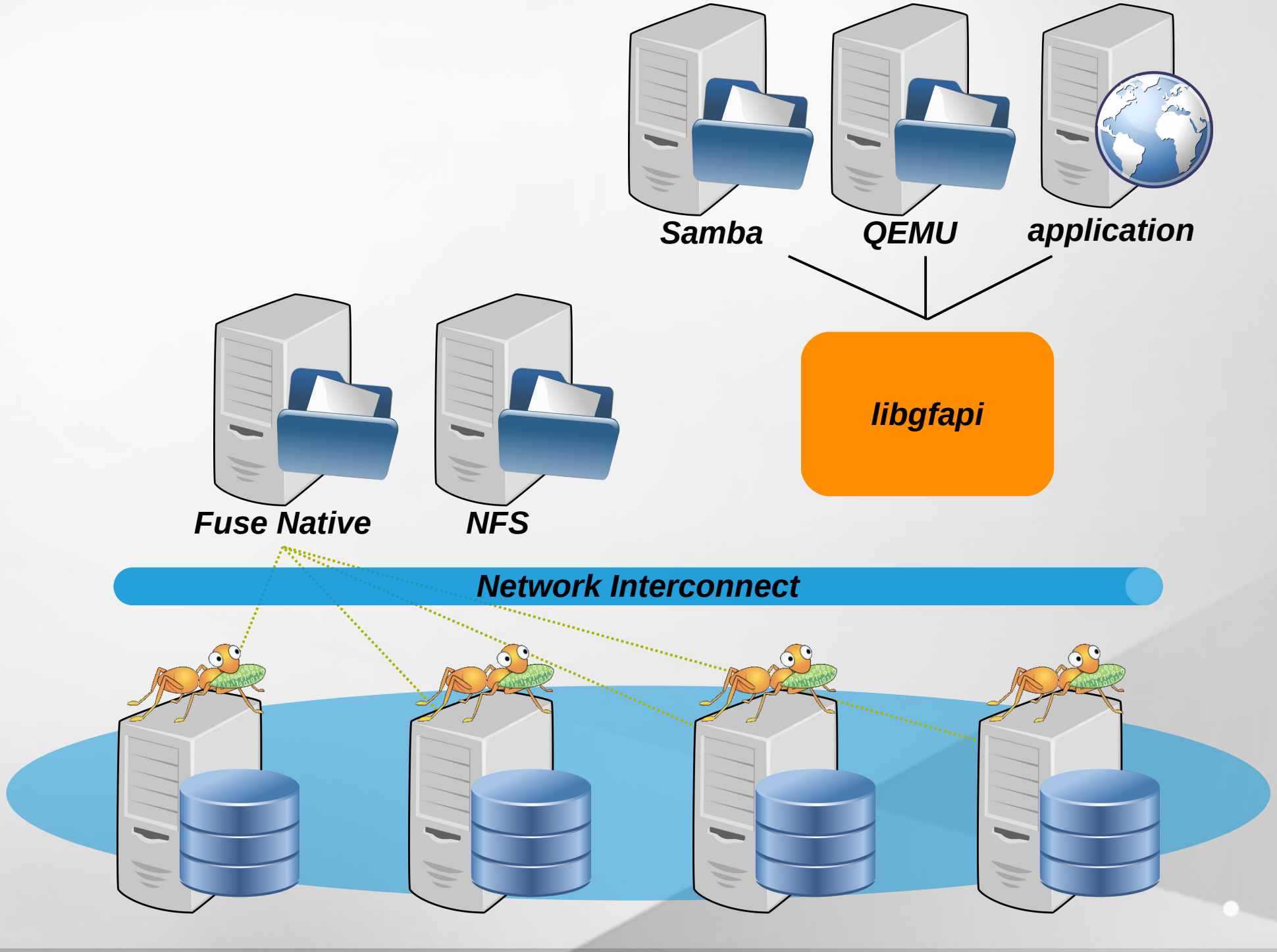
Technology Overview

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

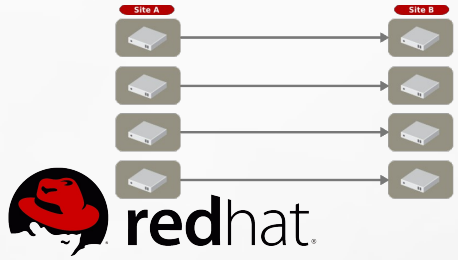
What is GlusterFS?

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

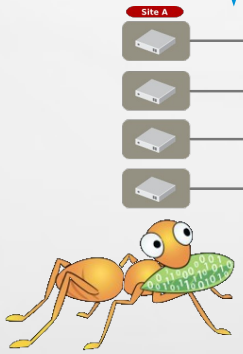

Samba
QEMU
application
libgfapi
Fuse Native
NFS
Network Interconnect

What is Red Hat Storage?

- Enterprise Implementation of GlusterFS
- Software Appliance
- Bare Metal Installation
- Built on RHEL + XFS
- Subscription Model
- Storage Software Appliance
  - Datacenter and Private Cloud Deployments
- Virtual Storage Appliance
  - Amazon Web Services Public Cloud Deployments



release



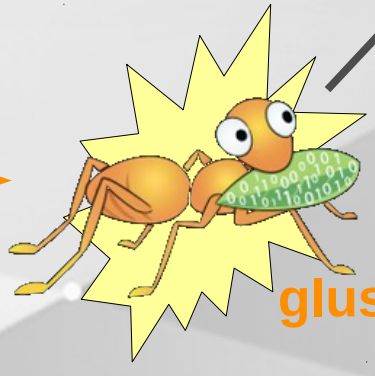
upstream

iterate

enhance

debug

release



glusterfs-3.4

RED HAT® STORAGE 2.1

GlusterFS vs. Traditional Solutions

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

Use Cases

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Common Solutions

- Media / Content Distribution Network (CDN)
- Backup / Archive / Disaster Recovery (DR)
- Large Scale File Server
- Home directories
- High Performance Computing (HPC)
- Infrastructure as a Service (IaaS) storage layer

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

Technology Stack

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Terminology

- Brick
  - A filesystem mountpoint
  - A unit of storage used as a GlusterFS building block
- Translator
  - Logic between the bits and the Global Namespace
  - Layered to provide GlusterFS functionality
- Volume
  - Bricks combined and passed through translators
- Peer
  - Server running the gluster daemon and sharing 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)
- XFS, EXT3/4, BTRFS
  - Extended attributes support required
  - *RHS: XFS required*

Gluster Components

- `glusterd`
  - Elastic volume management daemon
  - Runs on all export servers
  - Interfaced through `gluster` CLI
- `glusterfsd`
  - GlusterFS brick daemon
  - One process for each brick
  - Managed by `glusterd`

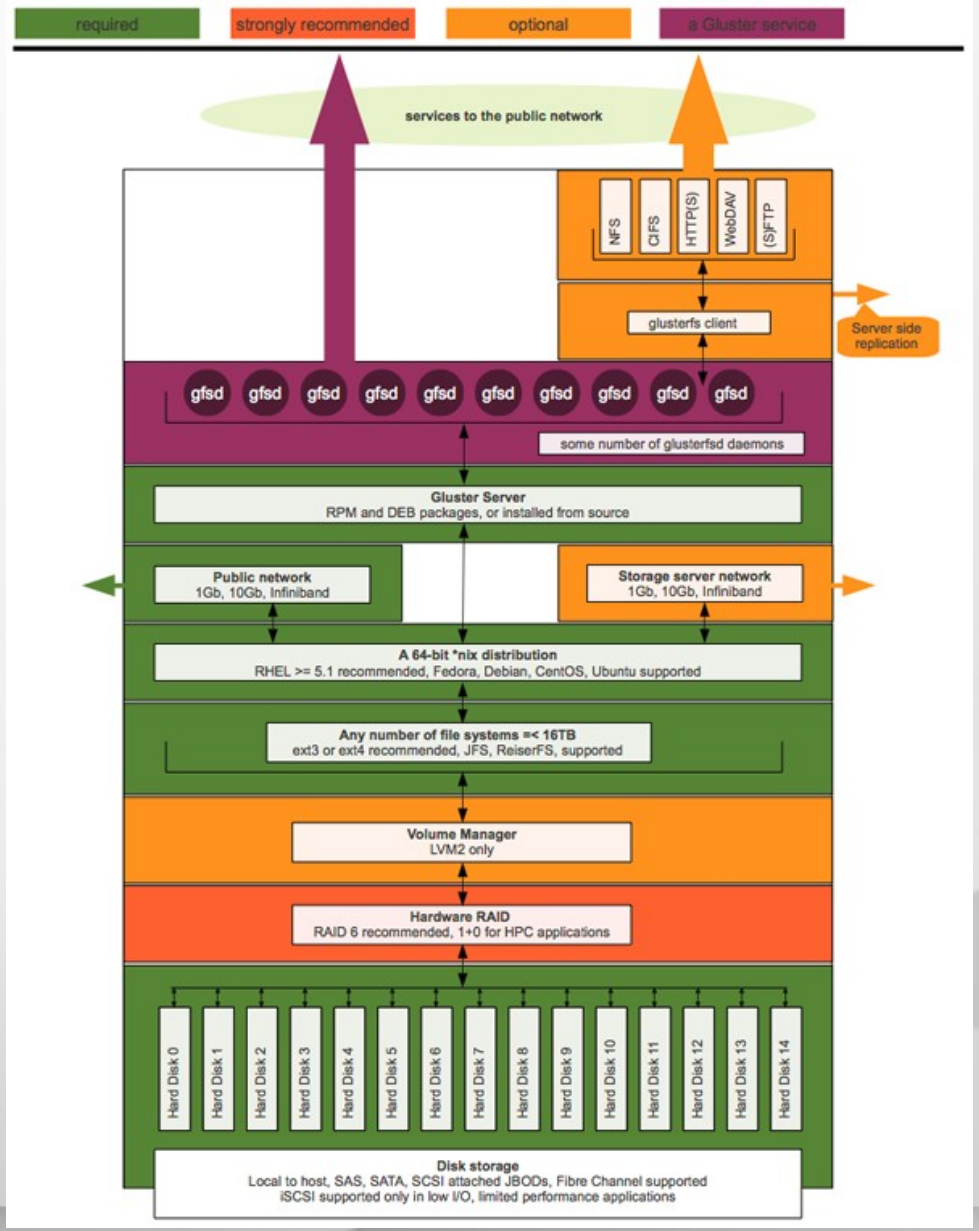
Gluster Components

- `glusterfs`
  - NFS server daemon
  - Self-heal daemon
  - FUSE client daemon
- `mount.glusterfs`
  - FUSE native mount tool
- `gluster`
  - Gluster Console Manager (CLI)

Data Access Overview

- GlusterFS Native Client
  - Filesystem in Userspace (FUSE)
- NFS
  - Built-in Service
- SMB/CIFS
  - Samba server required; NOW libgfapi-integrated!
- Gluster For OpenStack (G4O; aka UFO)
  - Simultaneous object-based access via Swift
- NEW! libgfapi flexible abstracted storage
  - Integrated with upstream Samba and Ganesha-NFS

Putting it All Together

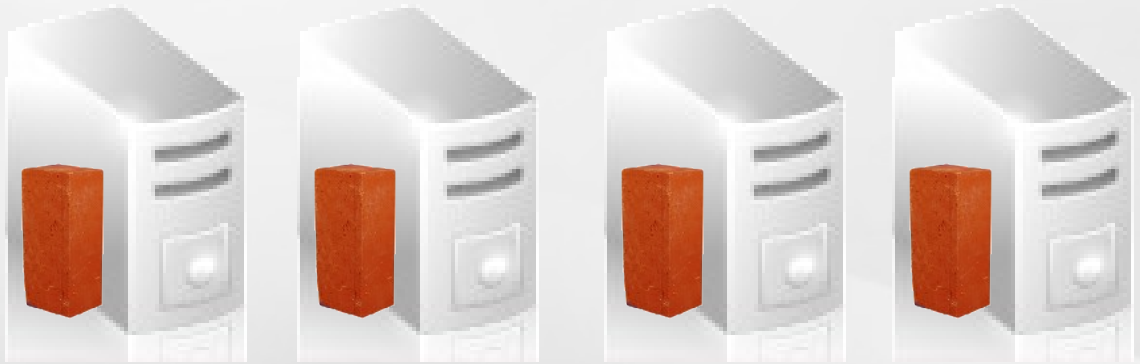


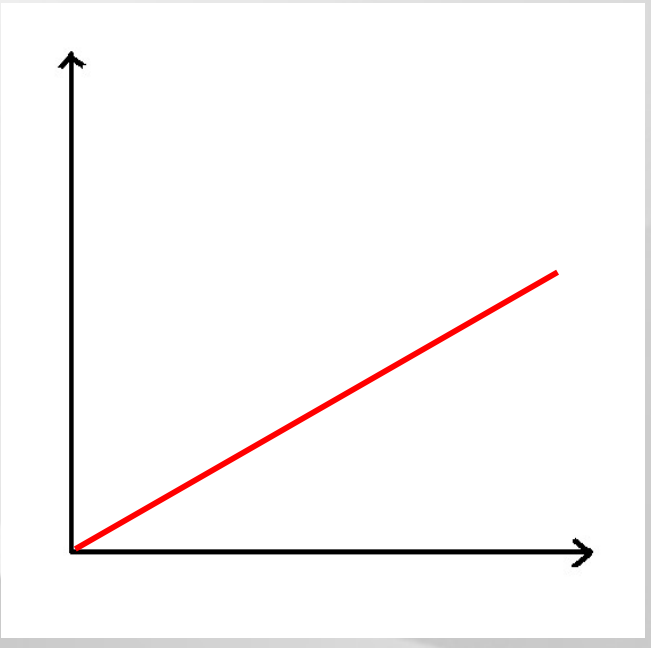
Scaling Up

- Add disks and filesystems to a server
- Expand a GlusterFS volume by adding bricks

Scaling Out

- Add GlusterFS nodes to trusted pool
- Add filesystems as new bricks





Under the Hood

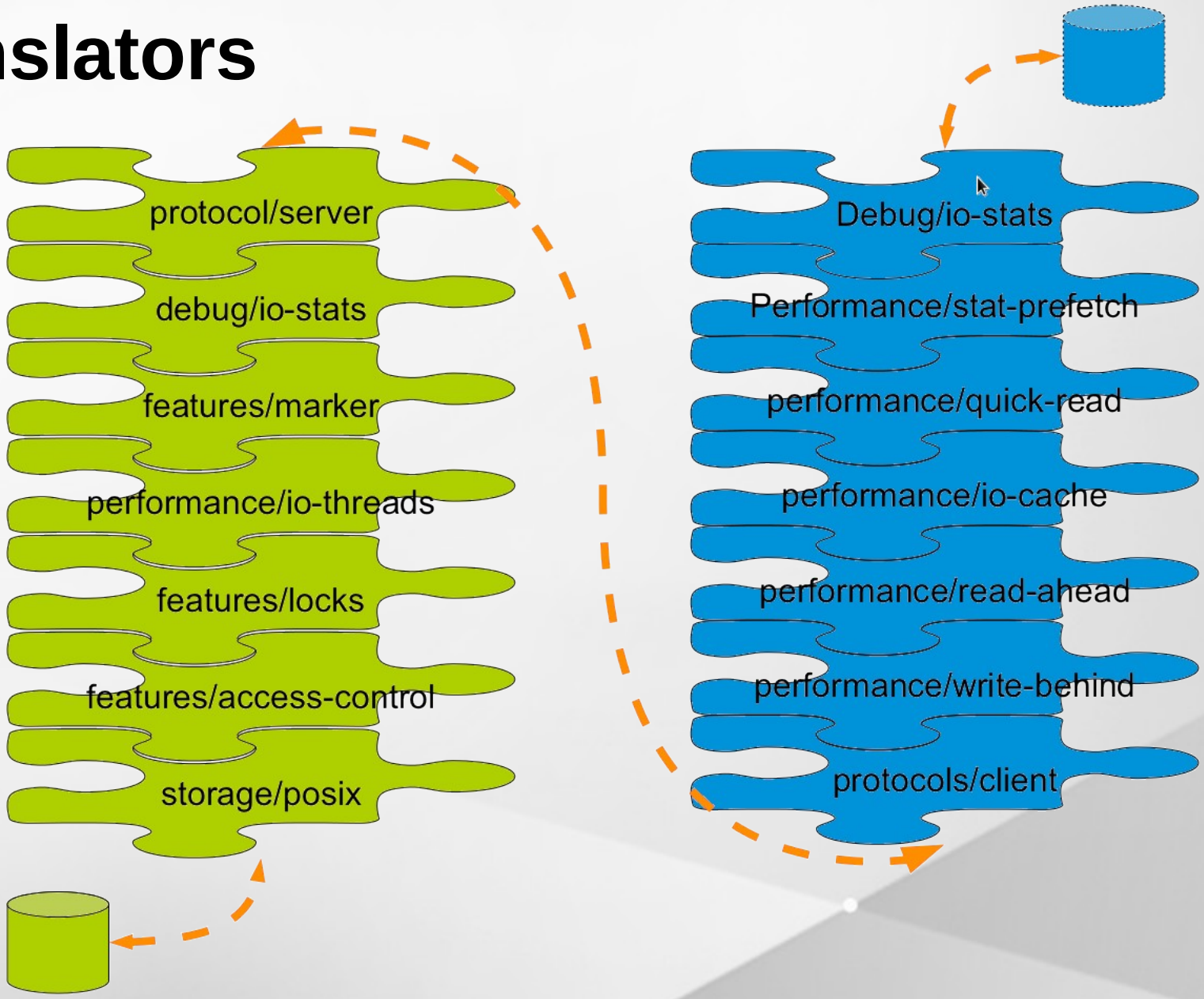
GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

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 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

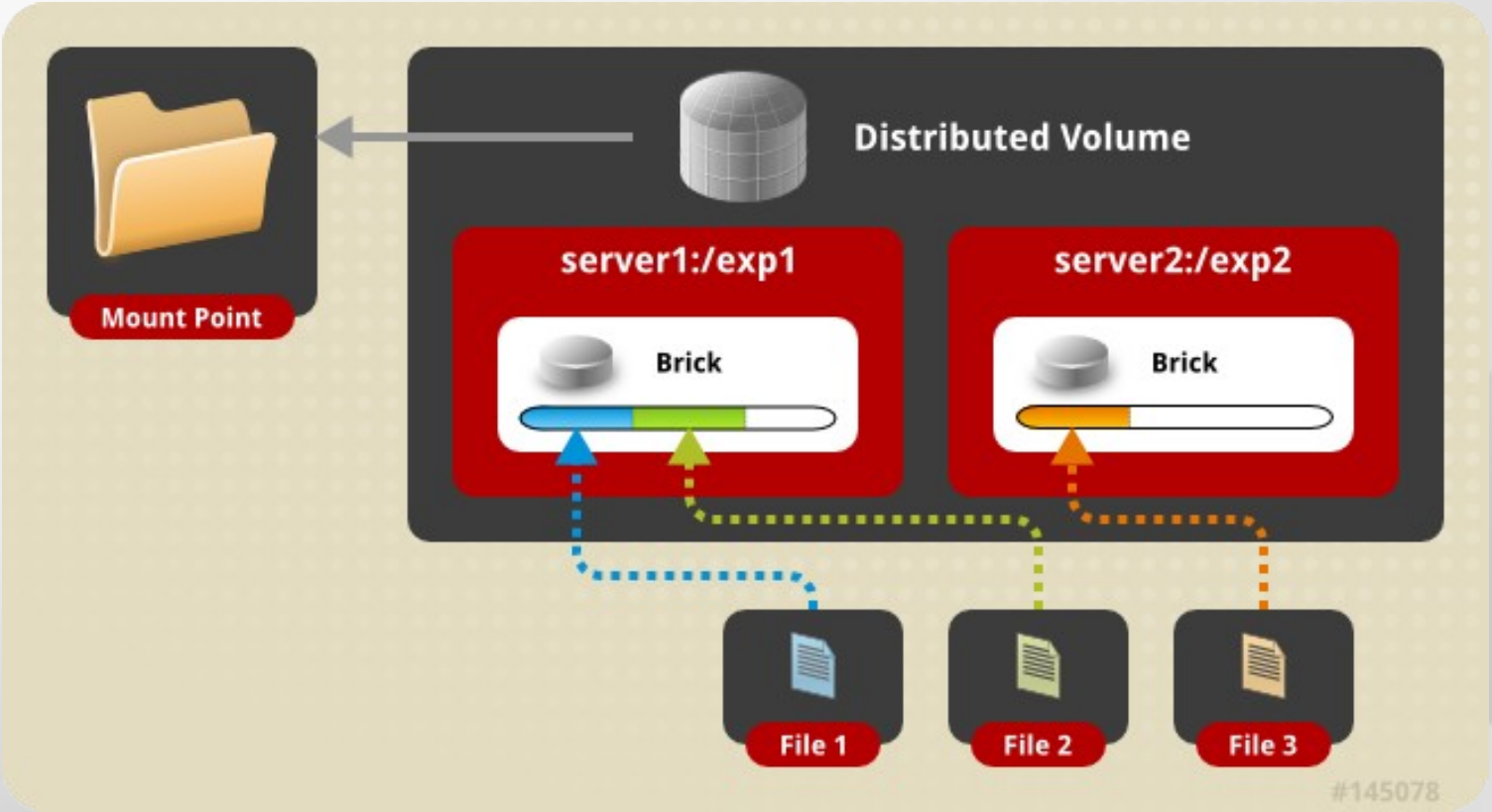
Basic Volumes

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

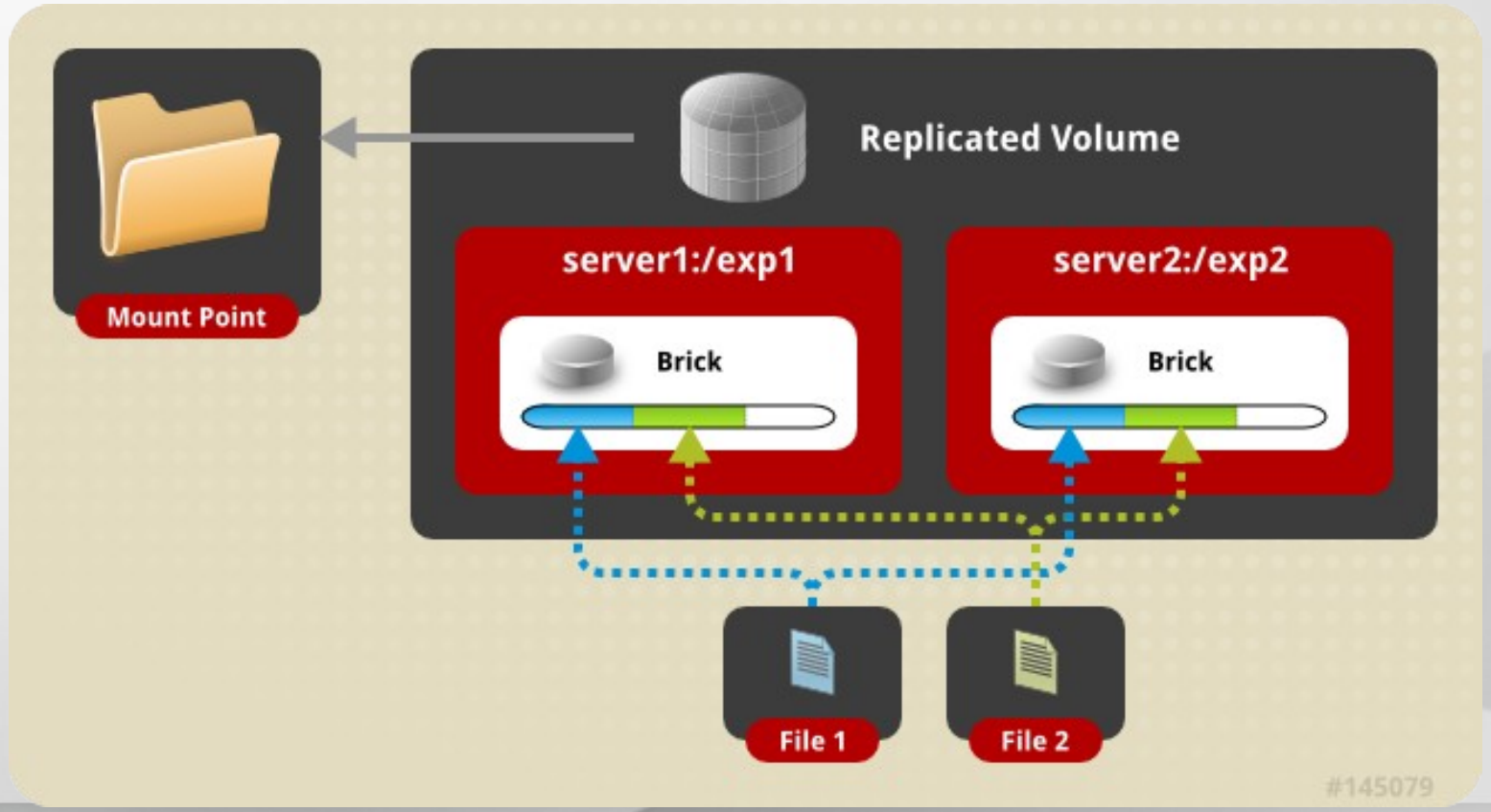
Distributed Volume

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



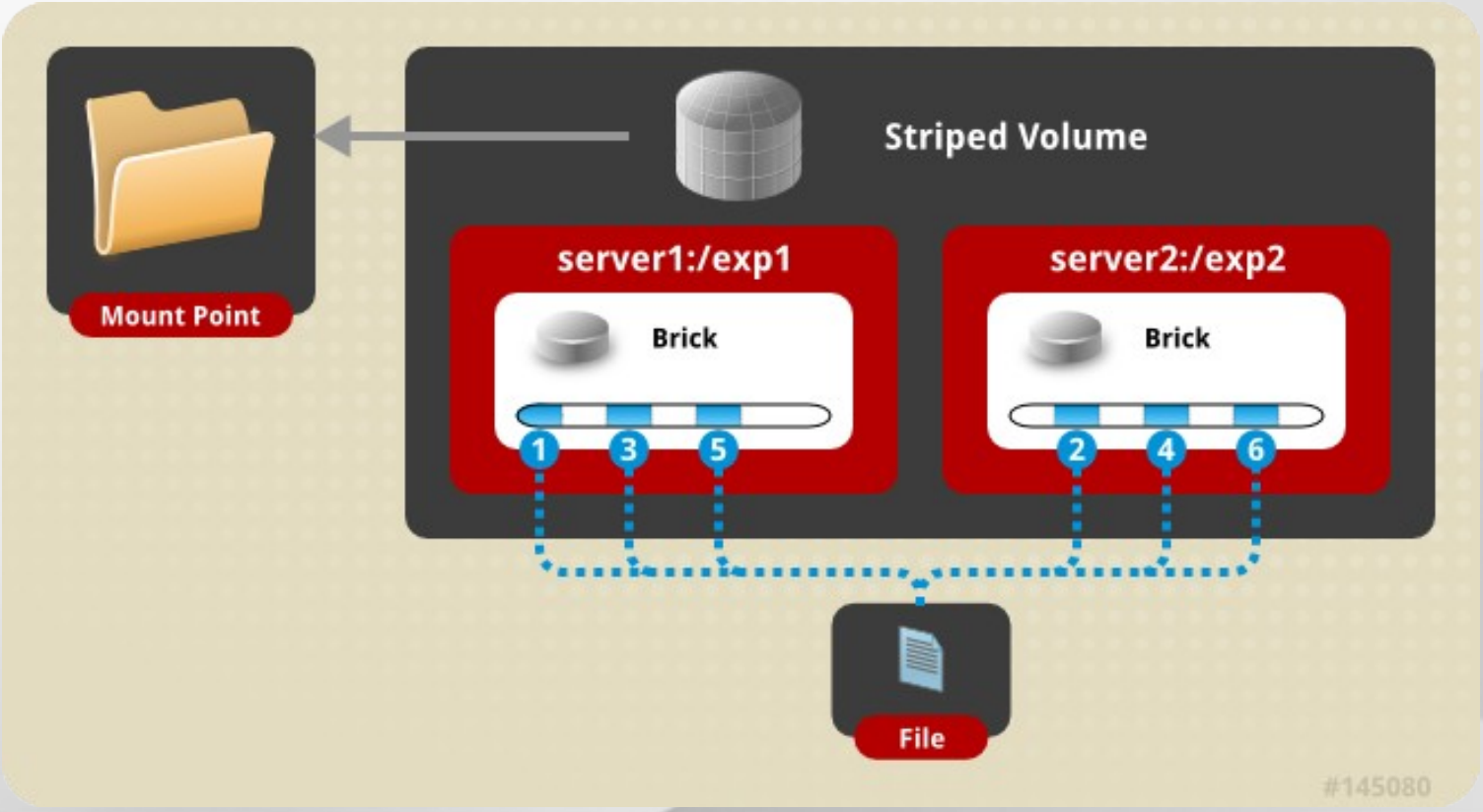
Replicated Volume

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



Striped Volumes

- Individual files split among bricks
- *Similar* to block-level RAID 0
- *Limited Use Cases* – HPC Pre/Post Processing



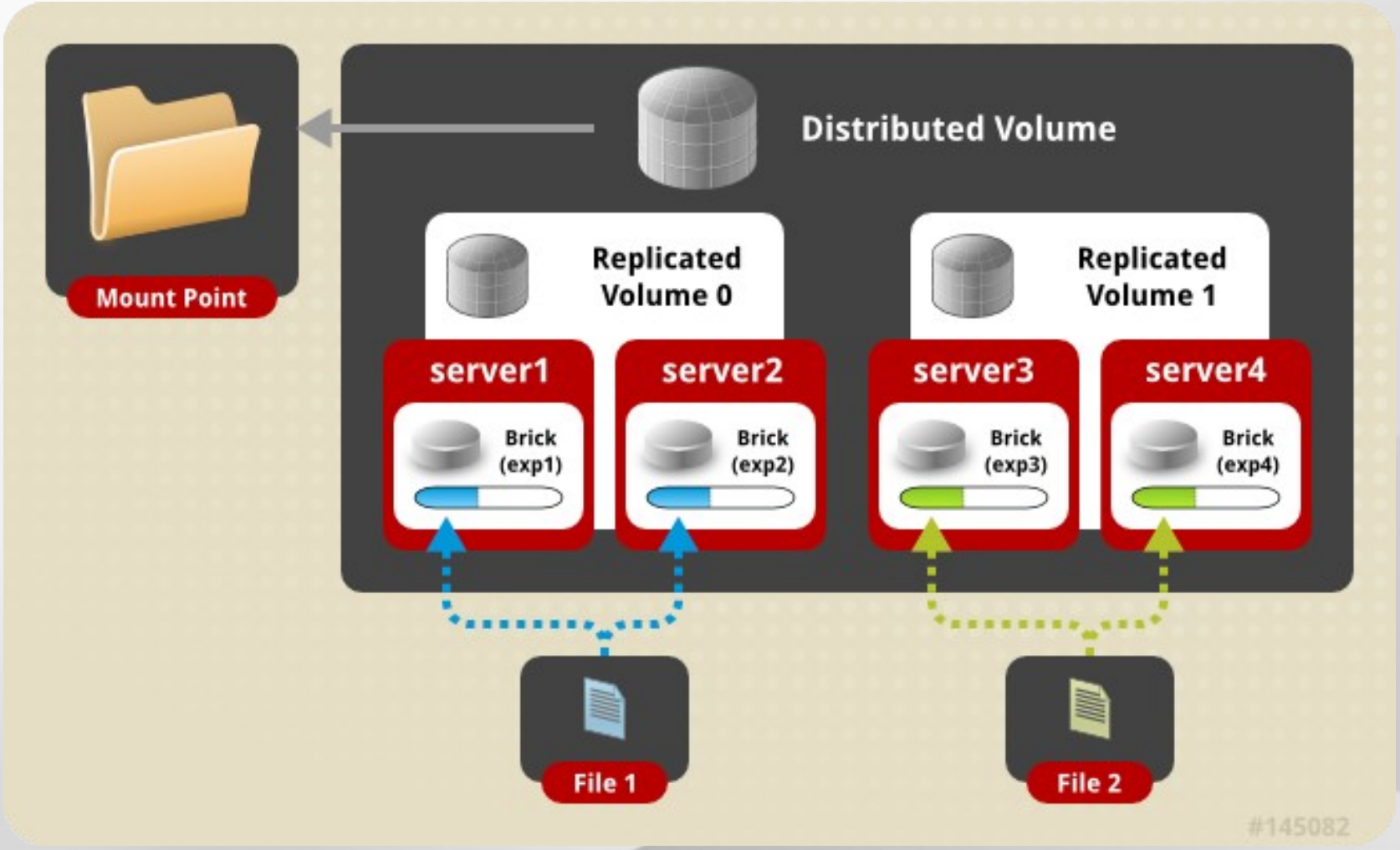
Layered Functionality

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

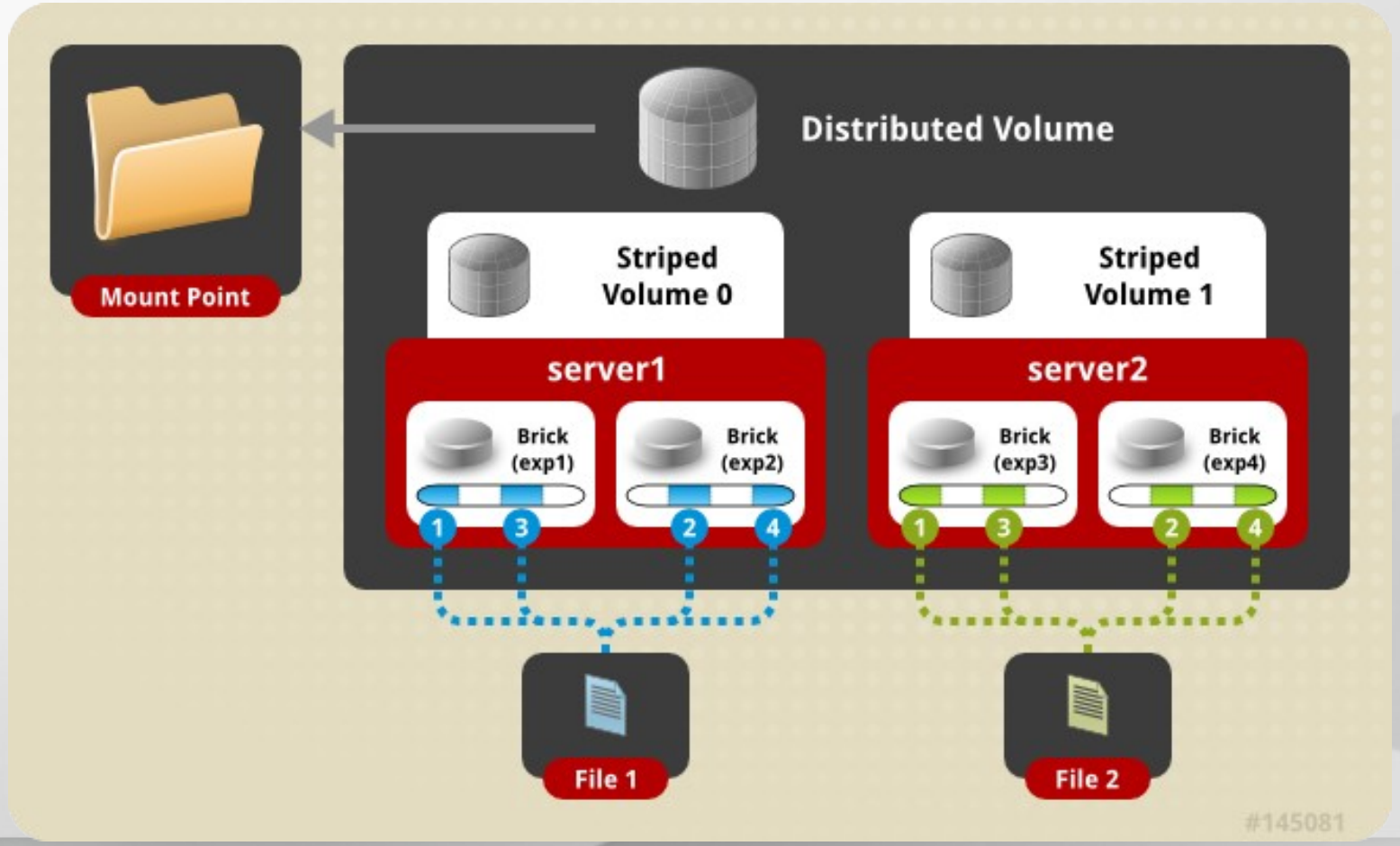
Distributed Replicated Volume

- Distributes files across replicated bricks



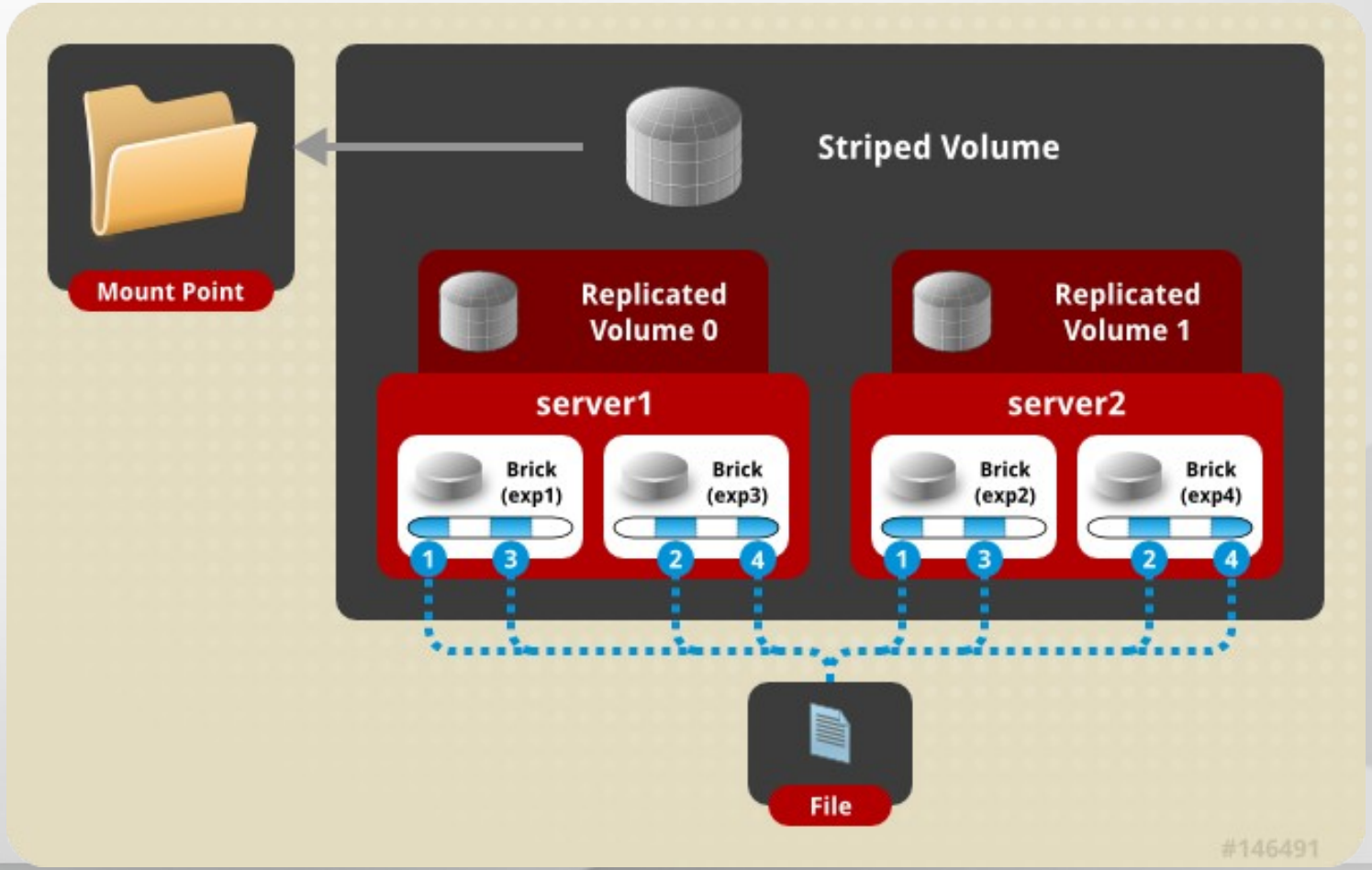
Distributed Striped Volume

- Files striped across two or more nodes
- Striping plus scalability



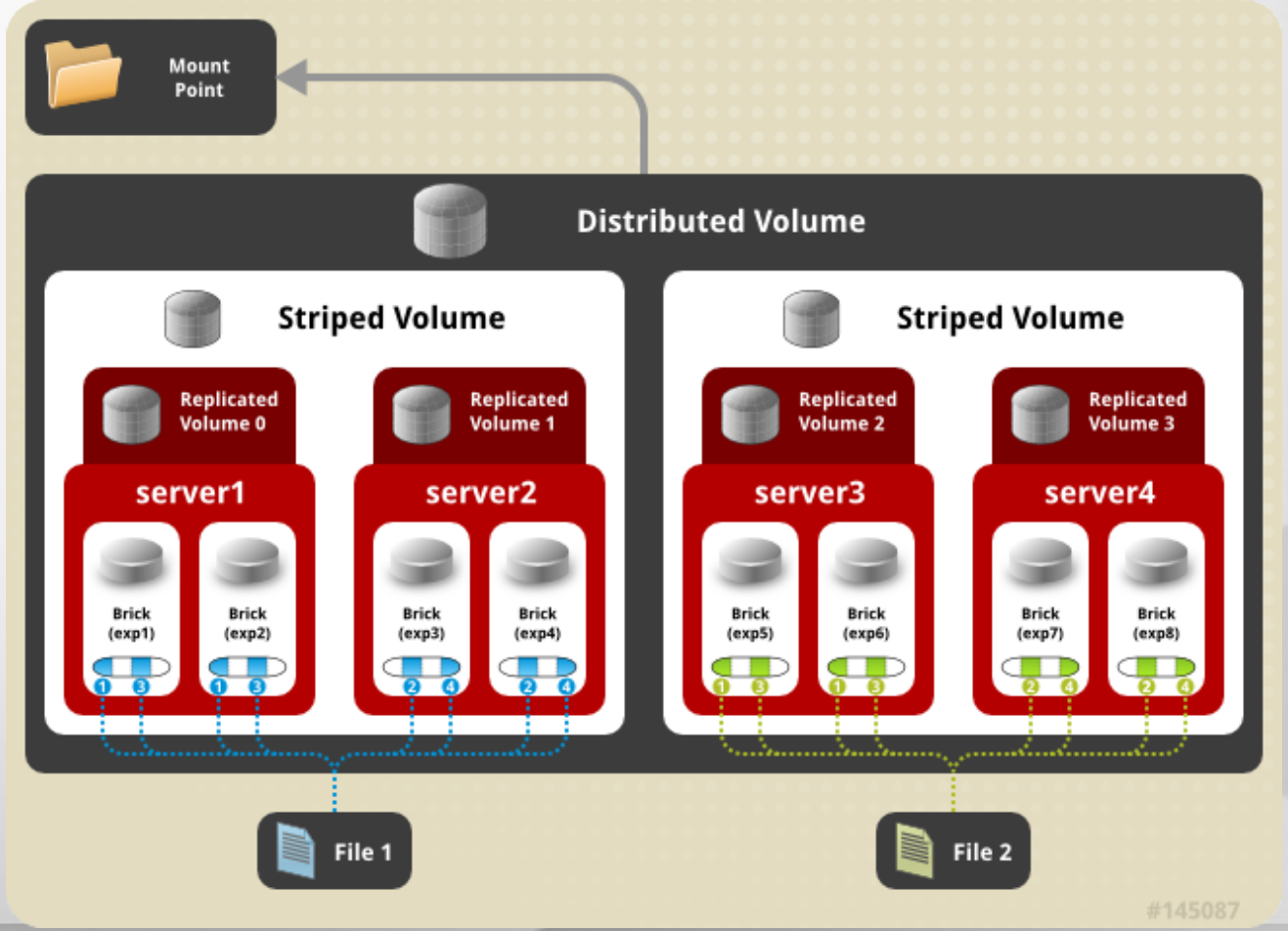
Striped Replicated Volume

- RHS 2.0 / GlusterFS 3.3+
- *Similar* to RAID 10 (1+0)



Distributed Striped Replicated Volume

- RHS 2.0 / GlusterFS 3.3+
- *Limited Use Cases* – Map Reduce



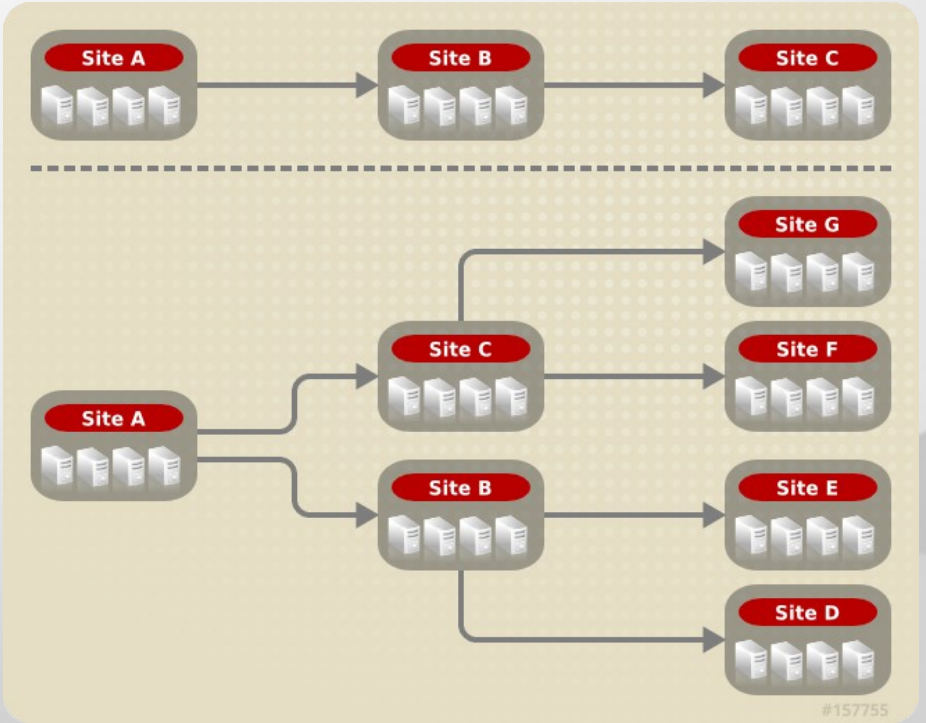
Asynchronous Replication

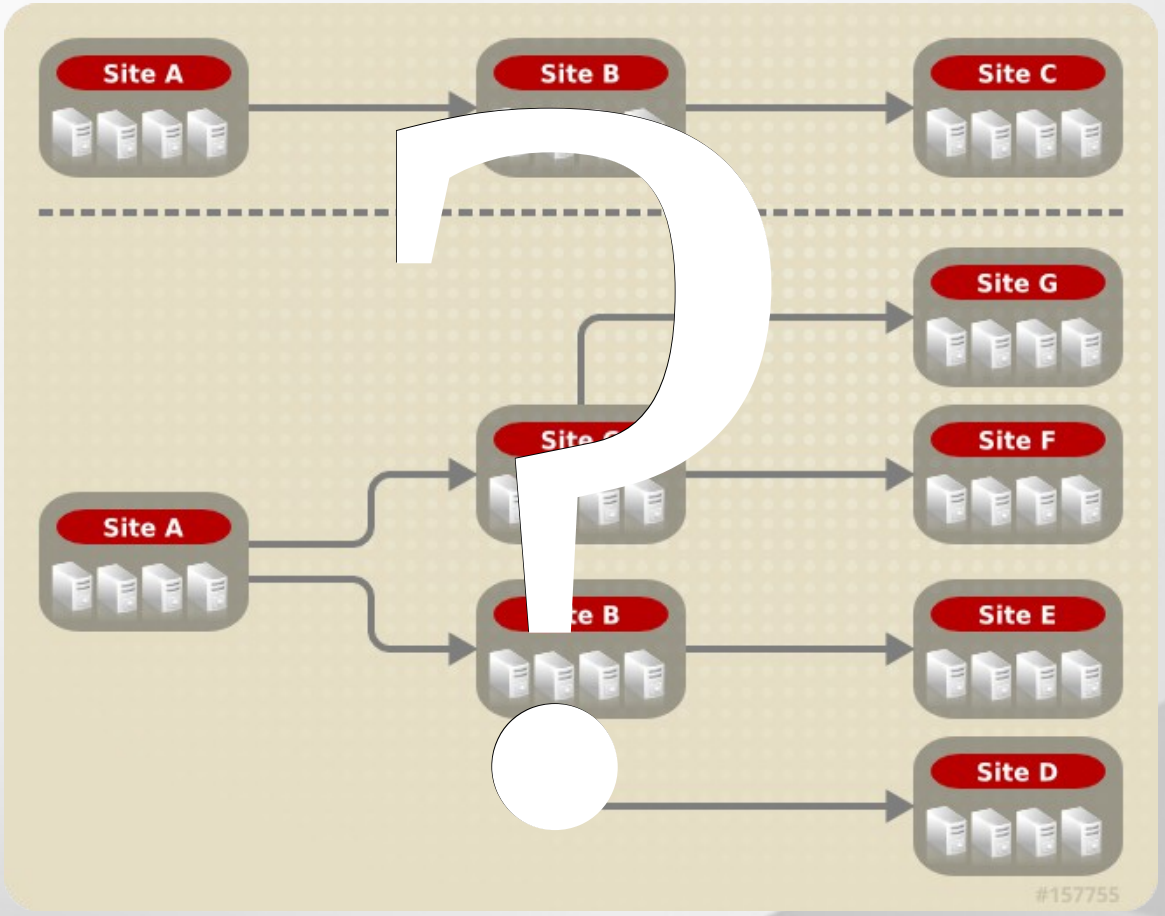
GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Geo Replication

- Asynchronous across LAN, WAN, or Internet
- Master-Slave model -- Cascading possible
- Continuous and incremental
- Data is passed between defined master and slave **only**





NEW! Distributed Geo-Replication

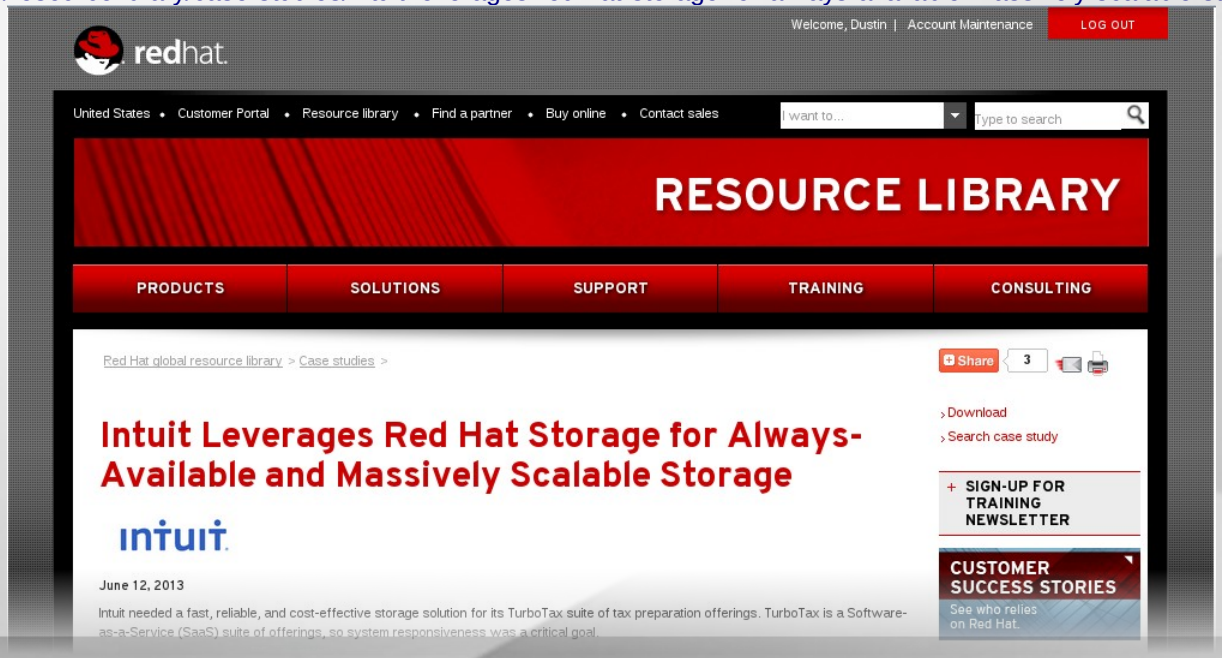
Distributed Geo-Replication

- Drastic performance improvements
  - Parallel transfers
  - Efficient source scanning
  - 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
  - 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

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

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

Demo Time!

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Do it!

Do it!

- Build a test environment in VMs in just minutes!
- Get the bits:
  - Fedora 19 has GlusterFS packages natively
  - RHS 2.1 ISO available on Red Hat Portal
  - Go upstream: www.gluster.org
  - Amazon Web Services (AWS)
    - Amazon Linux AMI includes GlusterFS packages
    - RHS AMI is available

Thank You!

Slides Available at: **http://people.redhat.com/dblack**

- **dustin@redhat.com**
- **storage-sales@redhat.com**
- **RHS:**

  www.redhat.com/storage/
- **GlusterFS:**

  www.gluster.org
- **TAM:**
  access.redhat.com/support/offerings/tam/

@Glusterorg

@RedHatStorage

Gluster

Red Hat Storage

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo