



Demystifying Gluster

GlusterFS for the SysAdmin

Dustin L. Black, RHCA

Sr. Technical Account Manager, Red Hat

2012-06-08

Agenda

- Technology Overview
- A Peek at Gluster Logic
- Scaling Up and Out
- Redundancy and Fault Tolerance
- Data Access
- General Administration
- Use Cases





Technology Overview

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

What is GlusterFS?

- POSIX-Compliant Distributed File System
- No Metadata Server
- Network Attached Storage (NAS)
- Heterogeneous Commodity Hardware
- Aggregated Storage and Memory
- Standards-Based – Clients, Applications, Networks
- Flexible and Agile Scaling
 - Capacity – Petabytes and beyond
 - Performance – Thousands of Clients
- Single Global Namespace



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



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





Technology Stack

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

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
- Node
 - Server running the gluster daemon and sharing volumes



Foundation Components

- Private Cloud (Datacenter)
 - Common Commodity x86_64 Servers
 - RHS: Hardware Compatibility List (HCL)
- Public Cloud
 - Amazon Web Services (AWS)
 - EC2 + EBS



Disk, LVM, and Filesystems

- Direct-Attached Storage (DAS)
- Just a Bunch Of Disks (JBOD)
- Hardware RAID
 - RHS: RAID 6 required
- Logical Volume Management (LVM)
- XFS, EXT3/4, BTRFS
 - Extended attributes support required



Data Access Overview

- Gluster Native Client
 - Filesystem in Userspace (FUSE)
- NFS
 - Built-in Service
- SMB/CIFS
 - Samba server 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`

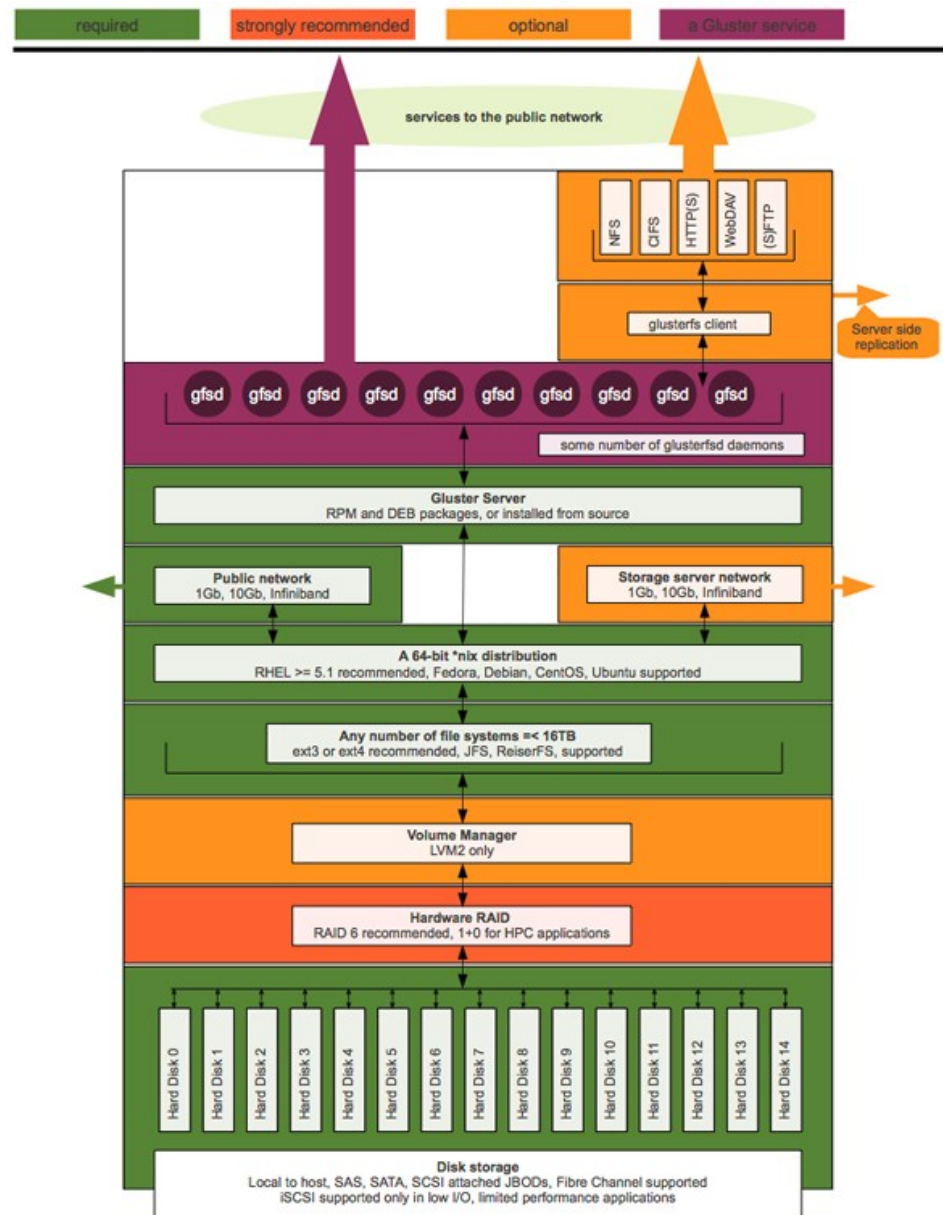


Gluster Components

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



Putting it All Together





Scaling

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

Scaling Up

- Add disks and filesystems to a node
- Expand a gluster volume by adding bricks

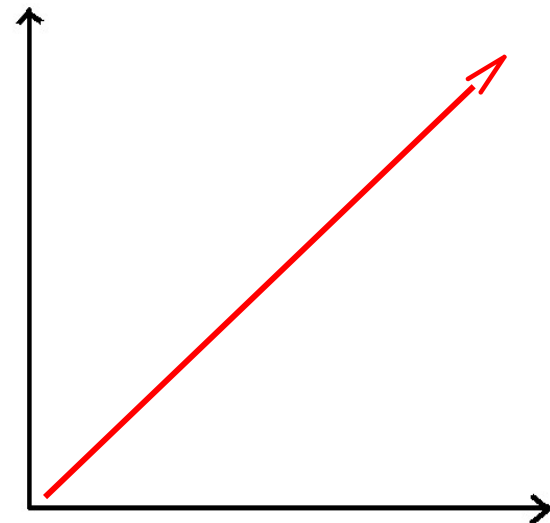
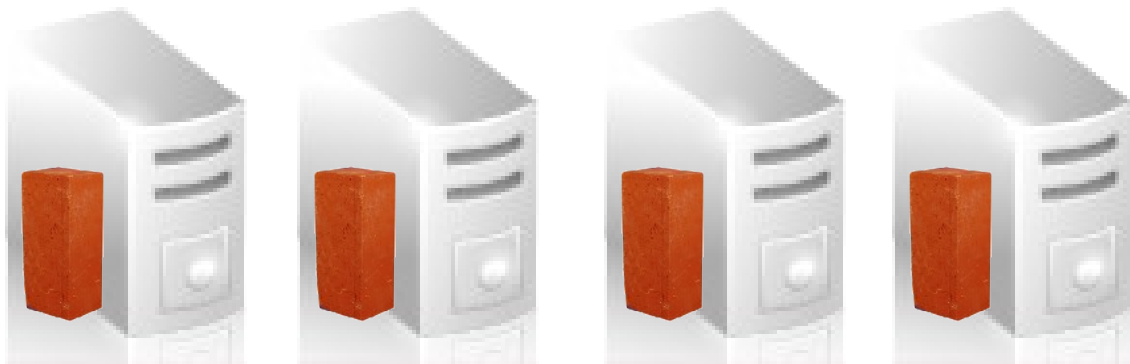


XFS



Scaling Out

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





Under the Hood

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

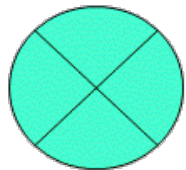
Elastic Hash Algorithm

- No central metadata
 - No Performance Bottleneck
 - Eliminates risk scenarios
- Location hashed intelligently on path and 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



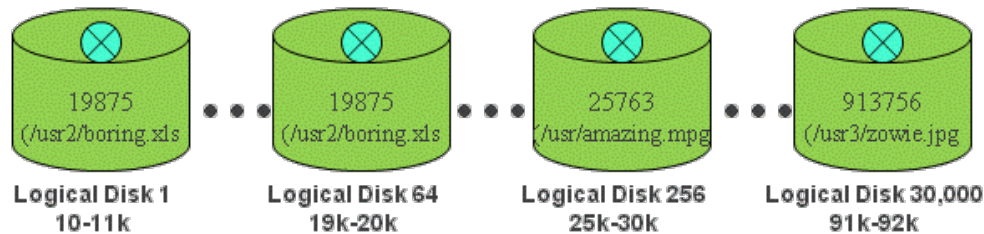
Elastic Hash Algorithm

Goal: Systematically locate files based solely on their name



Algorithm

- 1) Run path/filename through hash
- 2) Assign to logical disk based on numerical result
- 3) Separate logical storage from physical storage



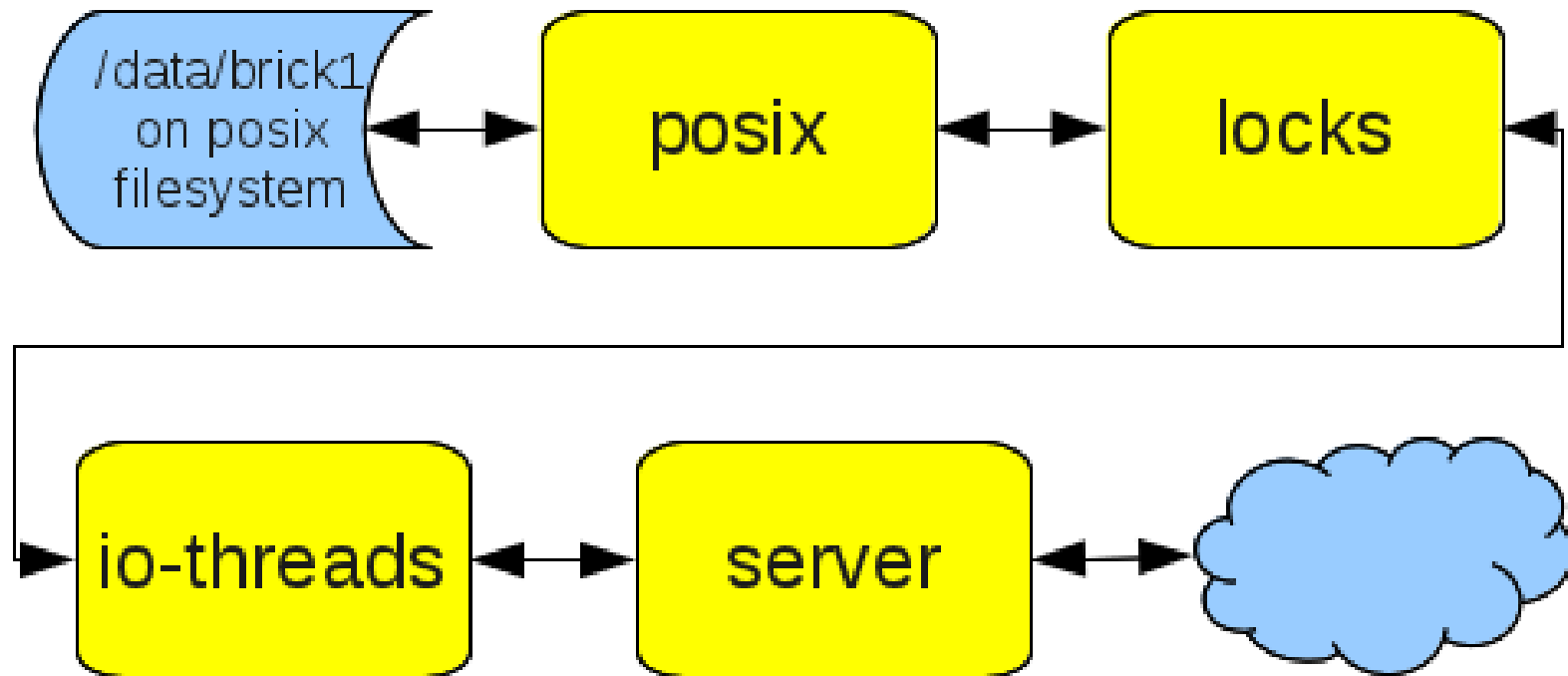
Gluster Mgt Functions: Add, subtract, replicate, heal, recover, etc.



- ◆ Provision huge # of virtual disks
- ◆ Use hash algorithm to assign systematically & in distributed fashion to *virtual* storage locations
- ◆ Virtualization lets you deal flexibly with physical disks
 - ◆ ... add, subtract, deal with different disk performance or capacity parameters, etc.
- ◆ **WE HAVE MET OUR GOAL!**



Translators





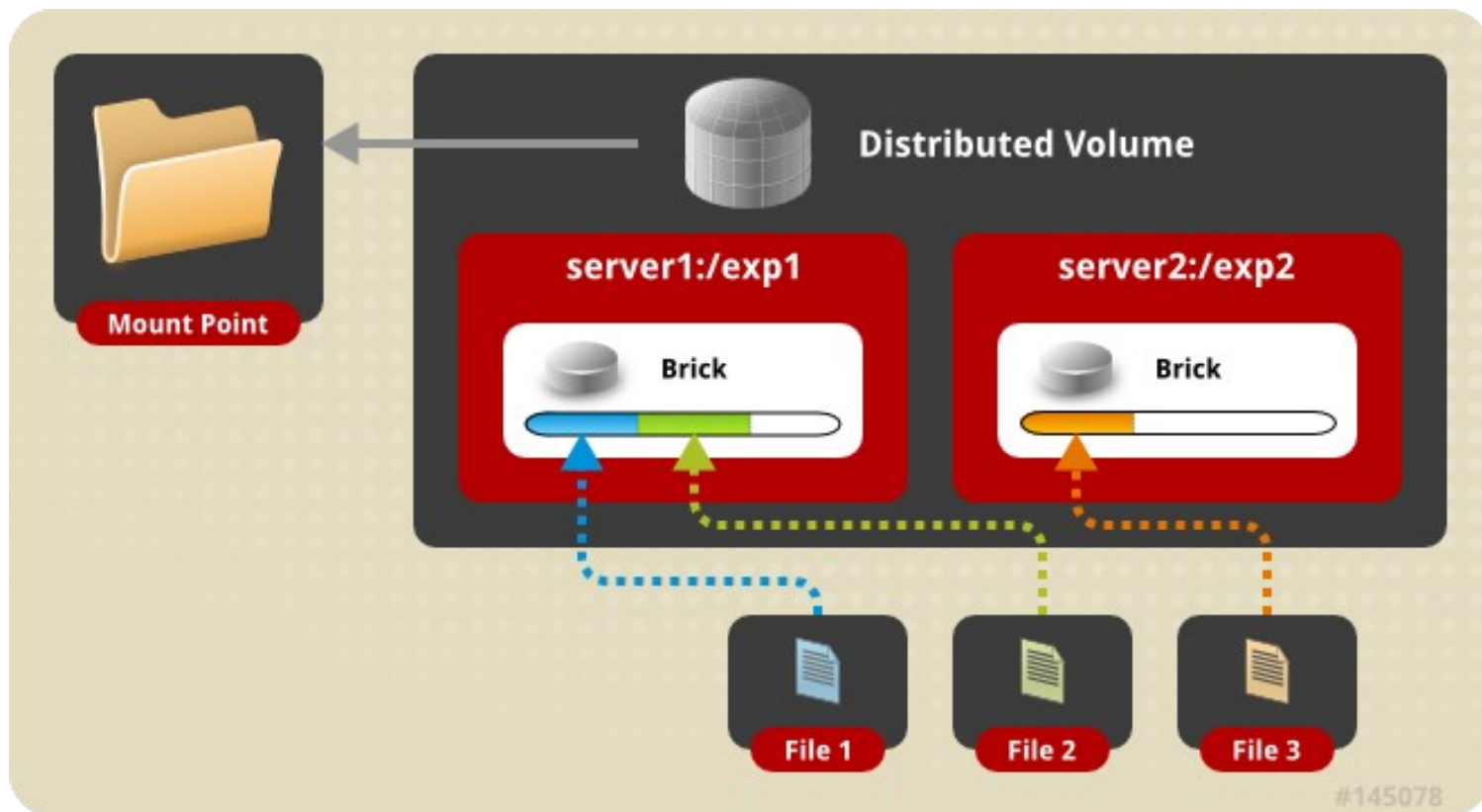
Distribution and Replication

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

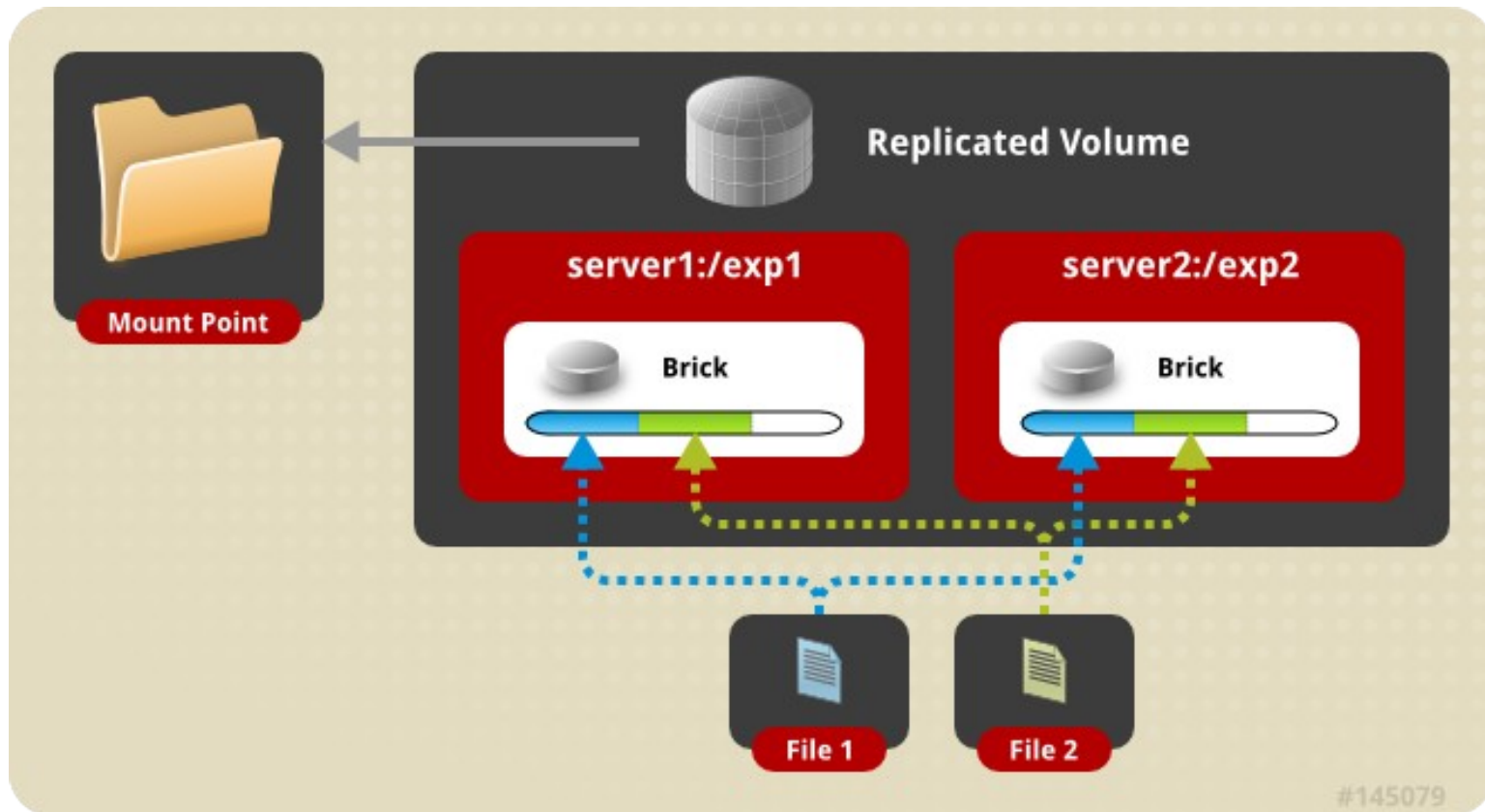
Distributed Volume

- Files “evenly” spread across bricks
- File-level RAID 0
- Server/Disk failure could be catastrophic



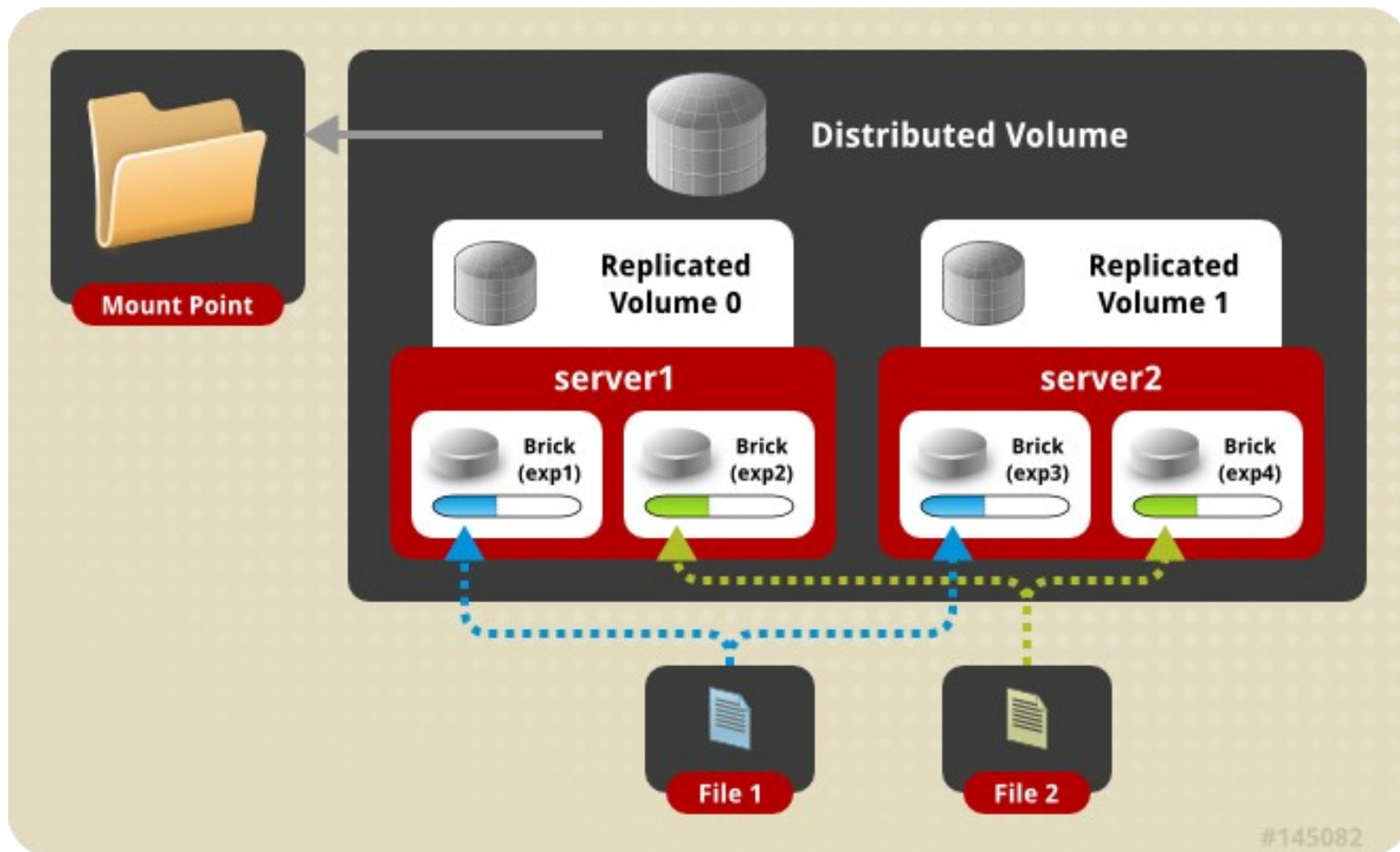
Replicated Volume

- Copies files to multiple bricks
- File-level RAID 1



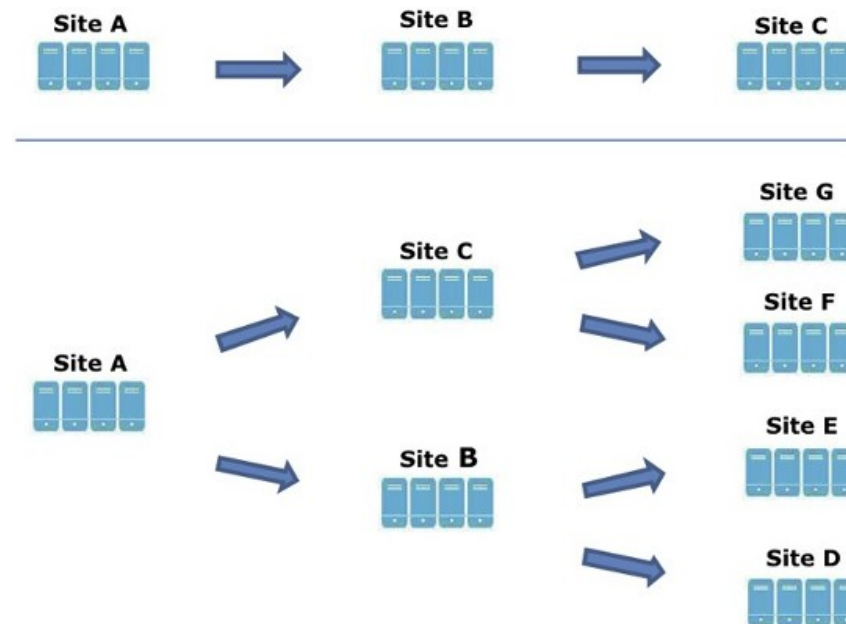
Distributed Replicated Volume

- Distributes files across replicated bricks
- RAID 1 plus improved read performance



Geo Replication

- Asynchronous across LAN, WAN, or Internet
- Master-Slave model -- Cascading possible
- Continuous and incremental
- Time should be synchronized on all master nodes



Replicated Volumes vs Geo-replication

Replicated Volumes	Geo-replication
Mirrors data across clusters	Mirrors data across geographically distributed clusters
Provides high-availability	Ensures backing up of data for disaster recovery
Synchronous replication (each and every file operation is sent across all the bricks)	Asynchronous replication (checks for the changes in files periodically and syncs them on detecting differences)





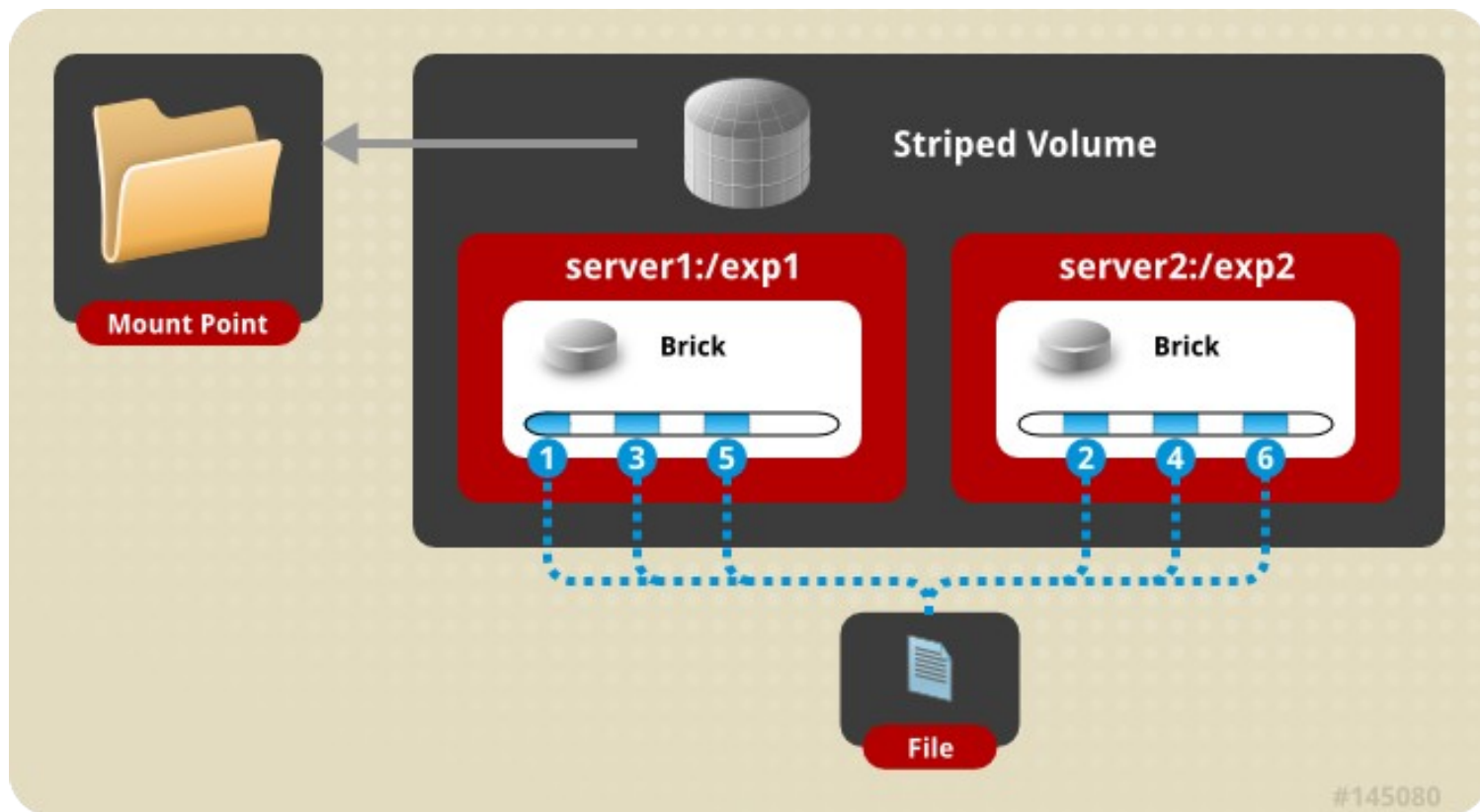
Striping and Beyond

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

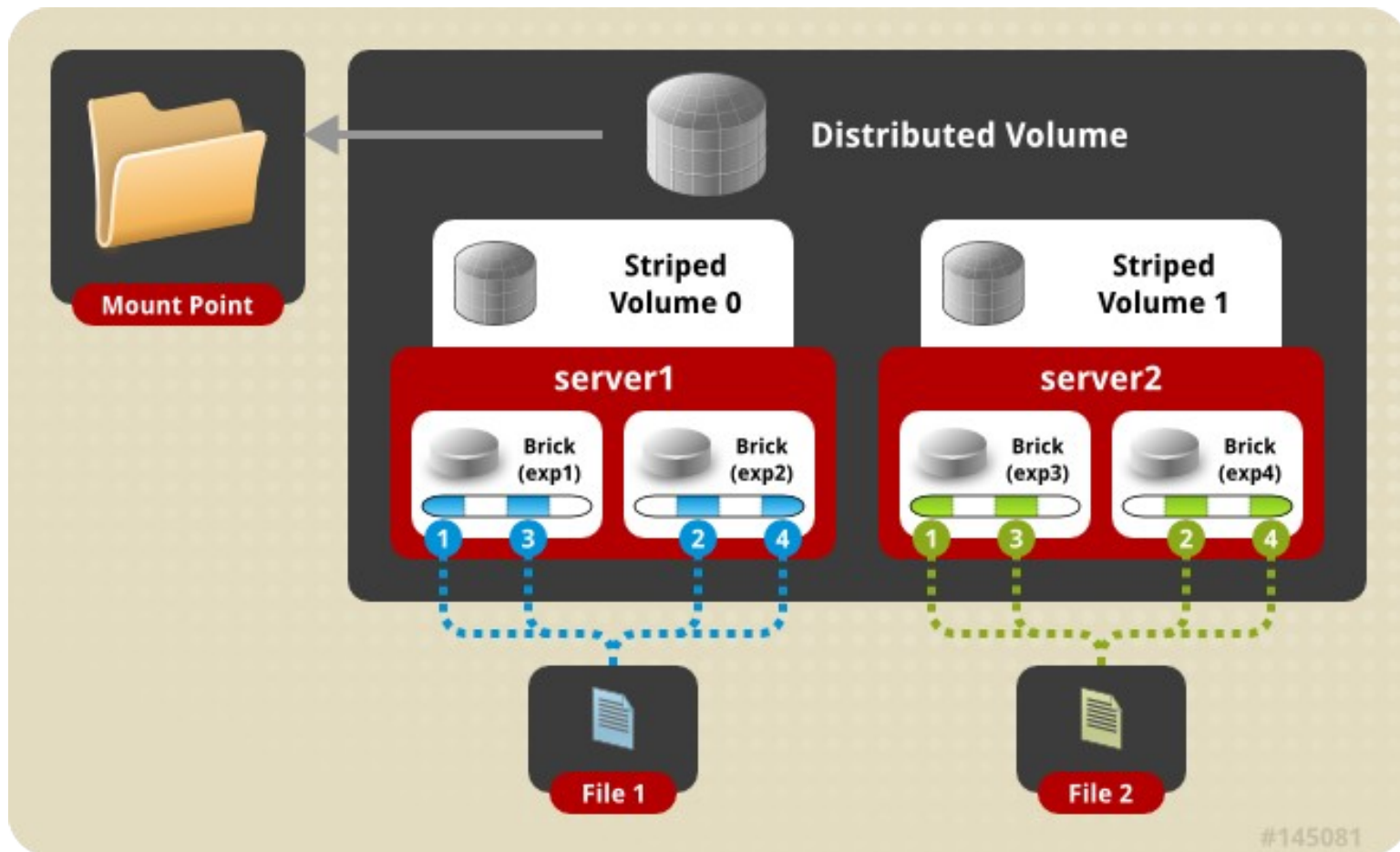
Striped Volumes

- Individual files split among bricks
- Similar to RAID 0
- Single file sizes can exceed size of a single brick.



Distributed Striped Volume

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





Data Access

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

Gluster Native Client (FUSE)

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



NFS

- Standard NFS v3 clients
 - Mount with vers=3 option
- Standard automounter is supported
- Mount to any node, or use a load balancer
- Gluster NFS server includes Network Lock Manager (NLM) to synchronize locks across clients
- Better performance for reading many small files



SMB/CIFS

- GlusterFS volume is first mounted with the Native Client on localhost
- Native mount point is then shared via Samba
- Must be setup on each node you wish to connect to via CIFS





General Administration

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

Preparing a Brick

```
# lvcreate -L 100G -n lv_brick1 vg_server1
# mkfs -t xfs -i size=512 /dev/vg_server1/lv_brick1
# mkdir /brick1
# mount /dev/vg_server1/lv_brick1 /brick1
# echo '/dev/vg_server1/lv_brick1 /brick1 xfs defaults 1 2' >> /etc/fstab
```



Adding Nodes (peers) and Volumes

```
gluster> peer probe server3
gluster> peer status
Number of Peers: 2

Hostname: server2
Uuid: 5e987bda-16dd-43c2-835b-08b7d55e94e5
State: Peer in Cluster (Connected)

Hostname: server3
Uuid: 1e0ca3aa-9ef7-4f66-8f15-cbc348f29ff7
State: Peer in Cluster (Connected)
```

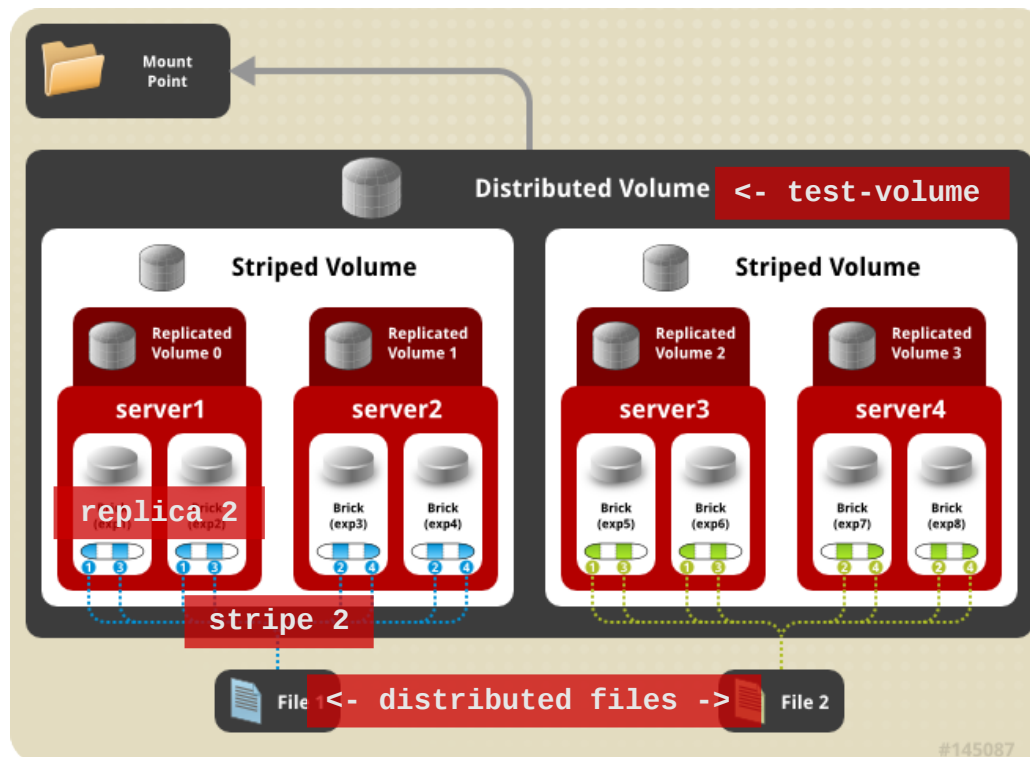
Distributed Volume

```
gluster> volume create my-dist-vol server2:/brick2 server3:/brick3
gluster> volume info my-dist-vol
Volume Name: my-dist-vol
Type: Distribute
Status: Created
Number of Bricks: 2
Transport-type: tcp
Bricks:
Brick1: server2:/brick2
Brick2: server3:/brick3
gluster> volume start my-dist-vol
```



Distributed Striped Replicated Volume

```
gluster> volume create test-volume stripe 2 replica 2 transport tcp \  
server1:/exp1 server1:/exp2 server2:/exp3 server2:/exp4 \  
server3:/exp5 server3:/exp6 server4:/exp7 server4:/exp8  
Multiple bricks of a replicate volume are present on the same server. This setup is not  
optimal.  
Do you still want to continue creating the volume? (y/n) y  
Creation of volume test-volume has been successful. Please start the volume to access  
data.
```



Distributed Striped Replicated Volume

```
gluster> volume create test-volume stripe 2 replica 2 transport tcp \  
server1:/exp1 server2:/exp3 server1:/exp2 server2:/exp4 \  
server3:/exp5 server4:/exp7 server3:/exp6 server4:/exp8  
Creation of volume test-volume has been successful. Please start the volume to access  
data.
```

```
gluster> volume info test-volume  
  
Volume Name: test-volume  
Type: Distributed-Striped-Replicate  
Volume ID: 8f8b8b59-d1a1-42fe-ae05-abe2537d0e2d  
Status: Created  
Number of Bricks: 2 x 2 x 2 = 8  
Transport-type: tcp  
Bricks:  
Brick1: server1:/exp1  
Brick2: server2:/exp3  
Brick3: server1:/exp2  
Brick4: server2:/exp4  
Brick5: server3:/exp5  
Brick6: server4:/exp7  
Brick7: server3:/exp6  
Brick8: server4:/exp8
```



Manipulating Bricks in a Volume

```
gluster> volume add-brick my-dist-vol server4:/brick4
```

```
gluster> volume remove-brick my-dist-vol server2:/brick2 start
gluster> volume remove-brick my-dist-vol server2:/brick2 status
  Node      Rebalanced-files      size      scanned      failures      status
  - - - - -  - - - - - - - - - - -  - - - - -  - - - - -  - - - - -
localhost      16      16777216      52           0      in progress
192.168.1.1     13      16723211      47           0      in progress
gluster> volume remove-brick my-dist-vol server2:/brick2 commit
```

```
gluster> volume rebalance my-dist-vol fix-layout start

gluster> volume rebalance my-dist-vol start
gluster> volume rebalance my-dist-vol status
  Node      Rebalanced-files      size      scanned      failures      status
  - - - - -  - - - - - - - - - - -  - - - - -  - - - - -  - - - - -
localhost      112      15674      170           0      completed
10.16.156.72   140      3423      321           2      completed
```



Migrating Data / Replacing Bricks

```
gluster> volume replace-brick my-dist-vol server3:/brick3 server5:/brick5 start
gluster> volume replace-brick my-dist-vol server3:/brick3 server5:/brick5 status
Current File = /usr/src/linux-headers-2.6.31-14/block/Makefile
Number of files migrated = 10567
Migration complete
gluster> volume replace-brick my-dist-vol server3:/brick3 server5:/brick5 commit
```



Volume Options

Auth

```
gluster> volume set my-dist-vol auth.allow 192.168.1.*  
gluster> volume set my-dist-vol auth.reject 10.*
```

NFS

```
gluster> volume set my-dist-vol nfs.volume-access read-only  
gluster> volume set my-dist-vol nfs.disable on
```

Other advanced options

```
gluster> volume set my-dist-vol features.read-only on  
gluster> volume set my-dist-vol performance.cache-size 67108864
```



Volume Top Command

```
gluster> volume top my-dist-vol read brick server3:/brick3 list-cnt 3
Brick:  server:/export/dir1
        =====Read file stats=====

read          filename
call count

116           /clients/client0/~dmtmp/SEED/LARGE.FIL
64            /clients/client0/~dmtmp/SEED/MEDIUM.FIL
54            /clients/client2/~dmtmp/SEED/LARGE.FIL
```

- Many top commands are available for analysis of files, directories, and bricks
- Read and write performance test commands available
 - Perform active dd tests and measure throughput



Volume Profiling

```
gluster> volume profile my-dist-vol start
```

```
gluster> volume profile my-dist-vol info
```

```
Brick: Test:/export/2
```

```
Cumulative Stats:
```

Block Size:	1b+	32b+	64b+
Read:	0	0	0
Write:	908	28	8

```
...
```

%-latency	Avg-latency	Min-Latency	Max-Latency	calls	Fop
4.82	1132.28	21.00	800970.00	4575	WRITE
5.70	156.47	9.00	665085.00	39163	REaddirP
11.35	315.02	9.00	1433947.00	38698	LOOKUP
11.88	1729.34	21.00	2569638.00	7382	FXATTROP
47.35	104235.02	2485.00	7789367.00	488	FSYNC

```
-----
```

```
Duration      : 335
```

```
BytesRead     : 94505058
```

```
BytesWritten  : 195571980
```



Geo-Replication

Remote GlusterFS Volume

```
gluster> volume geo-replication my-dist-vol slavehost1:my-dist-repl start
Starting geo-replication session between my-dist-vol & slavehost1:my-dist-repl has been
successful
gluster> volume geo-replication my-dist-vol status
MASTER                SLAVE                STATUS
-----
my-dist-vol           gluster://slavehost1:my-dist-repl           OK
```

Remote SSH

```
# ssh-keygen -f /var/lib/glusterd/geo-replication/secret.pem
# ssh-copy-id -i /var/lib/glusterd/geo-replication/secret.pem repluser@slavehost1

gluster> volume geo-replication my-dist-vol repluser@slavehost1:/repl_dir start
Starting geo-replication session between my-dist-vol & slavehost1:/repl_dir has been
successful
gluster> volume geo-replication my-dist-vol status
MASTER                SLAVE                STATUS
-----
my-dist-vol           ssh://repluser@slavehost1:/repl_dir         OK
```

```
gluster> volume info my-dist-vol
...
Options Reconfigured:
geo-replication.indexing: on
```





Use Cases

Demystifying Gluster

RHS and GlusterFS for the SysAdmin

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



CIC Electronic Signature Solutions

Hybrid Cloud: Electronic Signature Solutions



- Reduced time-to-market for new products
- Meeting all client SLAs
- Accelerating move to the cloud

- Challenge
 - Must leverage economics of the cloud
 - Storage performance in the cloud too slow
 - Need to meet demanding client SLA's
- Solution
 - Red Hat Storage Software Appliance
 - Amazon EC2 and Elastic Block Storage (EBS)
- Benefits
 - Faster development and delivery of new products
 - SLA's met with headroom to spare
 - Accelerated cloud migration
 - Scale-out for rapid and simple expansion
 - Data is highly available for 24/7 client access



Pandora Internet Radio

Private Cloud: Media Serving



- 1.2 PB of audio served per week
- 13 million files
- Over 50 GB/sec peak traffic

- Challenge
 - Explosive user & title growth
 - As many as 12 file formats for each song
 - 'Hot' content and long tail
- Solution
 - Three data centers, each with a six-node GlusterFS cluster
 - Replication for high availability
 - 250+ TB total capacity
- Benefits
 - Easily scale capacity
 - Centralized management; one administrator to manage day-to-day operations
 - No changes to application
 - Higher reliability



Brightcove

Private Cloud: Media Serving



brightcove™

- Over 1 PB currently in Gluster
- Separate 4 PB project in the works

- Challenge
 - Explosive customer & title growth
 - Massive video in multiple locations
 - Costs rising, esp. with HD formats
- Solution
 - Complete scale-out based on commodity DAS/JBOD and GlusterFS
 - Replication for high availability
 - 1PB total capacity
- Benefits
 - Easily scale capacity
 - Centralized management; one administrator to manage day-to-day operations
 - Higher reliability
 - Path to multi-site



Pattern Energy

High Performance Computing for Weather Prediction



- Rapid and advance weather predictions
- Maximizing energy assets
- Cost savings and avoidance

- Challenge
 - Need to deliver rapid advance weather predictions
 - Identify wind and solar abundance in advance
 - More effectively perform preventative maintenance and repair
- Solution
 - 32 HP compute nodes
 - Red Hat SSA for high throughput and availability
 - 20TB+ total capacity
- Benefits
 - Predicts solar and wind patterns 3 to 5 days in advance
 - Maximize energy production and repair times
 - Avoid costs of outsourcing weather predictions
 - Solution has paid for itself many times over





Thank You!

- dustin@redhat.com
- storage-sales@redhat.com
- RHS:
www.redhat.com/storage/
- GlusterFS:
www.gluster.org
- TAM:
access.redhat.com/support/offerings/tam/



 [@Glusterorg](https://twitter.com/Glusterorg)

 [@RedHatStorage](https://twitter.com/RedHatStorage)



 [Gluster](#)

 [Red Hat Storage](#)

Demystifying Gluster

RHS and GlusterFS for the SysAdmin