

# GlusterFS

Distributed Replicated Parallel File System



Martin Alfke <[martin.alfke@buero20.org](mailto:martin.alfke@buero20.org)>

# Agenda

- General Information on GlusterFS
- Architecture Overview
- GlusterFS Translators
- GlusterFS Configuration



# General Information

# General Information I

## File System Solutions

- \* Shared Disk File System
  - San or Block Access
  - Mostly used in HA setup (e.g. DRBD)
  
- \* Distributed File System
  - Network File System
    - NFS
    - SMB/CIFS
    - 9P
  
- \* Distributed replicated File System
  - Replication
  - HA and offline operation
    - Coda
    - MS DFS
    - MooseFS

# General Information II

## File System Solutions

- \* Distributed parallel File System
  - Setup across multiple servers
  - HPC
- \* Distributed replicated parallel File System
  - HPC and HA
    - Cosmos
    - MogileFS
    - GPFS (IBM)
    - GFS (Google)
    - Hadoop
    - GlusterFS

# Customer Platform

## Shared Storage Issues

- \* Bad performance of NFS kernel stack
- \* Limitations of concurrent NFS accesses
- \* Customer data already on NFS system

## Environment and Requirements

- \* Debian GNU/Linux Version 4 (etch) – 3 years old
- \* Most D f-t p FS need complex data migration
- \* Solution has to be expandable

# Why GlusterFS ?

## Decision basics

- \* Possibility to run NFS and GlusterFS in parallel
- \* No data migration necessary
- \* Easy setup
- \* Extendable (e.g. new storage nodes)
- \* min. Kernel 2.6.3x --> optimization for FUSE context switches !



# Architecture Overview



# GlusterFS Basics

## Hardware

- \* Any x86 Hardware
- \* Direct attached storage, RAID
- \* FC, Infiniband or iSCSI SAN
- \* Gigabit or 10 Gigabit network or Infiniband

## OS

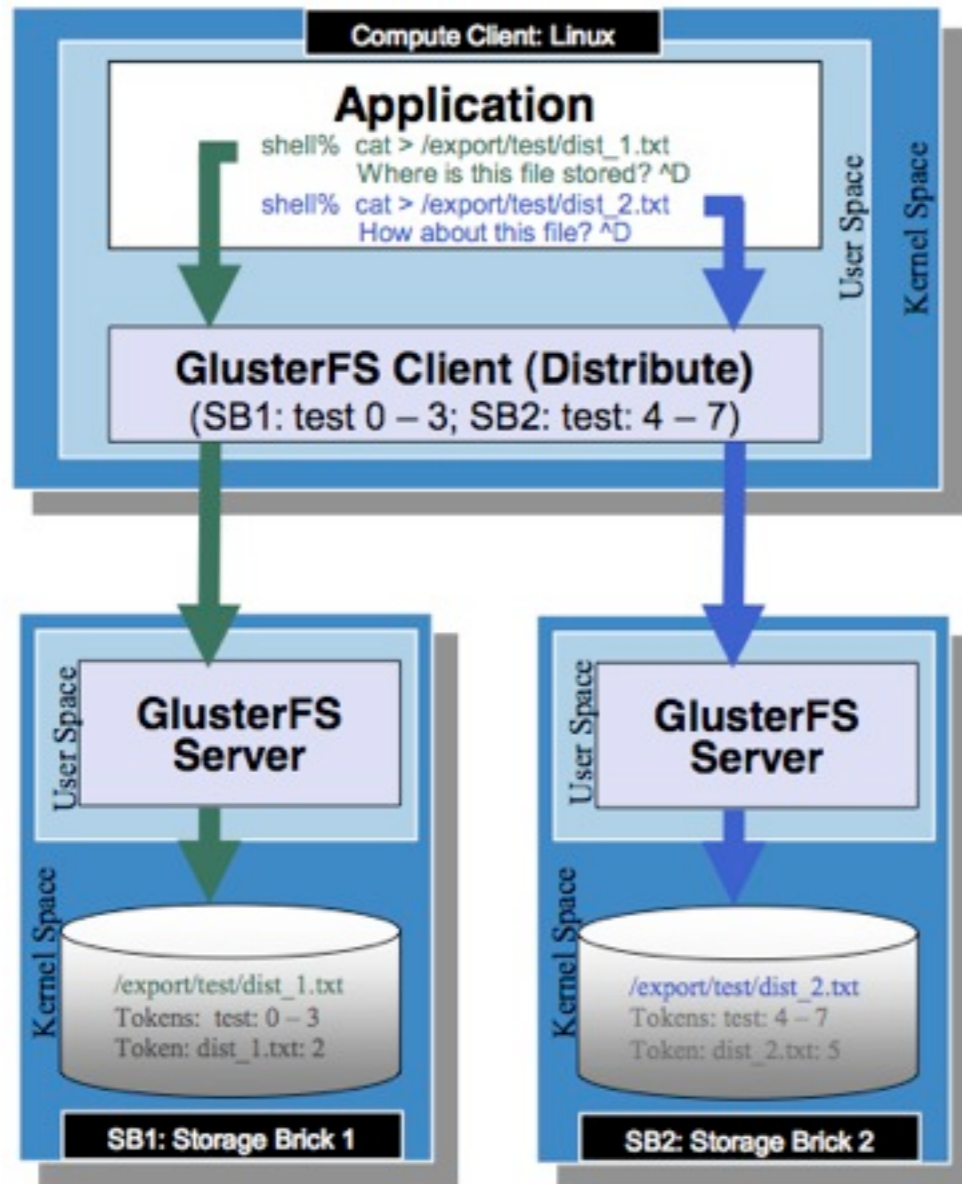
- \* Linux, Solaris, OpenSolaris, OS X, FreeBSD
- \* ext3 or ext4 Filesystem (tested)
- \* Other POSIX compliant filesystems should also work

## Architecture

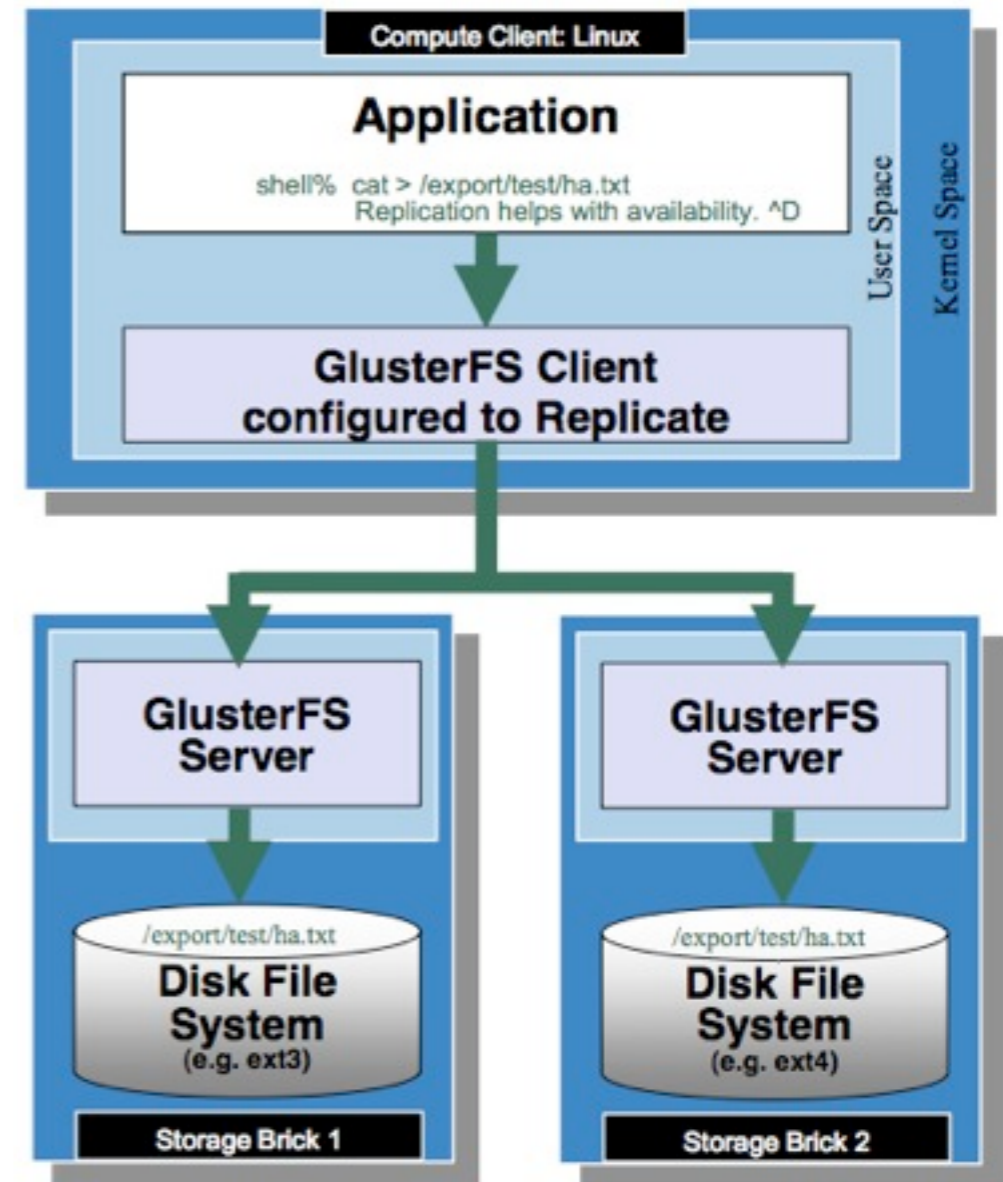
- \* No Meta-Data Server (fully distributed architecture - Elastic Hash)
- \* Replication (RAID 1)
- \* Distribution (RAID 0)
- \* FUSE (Standard)
- \* NFS (unfs3 - depreciated)
- \* SMB/CIFS
- \* DAV

# GlusterFS Architecture Overview I

## Distribution



## Replication



Images Copyright by Gluster.Inc.

# GlusterFS Architecture Overview II

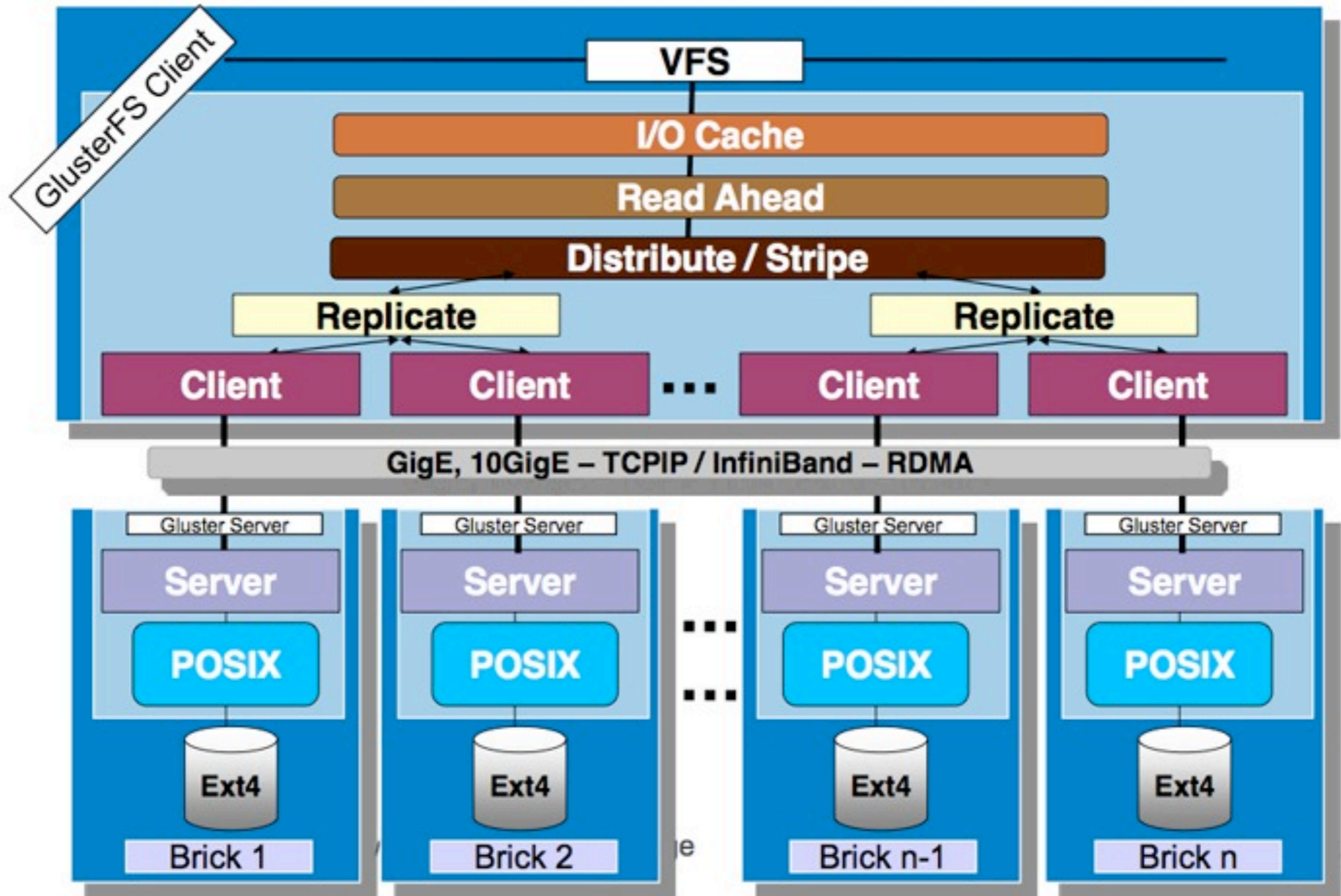


Image Copyright by Gluster.Inc.

yes, we are open!  
**büro 2.0**  
Open Source Bürogemeinschaft & Partnernetzwerk

# GlusterFS Architecture Overview III

## Recommended Server Setup

- \* GlusterFS daemons
- \* GlusterFS Server
- \* Network
- \* \*nix distribution
- \* FileSystem (ext3/ext4 or POSIX)
- \* Volume Manager (LVM2 only)
- \* HW RAID
- \* Disks

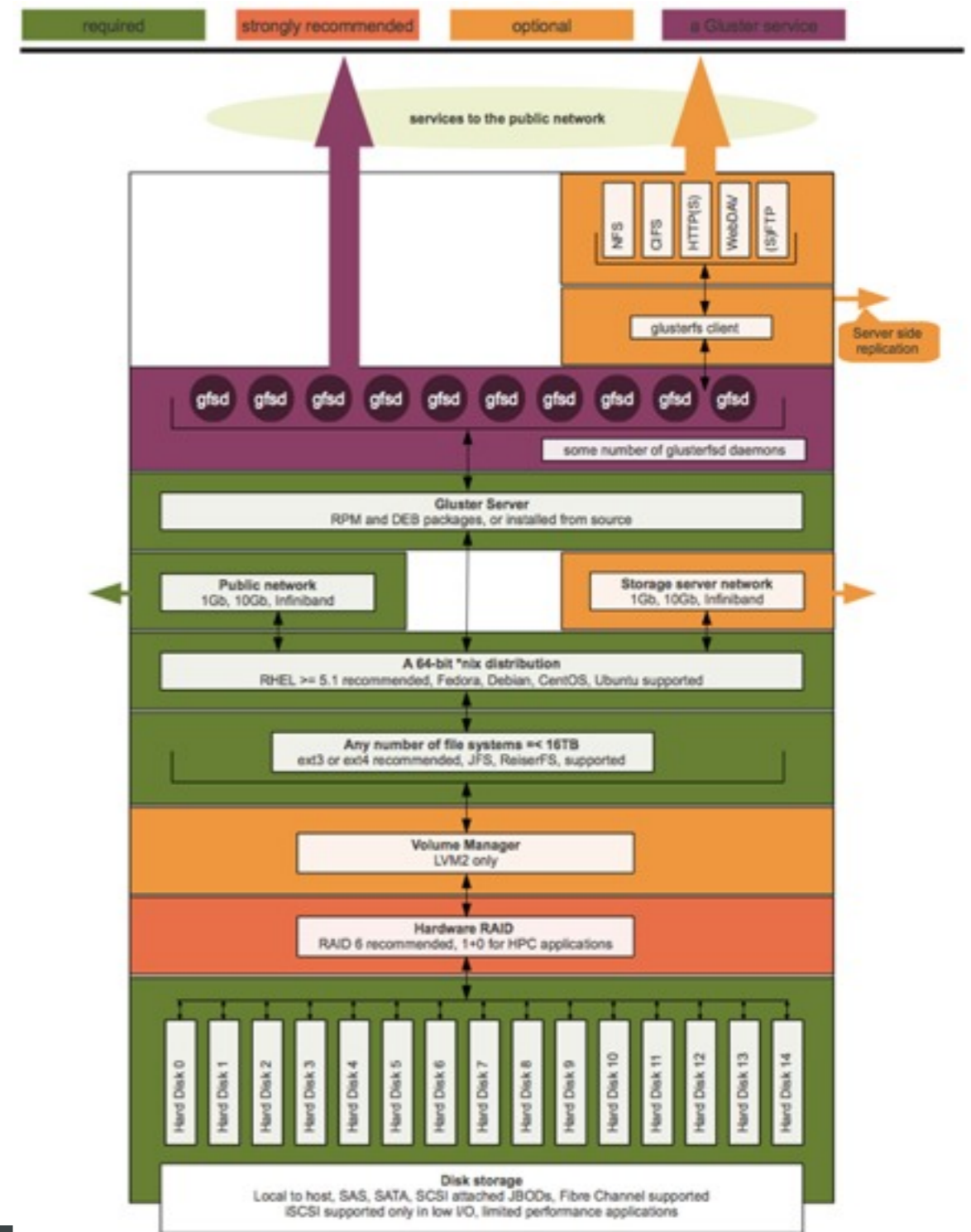


Image Copyright by Gluster, Inc.

yes, we are open!  
**büro 2.0**  
 Open Source Bürogemeinschaft & Partnernetzwerk



# GlusterFS Translators

# GlusterFS Translators

## GlusterFS modular extension

### \* storage

- posix – for underlying filesystem
- bdb – database storage system

### \* protocol

- server – required for server config (e.g. Infiniband or TCP)
- client – required for client config (e.g. Infiniband or TCP)

### \* cluster

- distribute – storage spread to multiple servers
- replicate – mirror content between servers
- stripe – striping between multiple servers
- unify – obsolete, use cluster/distribute
- nufa – HPC - higher preference for a local volume

### \* encryption

- rot-13 – sample code for encryption

# GlusterFS Translators

## \* performance

- readahead – for sequential read performance
- writebehind – aggregate written blocks
- io-threads – number of threads
- io-cache – read cache block size
- quickread – fetch small files in a single call
- stat-prefetch – prefetch stats in readdir
- symlink-cache – undocumented

# GlusterFS Translators


## \* features

- locks – POSIX locking
- filter – filtering on user id or group id
- access-control – undocumented
- path-converter – internal path converter
- quota – don't grow beyond disk space
- read-only – included in feature/filter
- trash – recycle bin (use on server)

## \* debug

- trace – trace glusterfs functions and system calls
- io-stats – collect performance data
- error-gen – undocumented





# GlusterFS Configuration

# GlusterFS Configuration Location

## Configuration files (path and names)

- \* Server
  - /etc/glusterfs/glusterfsd.vol
- \* Client
  - /etc/glusterfs/glusterfs.vol

## Possible location of config files for clients

- \* Local on disk
- \* Remote on glusterfs-Server

# GlusterFS Configuration Example

## 2 Servers - n Clients - replication

### \* Server 1 + 2

volume **posix**

type storage/posix  
option directory /data/export

end-volume

volume **locks**

type features/locks  
subvolumes **posix**

end-volume

volume **brick**

type performance/io-threads  
option thread-count 8  
subvolumes **locks**

end-volume

volume server

type protocol/server  
option transport-type tcp  
option auth.addr.**brick**.allow 192.168.0.102  
subvolumes **brick**

end-volume

# GlusterFS Configuration Example

## 2 Servers - n Clients - replication

### \* Clients

volume **remote1**

```
type protocol/client
option transport-type tcp
option remote-host server1.example.com
option remote-subvolume brick
```

end-volume

volume **remote2**

```
type protocol/client
option transport-type tcp
option remote-host server2.example.com
option remote-subvolume brick
```

end-volume

volume **replicate**

```
type cluster/replicate
subvolumes remote1 remote2
```

end-volume

volume **writebehind**

```
type performance/write-behind
option window-size 1MB
subvolumes replicate
```

end-volume

volume cache

```
type performance/io-cache
option cache-size 512MB
subvolumes writebehind
```

end-volume

# GlusterFS Configuration @Customer

## Servers - both

volume **posix**

type storage/posix  
option directory /data

End-volume

volume **iostats**

type debug/io-stats  
subvolumes **posix**

end-volume

volume **locks**

type features/locks  
subvolumes **iostats**

end-volume

volume **iothreads**

type performance/io-threads  
option thread-count 16  
subvolumes **locks**

end-volume

volume **writebehind**

type performance/write-behind  
option cache-size 64MB  
option flush-behind off  
subvolumes **iothreads**

end-volume

# GlusterFS Configuration @Customer

## Servers - both - continued

volume **brick**

type performance/io-cache  
option cache-size 2048MB  
option cache-timeout 5  
subvolumes **writebehind**

End-volume

volume **server-tcp**

type protocol/server  
option transport-type tcp  
option auth.addr.**brick**.allow \*  
option transport.socket.listen-port 6996  
option transport.socket.nodelay on  
subvolumes **brick**

end-volume

# GlusterFS Configuration @Customer

## Clients

volume **remote1**

```
type protocol/client
option transport-type tcp
option remote-host <IP server1>
option transport.socket.nodelay on
option remote-port 6996
option remote-subvolume brick
```

end-volume

volume **remote2**

```
type protocol/client
option transport-type tcp
option remote-host <IP server2>
option remote-subvolume brick
```

end-volume

volume **distribute**

```
type cluster/distribute
subvolumes remote1 remote2
```

end-volume

volume **iothreads**

```
type performance/io-threads
option thread-count 16
subvolumes distribute
```

end-volume

volume **writebehind**

```
type performance/write-behind
option cache-size 32MB
subvolumes iothreads
```

end-volume

volume **cache**

```
type performance/io-cache
option cache-size 256MB
option cache-timeout 10
subvolumes writebehind
```

end-volume



# GlusterFS Usage



# GlusterFS Usage

## Mounting GlusterFS on Clients

- \* manual mount

- \* server-side configuration:

- ```
mount -t glusterfs server_IP /mnt/glusterfs
```

- \* local (client) configuration

- ```
mount -t glusterfs /etc/glusterfs/glusterfs.vol /mnt/glusterfs
```

- \* automatic mount (fstab)

- \* server-side configuration:

- ```
server_IP /mnt/glusterfs glusterfs defaults,_netdev 0 0
```

- \* local(client) configuration:

- ```
/etc/glusterfs/glusterfs.vol /mnt/glusterfs glusterfs defaults,_netdev 0 0
```

# GlusterFS understood

## \*Links:

- <http://www.gluster.org/>
- [http://www.howtoforge.com/trip\\_search](http://www.howtoforge.com/trip_search)
- <http://www.gluster.org/docs/index.php/GlusterFS>
- [http://www.gluster.org/docs/index.php/GlusterFS\\_Volume\\_Specification](http://www.gluster.org/docs/index.php/GlusterFS_Volume_Specification)
- [http://www.gluster.org/docs/index.php/GlusterFS\\_Translators](http://www.gluster.org/docs/index.php/GlusterFS_Translators)
- [http://www.gluster.com/community/documentation/index.php/Translators\\_options](http://www.gluster.com/community/documentation/index.php/Translators_options)
- [http://www.gluster.com/community/documentation/index.php/GlusterFS\\_Technical\\_FAQ](http://www.gluster.com/community/documentation/index.php/GlusterFS_Technical_FAQ)

## \*Further information and examples:

- `apt-get install glusterfs-examples`

# GlusterFS

## Credits:

- \* Gluster Inc.
- \* Christian Gischewski <info@cgihome.de>
- \* Tobias Geiger <tobias.geiger@1und1.de>

Questions ?