

GlusterFS 分布式文件系统

2014/12/27

关于我

- ◆ 刘爱贵，中科院博士
- ◆ 研究方向：分布式存储
- ◆ 专注于存储技术的研究与开发
分布式存储资深理论研究与实践者
GlusterFS技术专家
- ◆ 博客：<http://blog.csdn.net/liuaogui>
- ◆ 微博：<http://weibo.com/liuag>
- ◆ Email：aigui.liu@gmail.com
- ◆ QQ：9187434



报告提纲

- ◆ GlusterFS 简介
- ◆ GlusterFS 原理剖析
- ◆ GlusterFS 应用场景
- ◆ GlusterFS 开放问题

(一)

GlusterFS 简介

GlusterFS 是什么？

GlusterFS is a unified, poly-protocol, scale-out filesystem serving many PBs of data.

- 用户空间设计，全局统一命名空间，堆栈式架构
- scale-out在线扩展，数百节点，数PB数据
- 一切皆文件，block+object+file



GlusterFS 标 签



GlusterFS 起 源

❖ How it all started

- Backgrounds in high performance, clustered computing
- Working at Lawrence Livermore National Labs
 - AB Periasamy & Hitesh Chellani design “Thunder”
 - One of the worlds fastest super computers
 - On Intel commodity hardware
 - Solved filesystem scalability and performance limitations
- Large customer in oil & gas persuaded them to focus on storage
- Gluster founded by Hitesh & AB to bring technology to market



Thunder

❖ Result: award winning technology



CRN
emerging vendor



GlusterFS 发展简史

GlusterFS = GNU Cluster File System



2011

2012

2013

- GlusterFS v3.3
- 对象存储，HDFS兼容
- 主动自修复
- 细粒度锁
- 复制优化

- GlusterFS v3.2.x
- 远程复制，监控，Quota
- Redhat 1.36亿\$收购Gluster

- GlusterFS v3.1
- 弹性云能力

2006-
2008-
2009-
2003

- GlusterFS v1.0 – v3.0
- 分布式文件系统，自修复
- 同步副本，条带，弹性哈希算法

创始人：Anand Babu Periasamy
目标：代替开源Lustre和商业产品GPFS

Story of Gluster



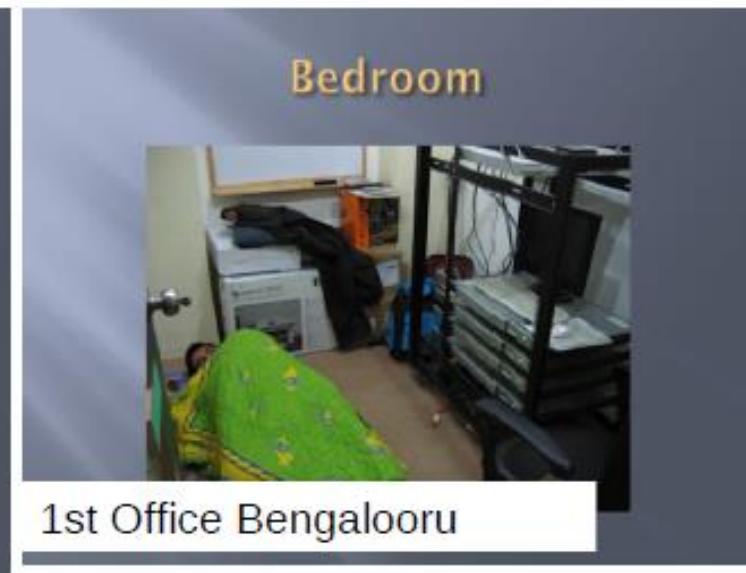
1st meeting room



1st Office US



1st Office Bengalooru



1st Office Bengalooru

GlusterFS 特点

- 最大特点是**简单**：架构、使用、管理
- 完全对称式架构，无元数据服务器
- 全UserSpace设计，Stack式扩展 (源自Hurd)
- Scale-out，高可用(无单点故障)
- 支持多种访问协议，支持RDMA

简单的分布式存储

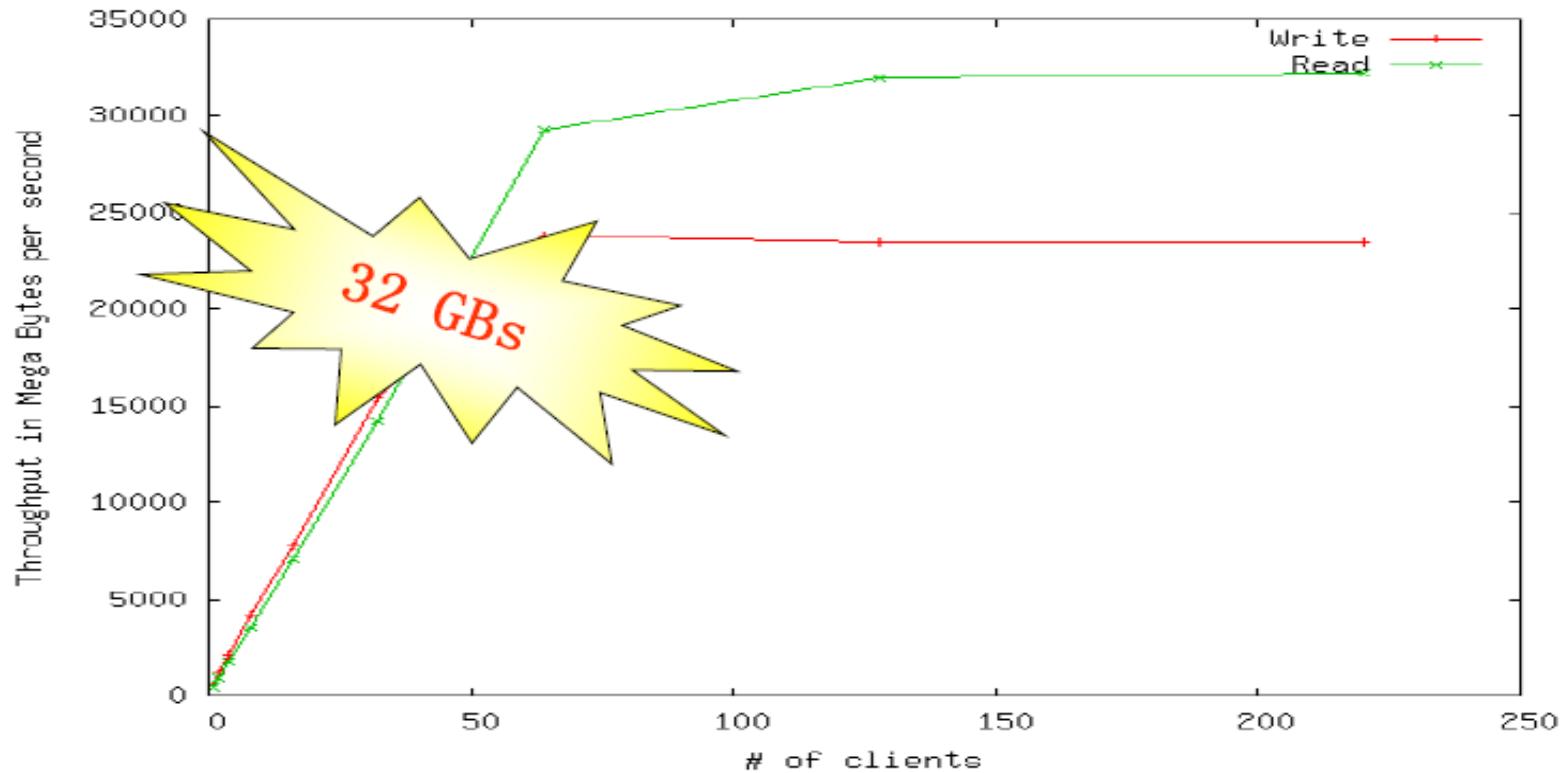
- 最简单配置和管理的分布式文件系统
- 使用gluster单一命令行工具管理
 - probe peer, create volume, start volume, mount
- 极其简便的系统管理
 - 集群关系，进程管理，端口映射，动态配置变更
 - online集群节点扩展/收缩
 - online集群参数变更
 - 系统升级

选择Gluster的理由

- 极其简便的管理和维护
- Block/File/Object统一存储
- 模块化扩展架构
- 支持IP/RDMA传输协议
- Data locality
- Compute/Virtualization透明存储系统

GlusterFS 高性能记录

Benchmark – 64 bricks with ib-verbs transport



Servers: 64 bricks clustered storage servers / bricks

Clients: Cluster of 220 servers

Interconnect: 10 Gbps InfiniBand interconnect; ib0verbs transport protocol

Method: 220 clients pounding the storage servers with multiple dd (disk-dump) instances

Size: Each clients reads / writes a 1 GB file with 1MB block size.

GlusterFS 社区部署



300,000+ 下载: 35,000/月, 每年增长300%+
1000+ 部署案例(45个国家)
2000+ 注册用户

GlusterFS 商业部署



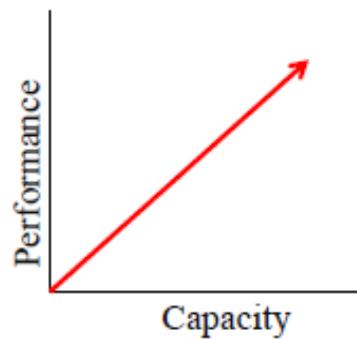
GlusterFS 中国用户

Confidential

(二)

GlusterFS 原理剖析

GlusterFS 架构设计目标



◆ Elasticity

- Flexibility adapt to growth/reduction
- Add, delete volumes & users
- Without disruption

◆ Scale linearly

- Multiple dimensions
 - Performance
 - Capacity
- Aggregated resources

◆ Eliminate metadata

- Improve file access speed

◆ Simplicity

- Ease of management
- No complex Kernel patches
- Run in user space

GlusterFS 架构特点

软件定义

无中心架构

全局命名空间

高性能

用户空间实现

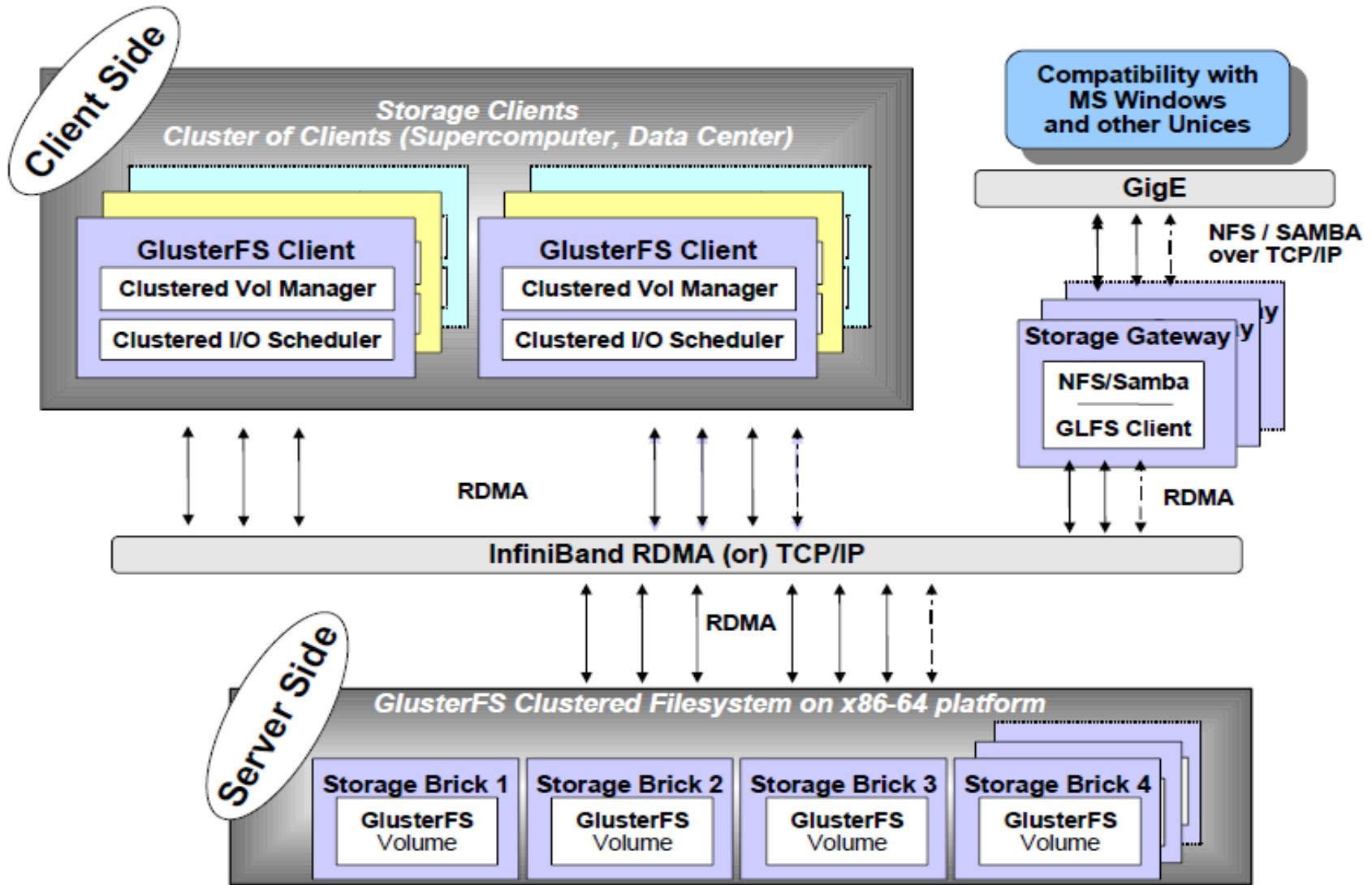
堆栈式设计

弹性横向扩展

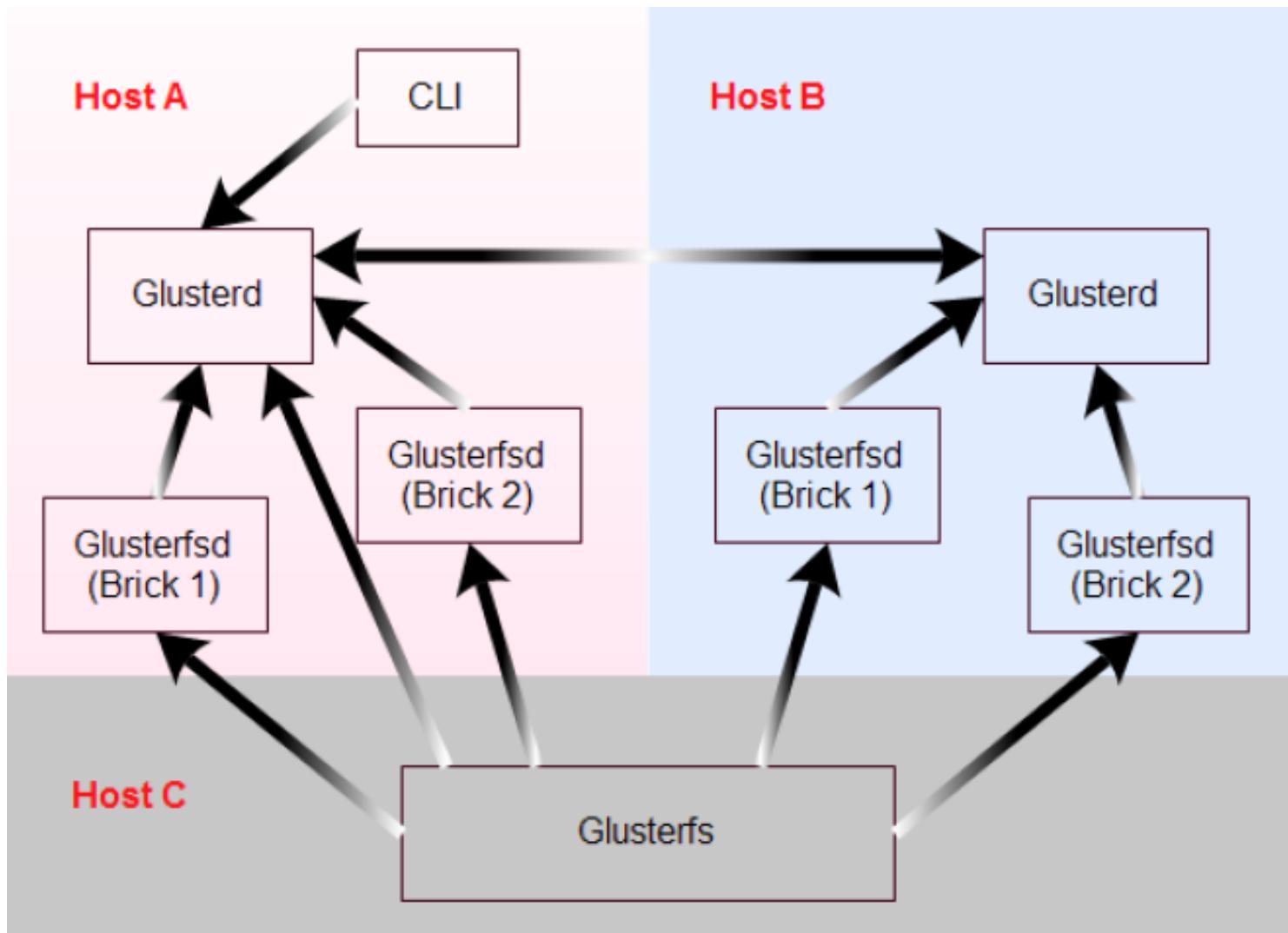
高速网络通信

数据自动修复

GlusterFS 总体架构

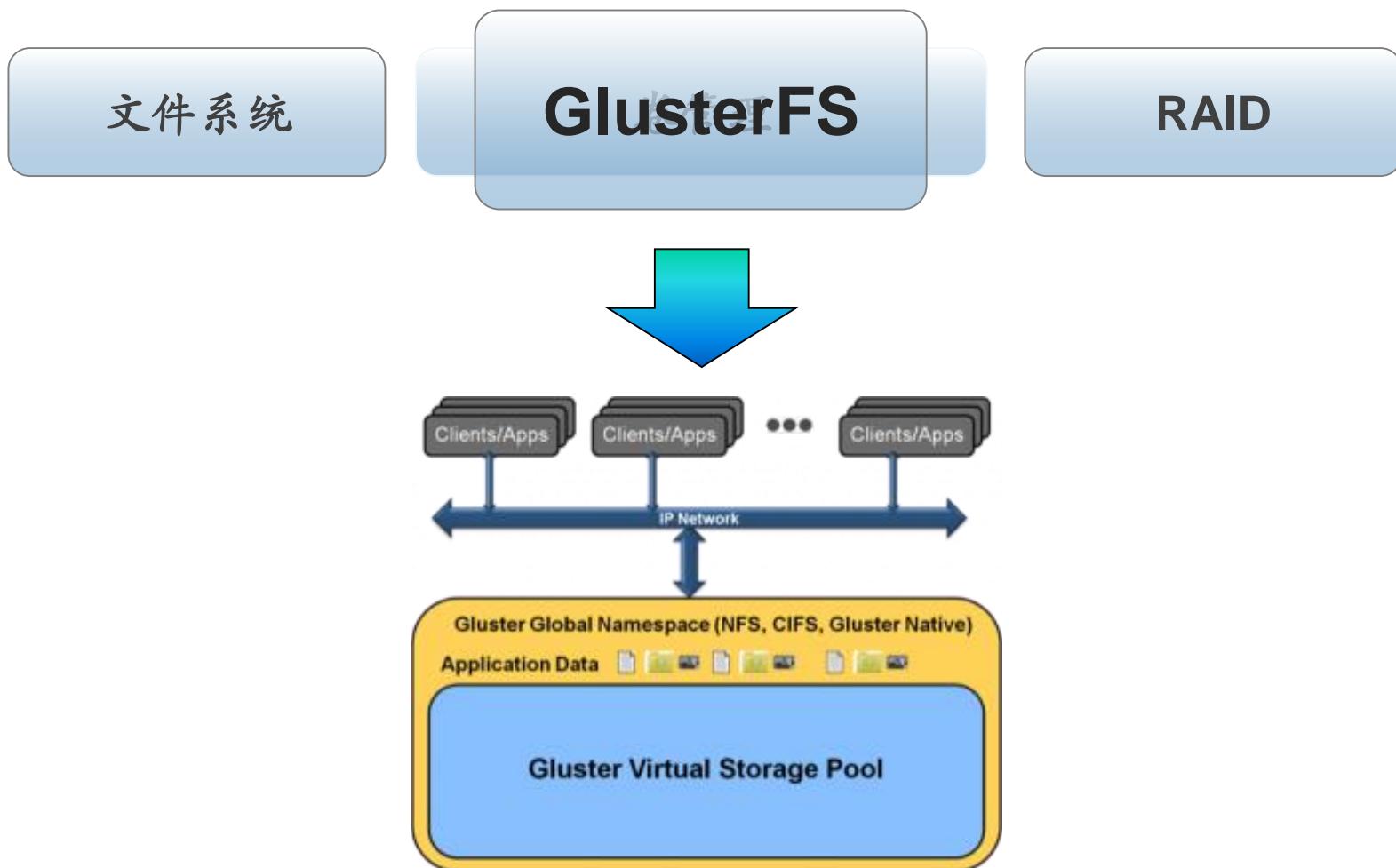


GlusterFS 服务进程

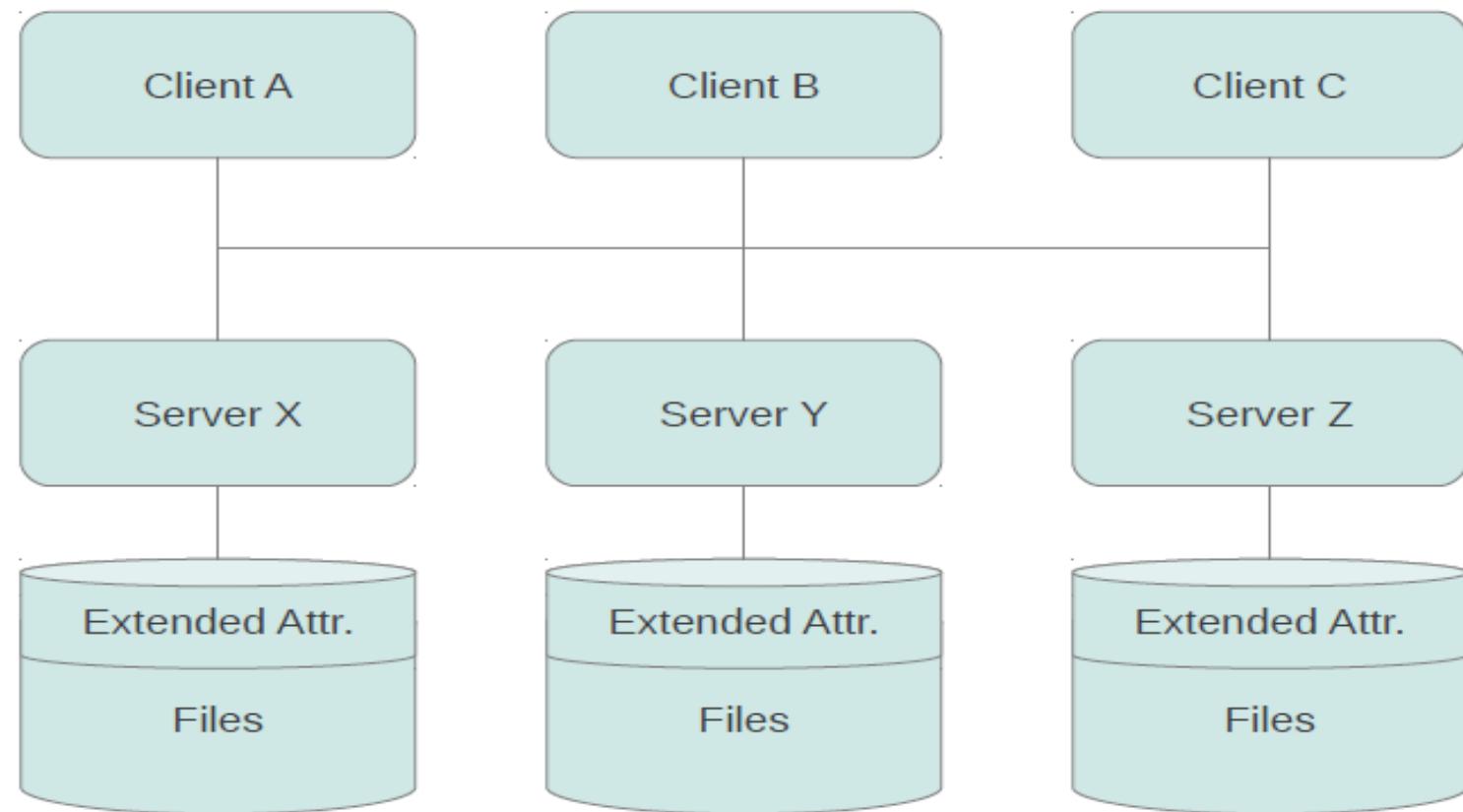


全局统一命名空间

通过分布式文件系统将物理分散的存储资源虚拟化成统一的存储池



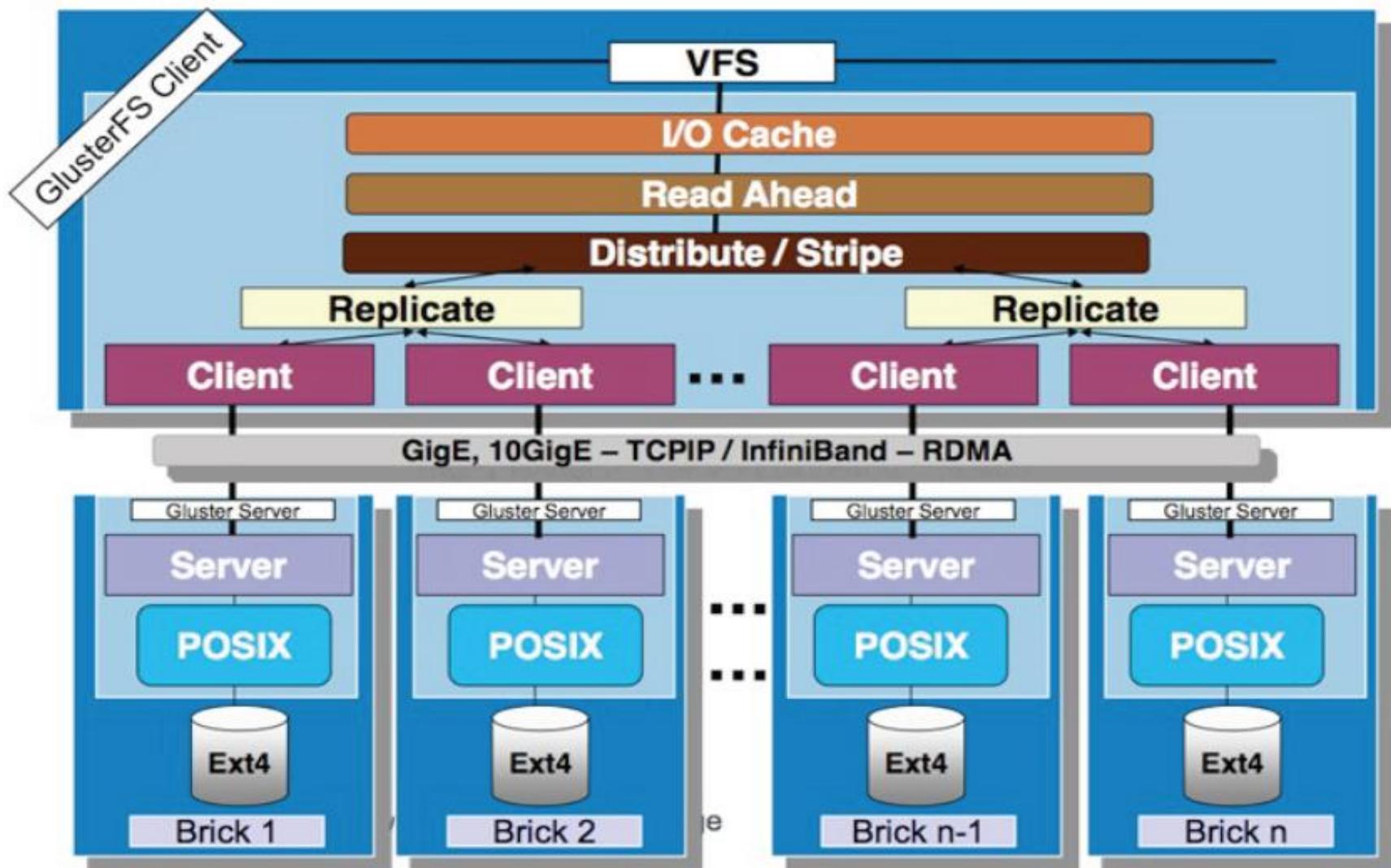
无集中元数据服务



Gluster 集群管理

- ◆ 集群管理模型
 - ◆ 全对称(如Corosync)
 - ◆ 缺点：规模小[<100] 优点：无需配置
 - ◆ 单独的控制集群(如Zookeeper)
 - ◆ 优点：规模大[>1000] 缺点：需要配置控制集群
- ◆ Gluster采用全对称式集群管理
 - ◆ Gluster节点之间的配置信息是完全一致的
 - ◆ 每个配置信息改动操作需要在多节点同步
 - ◆ 优化同步算法，可支持500+节点

GlusterFS 堆栈式软件架构



GlusterFS 基本概念

◆ 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

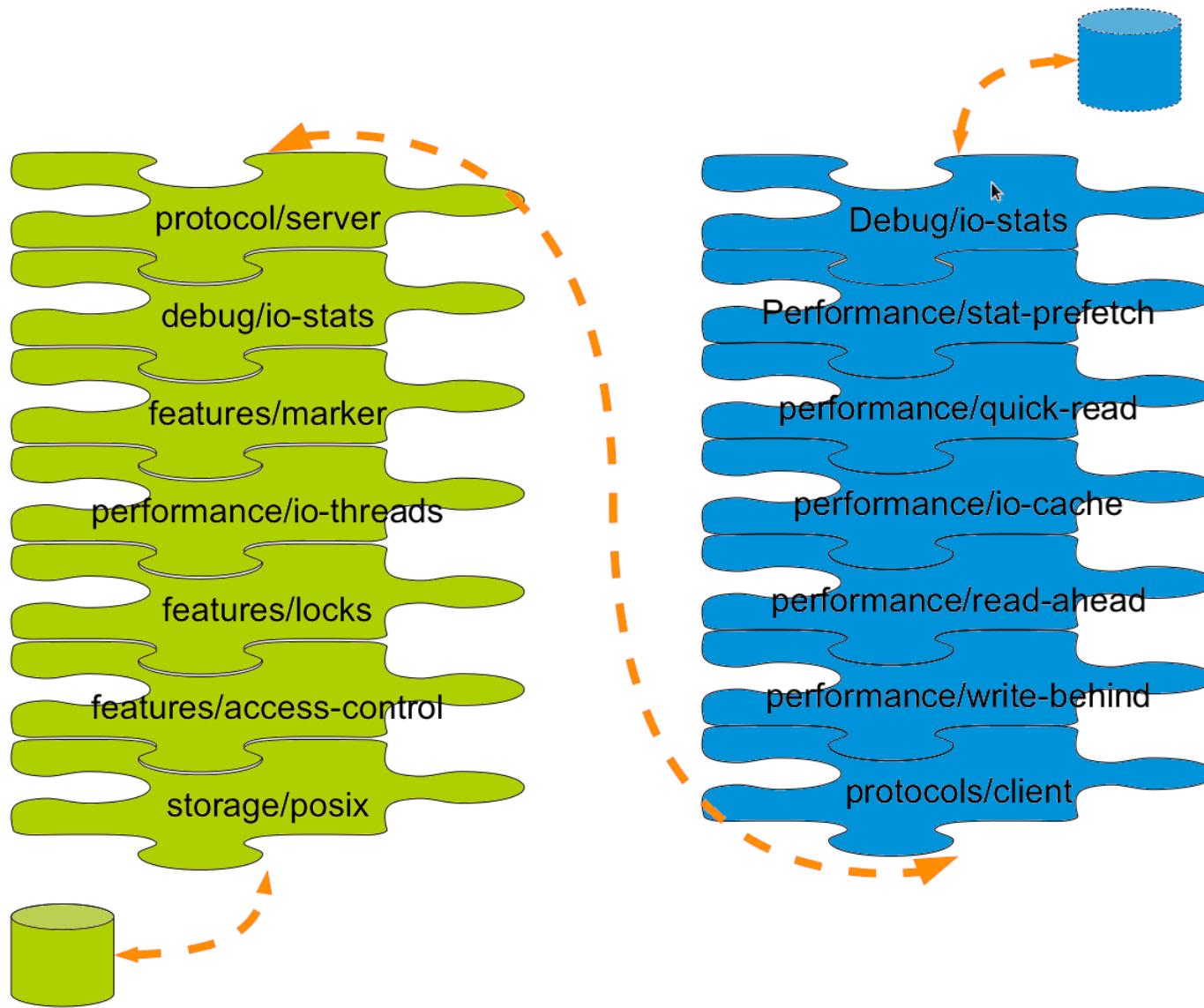
◆ Volume

- Bricks combined and passed through translators

◆ Node / Peer

- Server running the gluster daemon and sharing volumes

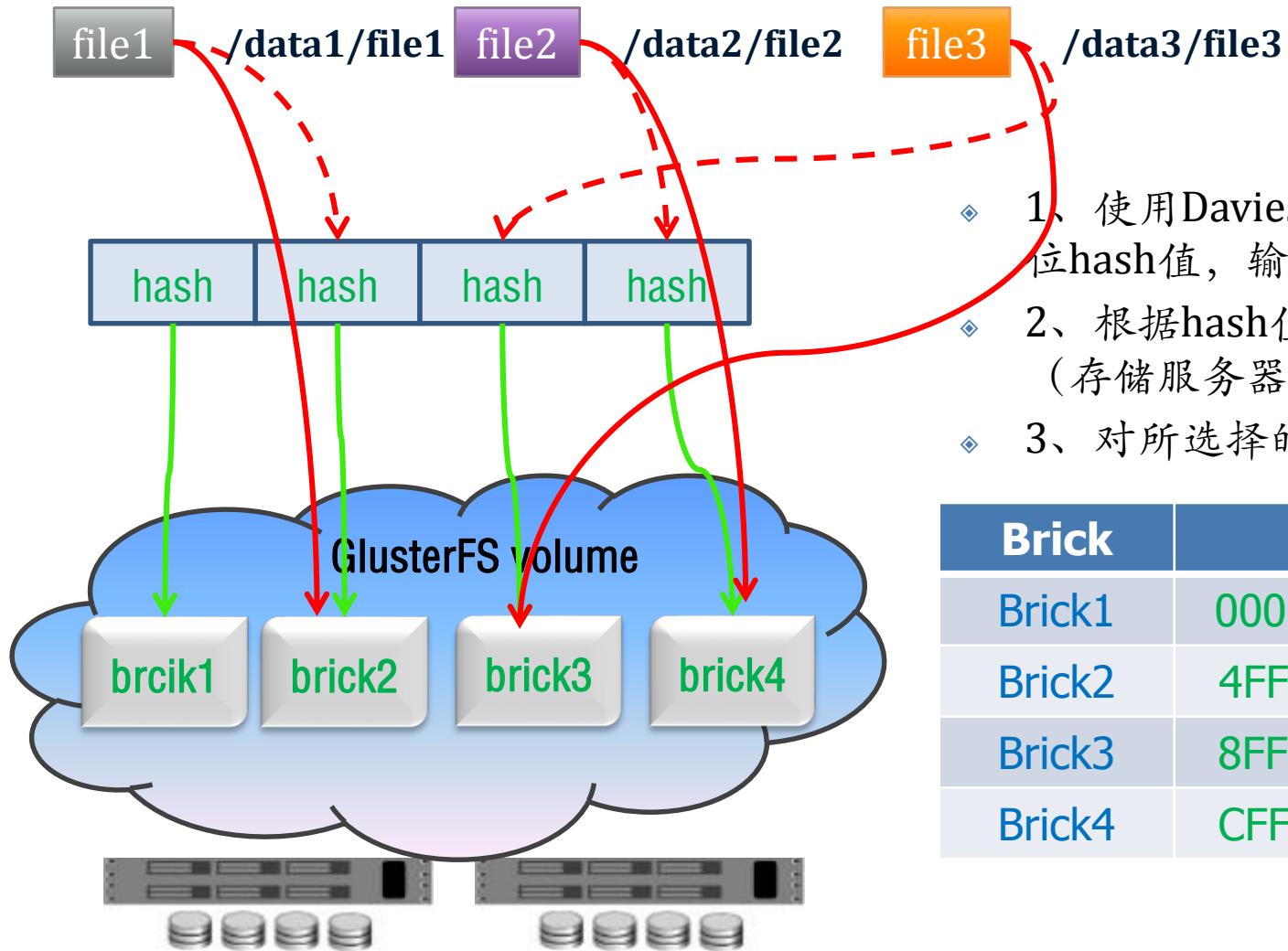
Translators



弹性哈希算法

- ◆ 无集中式元数据服务
 - 消除性能瓶颈，提高可靠性
- ◆ 采用Hash算法定位文件
 - 基于路径和文件名，一致性哈希DHT
- ◆ 弹性卷管理
 - 文件存储在逻辑卷中
 - 逻辑卷从物理存储池中划分
 - 逻辑可以在线进行扩容和缩减

弹性 Hash 算法流程



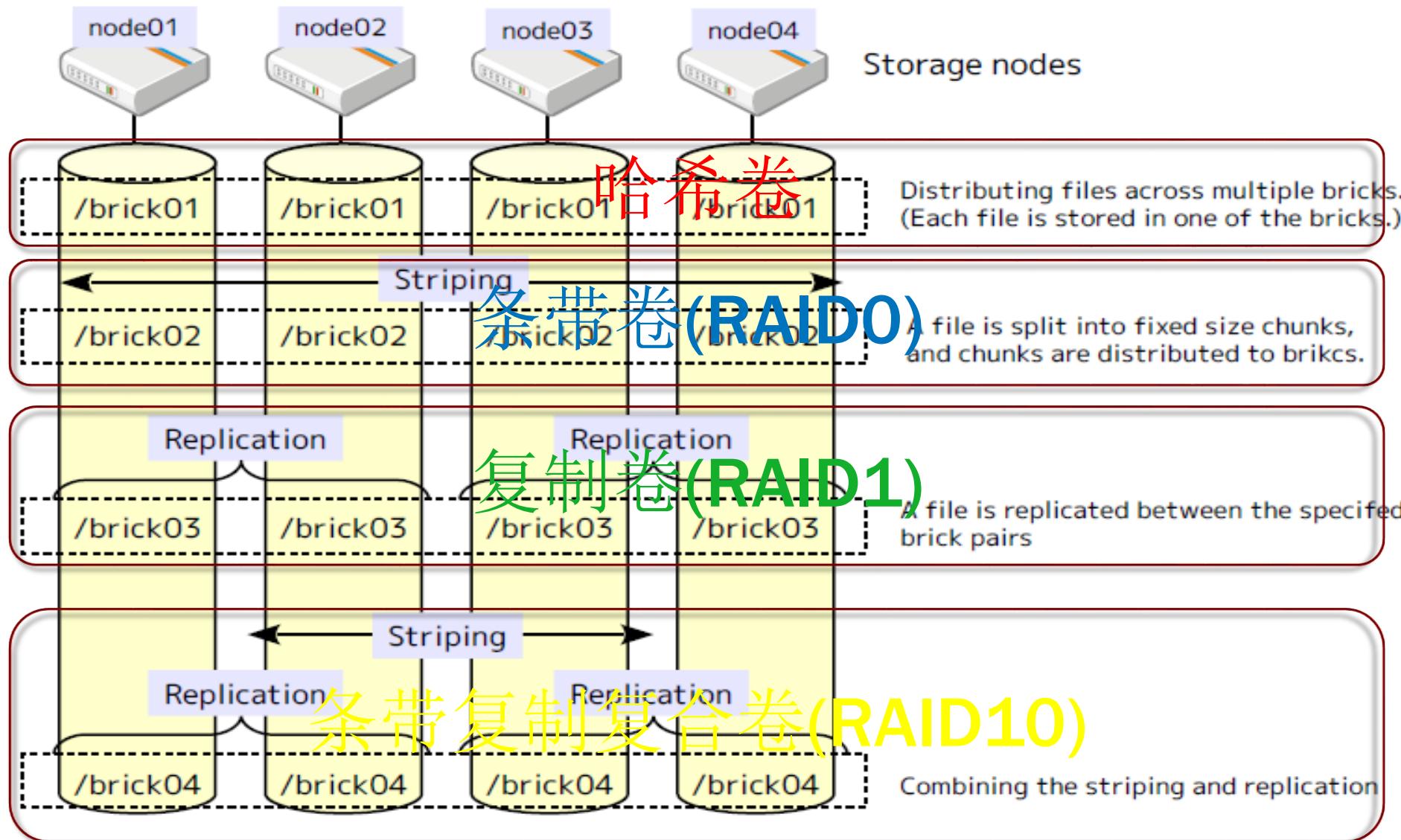
- ◆ 1、使用 Davies-Meyer 算法计算 32 位 hash 值，输入参数为文件名；
- ◆ 2、根据 hash 值在集群中选择子卷（存储服务器），进行文件定位；
- ◆ 3、对所选择的子卷进行数据访问。

Brick	Hash range
Brick1	00000000 ~ 3FFFFFFF
Brick2	4FFFFFFF ~ 7FFFFFFF
Brick3	8FFFFFFF ~ BFFFFFFF
Brick4	CFFFFFFF ~ FFFFFFFF

GlusterFS 卷类型

- ◆ 基本卷
 - ◆ 哈希卷 (Distributed Volume)
 - ◆ 复制卷 (Replicated Volume)
 - ◆ 条带卷 (Striped Volumes)
- ◆ 复合卷
 - ◆ 哈希复制卷(Distributed Replicated Volume)
 - ◆ 哈希条带卷 (Distributed Striped Volume)
 - ◆ 复制条带卷 (Replicated Striped Volume)
 - ◆ 哈希复制条带卷 (Distributed Replicated Striped Volume)

GlusterFS 卷数据分布



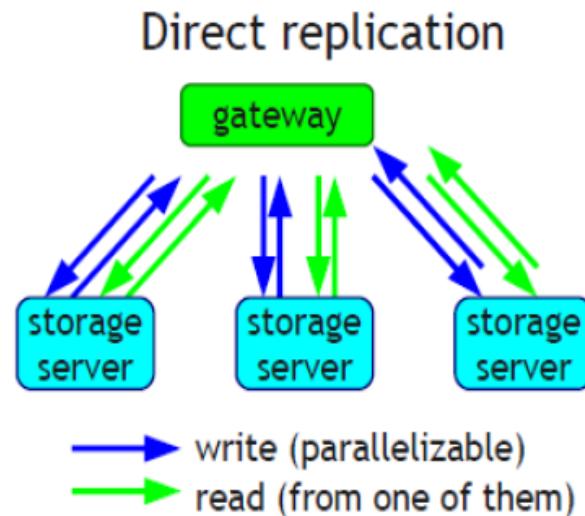
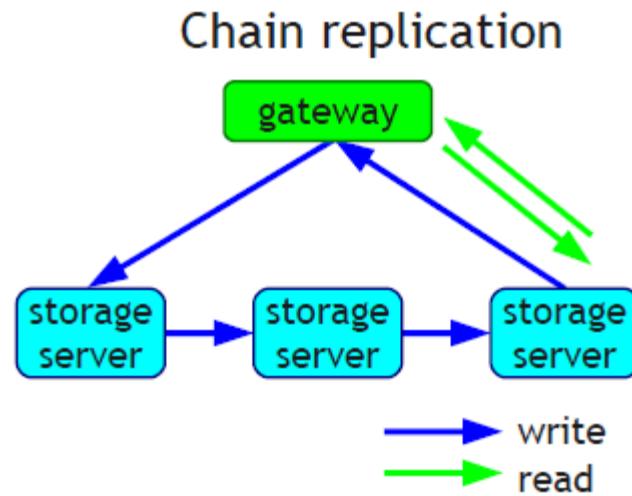
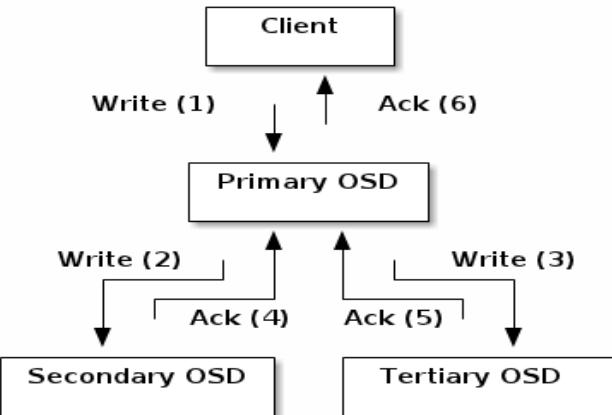
GlusterFS 命名空间

- ◆ 三种基本集群各由一个translator来实现，分别由自己独立的命名空间，使用自己的机制进行独立的维护和管理。
- ◆ 分布式集群，文件通过HASH算法分散到集群节点上，每个节点上的命名空间均不重叠，所有集群共同构成完整的命名空间，访问时使用HASH算法进行查找定位。
- ◆ 复制集群类似RAID1，所有节点命名空间均完全相同，每一个节点都可以表示完整的命名空间，访问时可以选择任意个节点。
- ◆ 条带集群与RAID0相似，所有节点具有相同的命名空间，但对象属性会有所不同，文件被分成数据块以Round Robin方式分布到所有节点上，访问时需要联动所有节点来获得完整的名字信息。

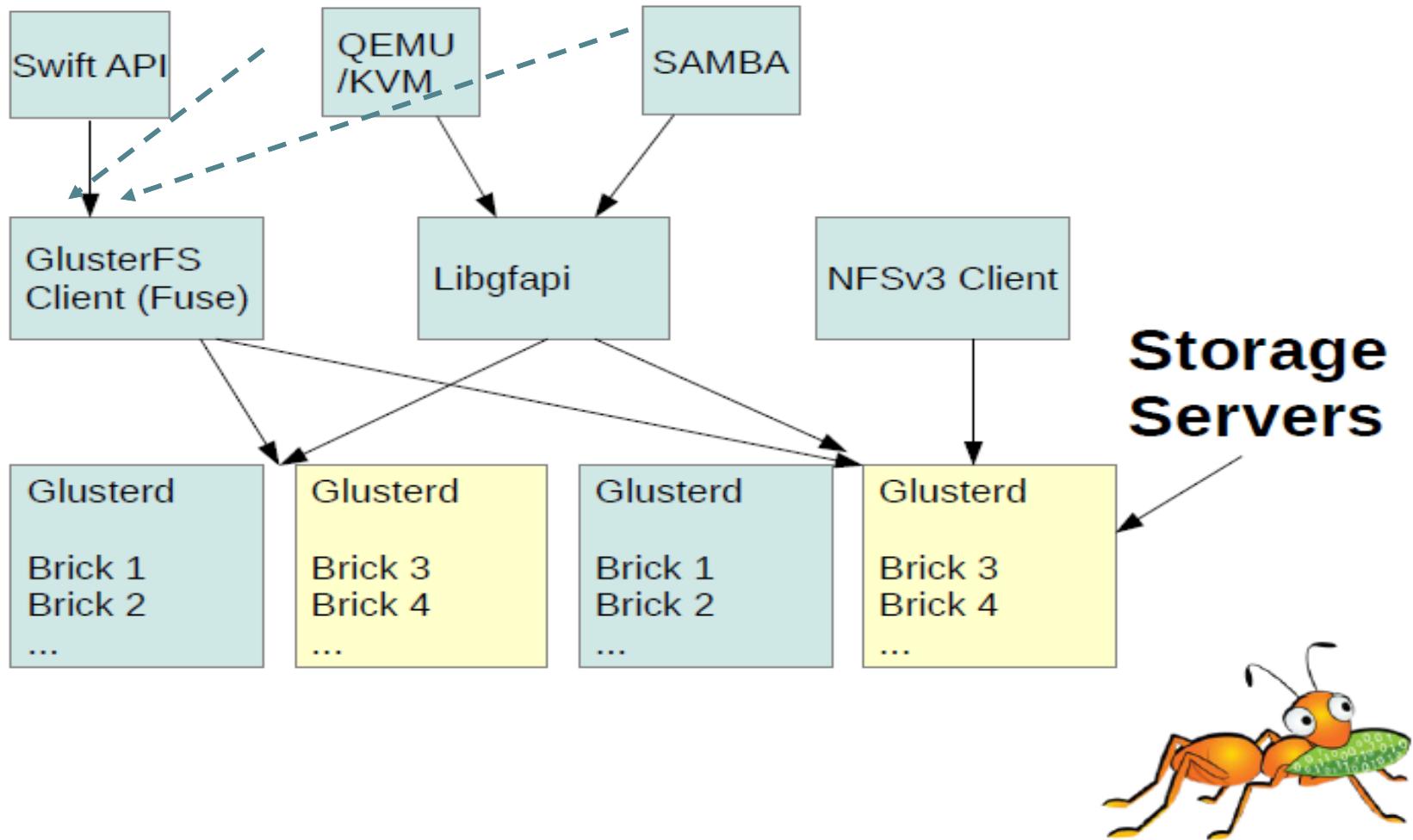
数据副本一致性模型

◆ 数据强一致性

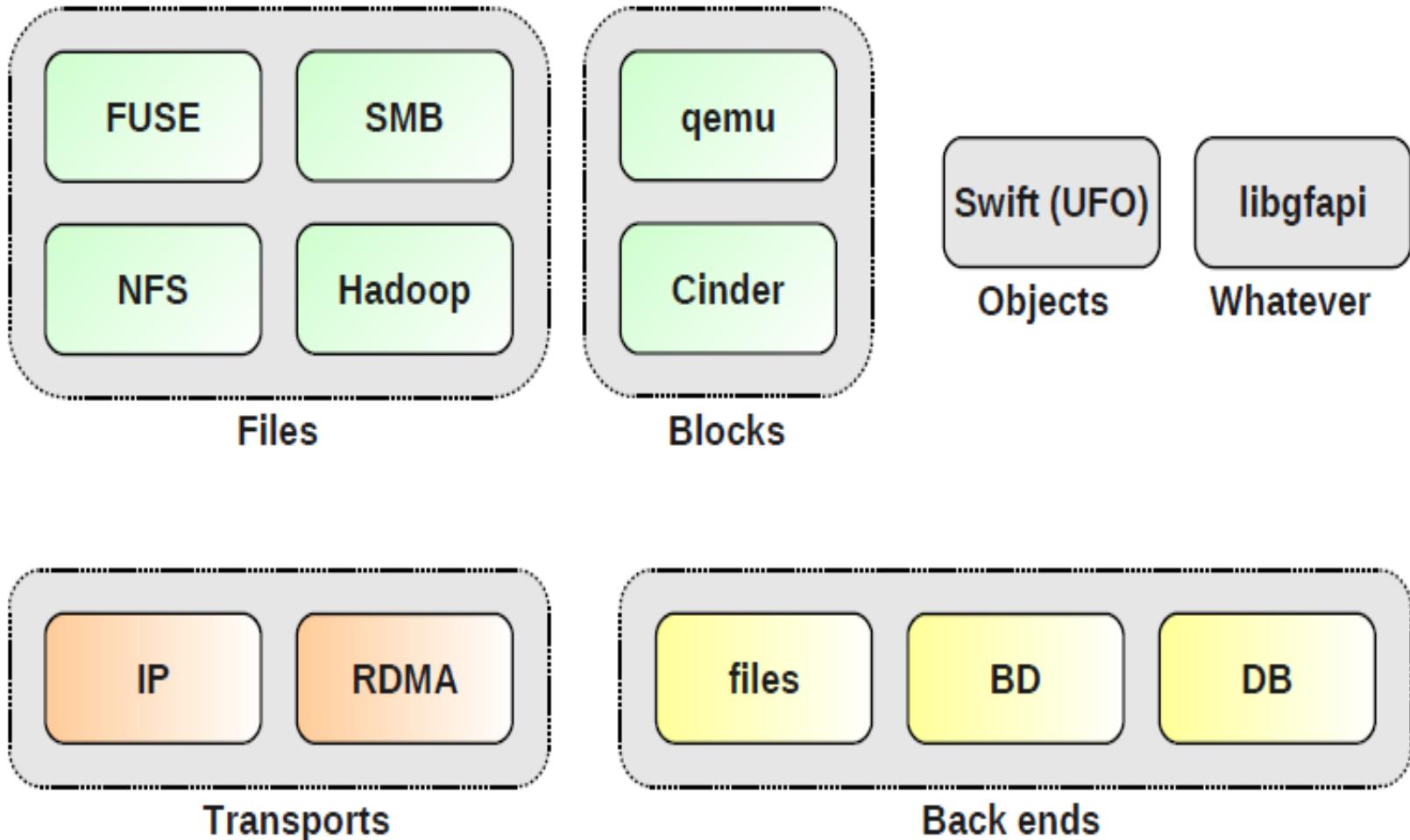
- ◆ Chain replication
- ◆ Direct replication (✓)
- ◆ Master-slave replication



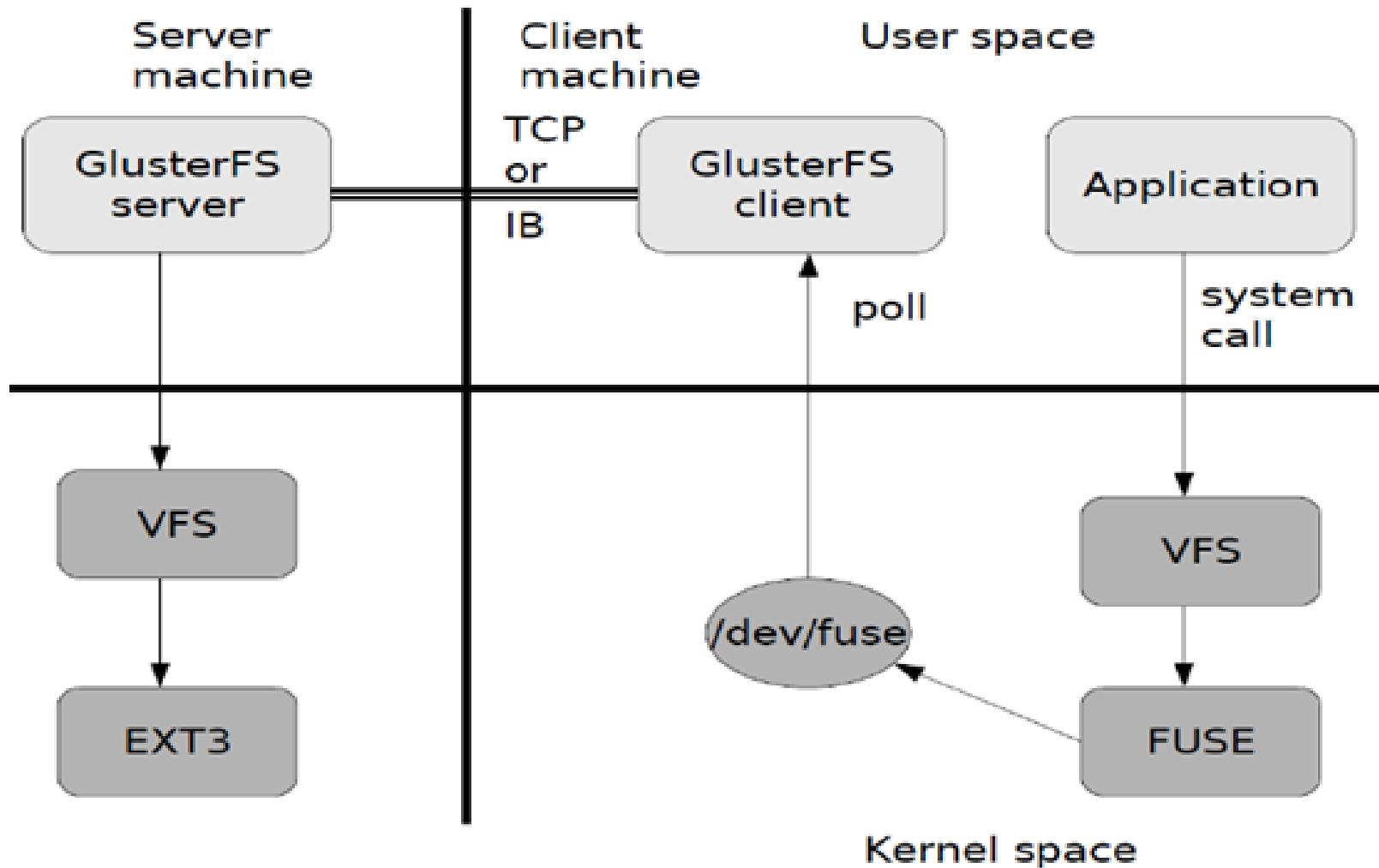
GlusterFS 系统交互



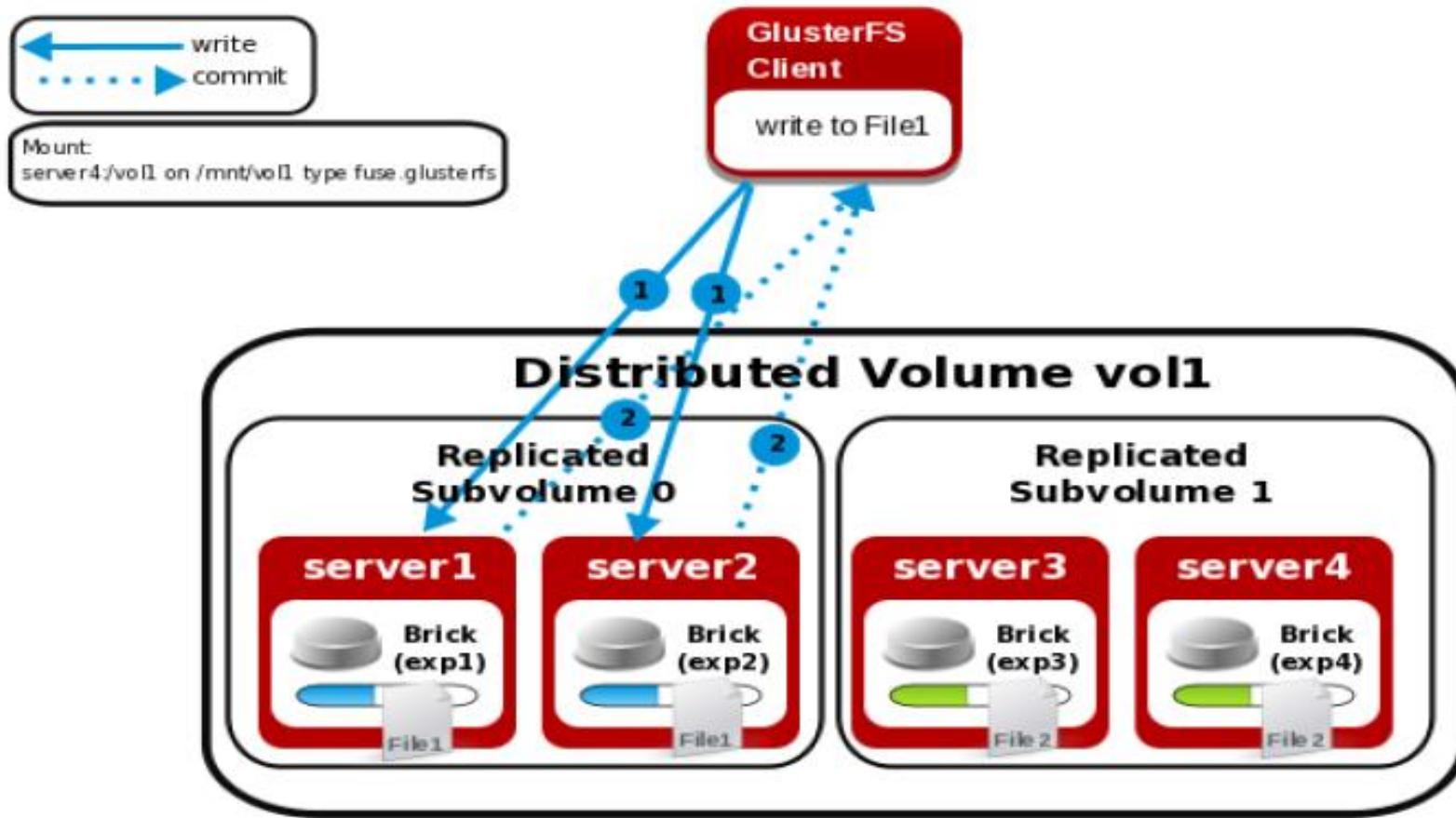
GlusterFS 访问接口



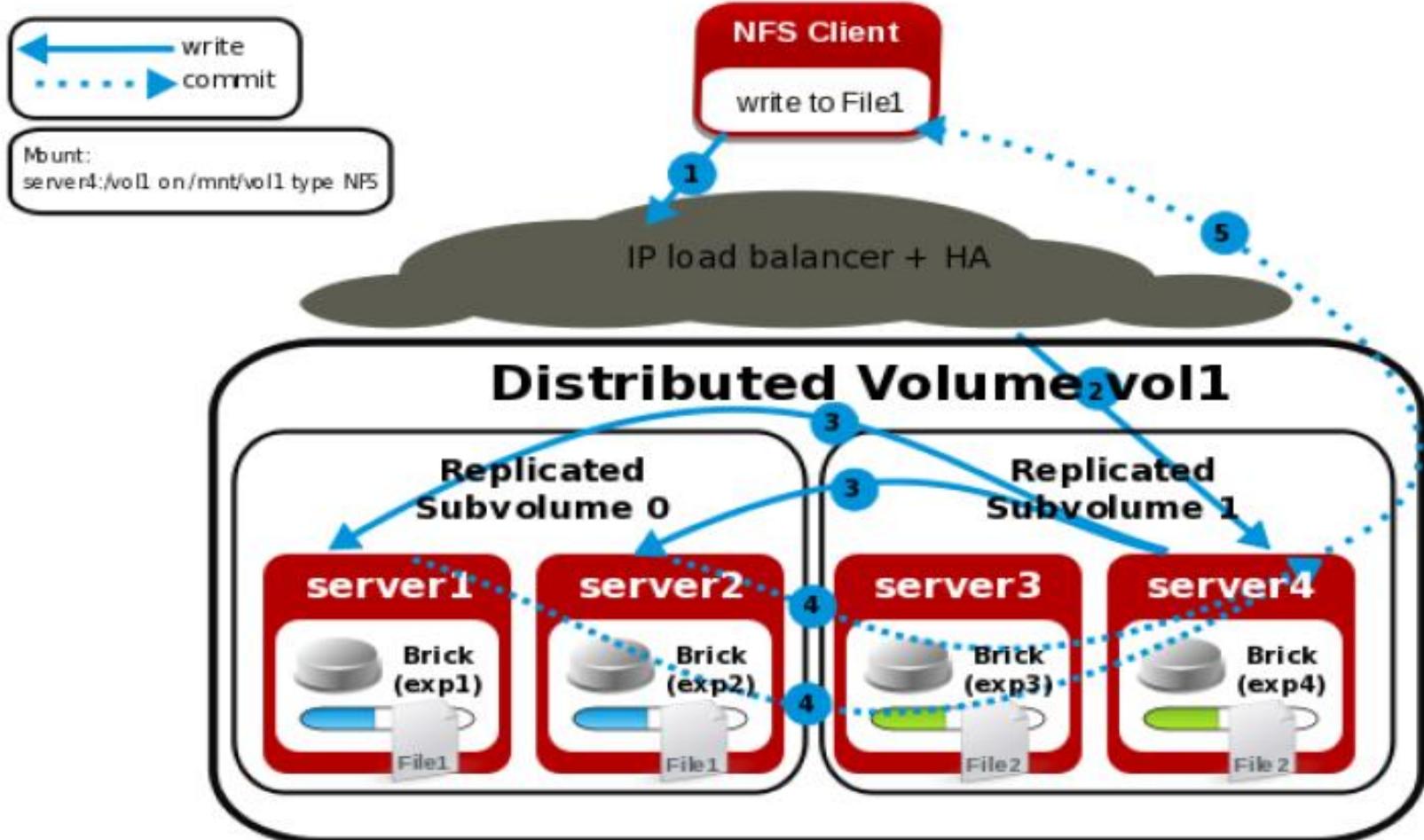
GlusterFS 数据流



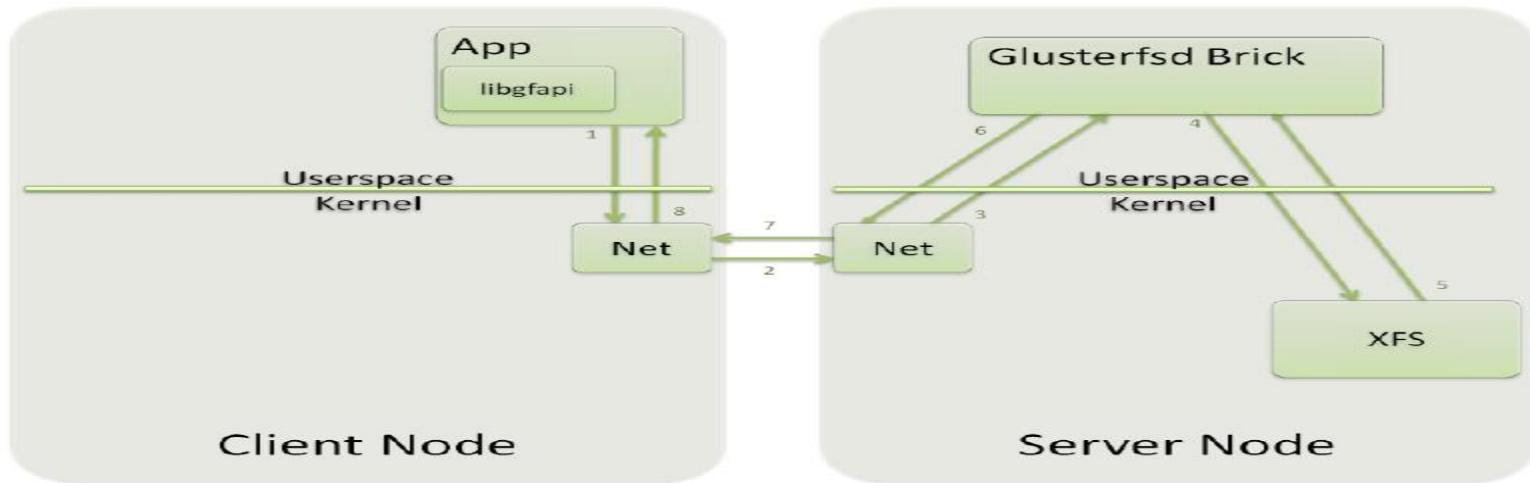
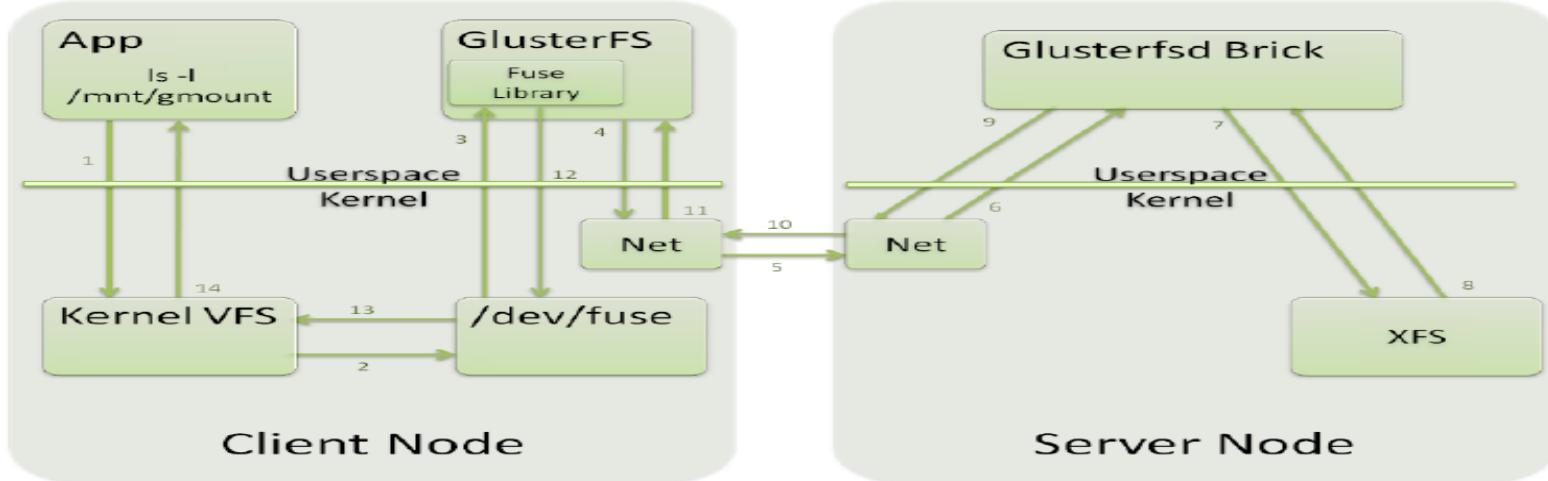
FUSE 访问



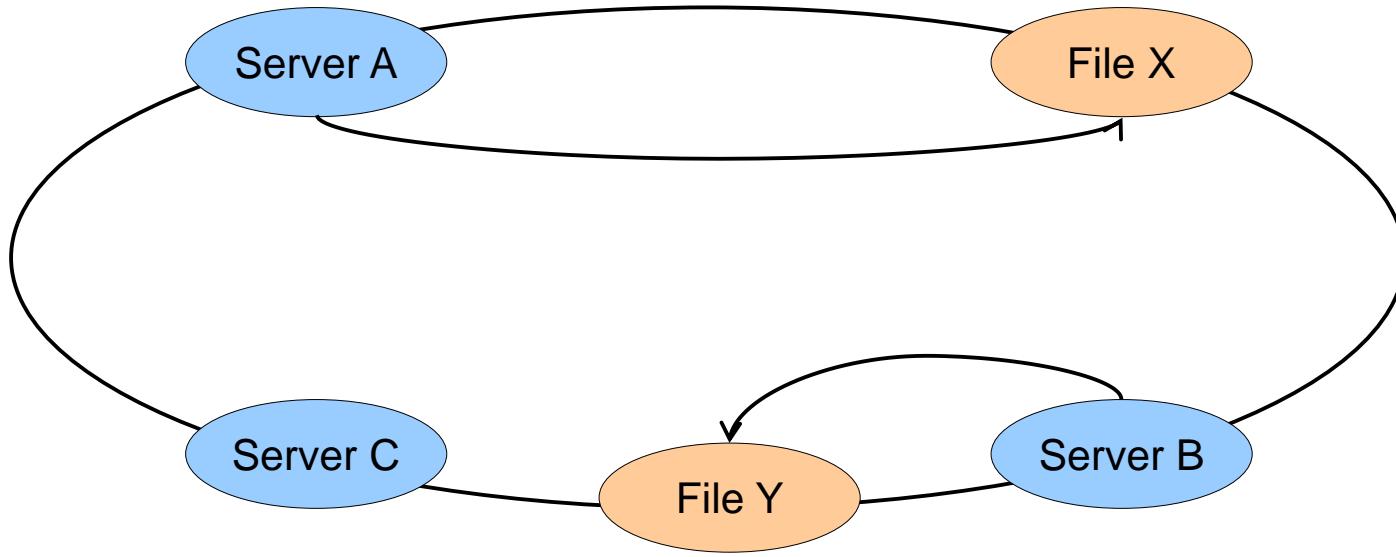
NFS/CIFS 访问



libgfapi 访问

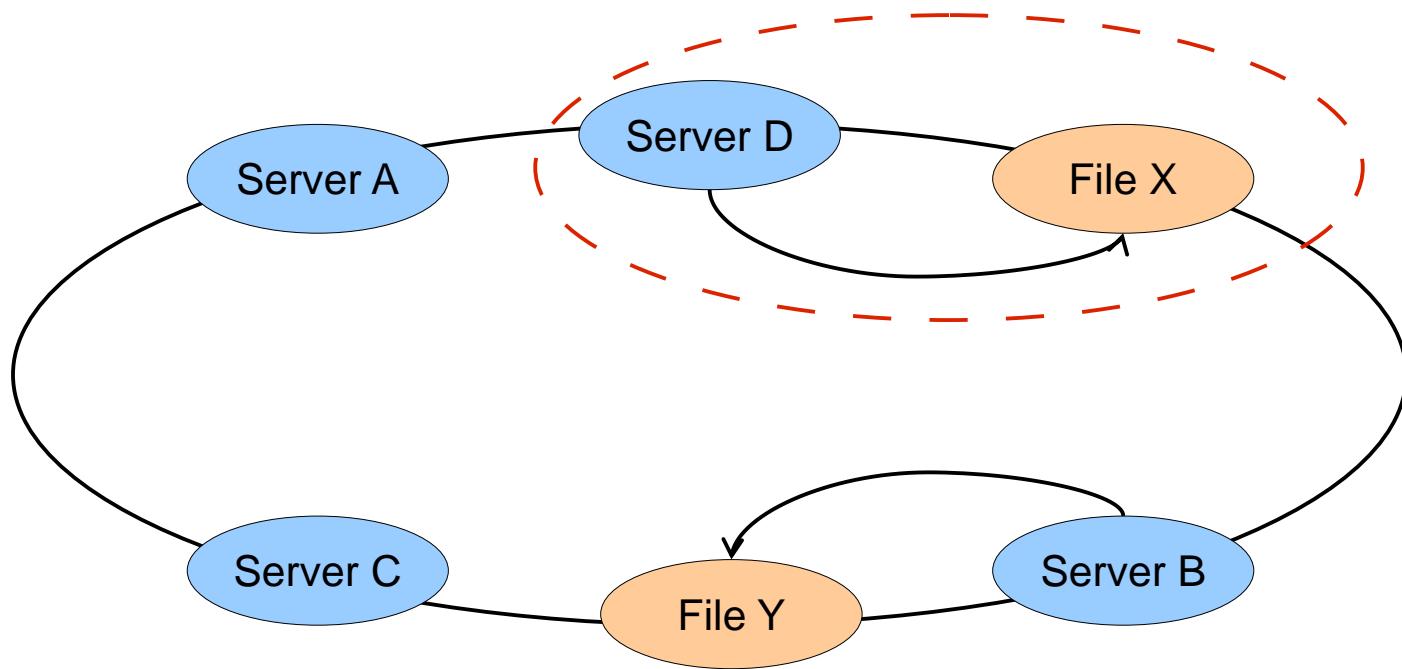


Distributed Hash Table (DHT)



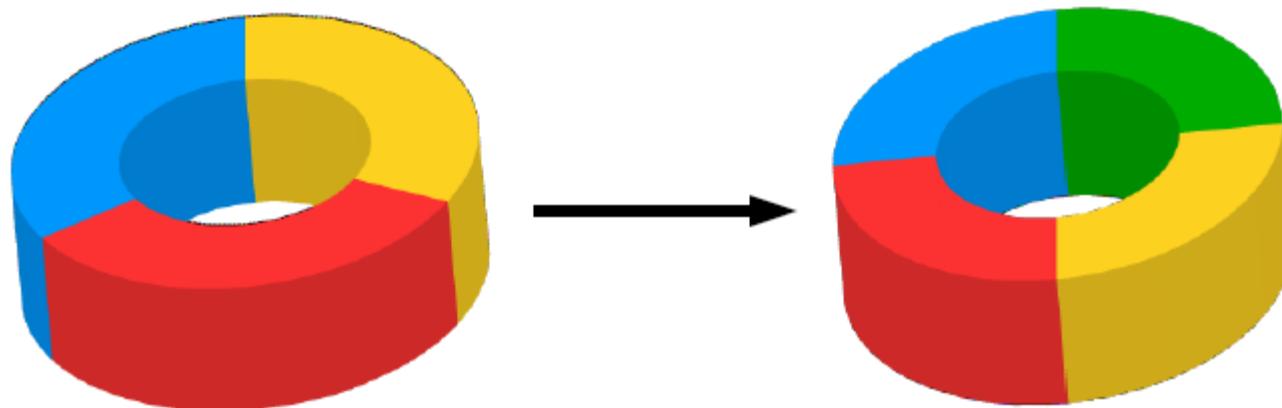
- GlusterFS弹性扩展的基础
- 确定目标hash和brick之间的映射关系

添加节点



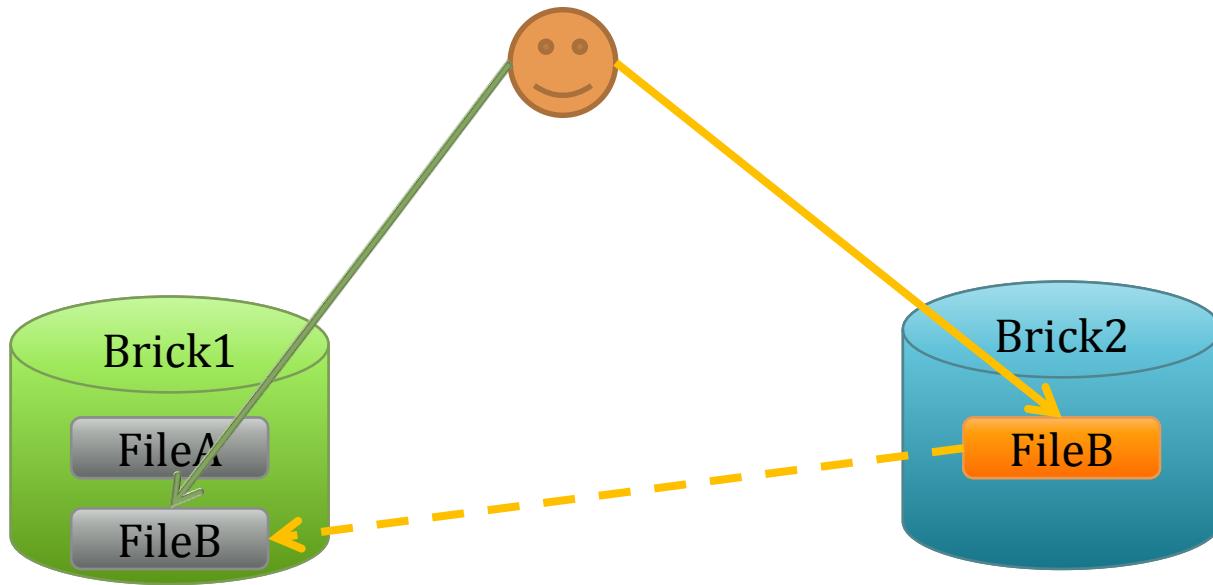
- 添加新节点，最小化数据重新分配
- 老数据分布模式不变，新数据分布到所有节点上
- 执行rebalance，数据重新分布

容量负载均衡



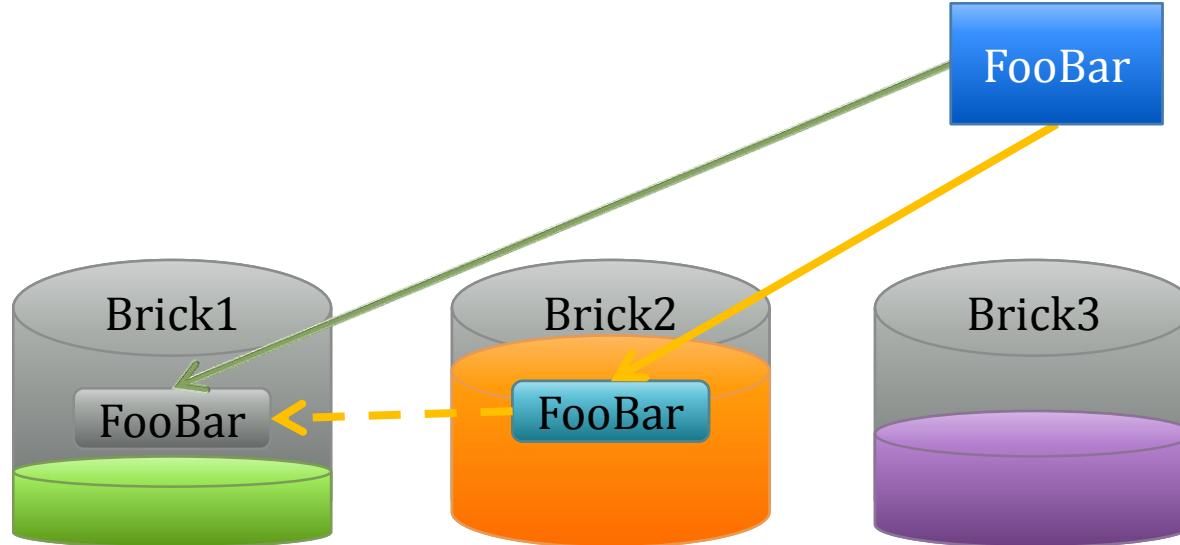
- Hash范围均衡分布，节点一变动全局
- 目标：优化数据分布，最小化数据迁移
- 数据迁移自动化、智能化、并行化

文件更名



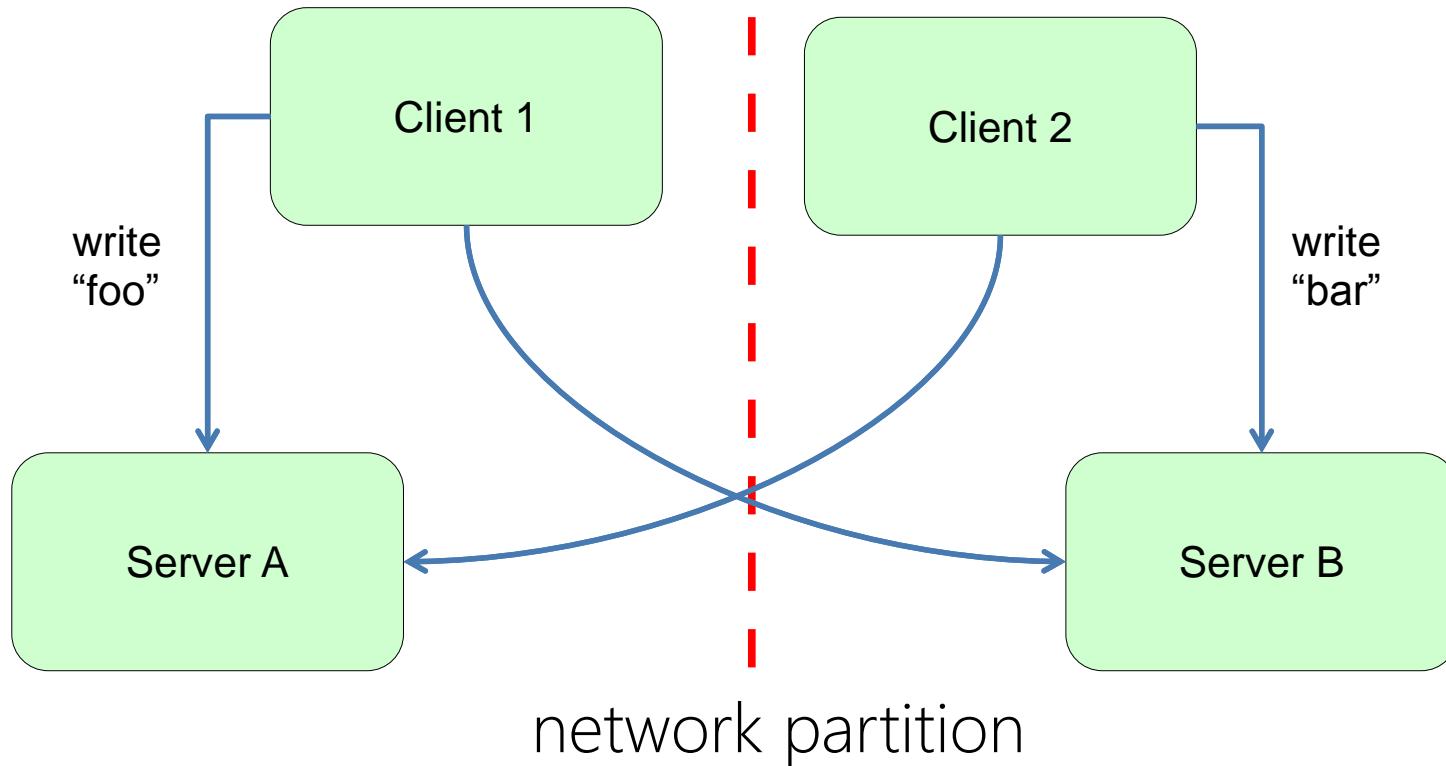
- 文件更名: FileA → FileB
- 原先的hash映射关系失效，大文件难以实时迁移
- 采用文件符号链接，访问时解析重定向

容量负载优先



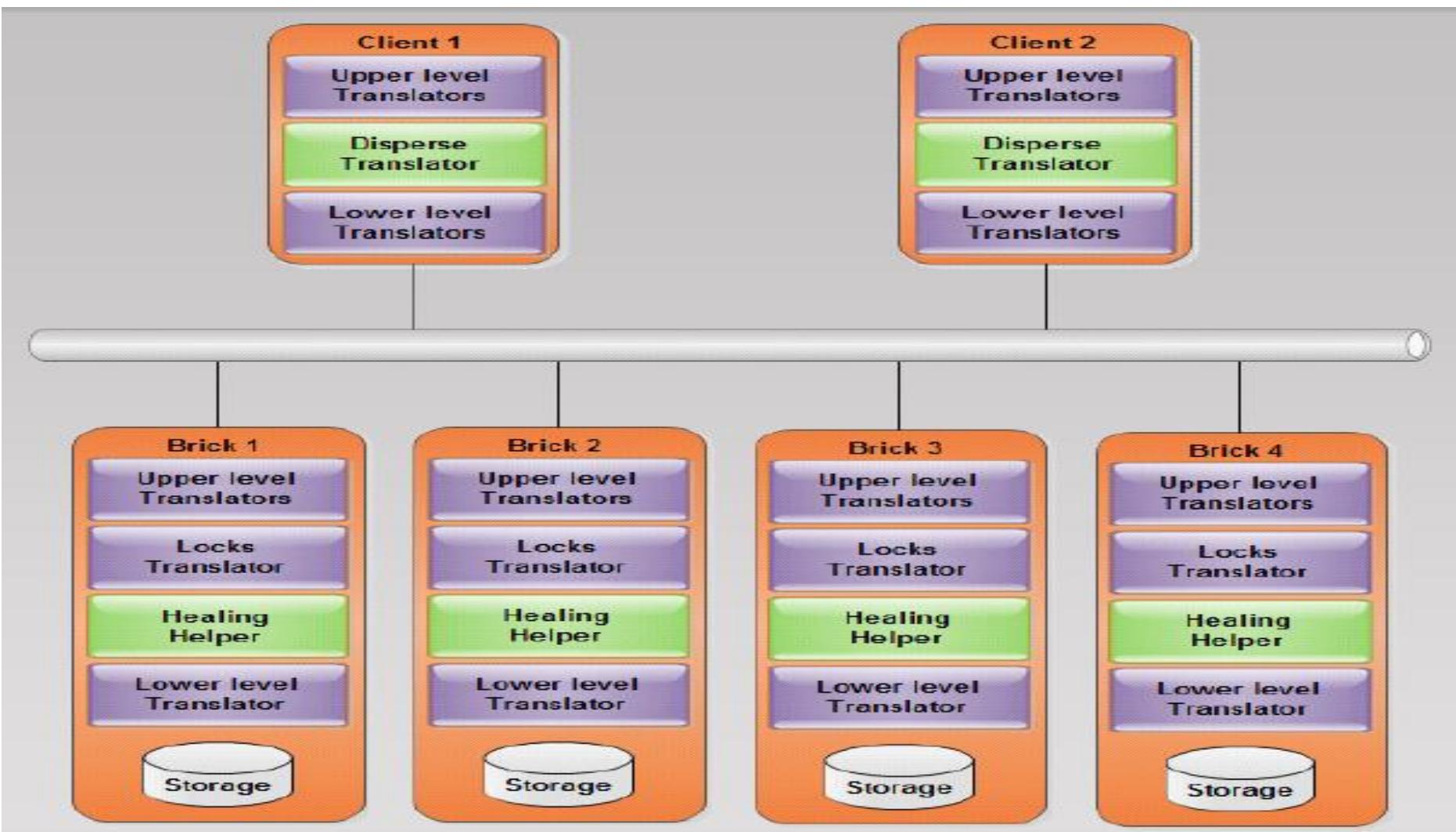
- 设置容量阈值，优先选择可用容量充足brick
- Hash目标brick上创建文件符号链接
- 访问时解析重定向

Split Brain

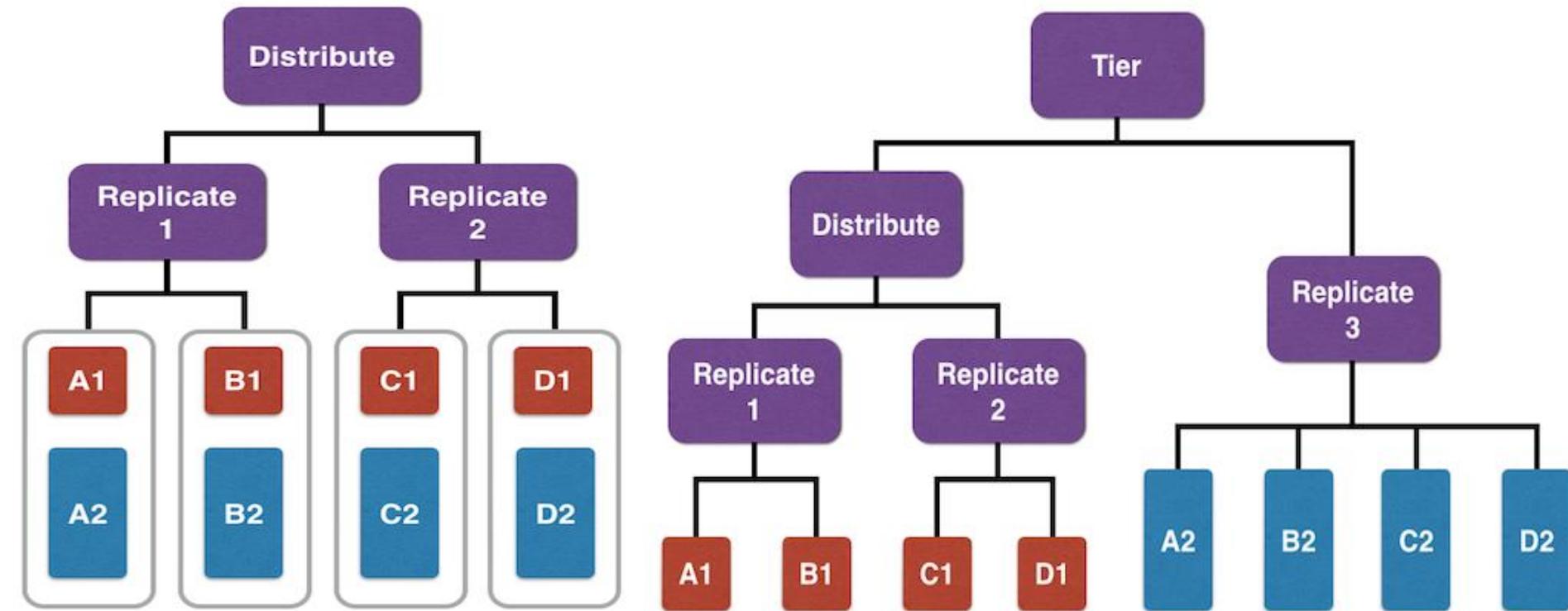


- 裂脑如何产生的？
- 解决方法： 1、报错处理； 2、Quorum方法($N=2?$)； 3、仲裁机制

Dispersed Volume (纠错码)



SSD Cache/Tier



(三)
GlusterFS 应用场景

GlusterFS 应用场景

- Media
 - 文档、图片、音频、视频
- Shared storage
 - 云存储、虚拟化存储、HPC
- Big data
 - 日志文件、RFID数据

配置参数调优

◆ Gluster volume set <卷> <参数>

参数项目	说明	缺省值	合法值
Auth.allow	IP访问授权	*(allow all)	Ip地址
Cluster.min-free-disk	剩余磁盘空间阈值	10%	百分比
Cluster.stripe-block-size	条带大小	128KB	字节
Network.frame-timeout	请求等待时间	1800s	0-1800
Network.ping-timeout	客户端等待时间	42s	0-42
Nfs.disabled	关闭NFS服务	Off	Off on
Performance.io-thread-count	IO线程数	16	0-65
Performance.cache-refresh-timeout	缓存校验周期	1s	0-61
Performance.cache-size	读缓存大小	32MB	字节

FUSE性能优化

- ◆ 更多的mountpoint，更多的并发访问
- ◆ Mount -o max_read 1048576
- ◆ 修改FUSE内核模块：每请求最大为128KB
fuse_i.h中，

```
#define FUSE_MAX_PAGES_PER_REQ 256
```

原先为32，每page为4KB， $256 * 4\text{KB} = 1\text{MB}$ 。重新编译FUSE模块，替换系统中的fuse.ko。

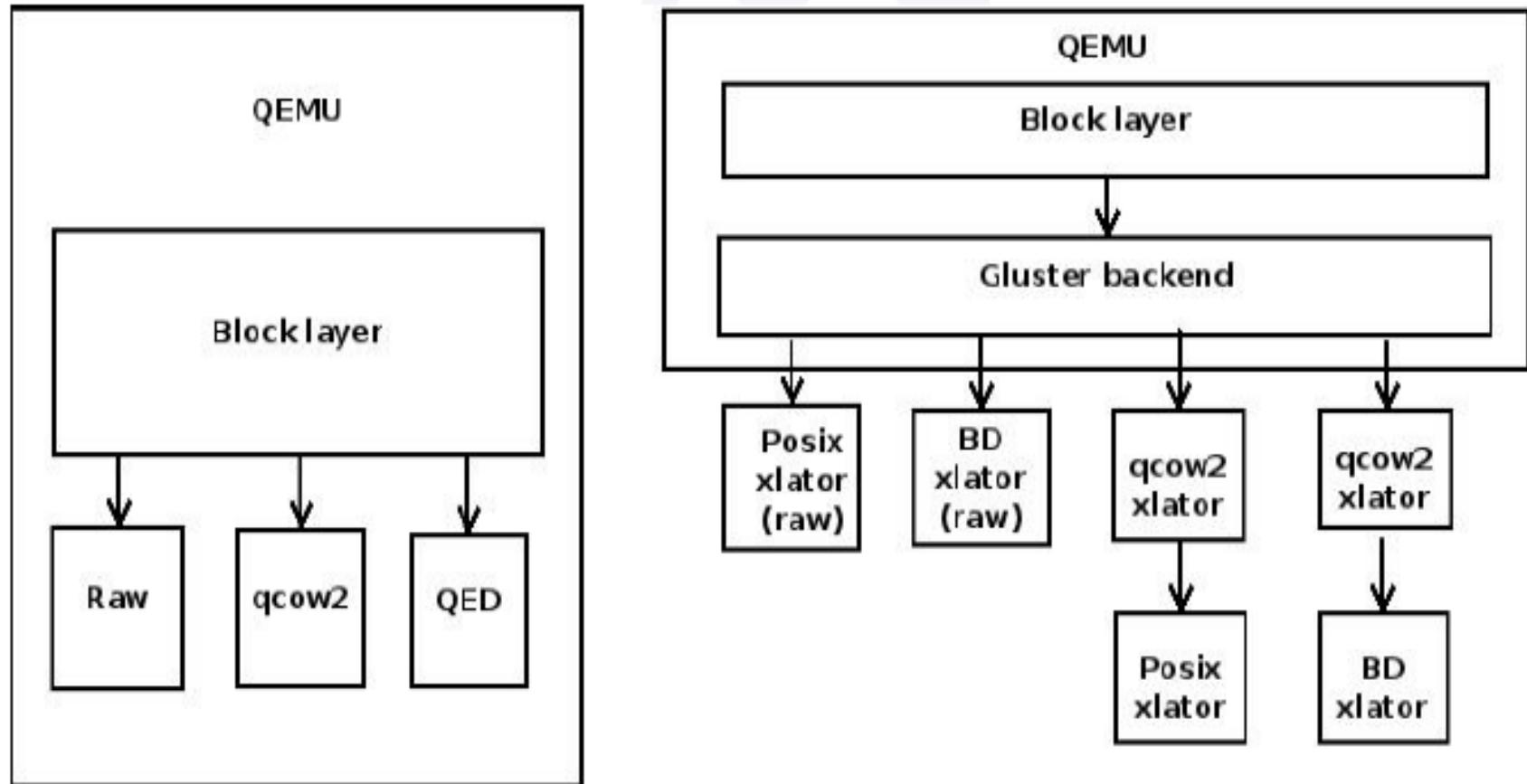
SSD系统优化

- ◆ I/O schedule算法
- ◆ CPU多核绑定
- ◆ 请求列队和最大请求数
- ◆ 禁用merge/rotational/read_ahead/barrier
- ◆ 分区/卷4KB对齐
- ◆ 文件系统开启SSD支持选项

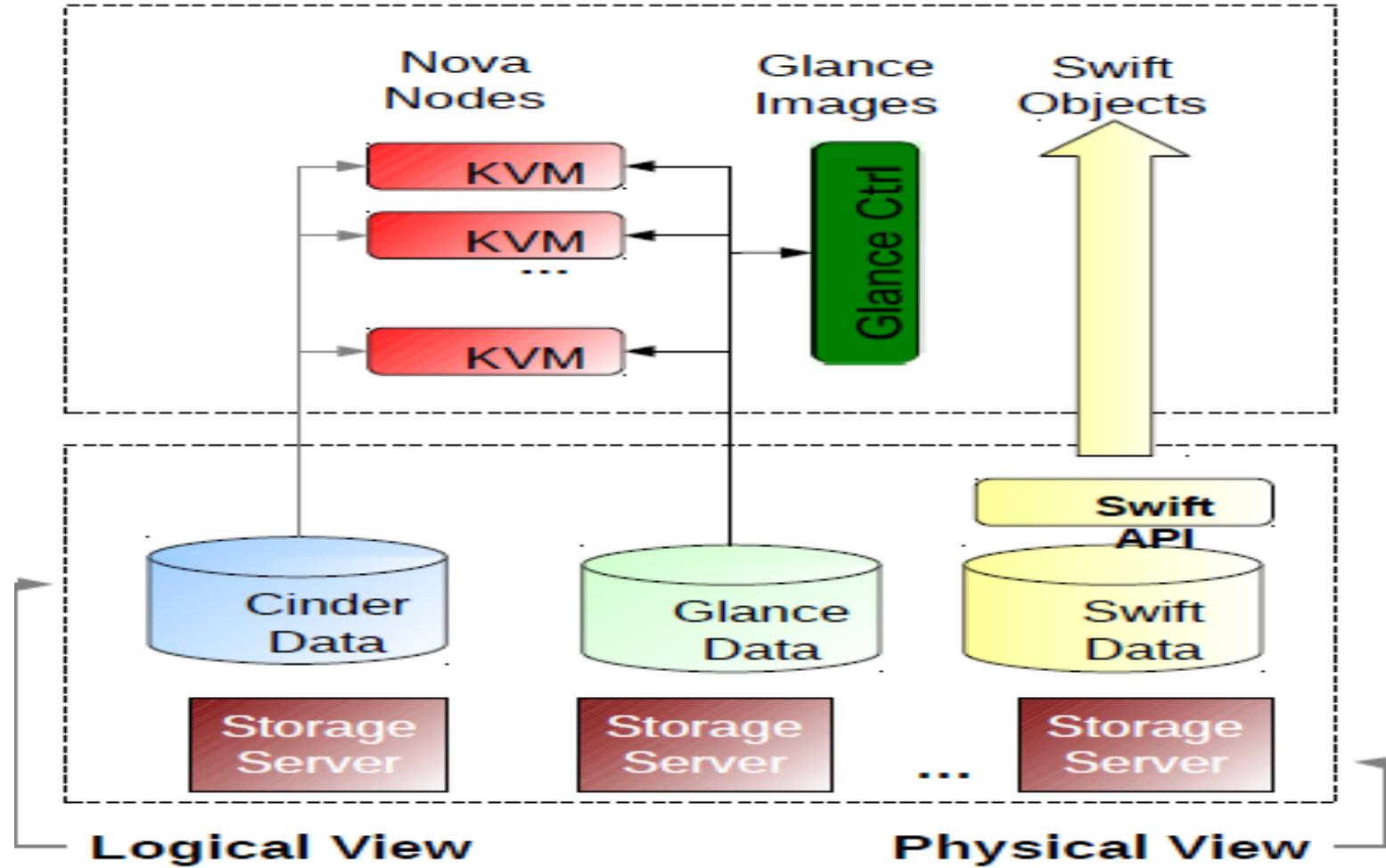
KVM 优 化

- ❖ 使用QEMU-GlusterFS(libgfapi)整合方案
- ❖ gluster volume set <volume> group virt
- ❖ tuned-adm profile rhs-virtualization
- ❖ KVM host: tuned-adm profile virtual-host
- ❖ Images和应用数据使用不同的volume
- ❖ 每个gluster节点不超过2个KVM Host (16 guest/host)
- ❖ 提高响应时间
 - ❖ 减少/sys/block/vda/queue/nr_request
 - ❖ Server/Guest: 128/8 (缺省值256/128)
- ❖ 提高读带宽
 - ❖ 提高/sys/block/vda/queue/read_ahead_kb
 - ❖ VM readahead: 4096 (缺省值128)

QEMU-GlusterFS 集成



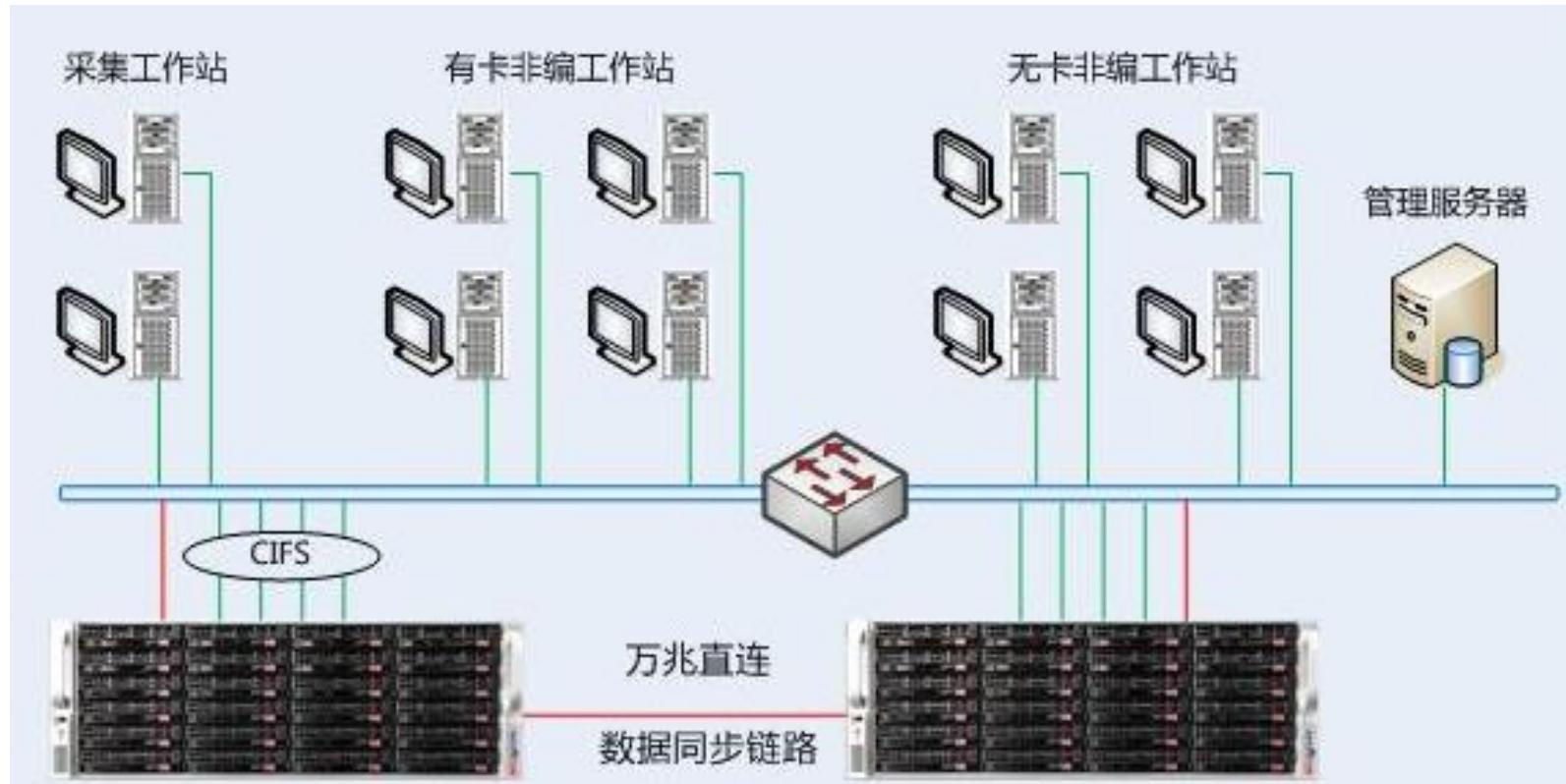
Openstack-GlusterFS集成



Havana 存儲 driver

- Coraid (AoE)
- Dell EqualLogic (iSCSI)
- EMC VMAX/VNX (iSCSI)
- GlusterFS (GlusterFS)
- HP 3PAR (iSCSI/FC)
- HP LeftHand (iSCSI)
- Hitachi HUS (iSCSI)
- Huawei HVS/T-series/Dorado (iSCSI/FC)
- IBM DS8000 (FC)
- IBM GPFS (GPFS)
- IBM Storwize family/SVC (iSCSI/FC)
- IBM XIV (iSCSI/FC)
- Local disk partitions
- LVM (iSCSI)
- NetApp (iSCSI/NFS)
- Nexenta (iSCSI)
- NFS (NFS)
- RBD (Ceph)
- Scality SOFS (scality)
- Sheepdog (sheepdog)
- Solaris (iSCSI)
- SolidFire (iSCSI)
- Windows Server 2012 (iSCSI)
- Zadara (iSCSI)

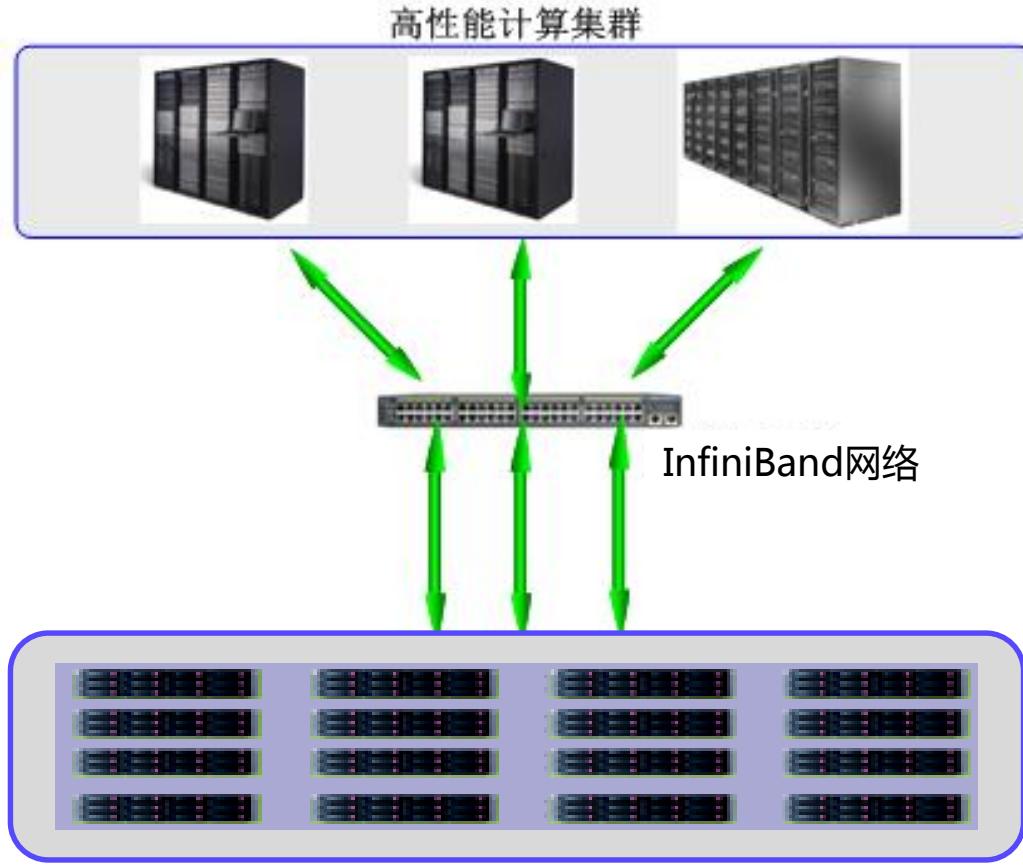
解决方案-广电非编



存储需求：

- 1、可用存储容量，根据业务发展弹性扩展；
- 2、聚合带宽，可扩展至数GB；
- 3、非编工作站通过CIFS访问，单客户端稳定的带宽
- 4、提供冗余可用性，保证业务不中断；

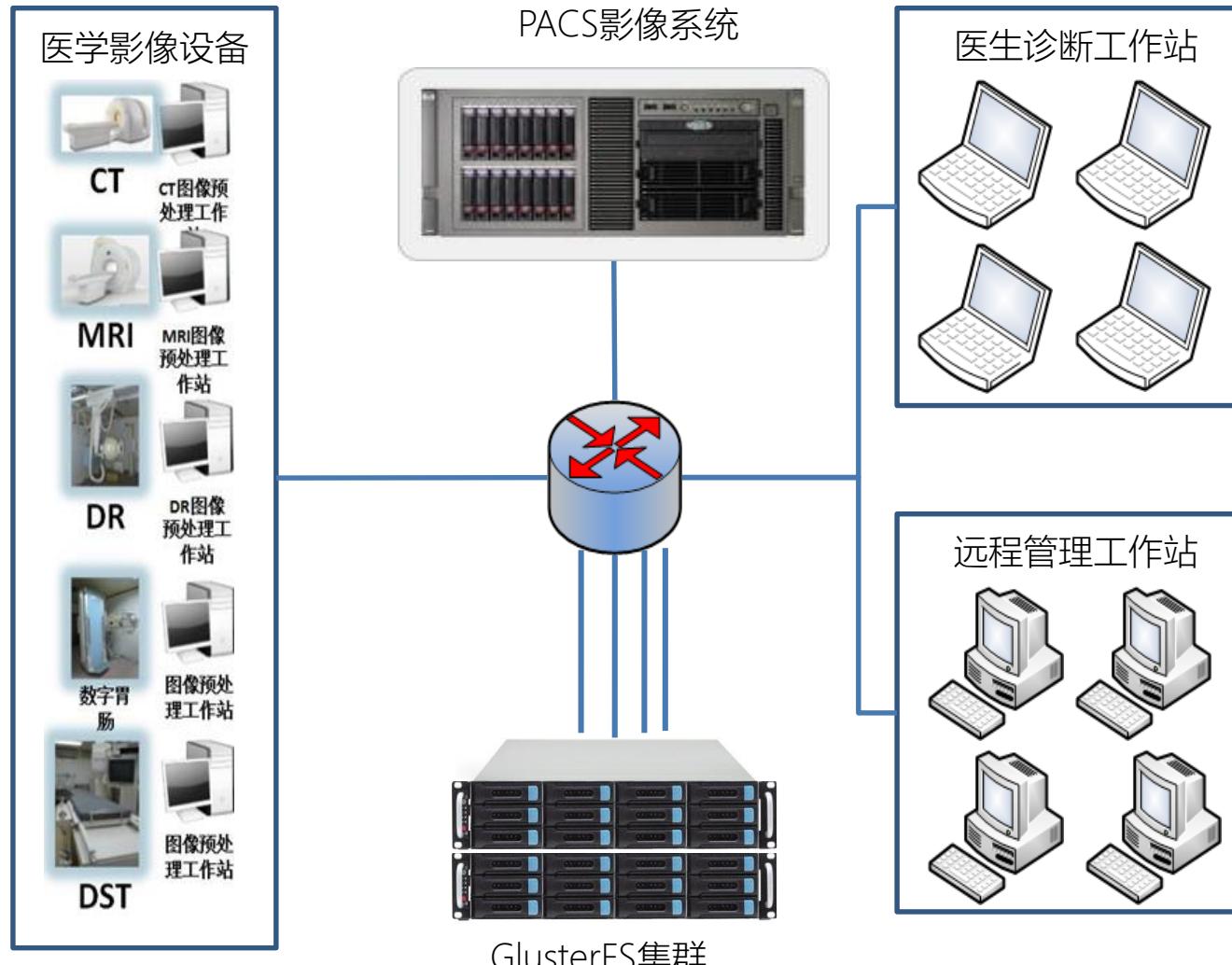
解决方案-HPC



存储需求

- 1、存储容量，可扩展至PB级
- 2、带宽数GB，单客户端可达GB级
- 3、超大文件，几GB至几十GB
- 4、支持高可用机制(副本或冗余)

解决方案-医院 PACS

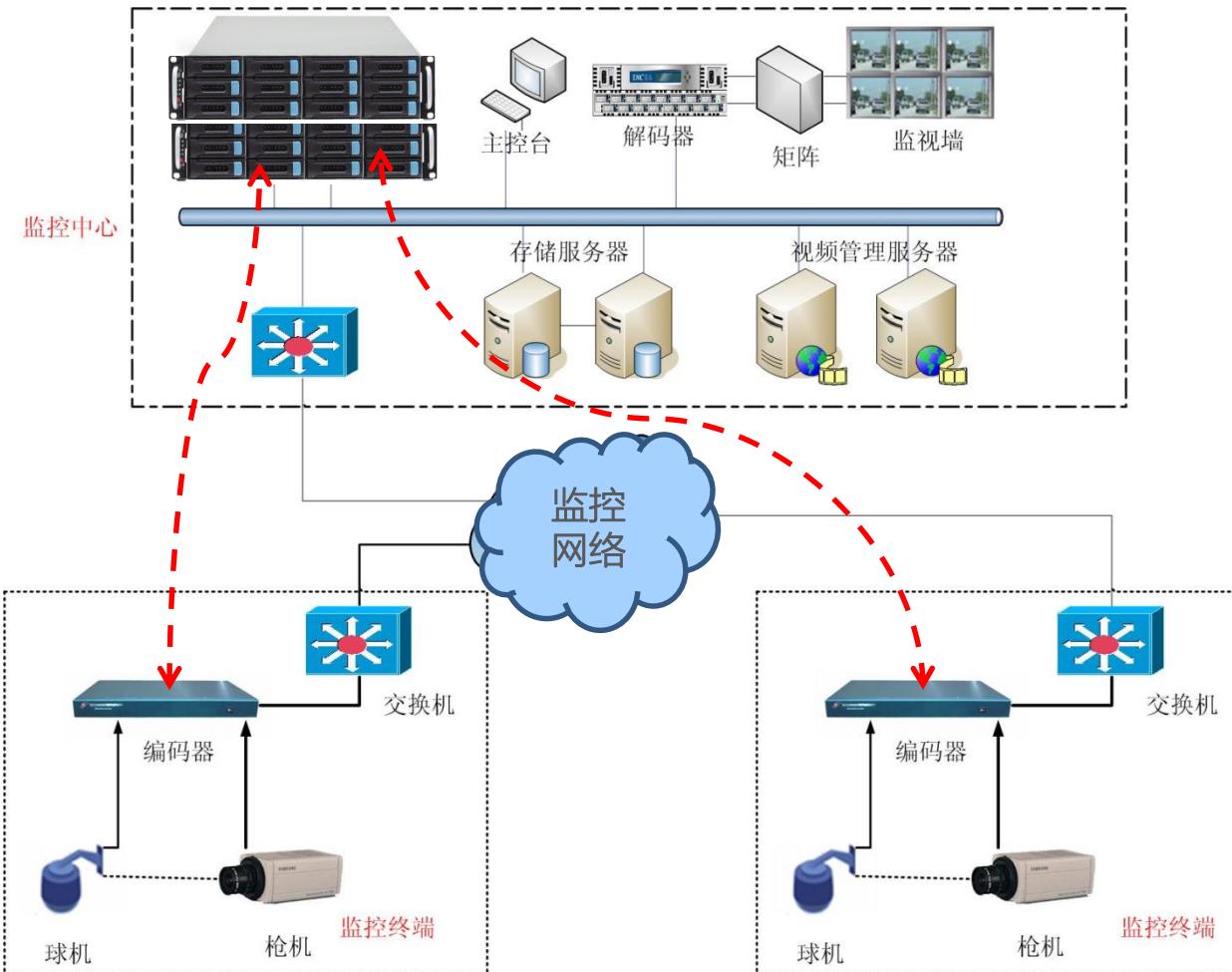


存储需求:

- 1、存储容量数十TB
- 2、文件数量百万级
- 3、文件大小MB级
- 4、并发多台工作站
- 5、通过FTP协议访问
- 6、提供高可用保证

解决方案-视频监控

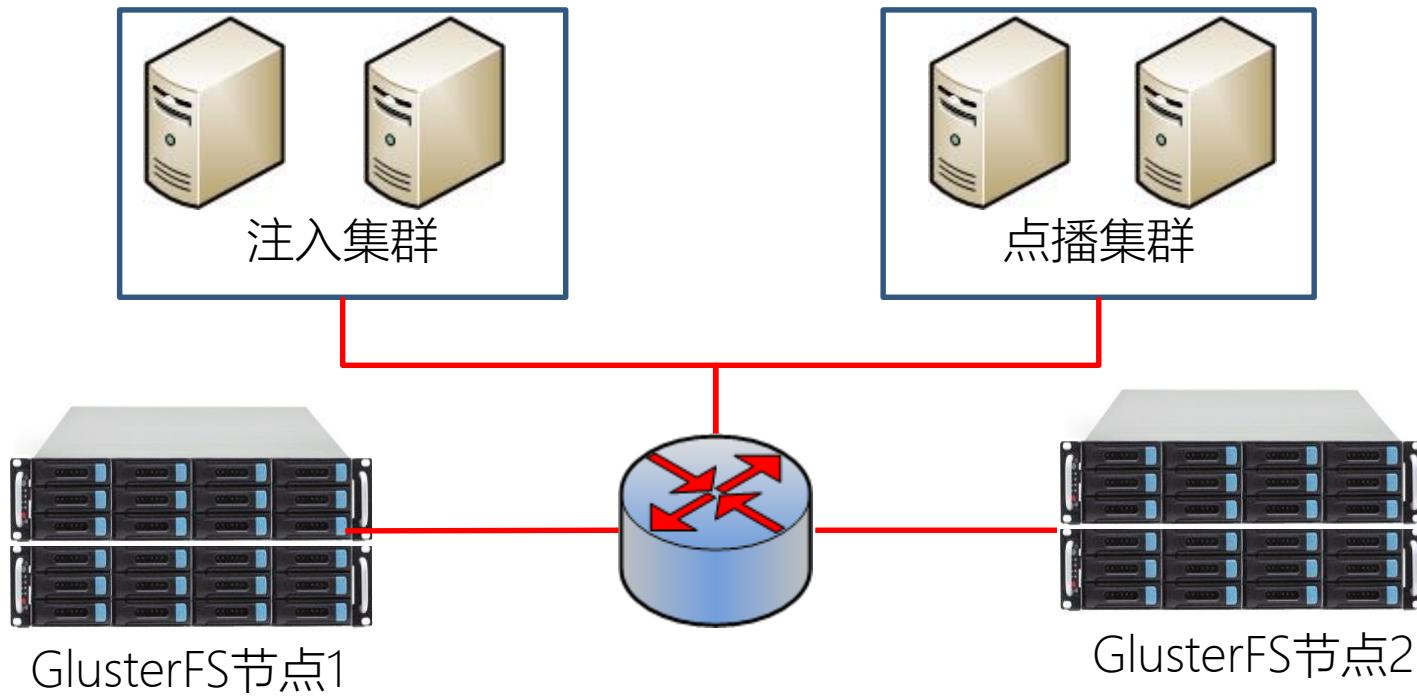
GlusterFS集群



存储需求：

- 1、实现集中管理，统一监控，可以进行7×24小时全天候的实时监控
- 2、提供大容量和高并发码流
- 3、综合成本低，基于以太网
- 4、扩展能力强，组网灵活
- 5、NAS易于使用管理

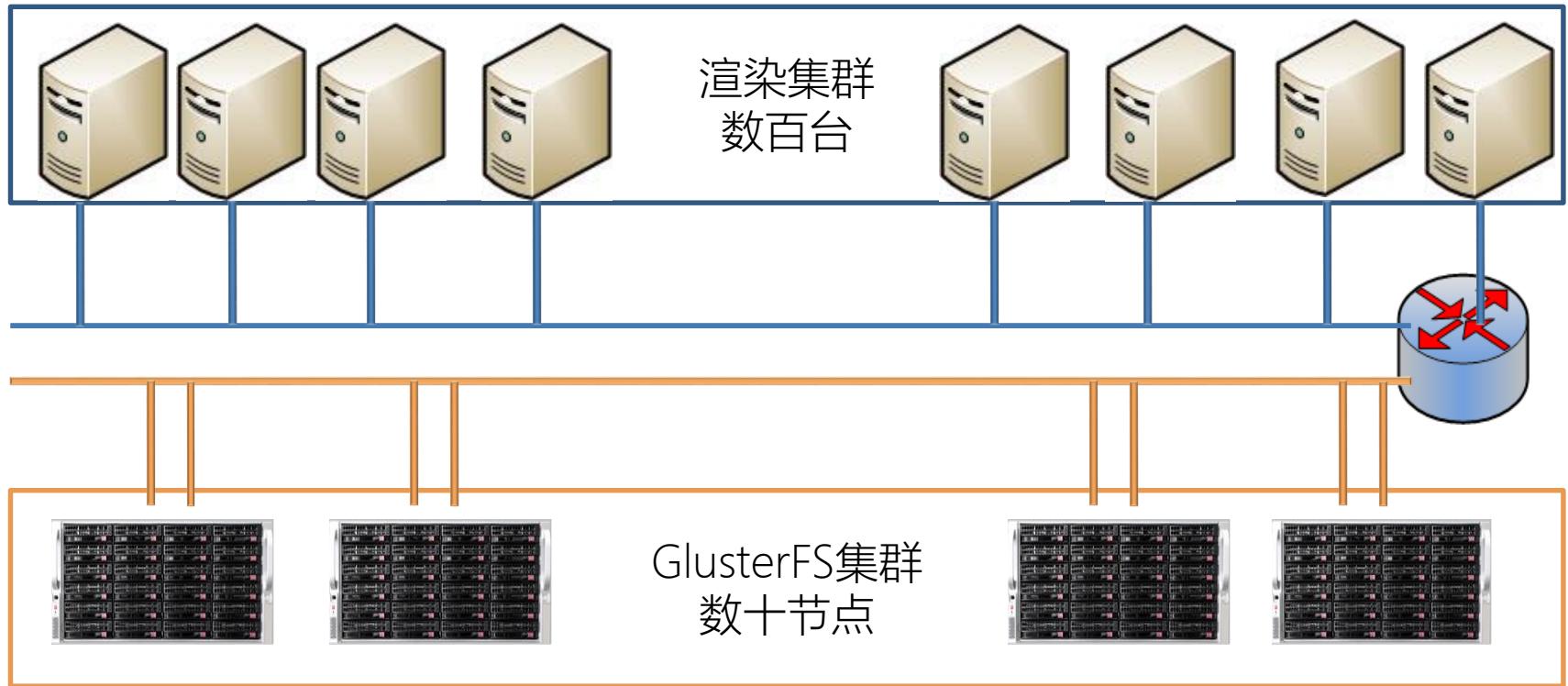
解决方案-视频点播



存储需求:

- 1、存储容量，可在线扩展至PB级
- 2、文件数量千万级，文件大小MB级以上
- 3、聚合带宽要求，NFS协议访问
- 4、单台注入和播出服务器带宽要求
- 5、并发视频流的服务能力要求

解决方案-视频渲染



存储需求：

- 1、存储系统可以平滑扩容，后期达到PB级；
- 2、文件数量达到数百万量级，多数为MB级文件；
- 3、百级/千级节点并发访问文件服务器；
- 4、保证数据可用性；

(四)
GlusterFS开放问题

GlusterFS 问题

- ❖ 元数据性能
- ❖ 海量小文件问题
- ❖ 集群管理模式
- ❖ 容量负载均衡
- ❖ 数据分布问题
- ❖ 数据可用性问题
- ❖ 数据安全问题
- ❖ Cache一致性问题
- ❖ 详细情况：

<http://blog.csdn.net/liuaigui/article/details/20941159>

3.7 开发计划

- ◆ 小文件性能优化
- ◆ SSD Cache/Tier
- ◆ 回收站功能
- ◆ 基于策略的Split-brain解决方法
- ◆ Rebalance性能改善

- ◆ 详细情况：
<http://www.gluster.org/community/documentation/index.php/Planning37>

4.0 开发计划

- ◆ 千级规模集群支持
- ◆ 弹性DHT 2.0
- ◆ Stripe 2.0
- ◆ 一致性客户端Cache
- ◆ 灵活副本

- ◆ 详细情况：

<http://blog.csdn.net/liuaigui/article/details/17314801>

GlusterFS未来发展

- ❖ Ceph一统天下？？？
- ❖ Lustre继续独占鳌头HPC
- ❖ MooseFS前途堪忧
- ❖ GlusterFS前途光明(道路曲折？)
 - ❖ 大道至**简**，Keep It as Simple and Stupid
 - ❖ **文件**存储，云存储，海量小文件
 - ❖ 弹性，扩展性，灵活性
 - ❖ RAS-P特征

Q & A