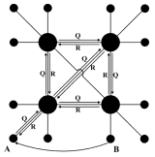
DATAGURU 炼数成金

Leader-us系列



大型分布式系统案例实战 第3周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

分布式系统之文件篇

- Linux文件系统的进化
- 传统的分布式文件系统
- 新型分布式文件系统
- 互联网领域中的小文件系统

Linux文件系统的进化

两种文件系统
日志型的与非日志型的

日志型文件系统

- ext3 许多流行的Linux发行版默认的文件系统
- ext4 由ext3增加许多显著特性和扩展进化而来的文件系统
- ReiserFS 全新设计的文件系统
- JFS IBM移植的UNIX文件系统
- XFS 为高性能和大文件设计的文件系统
- Btrfs 校检copy-on-write(写入时复制)文件系统

非日志型文件系统

ext2、FAT、VFAT、HPFS（OS/2）、NTFS、Sun的UFS等。

ext4 6 ----→ XFS 7

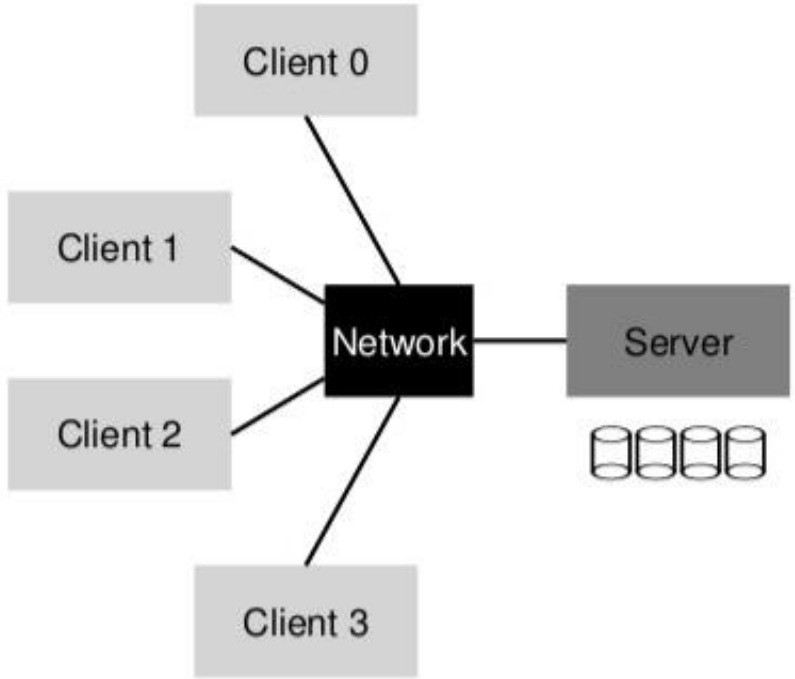
Linux文件系统的进化

当前文件系统中的几个强者

Facebook已经开始全线换用btrfs
红帽的Red 6使用ext4，7则使用了XFS
浙江省十二五重大科技专项资助项目研究了ZFS在基于Hadoop的视频存储系统中的应用

传统的分布式文件系统

经典的第一代分布式文件系统NFS



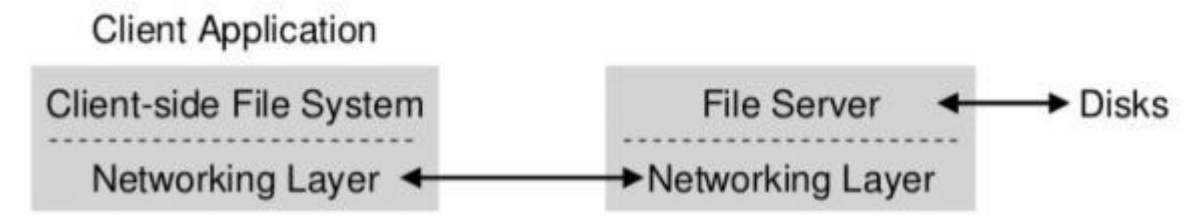
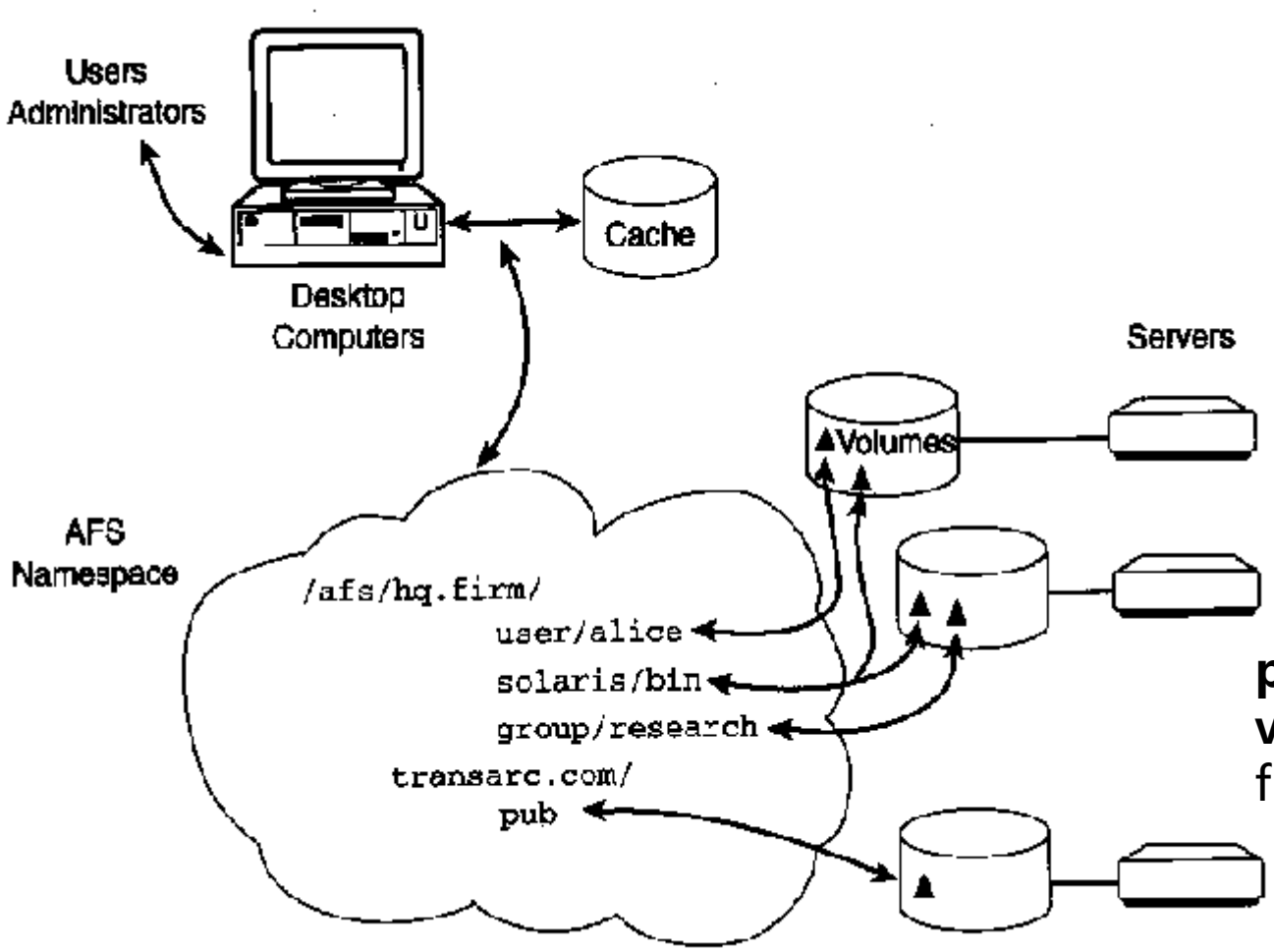


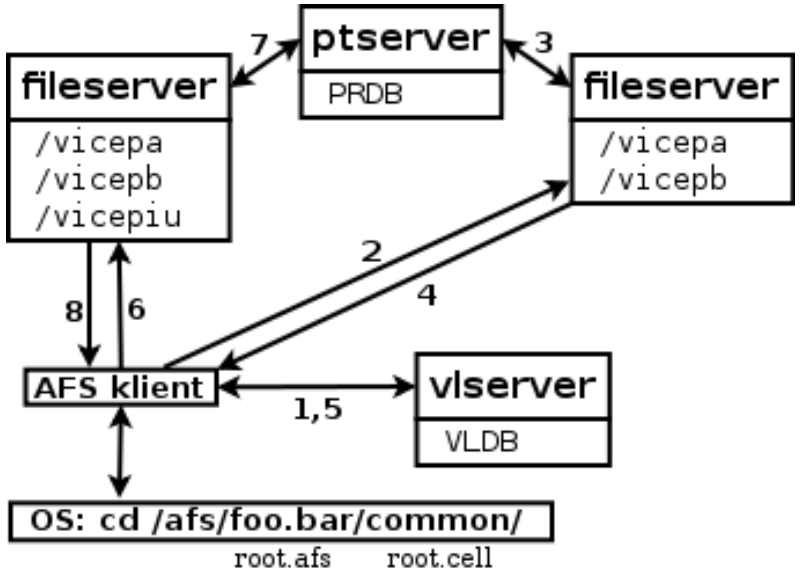
Figure 48.2: **Distributed File System Architecture**

一个分布式文件系统中有两种不同的软件：客户端文件系统和文件服务器。它们的行为共同决定分布式文件系统的行为

传统的分布式文件系统

对后来的分布式文件系统有重要影响的AFS

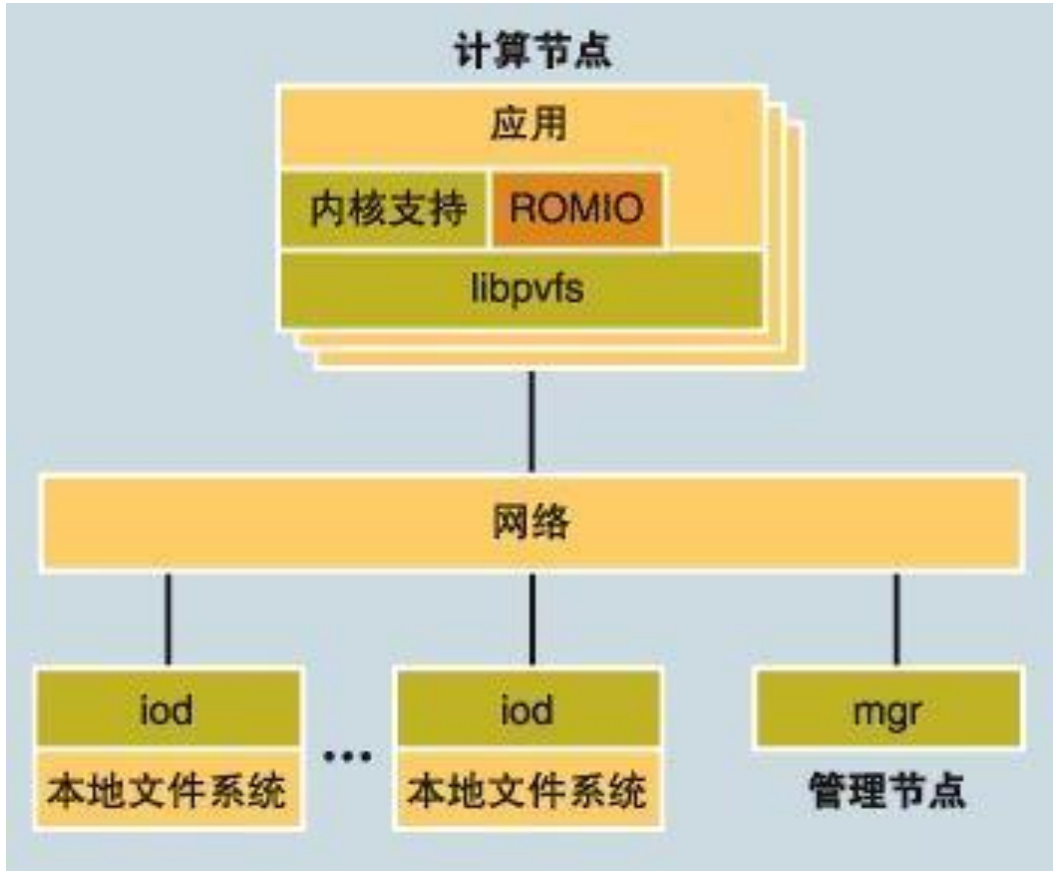




ptserver — 负责用户和鉴权
vlserver — 负责卷位置记录和查询
fileserver —负责处理位于 AFS 卷内的文件和目录

新型分布式文件系统

PVFS，Linux开源的并行虚拟文件系统



管理节点：即元数据服务器，负责管理所有的文件元数据信息；

I/O节点：运行I/O服务器，负责分布式文件系统中数据的存储和检索；

计算节点：处理应用访问，通过PVFS专有的libpvfs接口库，从底层访问PVFS服务器。

新型分布式文件系统

Lustre，专门针对高性能计算的基于对象存储的分布式文件系统

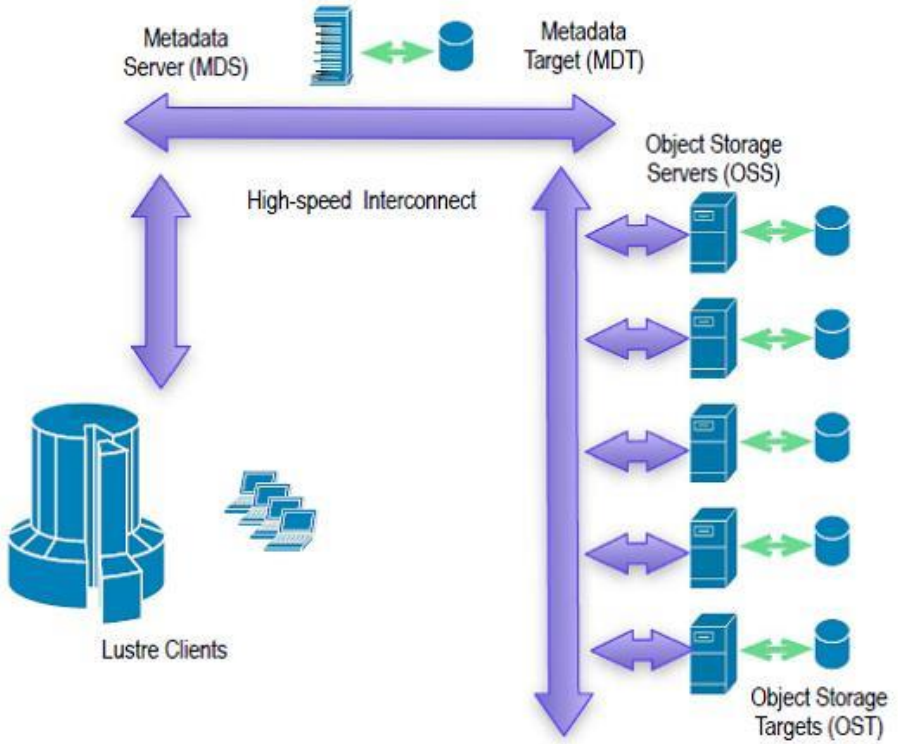
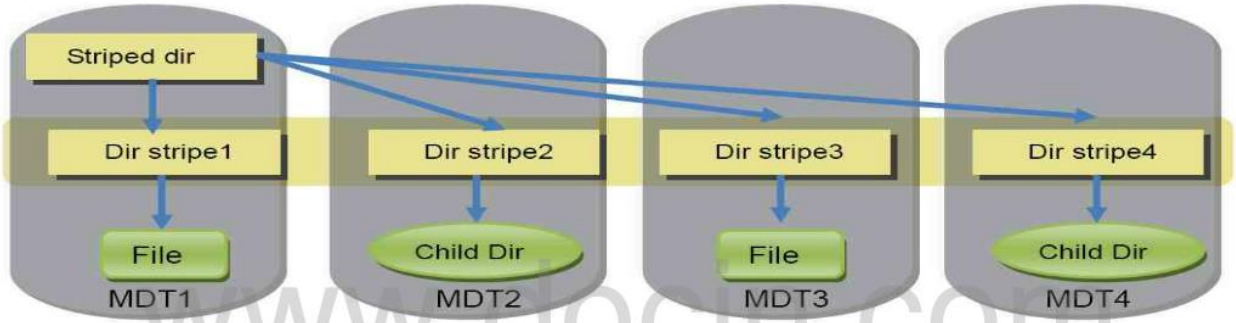


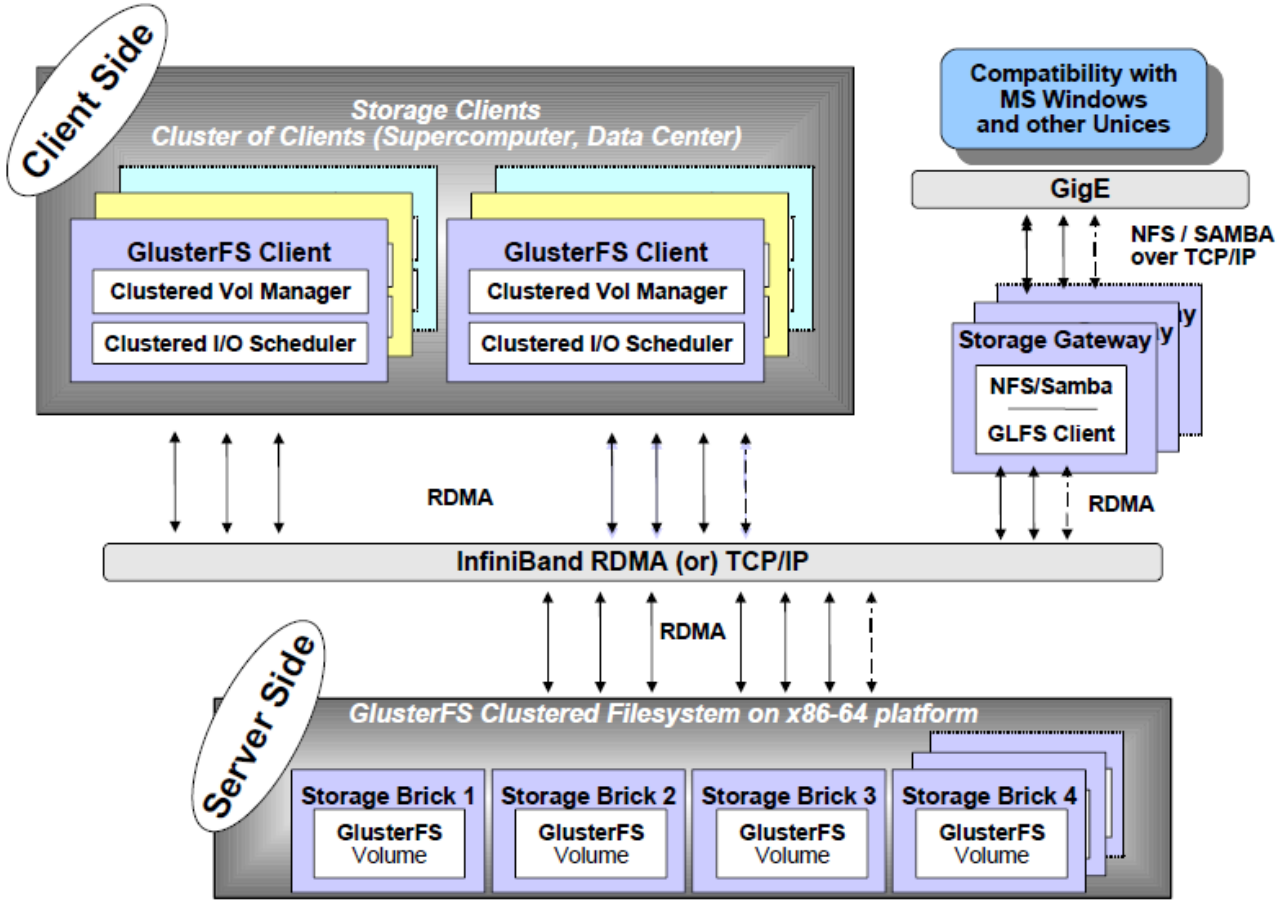
Figure 1: Lustre components.

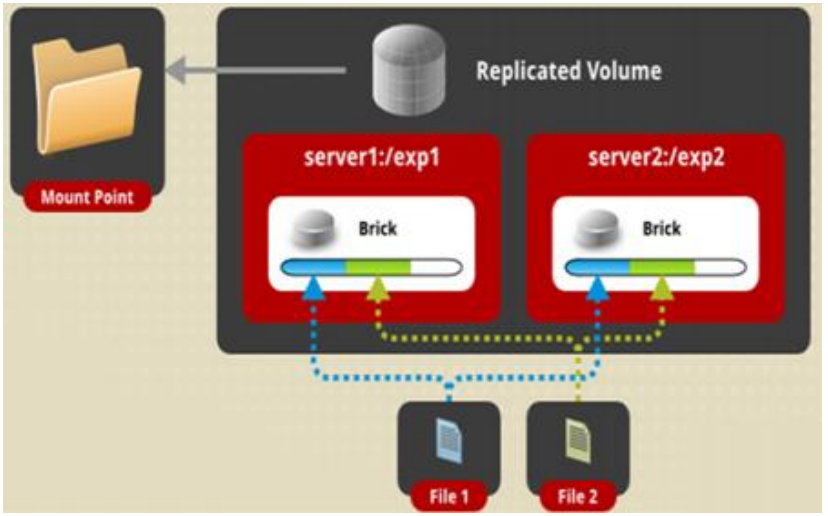
条带化



新型分布式文件系统

gluster fs:代替Lustre的开源的分布式文件系统





新型分布式文件系统

Google文件系统（GoogleFS）

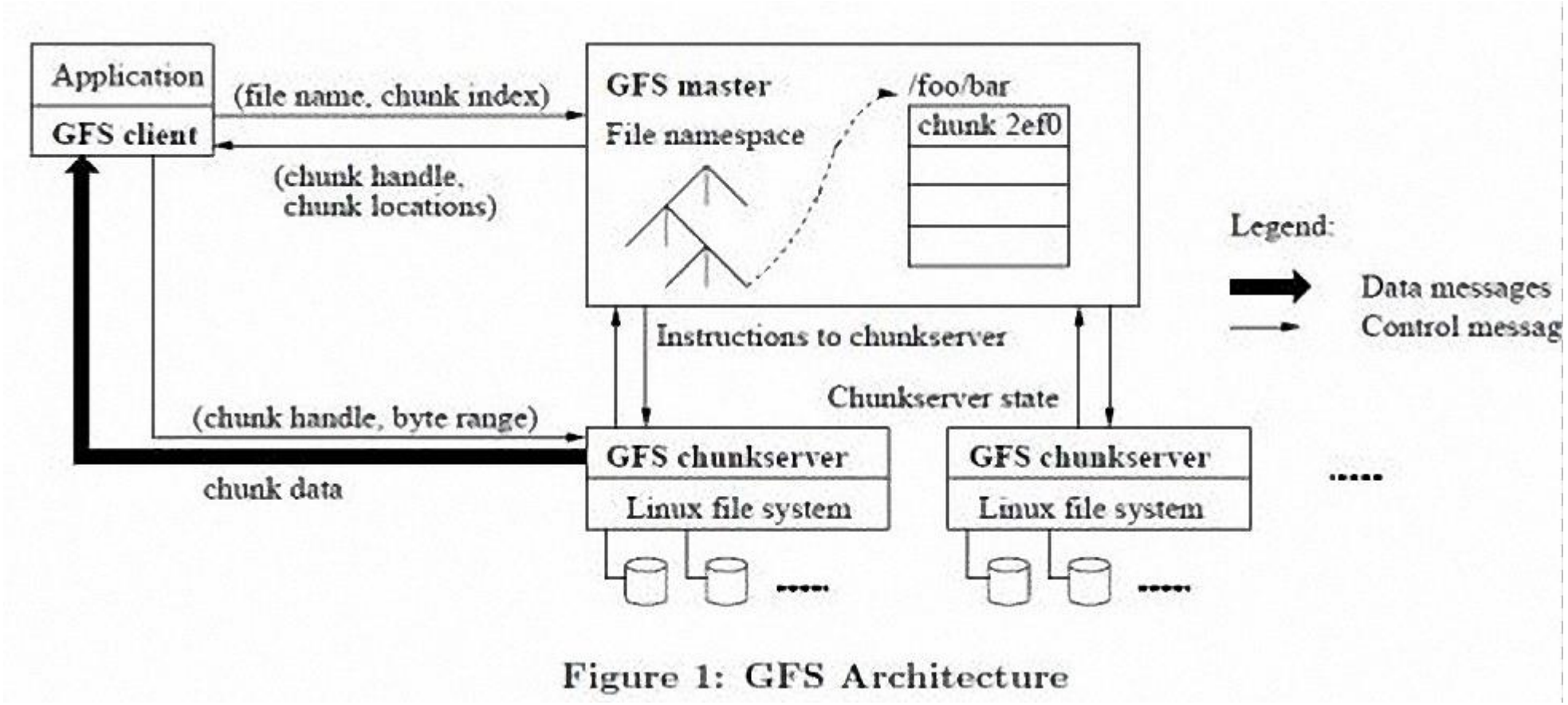
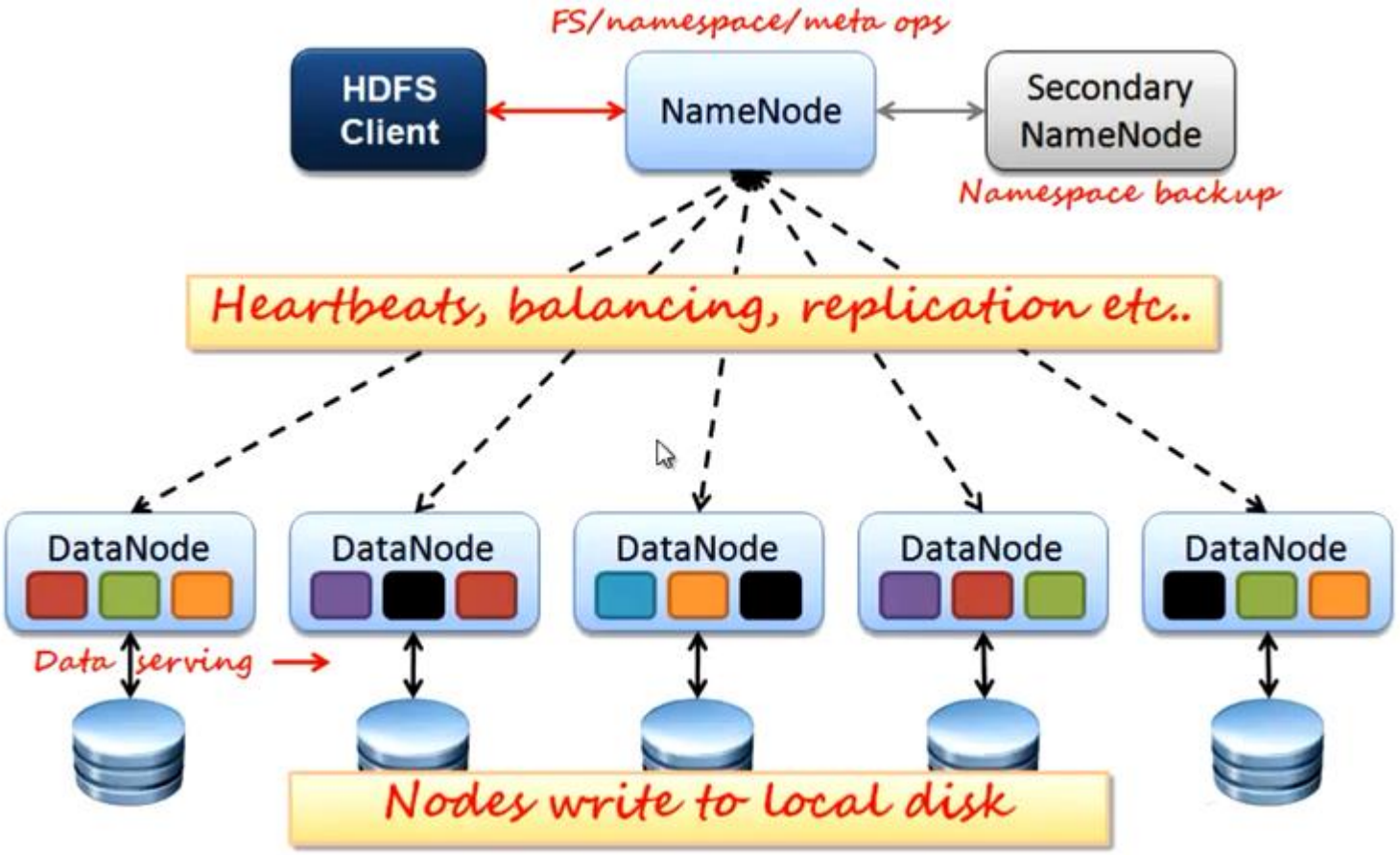


Figure 1: GFS Architecture

新型分布式文件系统

HDFS（Hadoop）

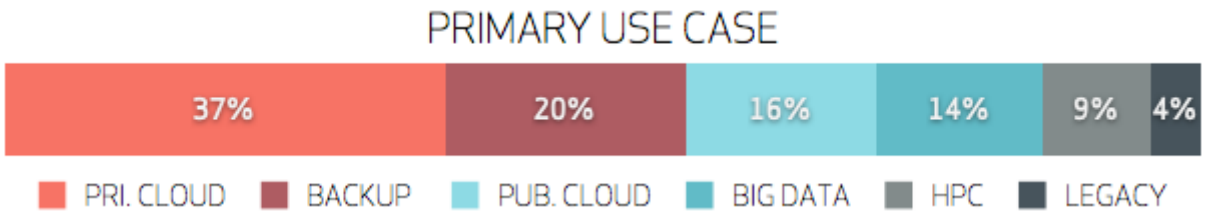


新型分布式文件系统

ceph：新一代的复合型分布式文件系统

Ceph是统一存储系统，支持三种接口。

- Object：有原生的API，而且也兼容Swift和S3的API
- Block：支持精简配置、快照、克隆
- File：Posix接口，支持快照



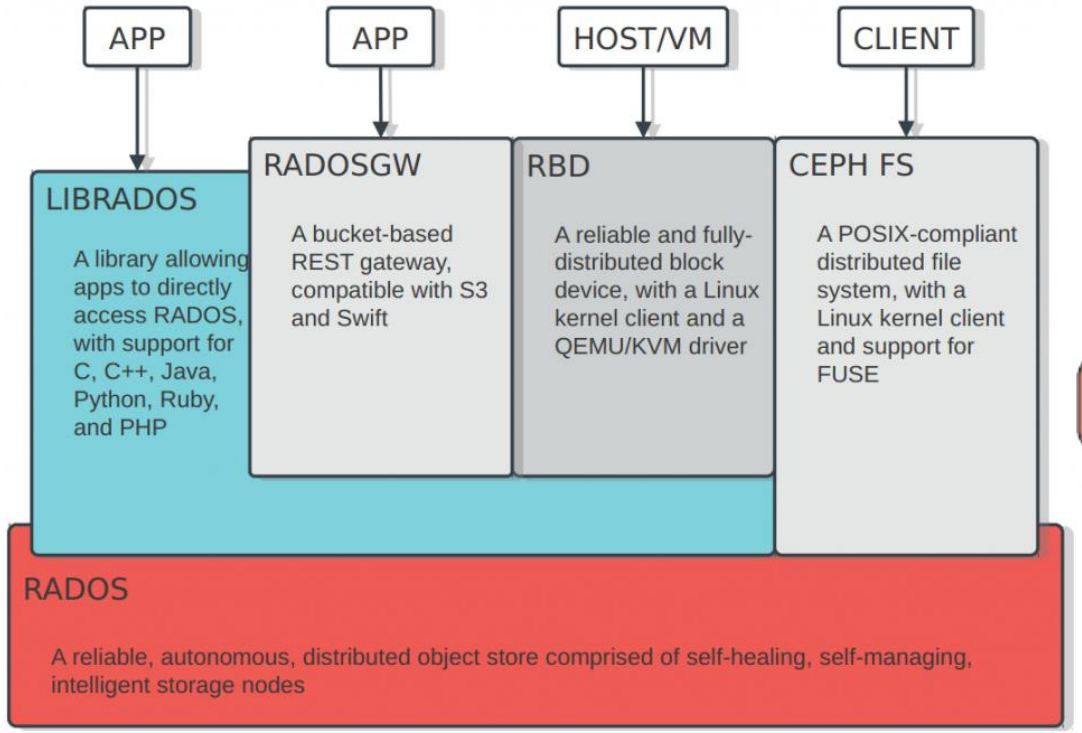
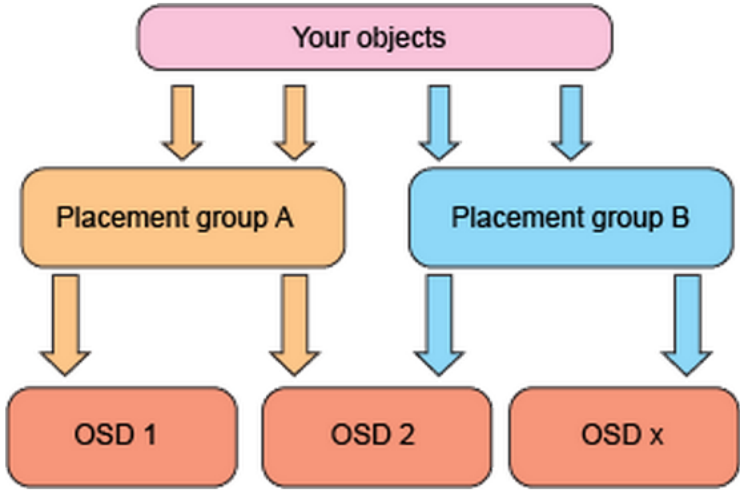
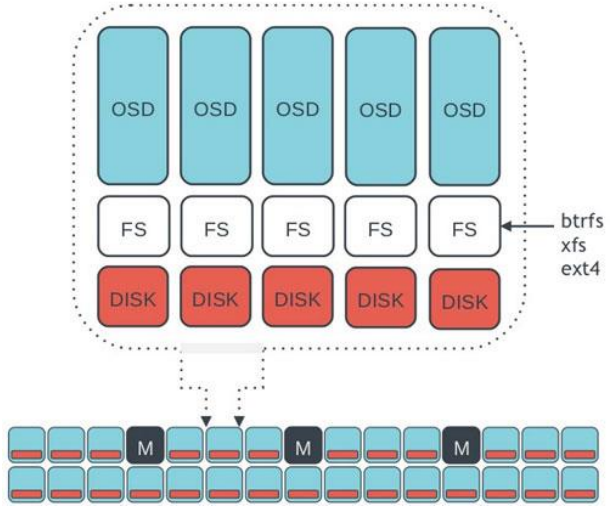


图 3. RADOS 位置分组

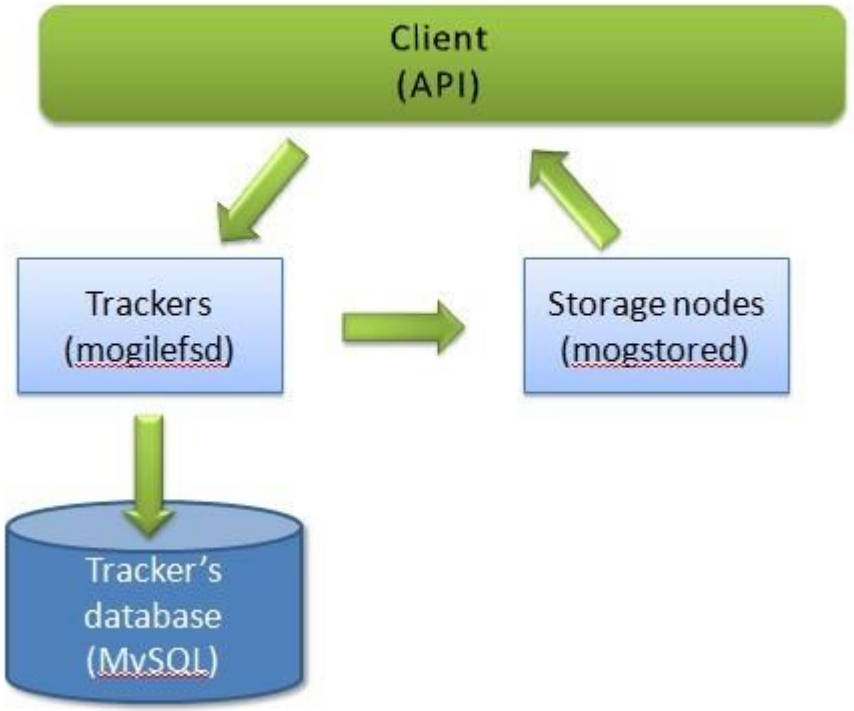


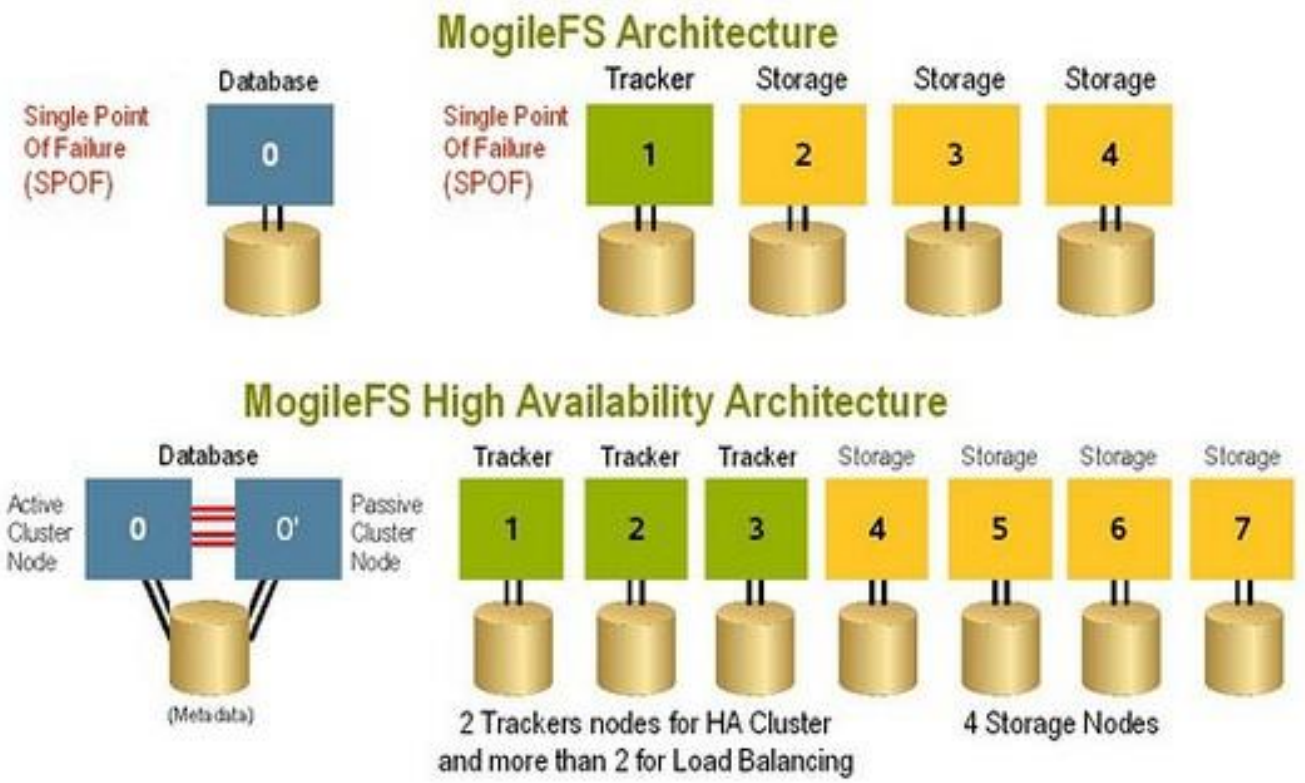


互联网领域中的小文件系统

MogileFS——影响最大的互联网小文件系统
FastDFS——穷人的解决方案（国产小有名气）
TFS——淘宝的HDFS Copy版本
GridFS——

MogileFS架构和原理



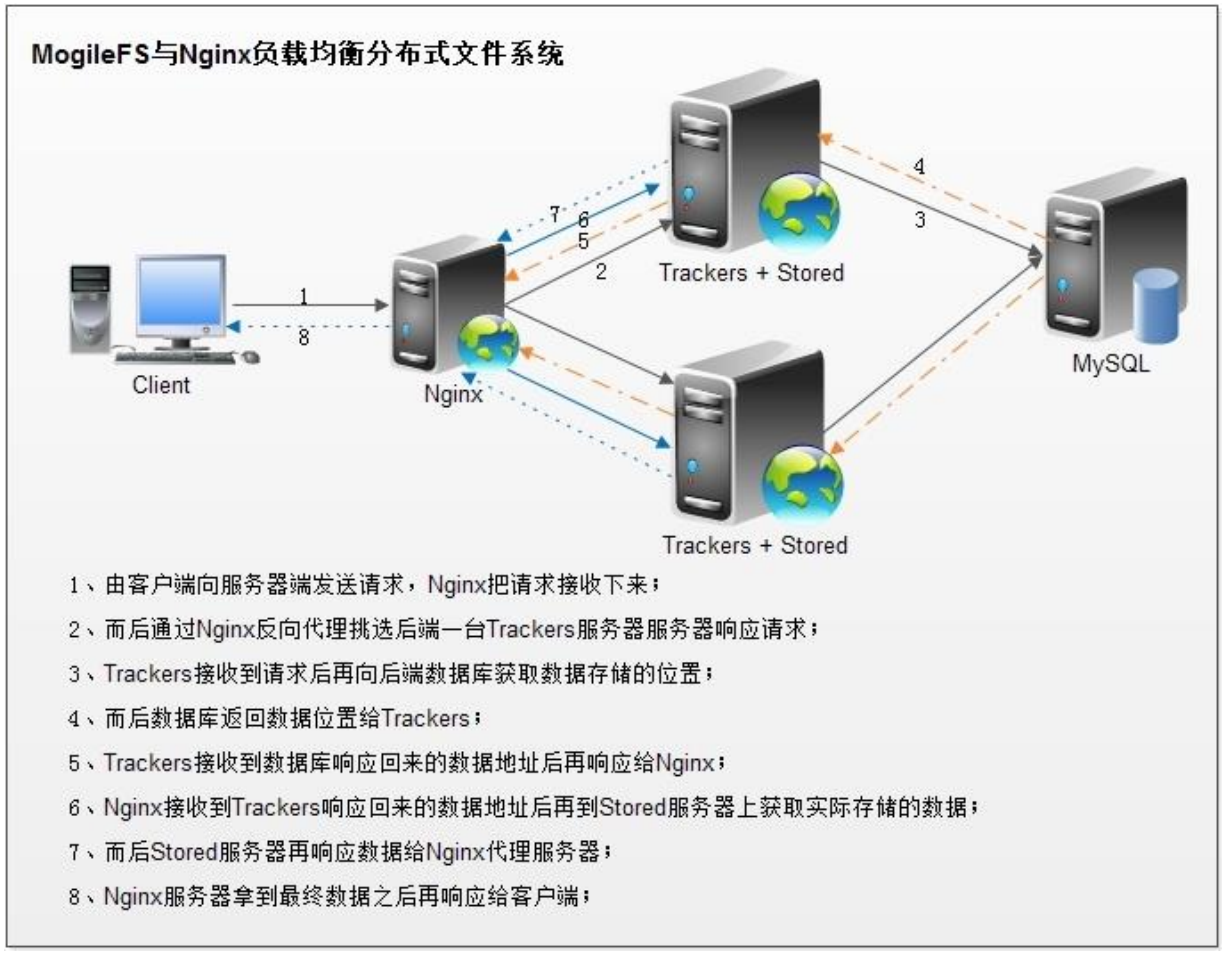


在MogileFS分布式文件存储系统中，文件通过 KEY 来引用，KEY 在同一个domain（存储域）下是唯一的，每个存储域可以定义不同的文件类别Class，可以针对不同的存储类别设置存储不同份数的文件副本。

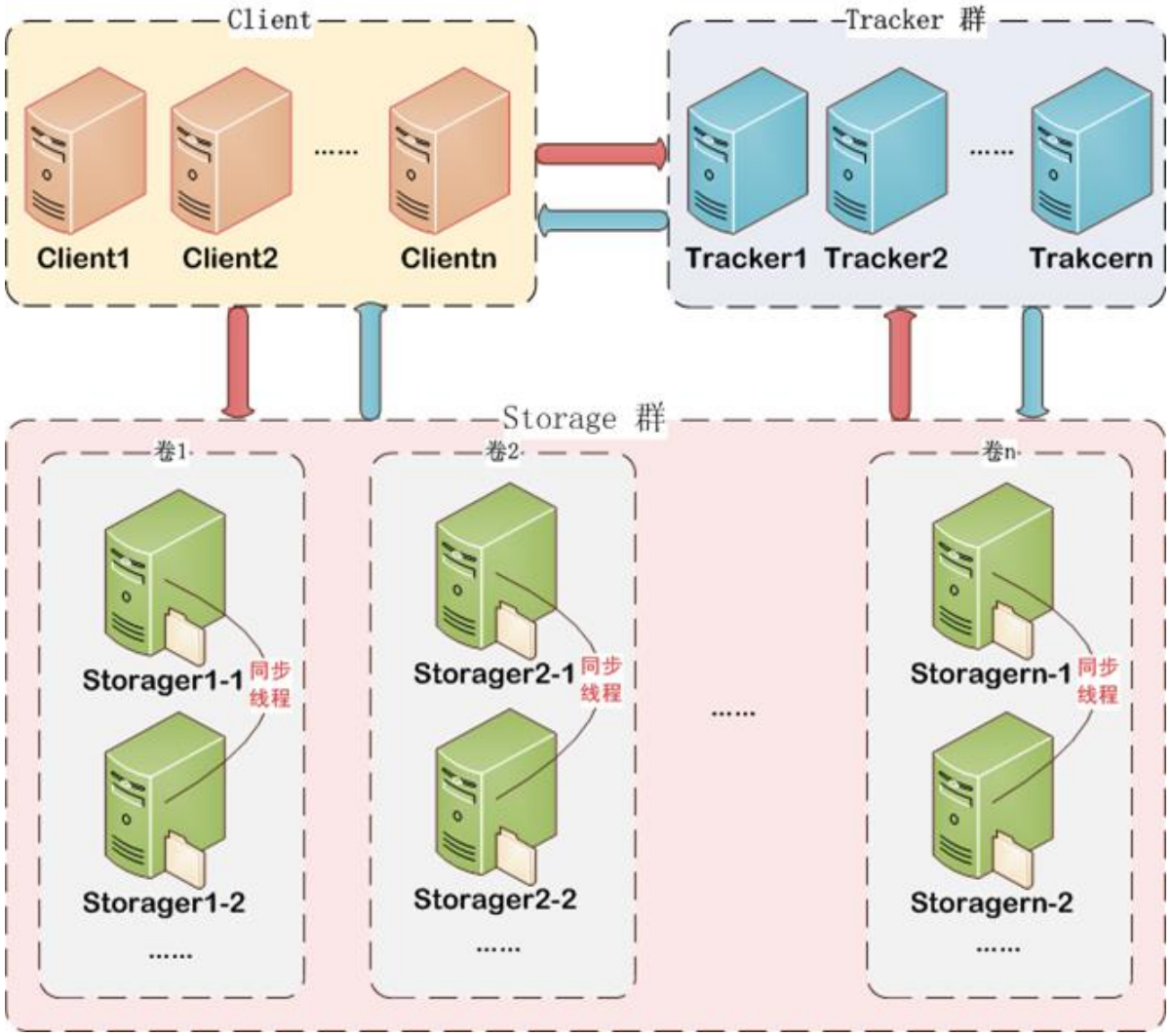
MogileFS的特性

•**应用层** — 不需要特殊的核心组件
•**无单点失败** — MogileFS分布式文件存储系统安装的三个组件（存储节点、跟踪器、跟踪用的数据库）,均可运行在多个 机器上,因此没有单点失败， 推荐至少两台机器。
•**自动的文件复制** — 基于不同的文件“分类”,文件可以被自动的复制到多个有足够存储空间的存储节点上,这样可以满足这个“类别”的最少复制要求.比如你有一个图片网站,你可以设置原始的JPEG图片需要复制 至少三份,但实际只有1or2份拷贝,如果丢失了数据,那么MogileFS分布式文件存储系统可以重新建立遗失的拷贝数
• **“比RAID好多了”** —RAID磁盘是冗余的,但主机不是,如果你整个机器坏了,那么文件也将不能访问. MogileFS分布式文件存储系统在不同的机器之间进行文件复制,因此文件始终是可用的.
•**不需要RAID** — 在MogileFS中的磁盘可以是做了RAID的也可以是没有,如果是为了安全性着想的话RAID没有必要买了,因为MogileFS分布式文件存储系统已经提供了.
•**简单的命名空间** –文件通过一个给定的key来确定,是一个全局的命名空间
•**不用共享任何东西** — MogileFS分布式文件存储系统不需要依靠昂贵的SAN来共享磁盘,每个机器只用维护好自己的磁盘.
•**传输中立,无特殊协议** — MogileFS分布式文件存储系统客户端可以通过NFS或HTTP来和MogileFS的存储节点来通信,但首先需要告知跟踪器一下.

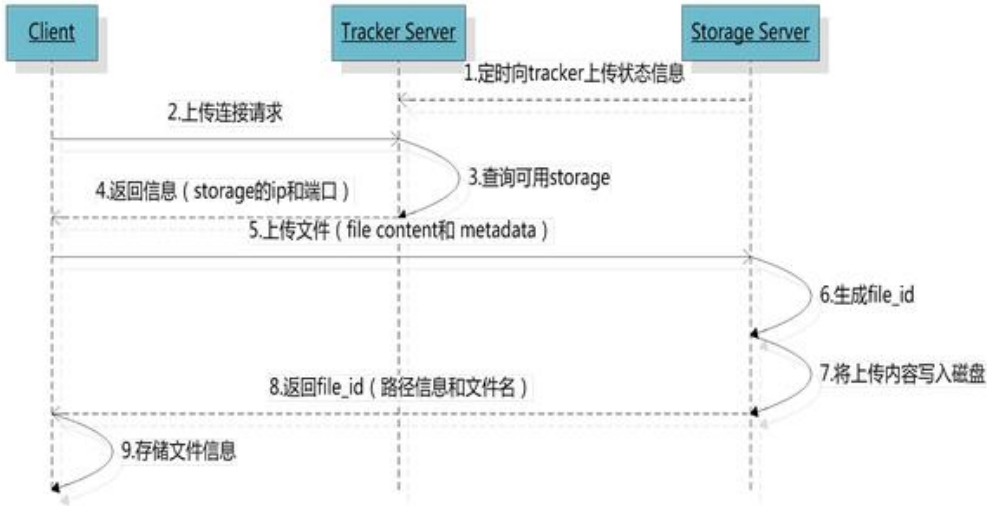
MogileFS+Nginx的负载均衡部署方案



FastDFS架构和原理



上传机制：

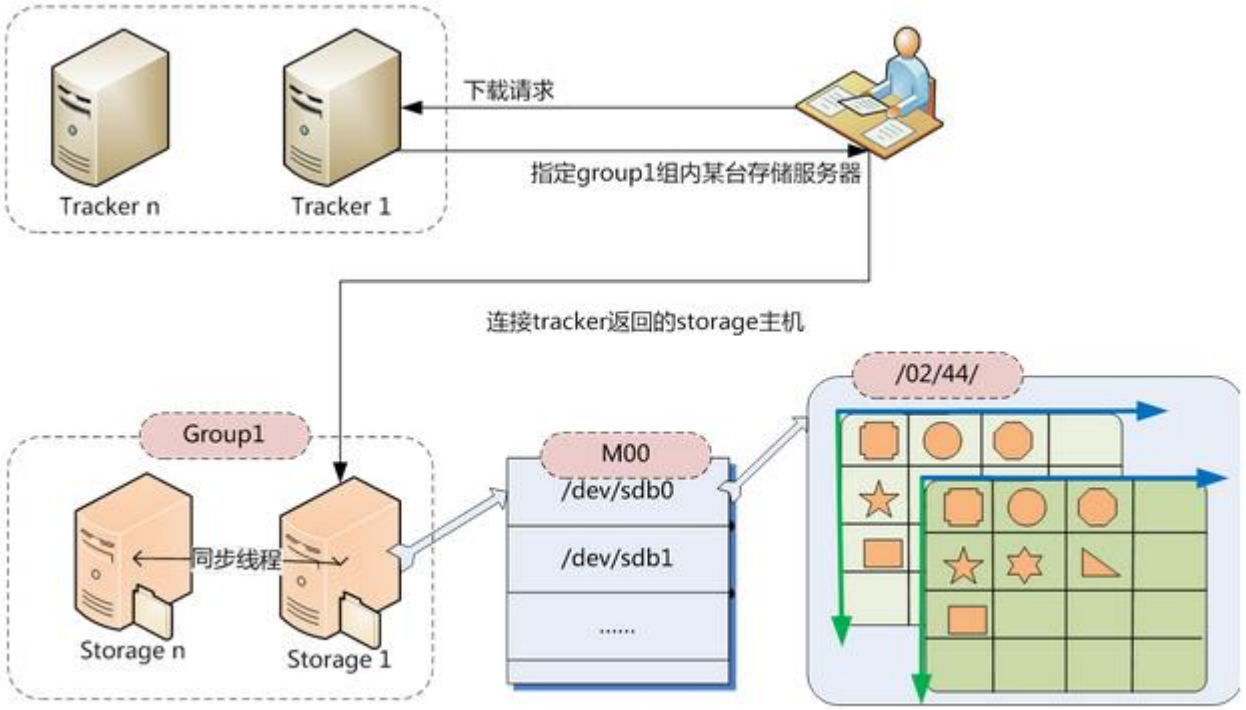


每个存储服务器都需要定时将自身的信息上报给tracker，这些信息就包括了本地同步时间（即，同步到的最新文件的时间戳）。而tracker根据各个存储服务器的上报情况，就能够知道刚刚上传的文件，在该存储组中是否已完成了同步

FastDFS的FID

精巧的FID：

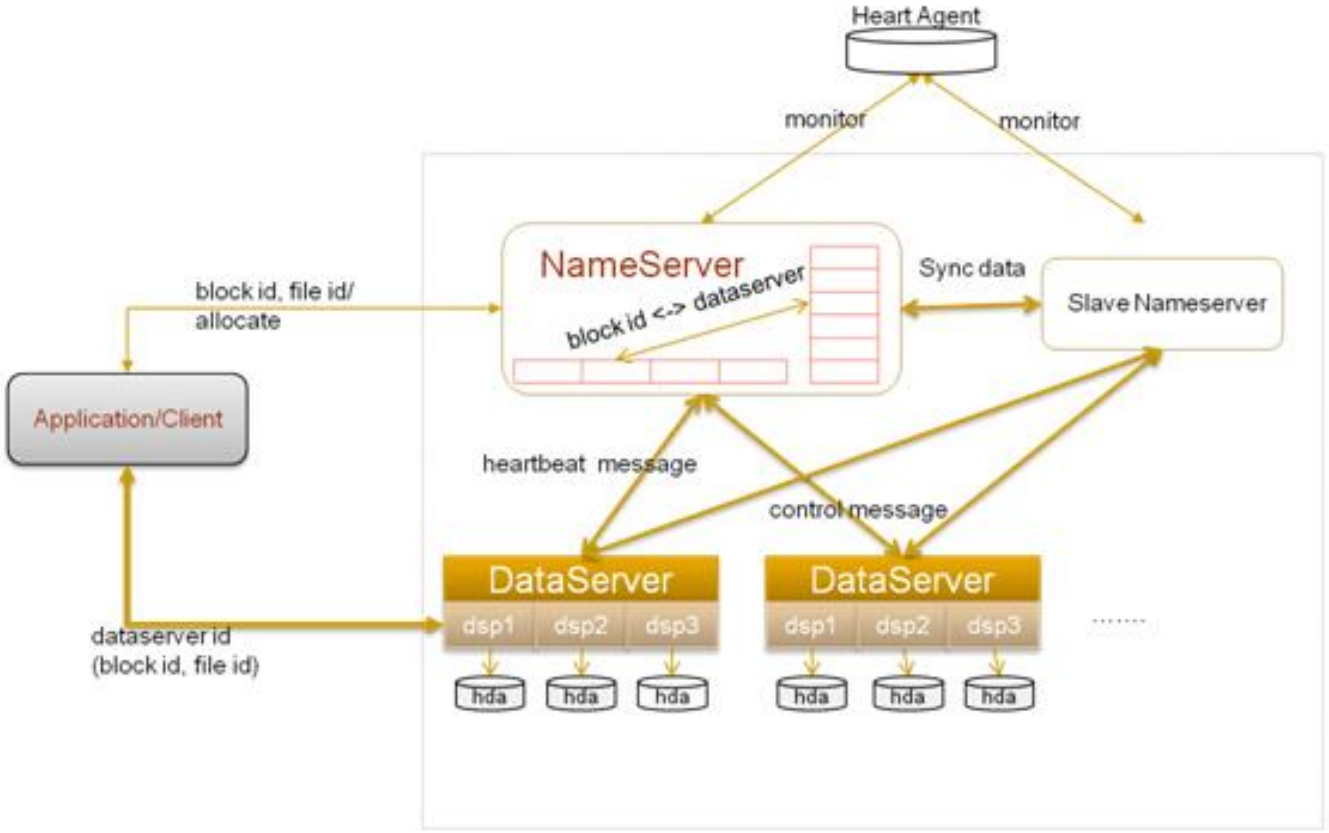
group1 /M00 /02/44/ wKgDrE34E8wAAAAAAAAGkEIYJK42378.sh



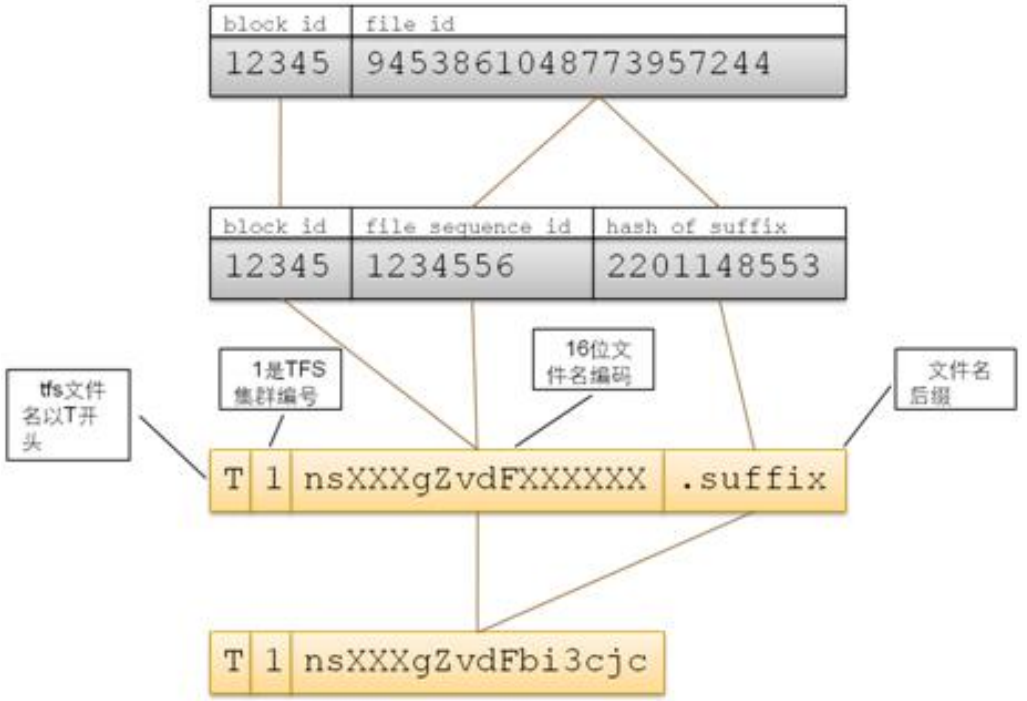
FastDFS的特性

- FastDFS和MogileFS相比，没有文件索引数据库，C语义开发，TCP Socket方式，整体性能更高
- 相对于MogileFS更为简单
- FastDFS的日志记录非常详细，便于排除问题
- 安装配置相对简单
- 目前只有一个人维护，是潜在的风险

TFS原理和架构



TFS文件名的结构

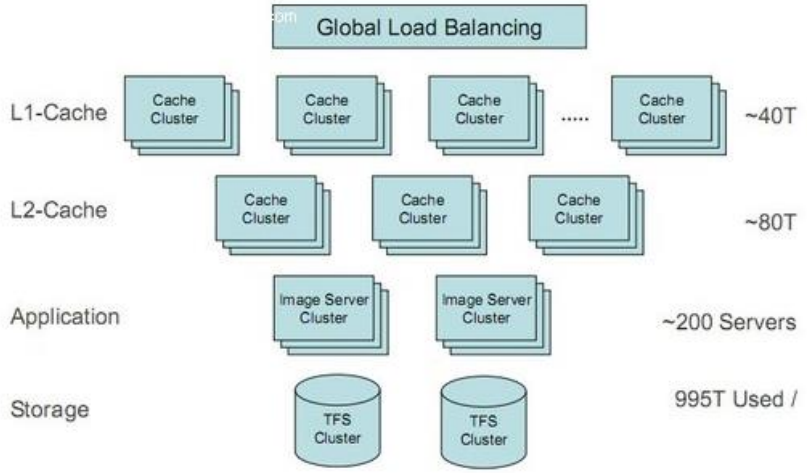


TFS的特性

- 总体参考HDFS架构和原理，细节方面则考虑的是小文件(<1M)的优化访问
- 在TFS的设计里面对应着是一个block同时只能有一个写或者更新操作。
- 随着写压力的增加，读文件的TPS会大幅下滑。

TFS的淘宝部署案例

淘宝网图片存储与处理系统全局拓扑



图片服务器前端还有一级和二级缓存服务器，尽量让图片在缓存中命中，最大程度的避免图片热点，实际上后端到达TFS的流量已经非常离散和平均。

Thanks

FAQ时间