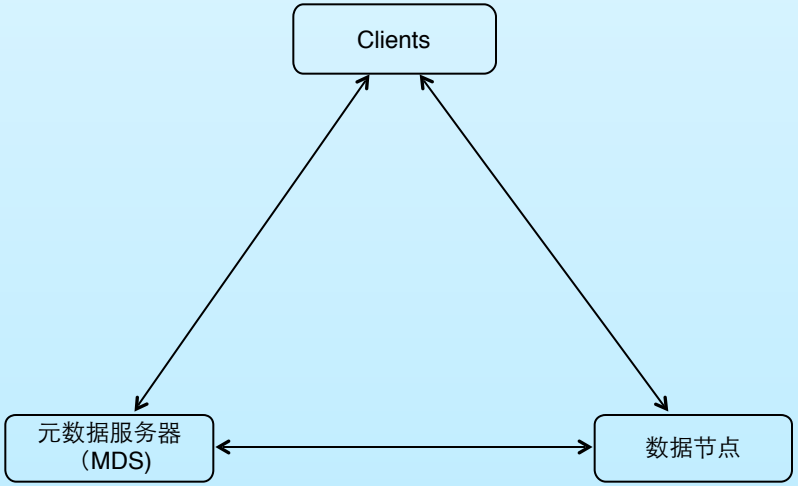
# 分布式文件系统

关键技术

# 分布式文件系统

非本地直连
通过网络连接

低成本*
高性能




现代高速网络互连技术
聚合/10GE/INIFIBAND

# 分类(1)

- C/S文件系统【NFS、CIFS】
  - 多客户端访问同一远程文件系统，文件系统本身不可扩展
  - 架构简单但两台服务器不能同时访问修改，扩展性差，性能有限
- 集群文件系统
- P2P文件系统
- 未来

# 分类(2)

- C/S文件系统
- 集群文件系统
  - 文件系统可扩展；非对称架构，有元数据系统
  - 常见三种架构管理
    - SAN共享存储架构（GPFS)：高性能，高成本，扩展性差
    - DAS直连存储架构：数据节点与数据节点混合部署；单点故障，采用副本保障可靠性／可用性
    - 并行文件系统架构： 元数据管理独立；客户端直接访问存储节点数据，高性能，低成本，扩展性好
- P2P文件系统
- 未来

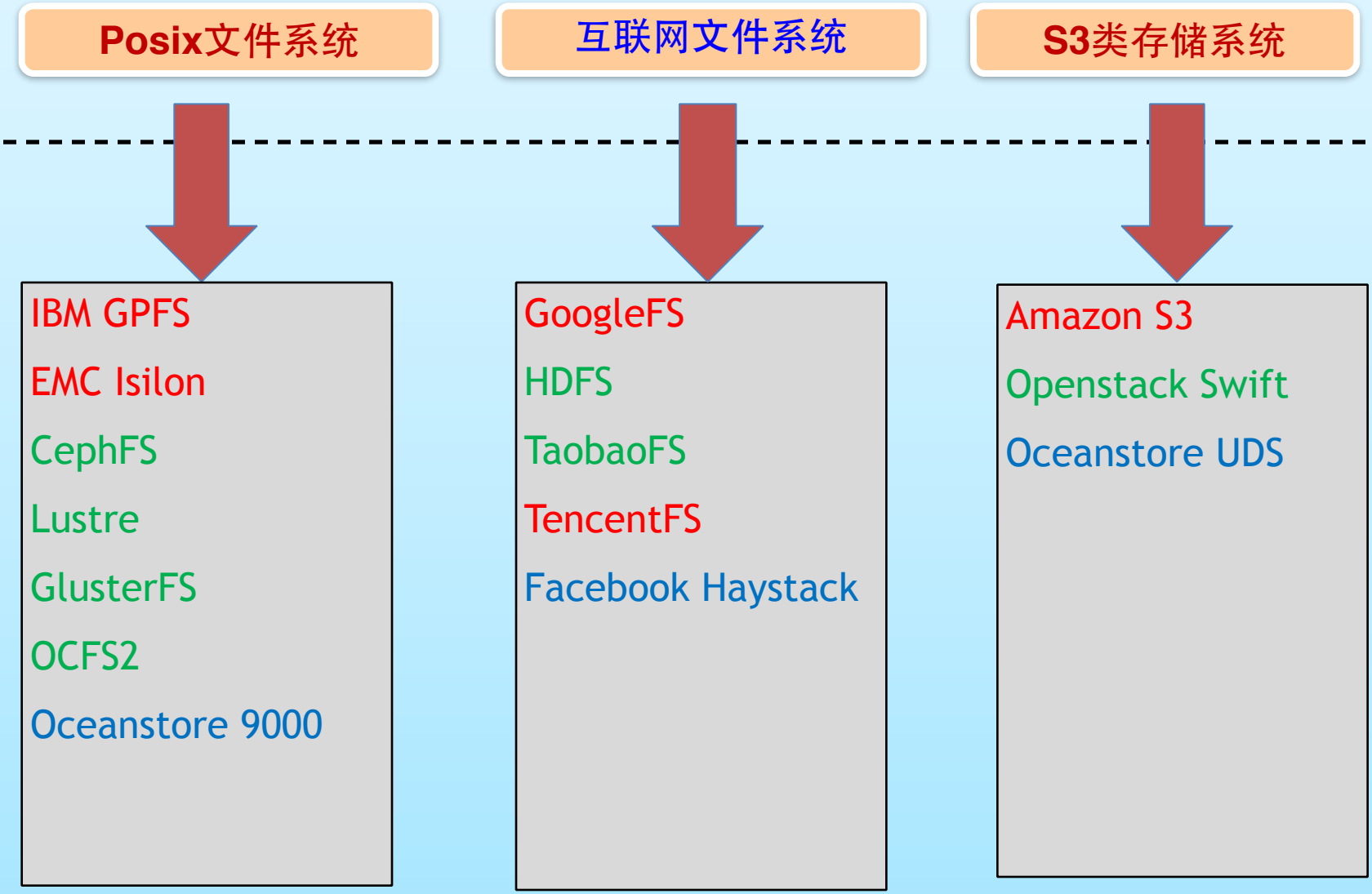
# 分类(3)

- C/S文件系统
- 集群文件系统
- P2P文件系统
  - 文件系统可扩展；对称架构，无元数据系统
  - 核心技术：去中心化访问
    - 快速资源定位技术<数据文件对应块寻址时间 us级>
      - 概率路由
      - Chord
      - Pastry
      - Tapetry
      - Byzantine Groups
- 未来
  - 介质(SSD/DRAM)、云、SDN

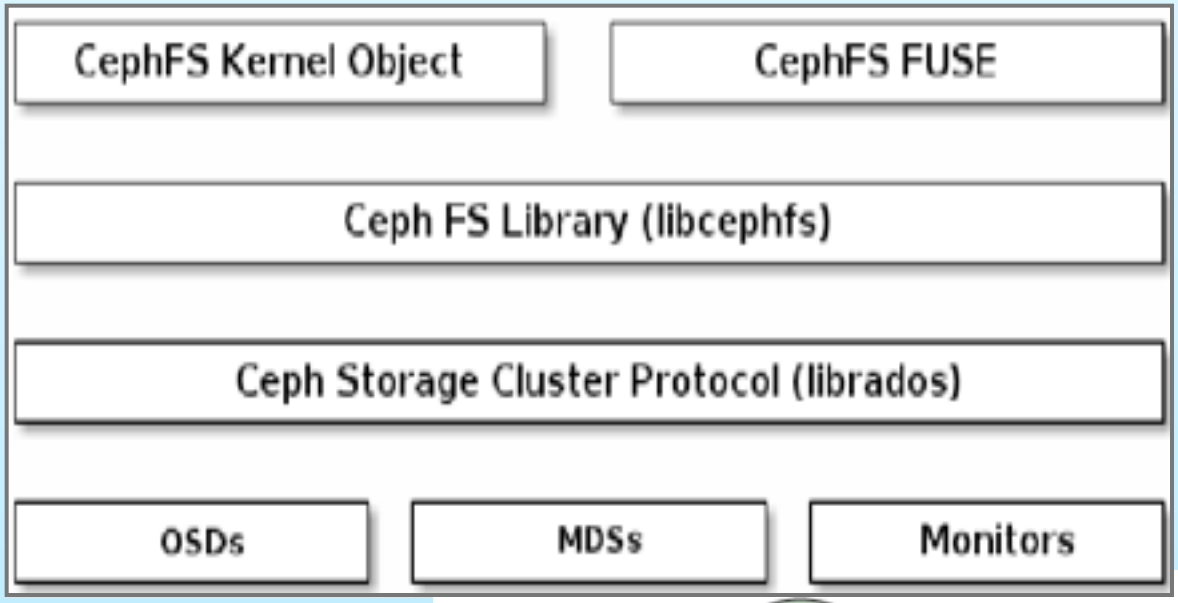
# 元数据管理(并行架构)

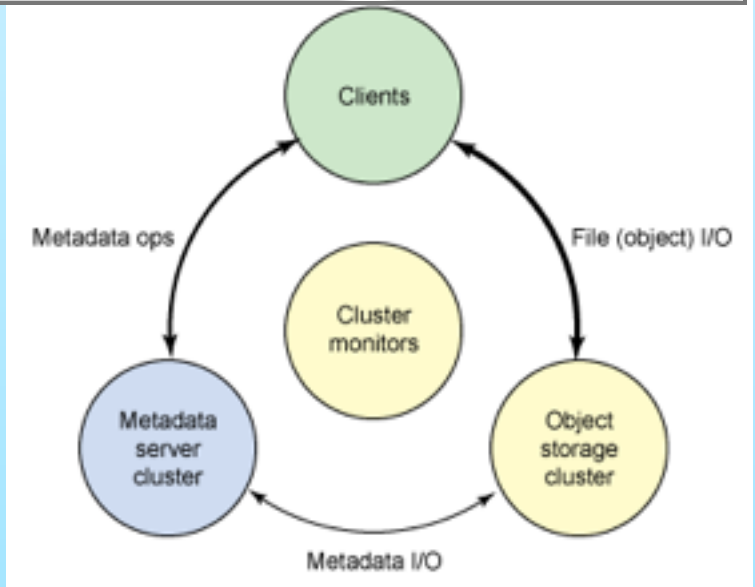
- 集中元数据管理
  - 元数据使用双控或者多控节点提供服务，存在单节点故障、扩展性问题
- 分布式元数据管理
  - 元数据采用全互联全冗余的组网机制，全对称分布式集群设计， 扩展性好
- 无元数据管理
  - 元数据服务使用动态子树逻辑分区执行，对变化工作负载进行动态调整、同时保留性能的位置，DC级扩展性

# 分布式文件系统


Posix文件系统
互联网文件系统
S3类存储系统
IBM GPFS
EMC Isilon
CephFS
Lustre
GlusterFS
OCFS2
Oceanstore 9000
GoogleFS
HDFS
TaobaoFS
TencentFS
Facebook Haystack
Amazon S3
Openstack Swift
Oceanstore UDS

# Ceph



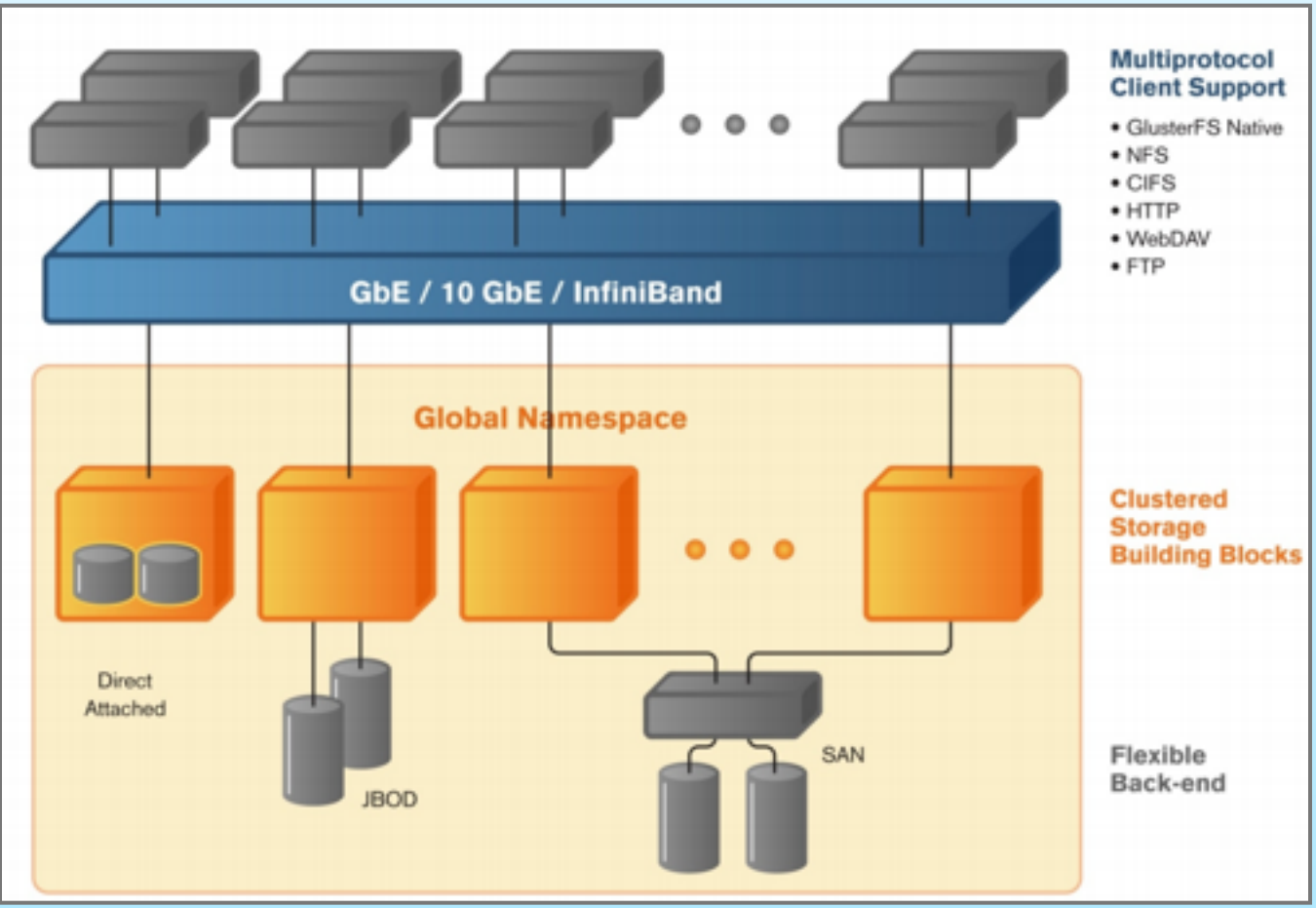




Ceph 特点：

1. Ceph同时支持块、文件、对象
2. 开源，融入Openstack体系
3. 多种接口：HDFS，NFS，Posix/Fuse
4. 动态子树分割
5. 支持快照

GPL

# GlusterFS

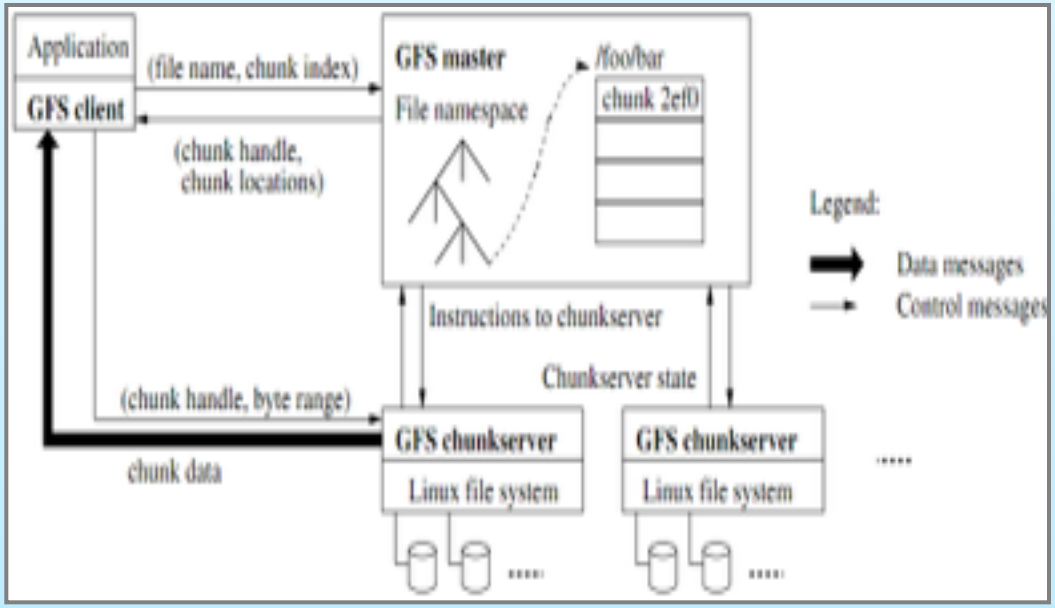


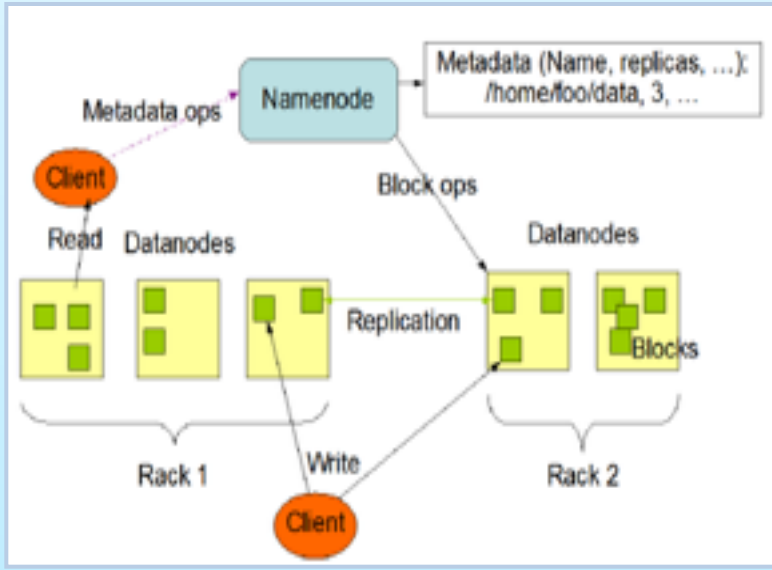

GlusterFS特点：

1. 多种接口，同时支持文件、对象
2. No metadata，Hash定位
3. 支持多种数据布局方式
4. 兼容多种异构存储

GPL

# GoogleFS/HDFS



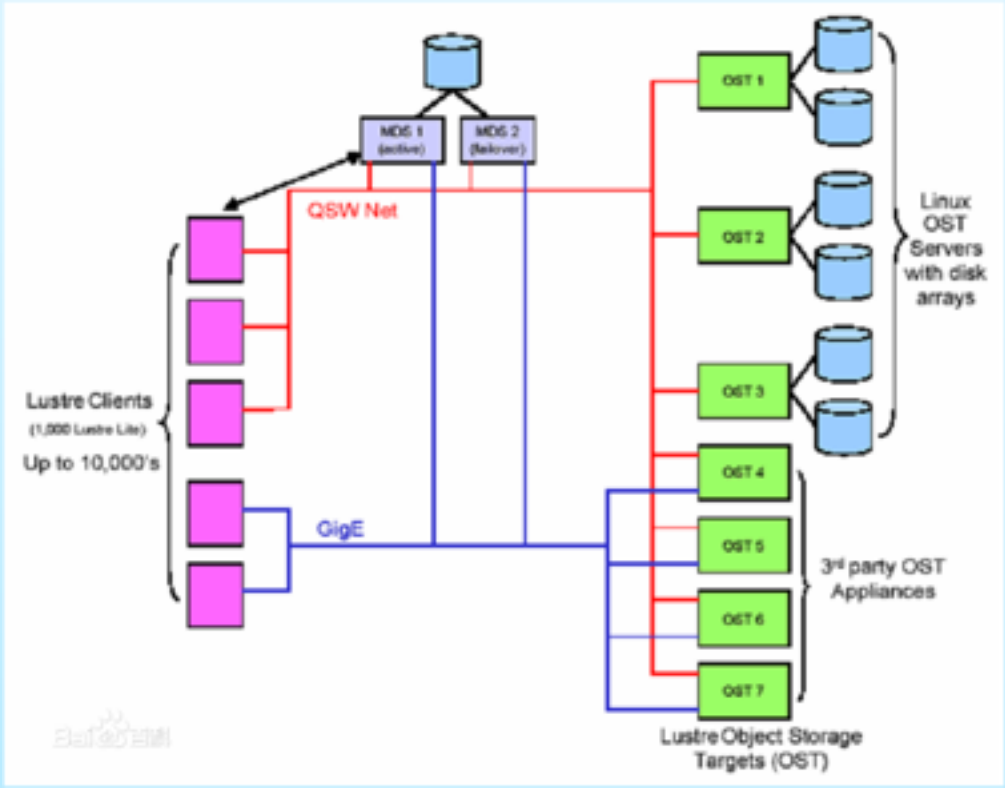




GoogleFS/HDFS特点：

1. 提供专用的私有API接口，传统业务无法运行
2. 互联网大数据的基础存储平台
3. 定位大数据场景，上层是BigTable/Hbase，MapReduce
4. 定位大文件，高带宽场景

Apache

# Lustre




唯一的命名空间

上万节点，

PB容量，

100GB/S

容错依靠RAID

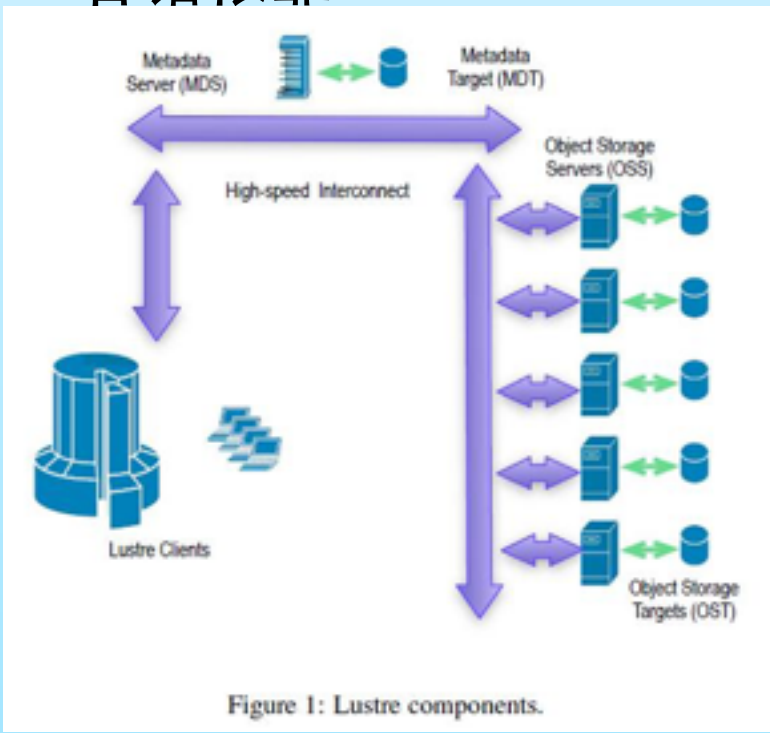
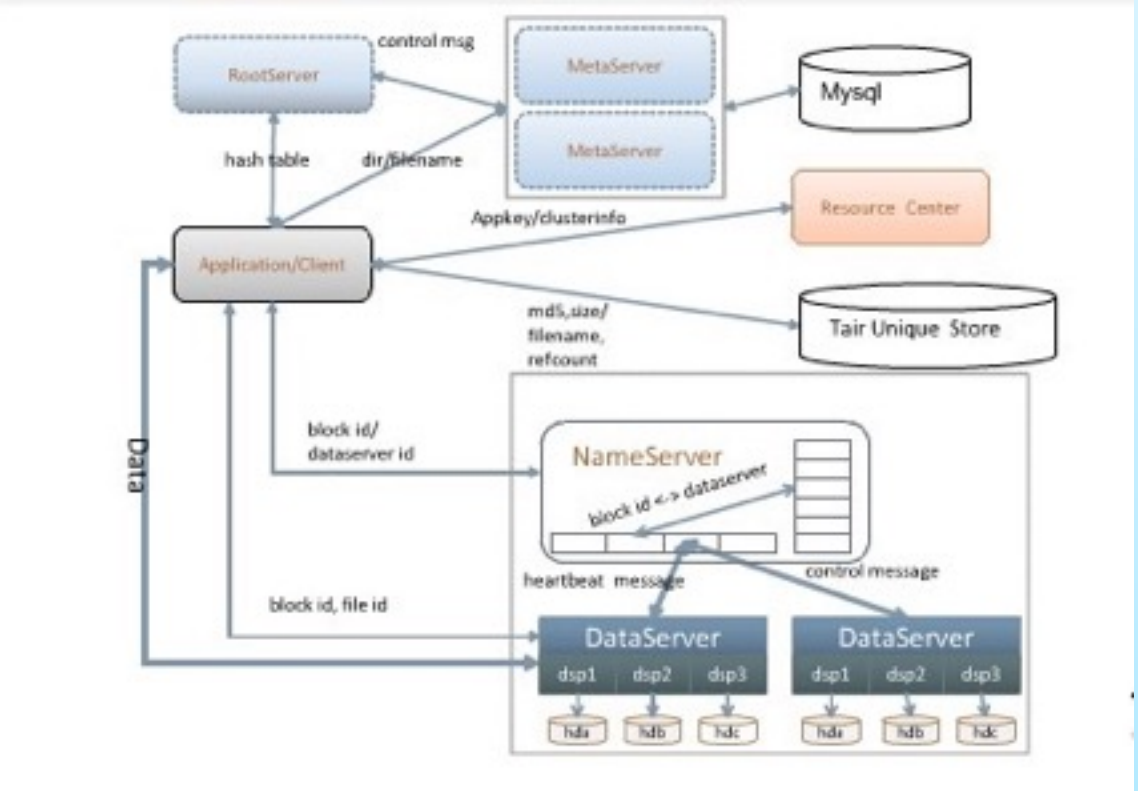



Figure 1: Lustre components.

GPL

# TaobaoFS

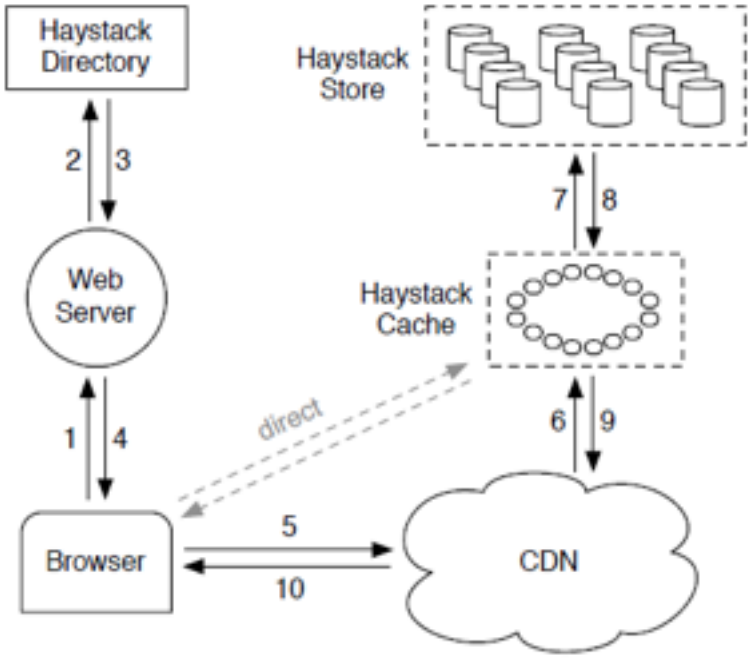


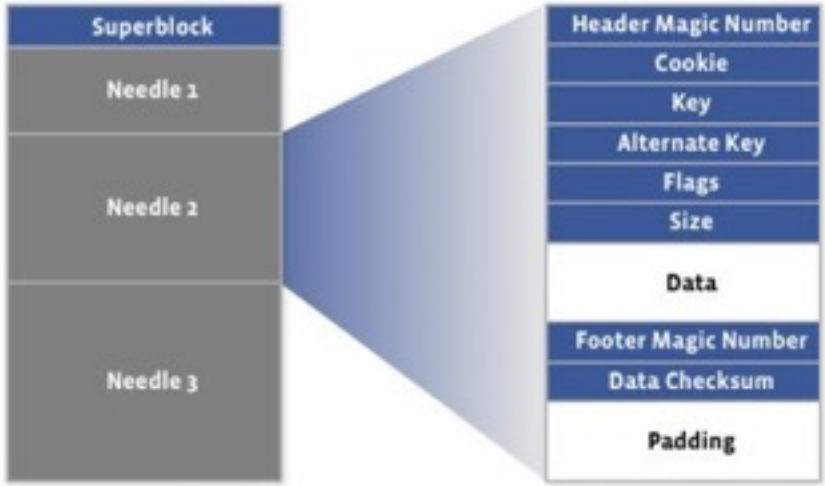

高可扩展
高可用
高性能
面向互联网服务
支持海量的非结构化数据

海量小文件／扁平化数据组织结构/有中心节点

GPL

# Facebook Haystack

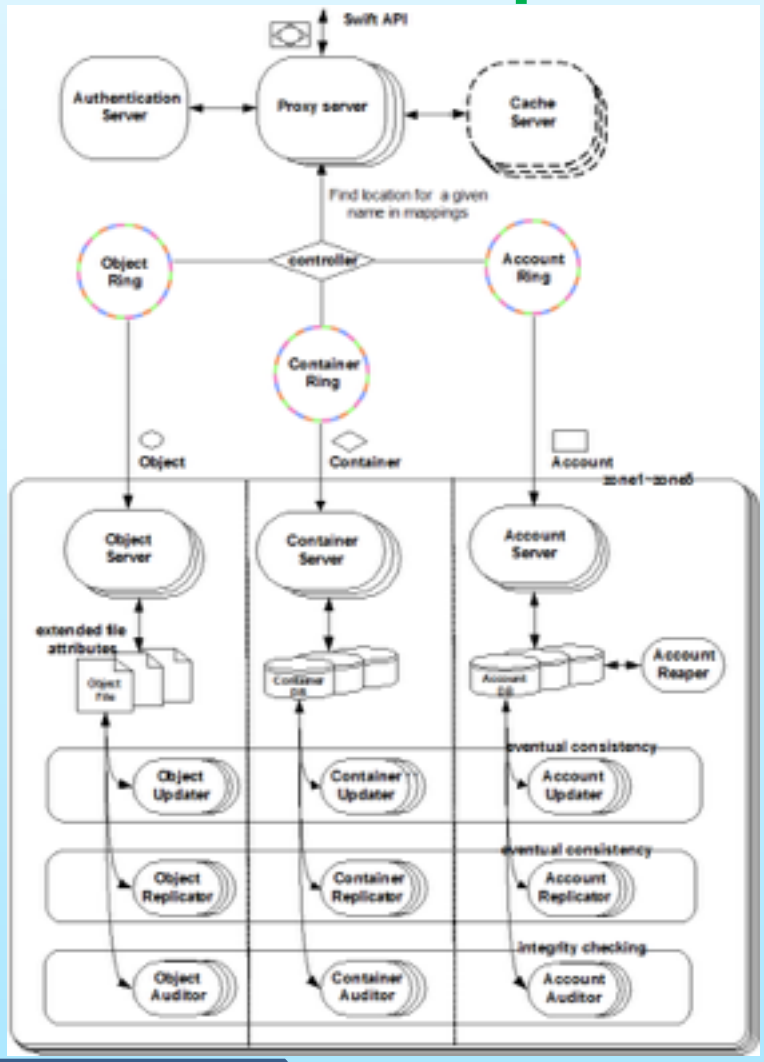


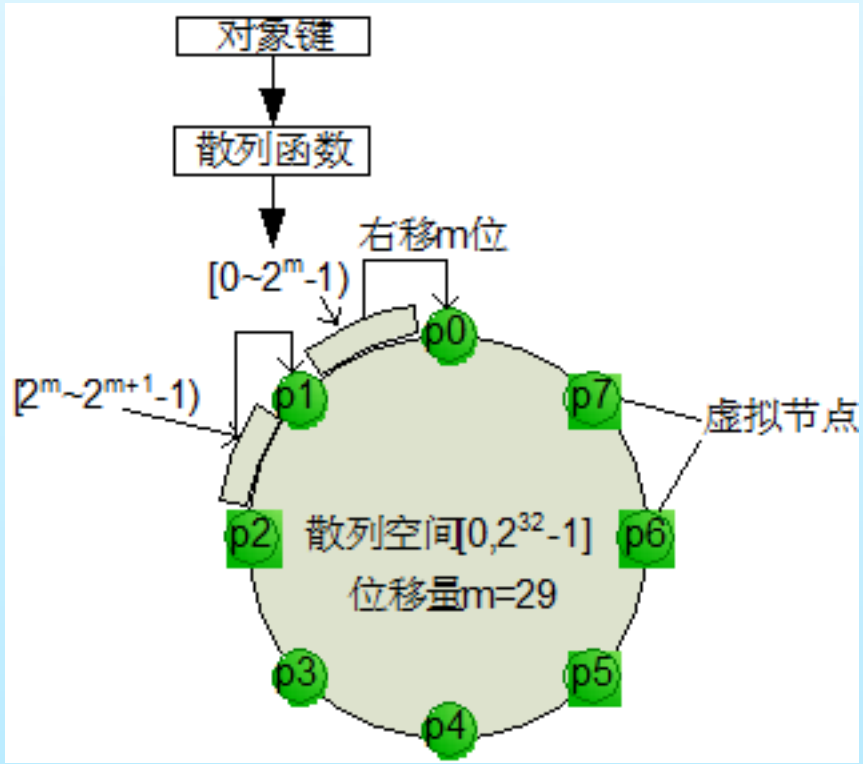




物理卷轴： 100GB+
一次写，多次读，有删除，无修改
图片存储专有“文件系统”

闭源?

# Openstack Swift




特点：

1. 提供对象接口
2. 弹性可伸缩
3. 高可用
4. 分布式对象存储
5. 大规模非结构化数据存储

Apache

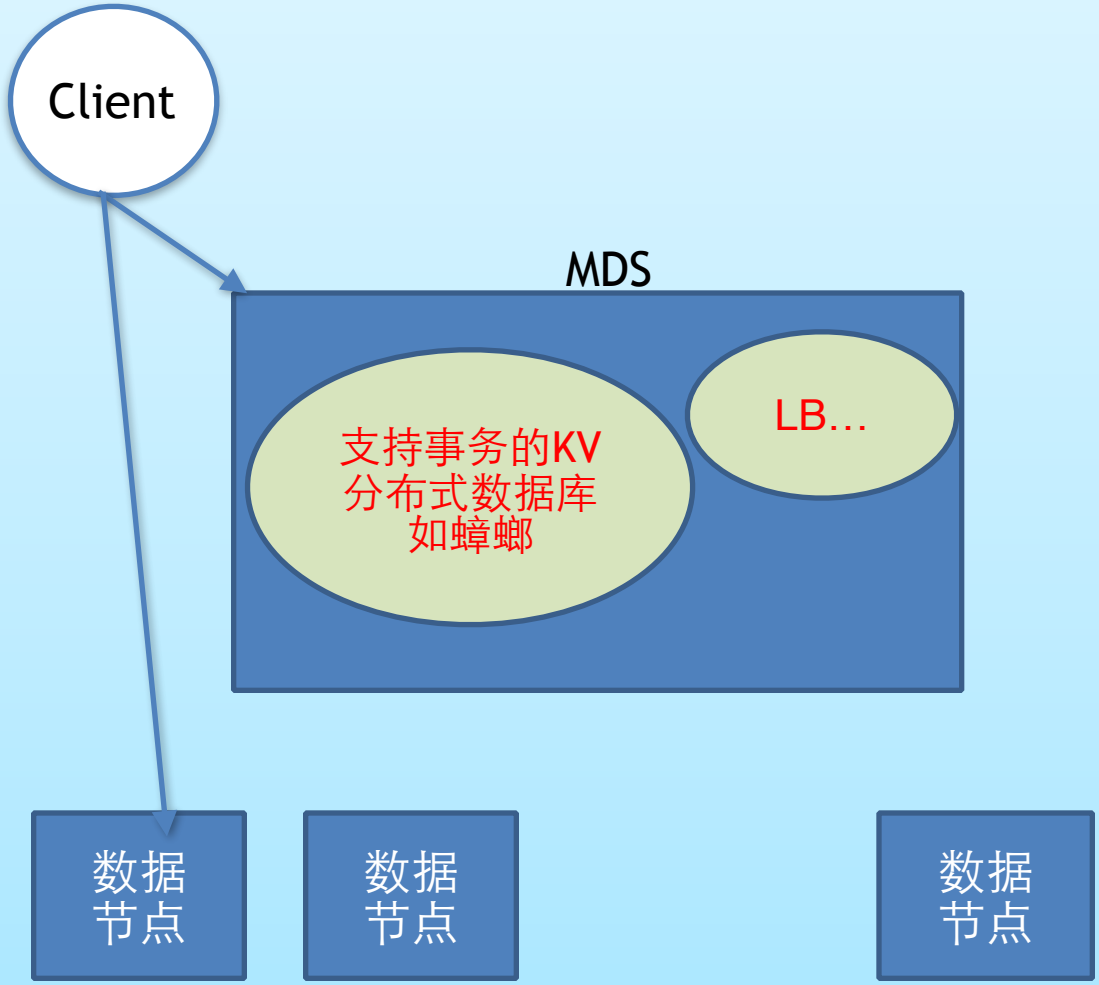
# 开源分布式文件系统

| 文件系统 | POSIX | 对象接口 | 块接口 | M/R支持 | 分布式 | 适用场景 | 社区控制 | 开源协议 |
|---|---|---|---|---|---|---|---|---|
| *Ceph* | 😀 | 😀 | 😀 | 😀 | 云 | 云 | | *GPL* |
| *GlusterFS* | 😀 | 😀 | | | 集群 | 大数据量离线应用 | Z RESEARCH | *GPL* |
| *HDFS* | 😞 | | | 😀 | 集群 | 大数据／大文件 | 社区 | *Apache* |
| *Luster* | 😀 | 😀 | | 😀 | 集群 | *HPCC* | *Intel* | *GPL* |
| *OCFS2* | 😀 | | | | 集群 | *RAC/*集群 | *Oracle* | *GPL* |
| *TaobaoFS* | 😞 | | 😀 | | 跨地域 | 互联网／小文件 | 阿里 | *GPL* |
| *Haystack* | | | | | 跨地域 | 互联网／图片 | *Facebook* | ? |
| *Swift* | 😞 | 😀 | 😞 | 😞 | 云 | 云/互联网 | 社区 | *Apache* |

# 白蚁分布式文件系统

- 支持POSIX接口和语义
- 支持对象接口
- 支持快照
- 支持云环境部署
- 支持Docker等轻量虚拟化环境
- 支持跨数据中心部署
- 高可靠，大容量，高性能，低价格*
- Apache商业友好协议
- 社区运作

# 架构

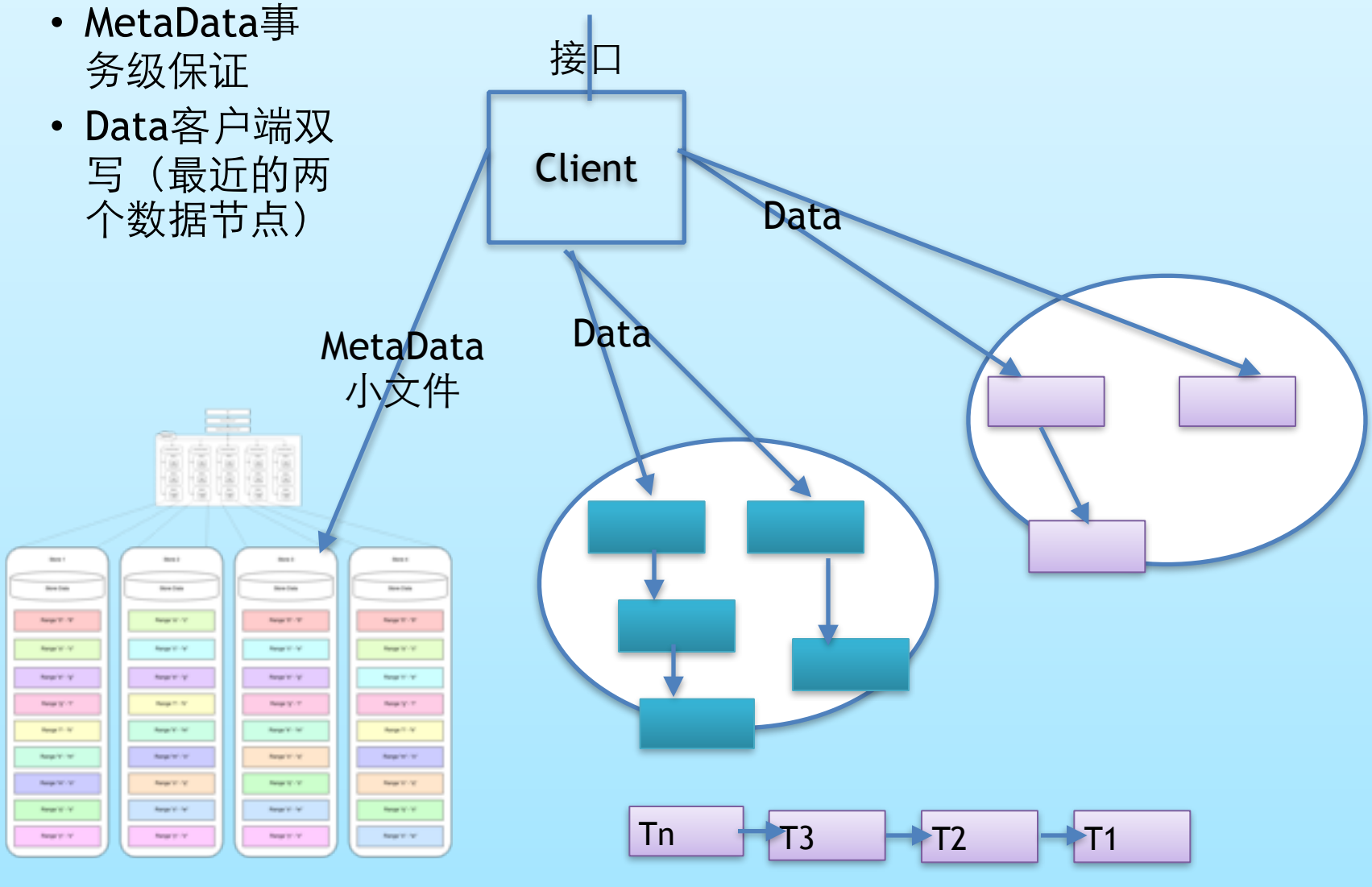


# 特点

- 并行文件系统架构
- 扁平化存储元数据
- 事务级元数据管理
- 用户定义复制管理
- 本地高速缓存优化+CDN

# 主要流程

- MetaData事务级保证
- Data客户端双写（最近的两个数据节点）



# 欢迎加入

- 社区开发

微信号：18500988099