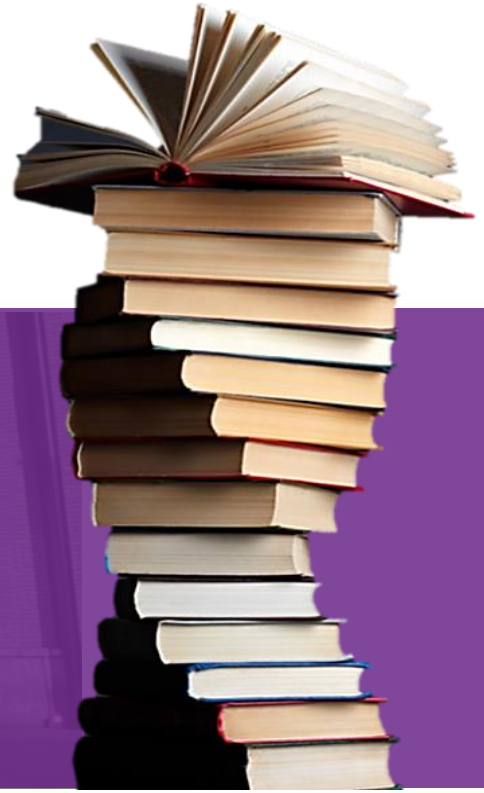
优速快递

信息技术中心



分布式文件系统方案

杜春生
2018.08.18

目录

Contents

方案背景

随着公司业务的发展，对于大量的图片、文档、音视频等文件的存储变的相当复杂，文件的管理、存储的扩容也越来越复杂，并且维护的成本会成倍增长。为了降低成本的同时提高系统高可用，特意设计一套适合我们自己的分布式文件系统（简称UDFS）。

■ 方案目标

◆接入简单

采用http、dubbo协议接入简单

◆数据迁移/数据备份

当前数据迁移，历史数据备份

◆冗余备份/负载均衡

存储2份数据保证数据可靠，负载均衡提高系统吞吐量

◆冷热数据路由

系统自动路由查找冷数据

◆高可用/高并发

◆易扩展

易横向、纵向扩展

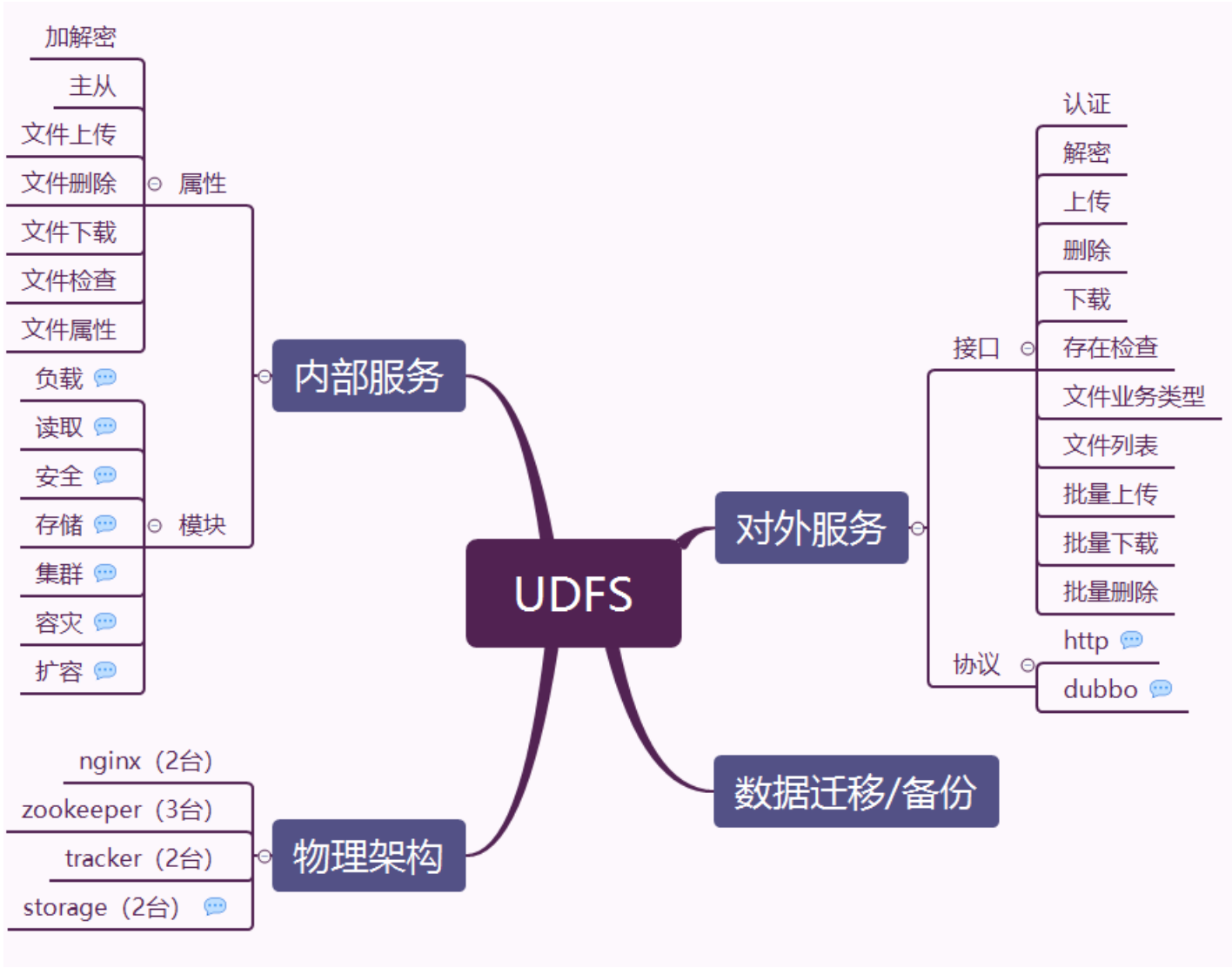
◆安全性

访问认证、特殊文件加密

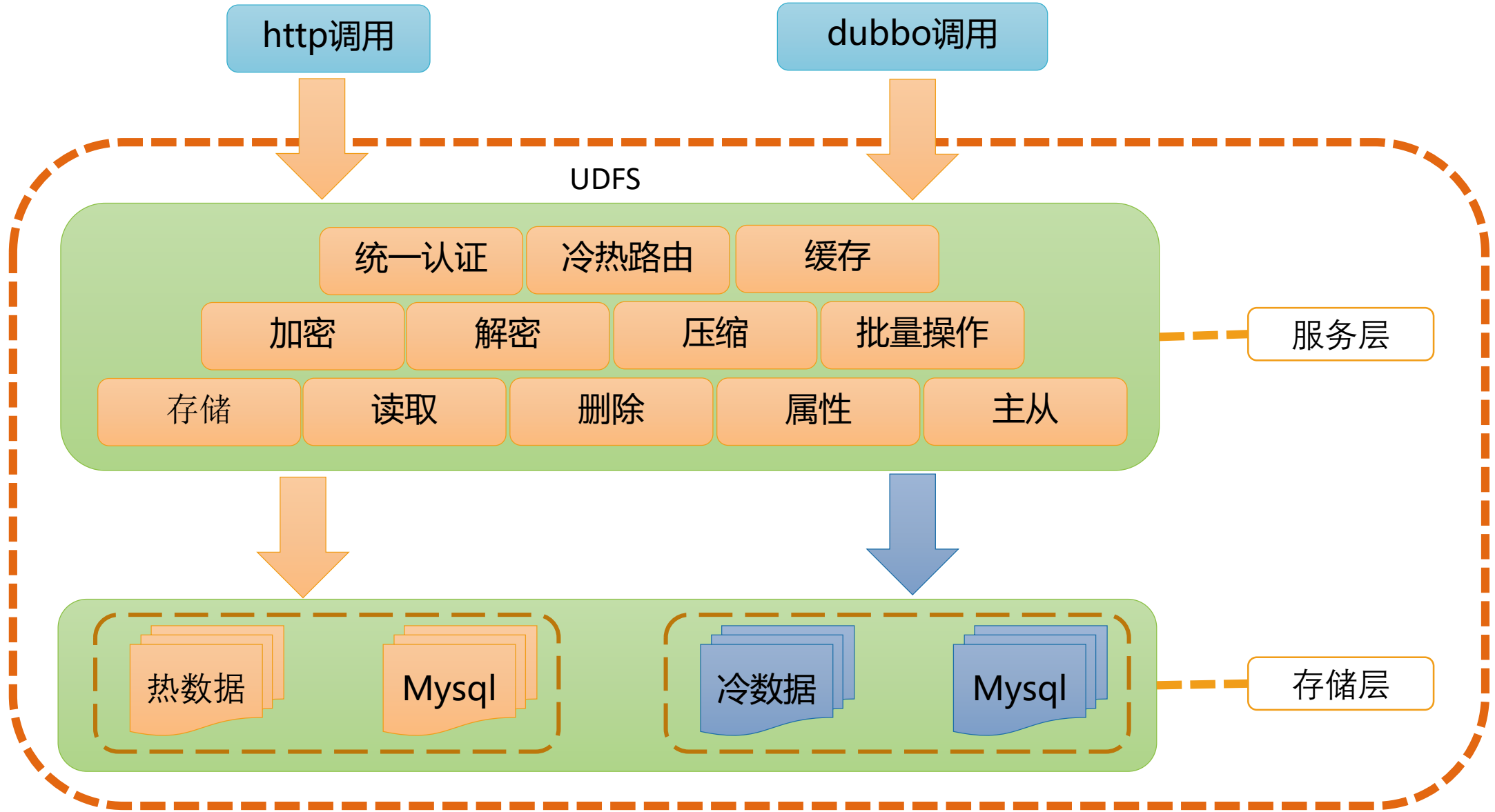
◆降低成本

采用普通服务器存储，降低资源成本

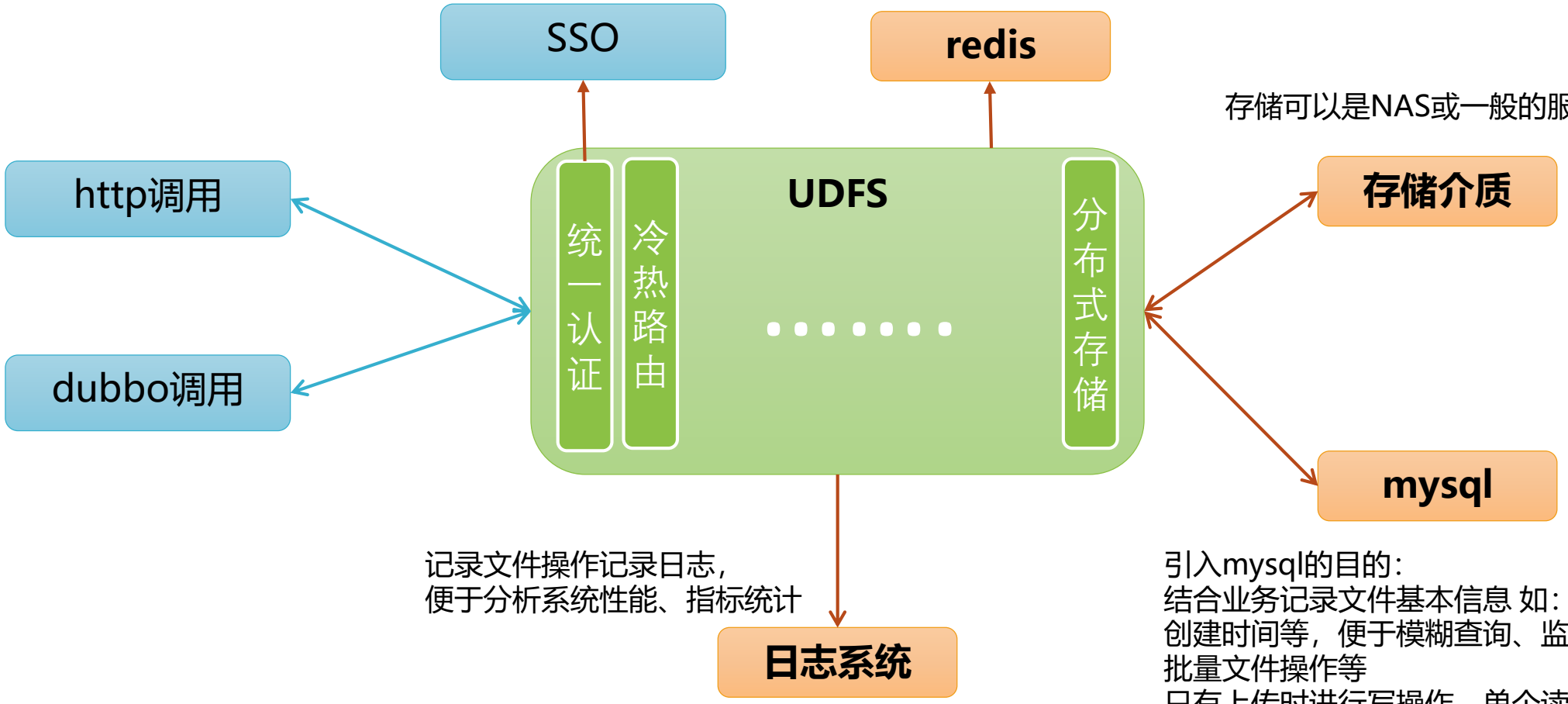
方案设计-主要功能模块


UDFS
内部服务
属性
加解密
主从
文件上传
文件删除
文件下载
文件检查
文件属性
模块
负载
读取
安全
存储
集群
容灾
扩容
对外服务
接口
认证
解密
上传
删除
下载
存在检查
文件业务类型
文件列表
批量上传
批量下载
批量删除
协议
http
dubbo
物理架构
nginx（2台）
zookeeper（3台）
tracker（2台）
storage（2台）
数据迁移/备份

方案架构


http调用
dubbo调用
UDFS
统一认证
冷热路由
缓存
加密
解密
压缩
批量操作
存储
读取
删除
属性
主从
服务层
热数据
Mysql
冷数据
Mysql
存储层

方案架构-调用图


SSO
redis
存储可以是NAS或一般的服务器硬盘
http调用
dubbo调用
统一认证
冷热路由
UDFS
分布式存储
存储介质
mysql
记录文件操作记录日志，
便于分析系统性能、指标统计
日志系统
引入mysql的目的：
结合业务记录文件基本信息 如：订单id、fileid、
创建时间等，便于模糊查询、监控、
批量文件操作等
只有上传时进行写操作，单个读取时一般不用访问
根据使用场景也可忽略

技术选型-分布式文件系统类型

- 通用分布式文件系统

  和传统的本地文件系统（如ext3、NTFS等）相对应，应用端可以mount使用。典型代表：lustre、MooseFS

- 专用分布式文件系统

  基于google FS的思想，文件上传后不能修改。不能mount使用，需要使用专有API对文件进行访问，也可称作分布式文件存储服务。典型代表：MogileFS、FastDFS、TFS

指标	通用分布式 文件系统	专用分布式 文件系统
开发者友好性	较好	较差
系统复杂性	较高	较低
系统性能	一般	较高

技术选型-常用分布式文件系统

框架	适合类型	文件分布	系统性能	复杂度	备份机制	通讯协议	社区支持	开发语言
FastDFS	4KB~500MB	小文件合并存储不分片处理	很高	简单	组内冗余备份	Api HTTP	国内用户群	C语言
TFS	所有文件	小文件合并，以block组织分片		复杂	Block存储多份,主辅灾备	API http	少	C++
MFS	大于64K	分片存储	Master占内存多		多点备份动态冗余	使用fuse挂在	较多	Perl
HDFS	大文件	大文件分片分块存储		简单	多副本	原生api	较多	java
Ceph	对象文件块	OSD一主多从		复杂	多副本	原生api	较少	C++
MogileFS	海量小图片		高	复杂	动态冗余	原生api	文档少	Perl
ClusterFS	大文件			简单	镜像		多	C

■ 技术选型-fastdfs&hadoop

属性	FastDFS	Hadoop	描述
实时性	强	弱	hadoop主要在离线分析
文件大小	中小文件	超大文件（几百MB、GB甚至TB级）	目前需求来说都是中小文件
小文件管理	效率高	效率低	hadoop为了提高对小文件的性能管理，必须要定时的进行归档处理，甚至还要压缩， 归档后的文件存在无法删除
运维成本	较低	较高	Hadoop时常有新问题出现，有一些是自身的bug造成，出现问题分析或解决问题周期长
开发成本	较低	较高	FastDFS有大量的开发文档，bbs比较活跃，hadoop版本及组件太多，需要经常调优
数据冗余	最低2份	最低3份	根据目前的量来说，在节约成本的情况下，FastDfs还是占有优势。

注：主要是从时效、安全、便捷等方面，对于只限于文件存储方面做的比较

技术选型- FastDfs

对于我们当前业务场景建议使用FastDfs进行分布式文件存储

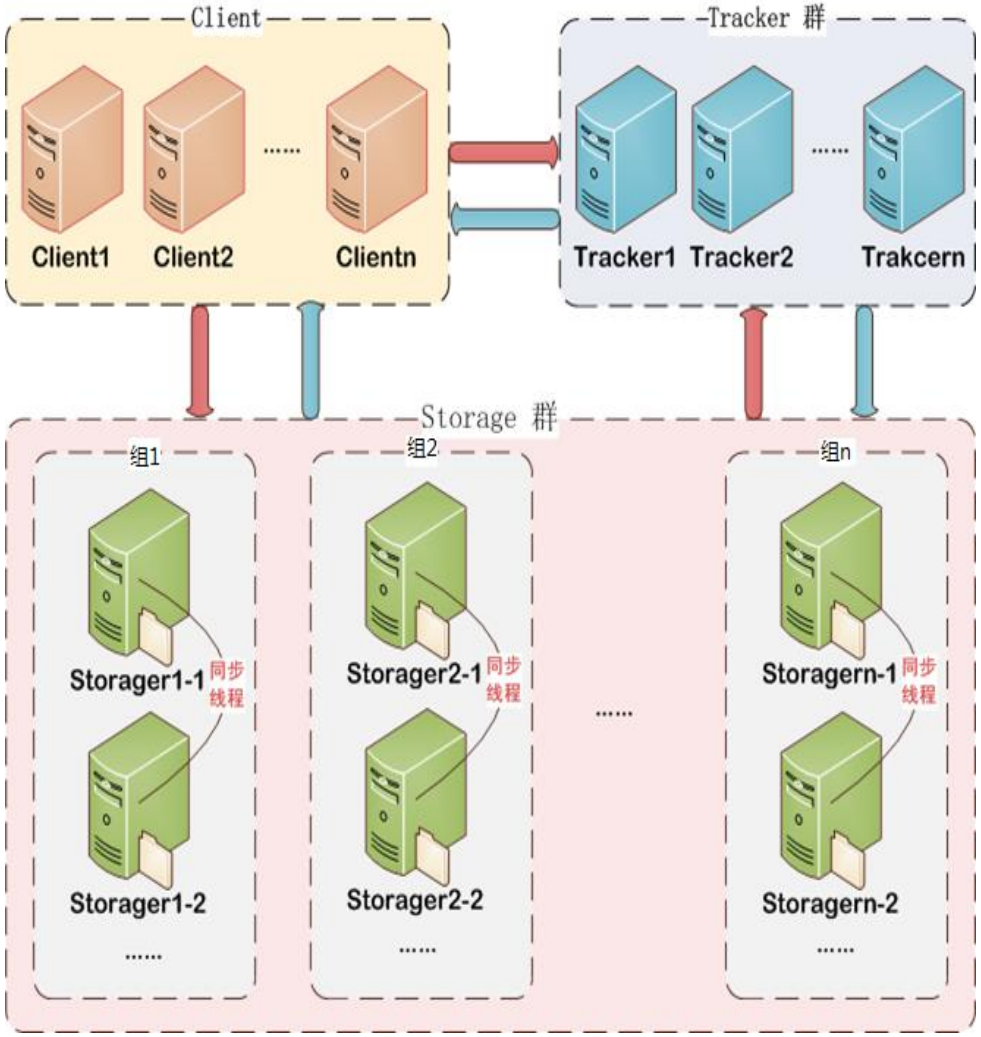
- 对于目前文件大多在100k左右完全适合FastDfs场景。
- FastDfs部署、运维来说简单快捷，版本稳定变动较少
- FastDFS是一个开源的轻量级分布式文件系统，特别适合以文件为载体的在线服务
- 架构简单
- 有冗余备份和负载均衡设计
- 扩充容量方便
- 存储介质一般硬盘即可
- 已经有一定数量的客户，文件存储规模大的达到了**PB**级

技术选型-FastDfs使用情况

至少有25家公司在使用FastDFS，其中有好几家是做网盘的公司。

- UC （http://www.uc.cn/，存储容量超过10TB）
- 支付宝（http://www.alipay.com/）
- 京东商城（http://www.360buy.com/）
- 淘淘搜（http://www.taotaosou.com/）
- 飞信（http://feixin.1008**/）
- 赶集网（http://www.ganji.com/）
- 淘米网（http://www.61.com/）
- 58同城（http://www.58.com/）
- 搜房网（http://www.soufun.com/）
- 你我贷（http://niwodai.com，存储容量超100TB）

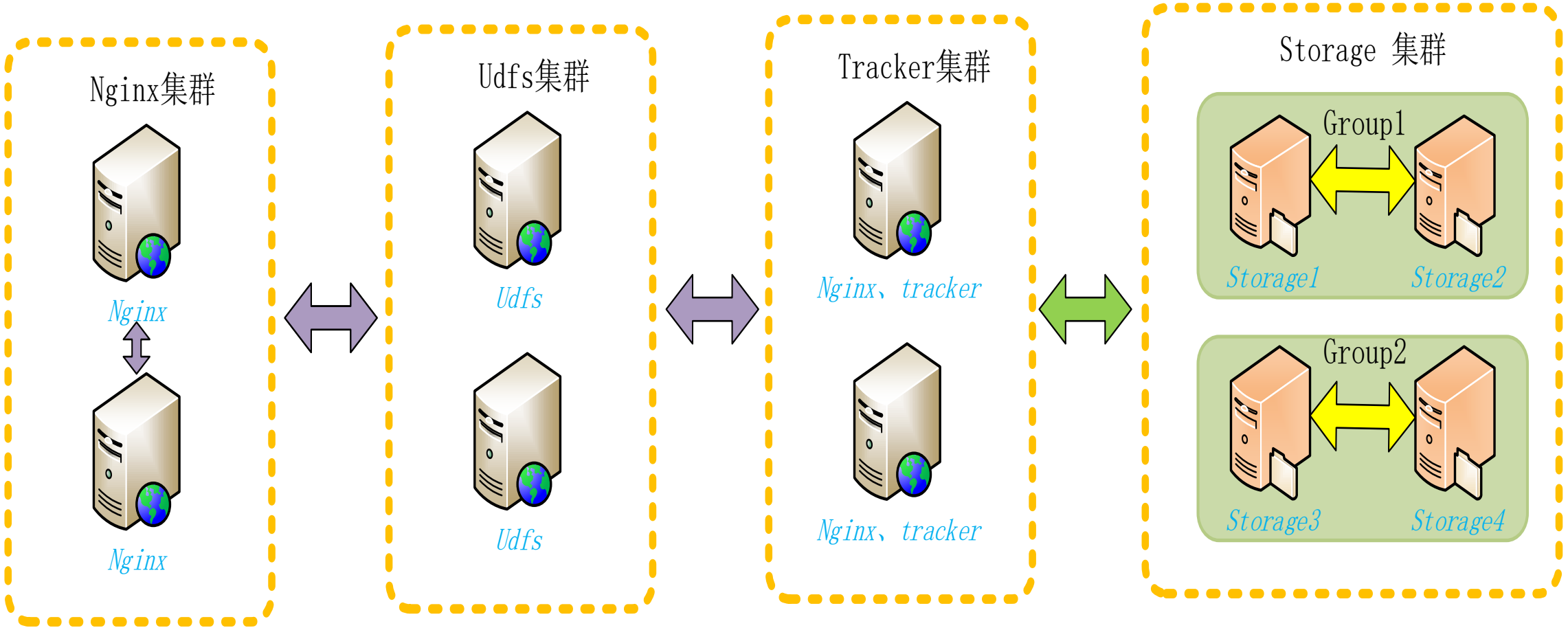
技术选型- FastDfs系统架构



- Tracker对等性部署不存在SPOF
- Client端使用FastDFS提供的Jar包可直接与Tracker或Storage集群进行文件存取操作
- Tracker集群可水平扩展达到 高可用及高性能要求
- Storage集群存储采用分组及组内文件同步复制达到分布式存储及文件冗余要求
- 实现横向、纵向扩展
- 采用binlog进行文件同步
- 文件id命名规则(组/磁盘/目录/文件名)

group/M00/09/BE/rBBZolgj6OiAY6cHAAG019shnqU964.jpg

技术选型- Udfs整体部署结构



最小配置storage（2台）tracker（2台）Udfs(2台) Nginx目前有集群可以共用或F5做负载

技术选型-硬件资源

数据存储周期

- 热数据超过3个月迁移到冷数据
- 冷数据保留1年（当前是一年，后续可以根据业务需要来定）

存储方案

目前我们每天量在80G左右，按照热数据存放3个月的话应该是3个T以内，考虑未来业务增长按照2倍量（7T）规划制定硬件配置

存储硬件配置

- 热数据配置（保留3个月）

  物理机 2台

  24槽位（15K 900GB）

  2U E5-2620*2 32GB内存

- 冷数据配置

  物理机 4台

  16槽位（7.2K 4T）

  2U E5-2620*2 32GB内存

开发计划

费用类别	部门	岗位	级别	人/月
人员成本	内部	开发	架构师	1.25
			高级开发	0
			中级开发	0
		需求	高级需求	0
			中级需求	0.25
		测试	高级测试	0.75
	外部	测试	中级测试	
	人力合计	-	-	
硬件成本	服务器硬件	-	-	
总成本				

- 需求整理0.25人月
- fastdfs测试环境搭建（环境准备、参数配置、部署手册等） 0.25 人月
- UDFS服务开发 （服务端+客户端+开发测试） 0.5人月
- UDFS性能测试及调优 0.5*2人月
- UDFS功能测试0.25人月

测试要点

功能测试

- 文件的常规操作
- 冷热数据路由测试
- 单点故障测试
- 数据同步测试
- 扩容测试
- 系统接入兼容性测试

性能测试

- 上传文件测试
- 下载文件测试
- 删除文件测试

数据迁移

迁移条件

- 热数据超过3个月迁移到冷数据
- 冷数据保留1年（当前是一年，后续可以根据业务需要来定）

迁移方案

- 根据时间戳迁移

  1、可操作系统层面迁移（shell脚本）

  2、可以udfs层面迁移（java脚本）

QA

优速快递

优速快递

大包裹　用优速