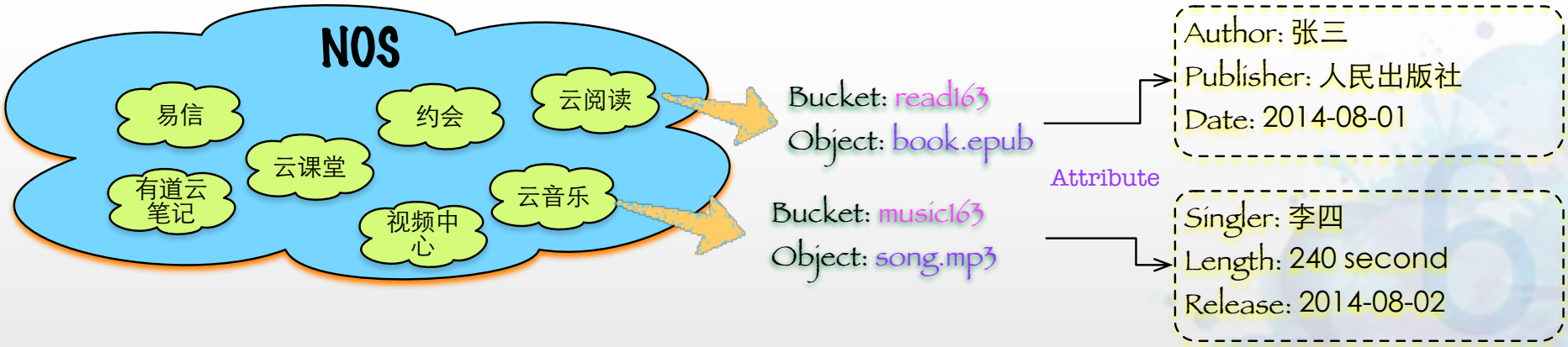
SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

网易云对象存储关键技术解析

网易杭州研究院-来东敏

网易云对象存储是什么

- NetEase Object Storage（NOS）一个海量Key-Value系统
  - Key：最大支持1K字节
  - Value：最大支持1TB二进制文件，比如图片、视频等静态文件
  - Attribute：最多10个键值对
- 容器（Bucket）：命名空间

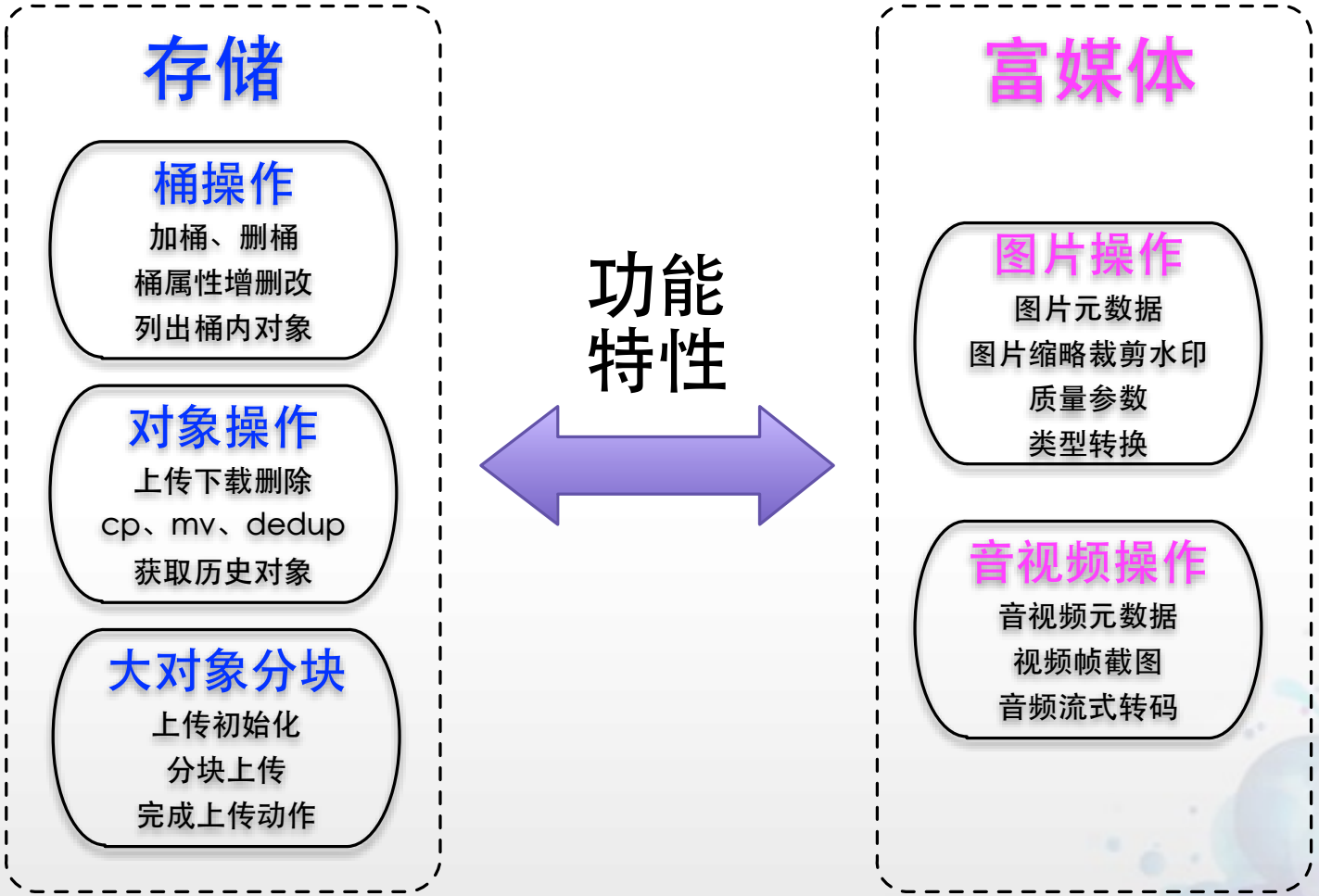


一些数据

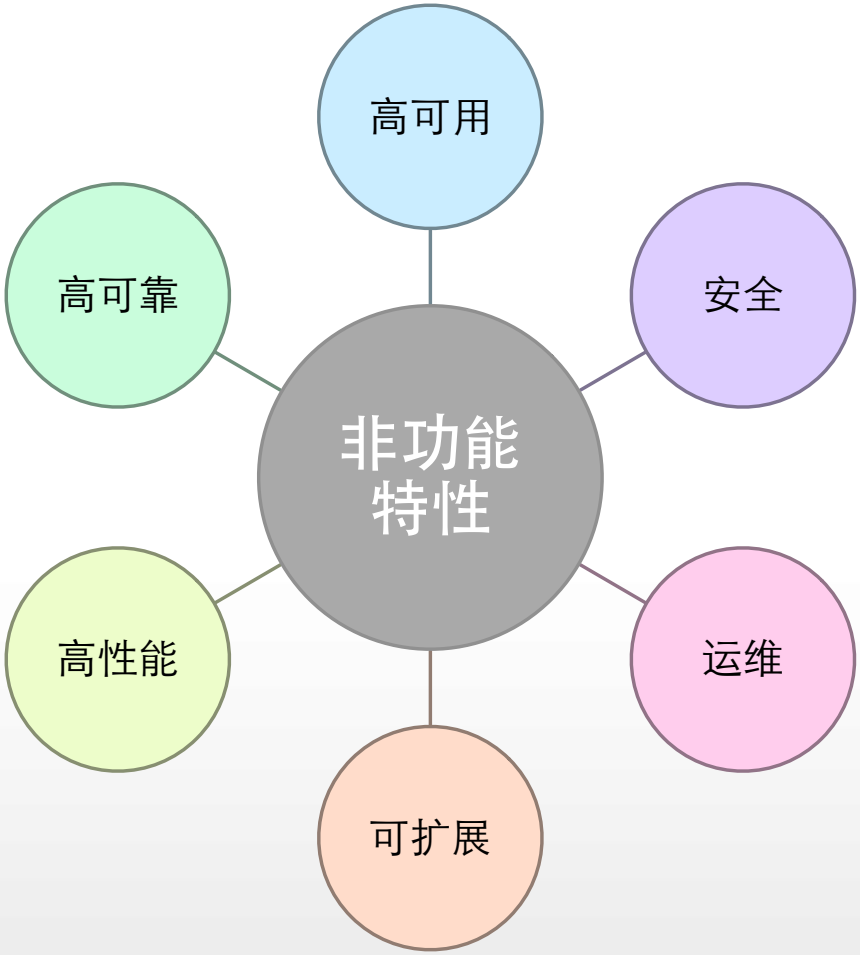
- 上线时间：2012.10
- 产品数量：30+
- 桶的数量：100+
- 对象数量：700,000,000
- 逻辑存储：300T
- 常规日增：1T
- 常规流量：3Gbps


网易看游戏 Beta
網易 网易云 beta cloud.netease.com
网易云课堂 BETA
163 网易免费邮 mail.163.com 中文邮箱第一品牌
易信 让沟通更有趣
網易宝 乐生活 易支付 epay.163.com
花田 我们相遇的地方！
網易 www.163.com
有道云笔记 note.youdao.com
网易视频 V.163.COM
同城约会 yuehui.163.com
网易应用中心 m.163.com
网易云阅读
网易印像派 亲密时刻 永久珍藏 YXP.163.COM
網易公开课
网易安全中心 NetEase Security Center
WORLD OF WARCRAFT 魔兽世界 熊猫人之谜
炉石传说 魔兽英雄传
網易博客 blog.163.com 新锐人文生活
天谕 无束缚3D幻想网游
網易房产 house.163.com
网易云音乐 Beta
SACC 2014中国系统架构师大会 SYSTEM ARCHITECT CONFERENCE CHINA 2014
SequeMedia ChinaUnix ITPUB

不只是Key-Value (1)



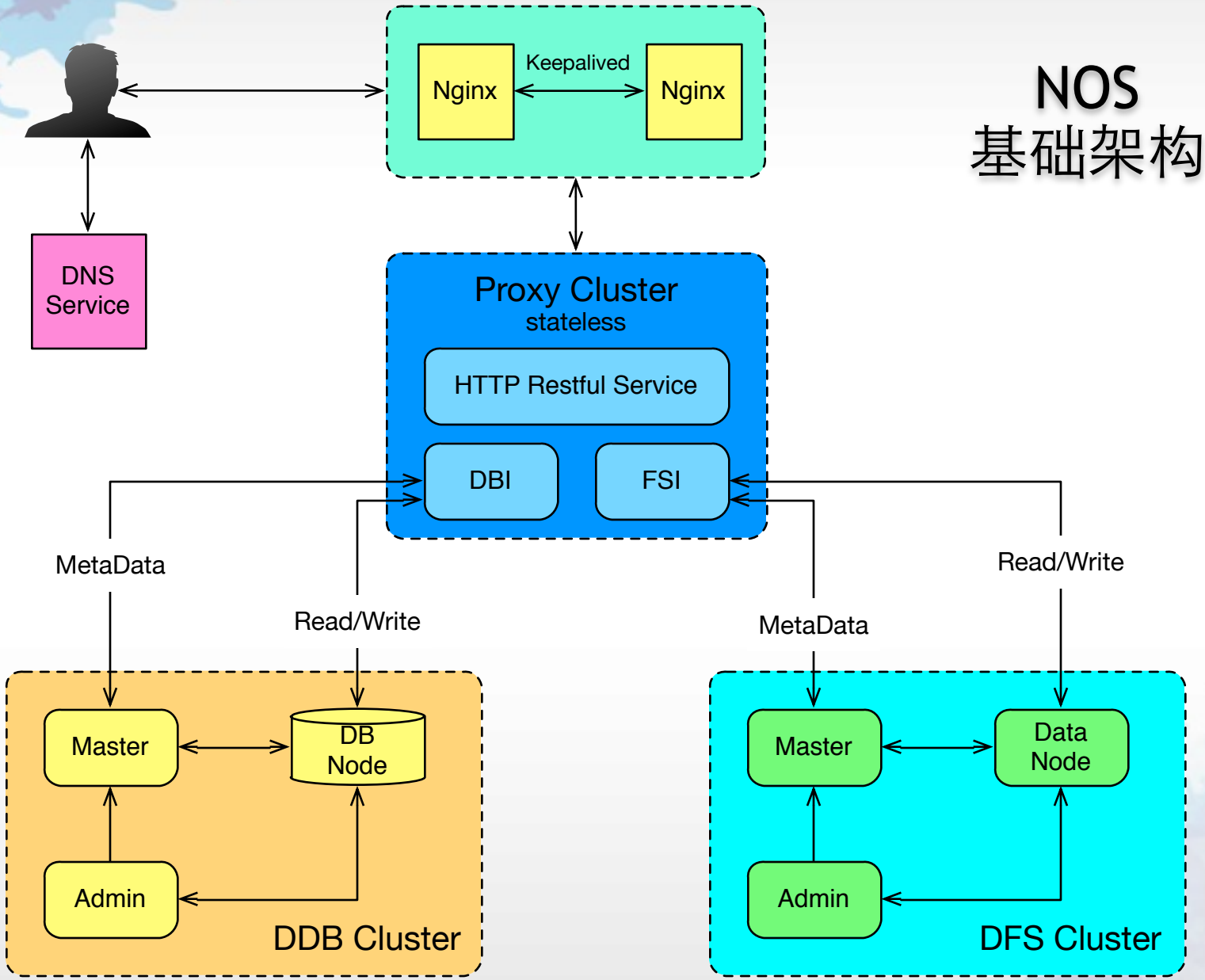
不只是Key-Value (2)



关键技术

1. 元数据存储组件：DDB
2. 数据存储组件：DFS
3. 列出桶内对象：ListObject
4. 基于NOS的用户态文件系统：NOSFS
5. 富媒体服务框架：Tobie
6. 多租户流量隔离：LimitServer

NOS
基础架构

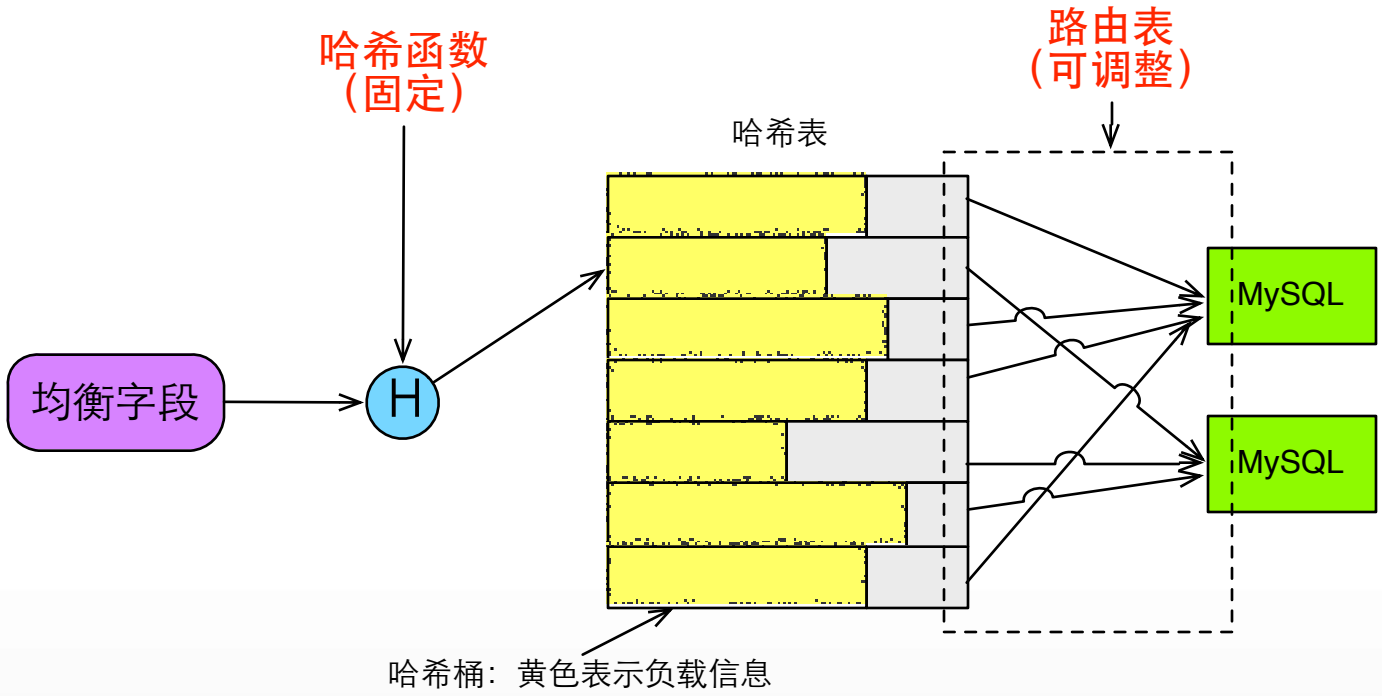

Nginx
Keepalived
Nginx
DNS
Service
Proxy Cluster
stateless
HTTP Restful Service
DBI
FSI
MetaData
Read/Write
Read/Write
MetaData
Master
DB
Node
Admin
DDB Cluster
Master
Data
Node
Admin
DFS Cluster

DDB介绍

- 分布式数据库系统（Distributed DataBase, 简称 DDB）是网易杭研后台技术中心研发的分布式关系数据库平台

- 主要目标是解决以下问题
  - 海量结构化数据存储
  - 高并发高吞吐数据访问

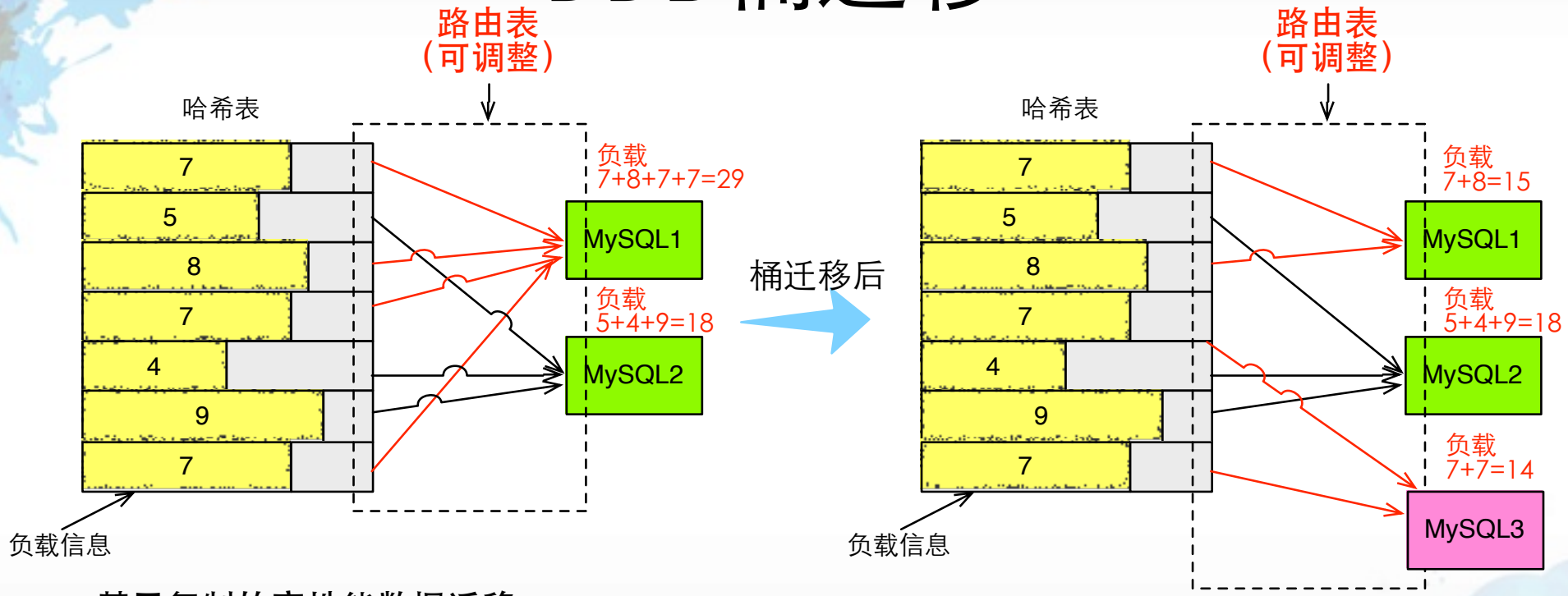
DDB分区策略



均衡字段：用于计算哈希值的字段

两级映射：结合哈希的高效性和路由表的可管理性

DDB桶迁移



- 基于复制的高性能数据迁移
  - 方法：在目的节点建立待迁移桶的数据复制，迁移完成后修改路由表
  - 特点：在线复制，性能不错
- 路由表版本号
  - 为路由表设置递增的版本号，迁移时增加源节点的路由表版本号
  - 客户端请求源节点，发现路由表版本不匹配，同步路由表后正确路由至目的节点

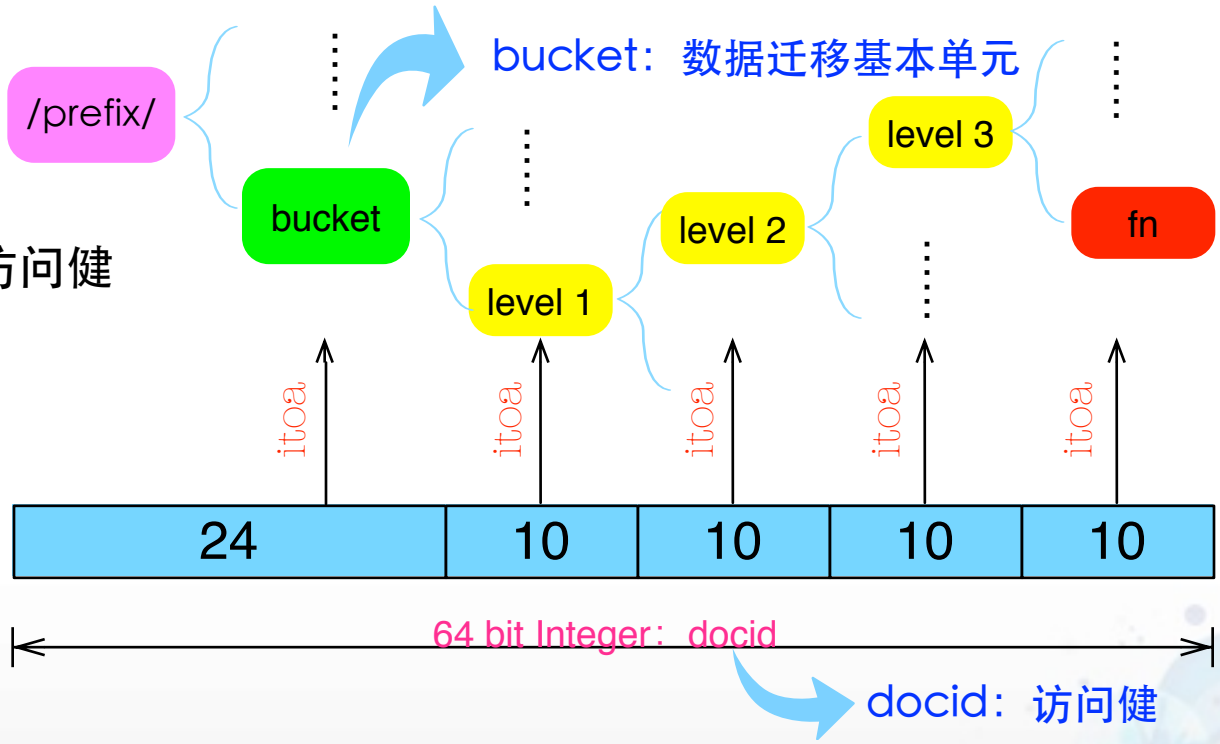
DFS介绍

- 分布式文件系统（Distributed FileSystem，简称 DFB）是网易杭研后台技术中心研发的分布式非结构化数据存储平台

- 主要目标是解决以下问题
  - 海量非结构化数据存储
  - 高并发高吞吐数据访问

DFS访问健

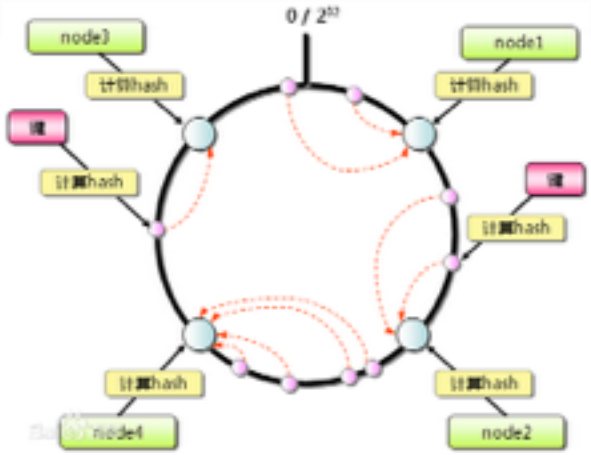
- 64位长整形docid作为访问健
- 对外：扁平的名字空间
- 对内：分层文件索引

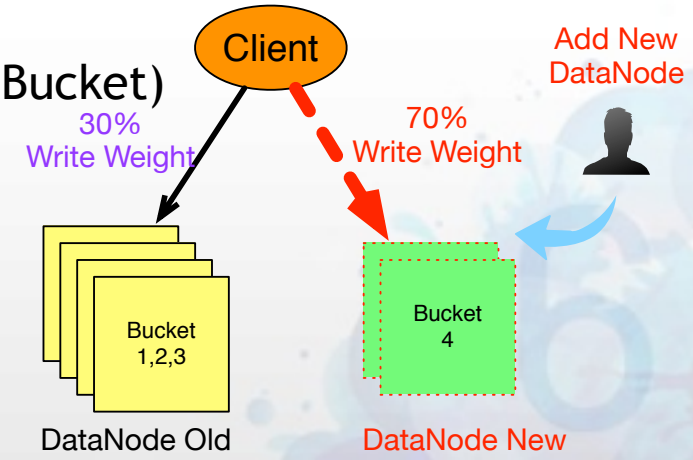


Eg: 21990232555553 => /prefix/2/0/0/0/1

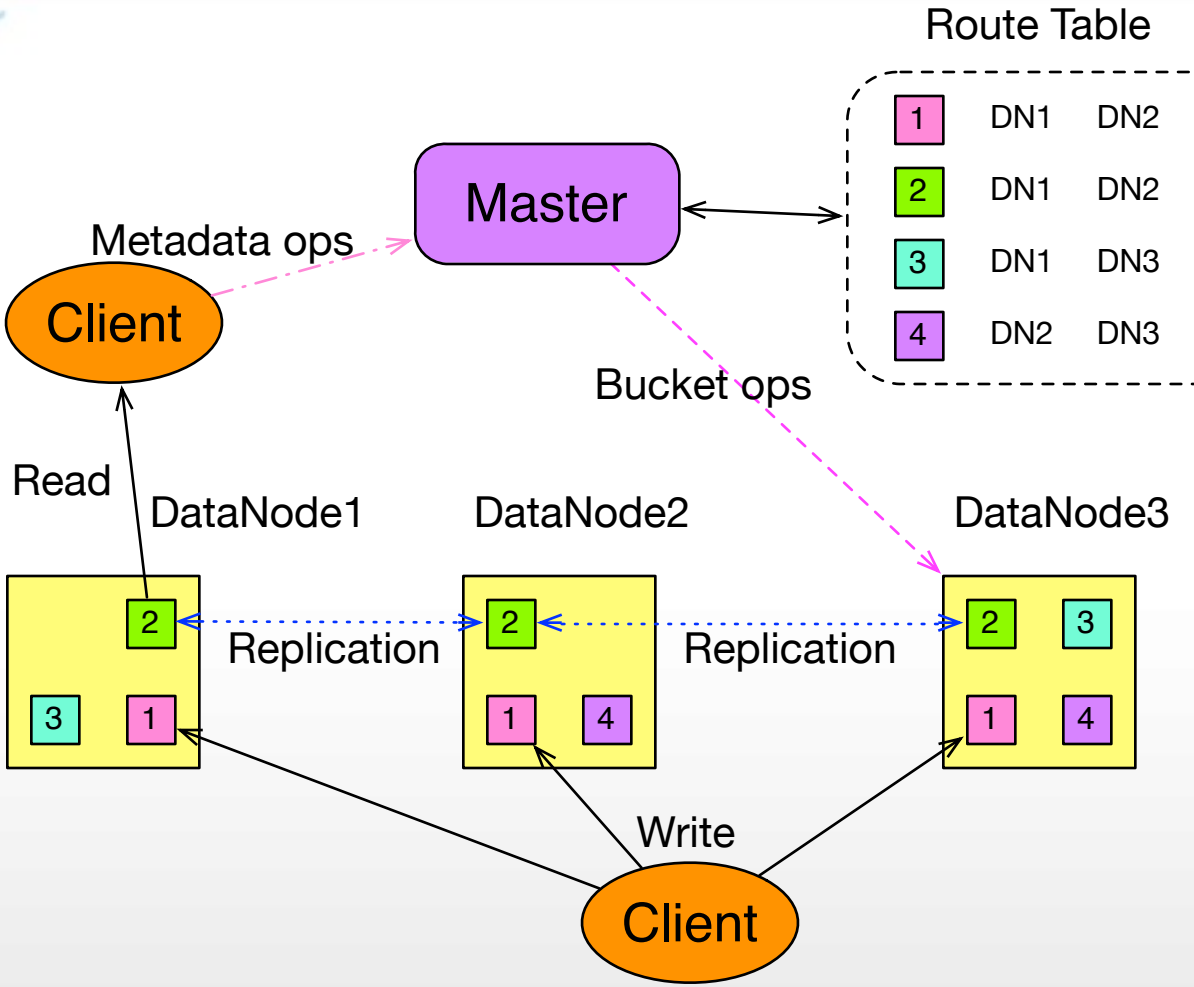
DFS分区策略

- 虚拟节点一致性哈希，问题：扩容数据迁移
- 对于PB级大数据量系统，措施：无迁移扩容
- DFS方案：分区(Bucket)预留 + 系统生成访问健(Docid)
  - 预先规划Bucket(2^24)，不立即启用
  - 访问键(Docid)由系统分配而不是应用程序指定
- 增加一个物理节点时
  - 启用一批分区(Bucket)给该物理节点
  - 系统生成访问健(Docid)时，使用新分配的分区(Bucket)
- 适用范围
  - 访问键(Docid)由系统分配
  - 负载在初始分配完成后基本不变





DFS系统构架



特点1：轻量级Master

特点2：缓存路由表

特点3：并发写

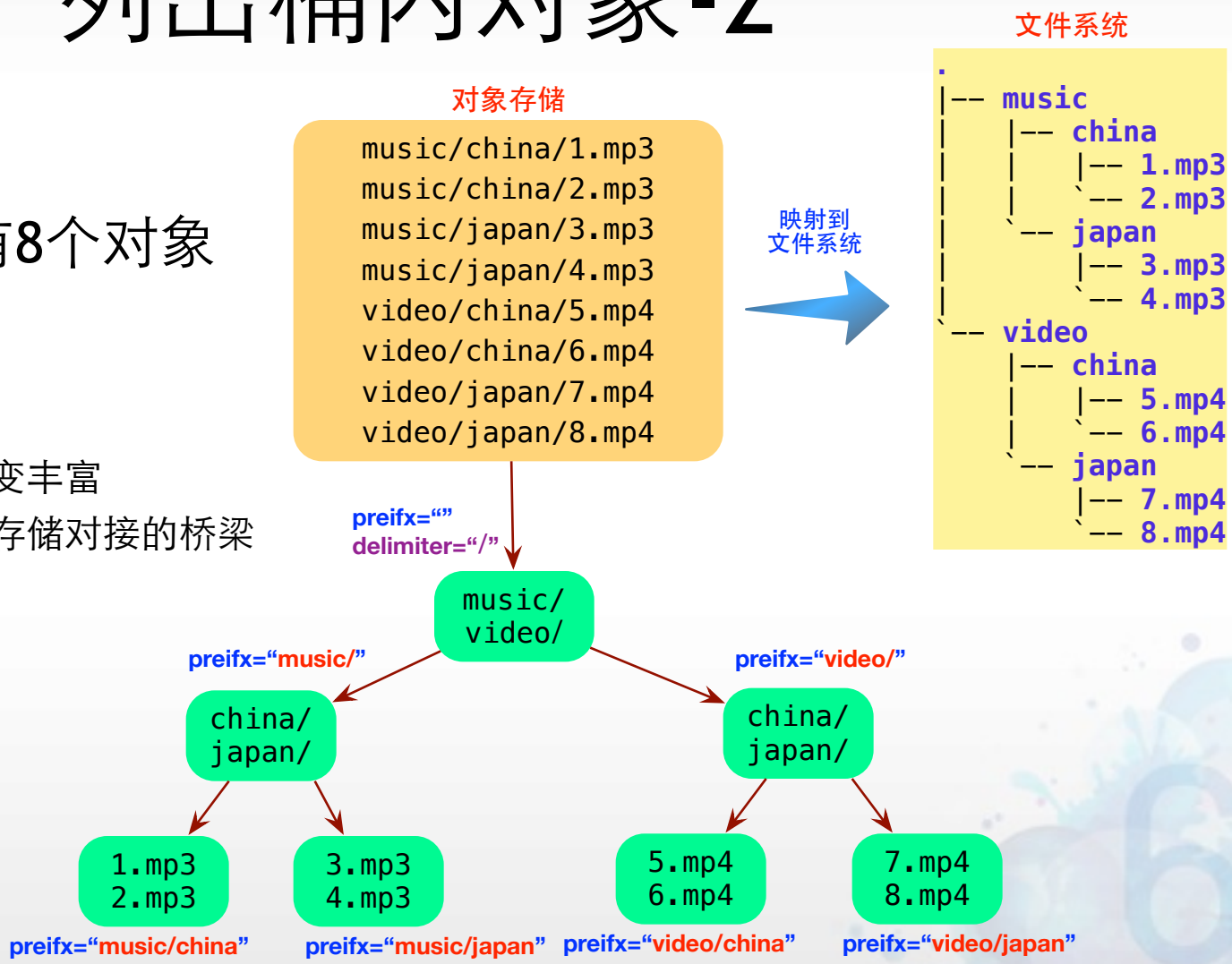
特点4：docid预分配

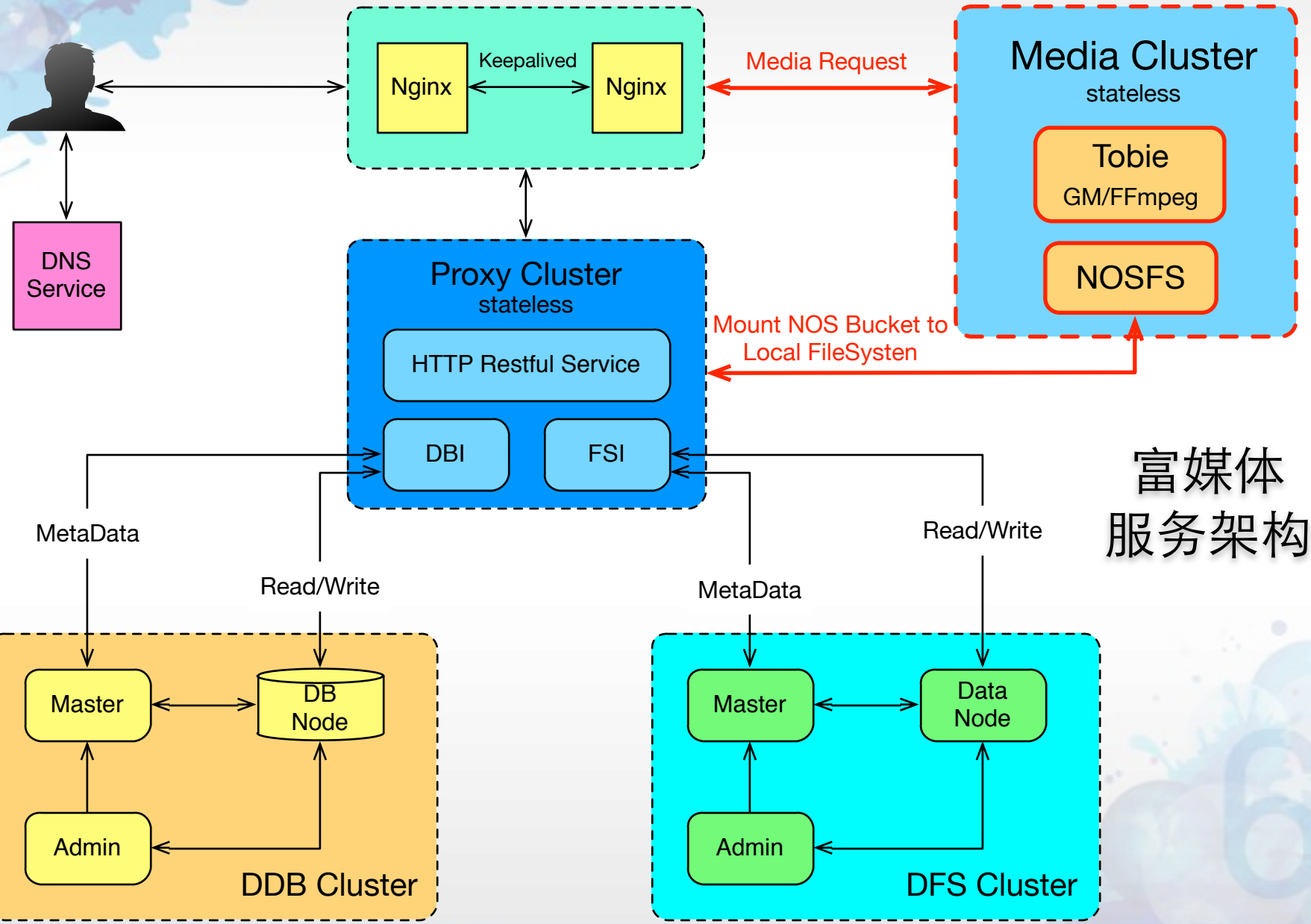
列出桶内对象-1

- 功能描述
  - 根据简单的检索条件，返回对象列表的子集，接口：**ListObject**
- 相关参数
  - Prefix：对象Key的前缀，可以使用前缀对桶内对象分组；
  - Delimiter：检索分隔符，用于做类似groupby的操作；
  - Max-keys：符合条件的对象Key列表数量；
  - Marker：字典序起始标记，只返回该标记后的对象Key；

列出桶内对象-2

- 桶music163有8个对象
- 现实意义
  - 命名空间从扁平变丰富
  - 对象存储和文件存储对接的桥梁



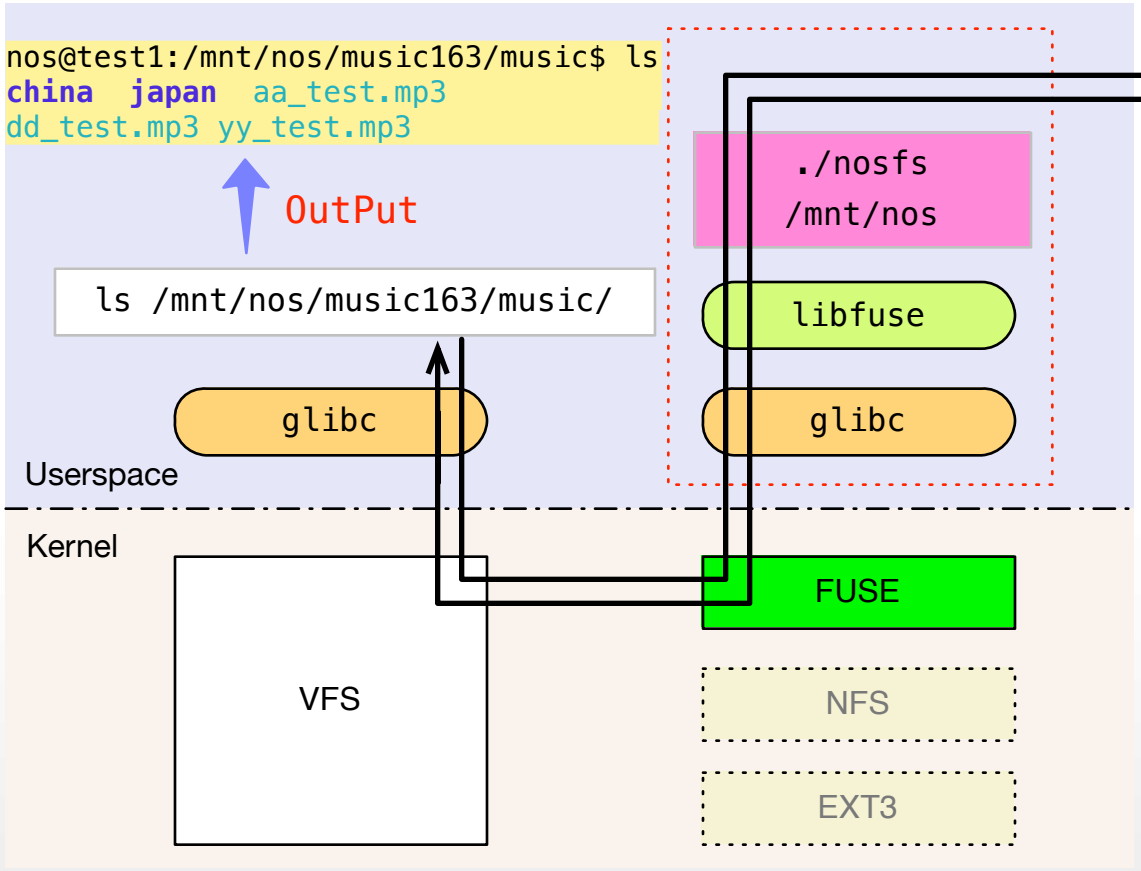

Nginx
Keepalived
Nginx
Media Request
Media Cluster
stateless
Tobie
GM/FFmpeg
NOSFS
DNS
Service
Proxy Cluster
stateless
HTTP Restful Service
DBI
FSI
Mount NOS Bucket to
Local FileSysten
MetaData
Read/Write
Read/Write
MetaData
Master
DB
Node
Admin
DDB Cluster
Master
Data
Node
Admin
DFS Cluster

富媒体服务架构

NOSFS-1

- 相关描述
  - 基于FUSE的用户态文件系统
  - 通过NOSFS将NOS桶挂载成本地文件系统
  - 应用场景：富媒体服务、数据备份等
- 性能优化
  - 读性能较差：预读
  - 数据缓存、元数据缓存
- 限制和坑
  - 不支持随机写，不支持文件夹移动
  - 锁父目录问题：通过虚拟目录绕开
  - List性能问题：通过编码对象路径绕开

NOSFS-2

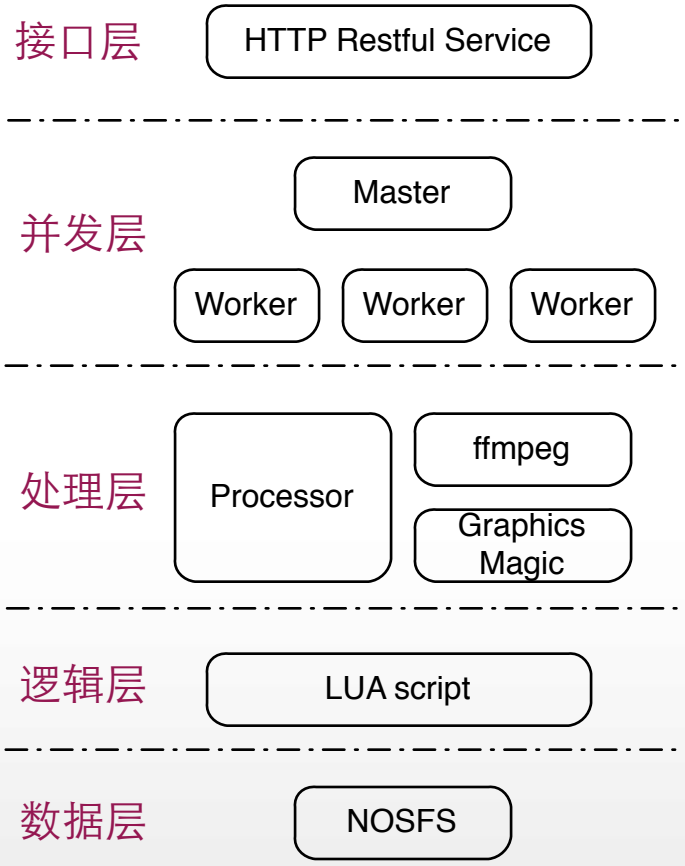

nos@test1:/mnt/nos/music163/music$ ls
china japan aa_test.mp3
dd_test.mp3 yy_test.mp3
OutPut
ls /mnt/nos/music163/music/
./nosfs
/mnt/nos
libfuse
glibc
glibc
Userspace
Kernel
FUSE
VFS
NFS
EXT3
HTTP
NOS
1 music/aa_test.mp3
2 music/china/1.mp3
3 music/china/2.mp3
4 …
5 music/china/n.mp3
6 music/dd_test.mp3
7 music/japan/1.mp3
8 music/japan/2.mp3
9 …
10 music/japan/k.mp3
11 music/yy_test.mp3

富媒体服务-1

- 实现功能
  - 图片操作：缩略、裁剪、水印等；
  - 音频操作：流式转码、元数据等；
  - 视频操作：帧截图、元数据等；
  - 其他功能：管道链式处理；

- 相关技术
  - 服务接口：Libevent HTTP
  - 媒体类库：ffmpeg & GraphicsMagick
  - 父子进程：媒体库内存泄露问题、健壮性
  - 过载保护：请求队列
  - C++内嵌Lua：比同类产品快2~3倍
    - C++：网络接口，媒体库调用【高性能】
    - Lua：参数解析，调用链生成【灵活性】

富媒体服务-2



基于 Libevent 实现 HTTP 服务端

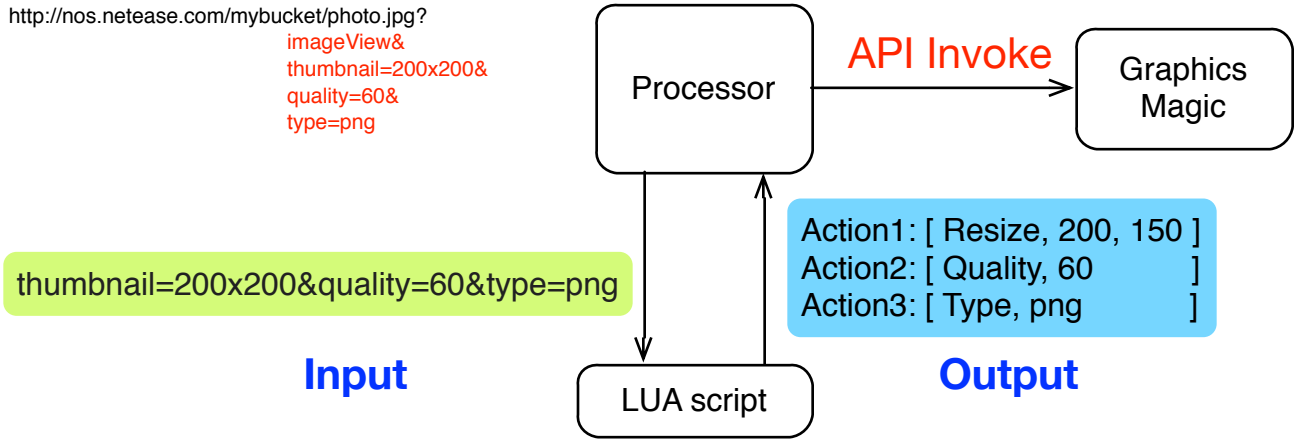
可配 Worker 数量
可配处理若干请求后 Worker 退出

C++ API调用保证高性能
Processor 粘合处理逻辑

接收请求字符串
返回处理序列

mount 到 NOS 桶
本地请求转化为 NOS 接口请求

富媒体服务-3



- Lua扩展示例
  - 按总像素缩略：用户指定总像素，如：pixel=160000
  - 裁剪缩略图片：对缩略后的图片进行裁剪以满足尺寸要求
  - 人脸智能识别：Lua调用人脸识别程序获取图片中人脸位置
  - 图片黑边检测：Lua调用黑边检测程序获取黑边位置

流量隔离-1

- 资源共享和隔离需求并存
  - 处理方式：运维手段而不是常规手段；
  - 处理流程：收到流量统计报警后紧急限流；
- 资源类型
  - 存储：底层对象存储；
  - 计算：富媒体处理请求；
  - 流量：入口流量和请求；
- 应用场景
  - 网络攻击：恶意用户发起大量请求，占用大量网络连接和流量；
  - 突发流量：产品突然火了，或者内网用户下载操作系统镜像；

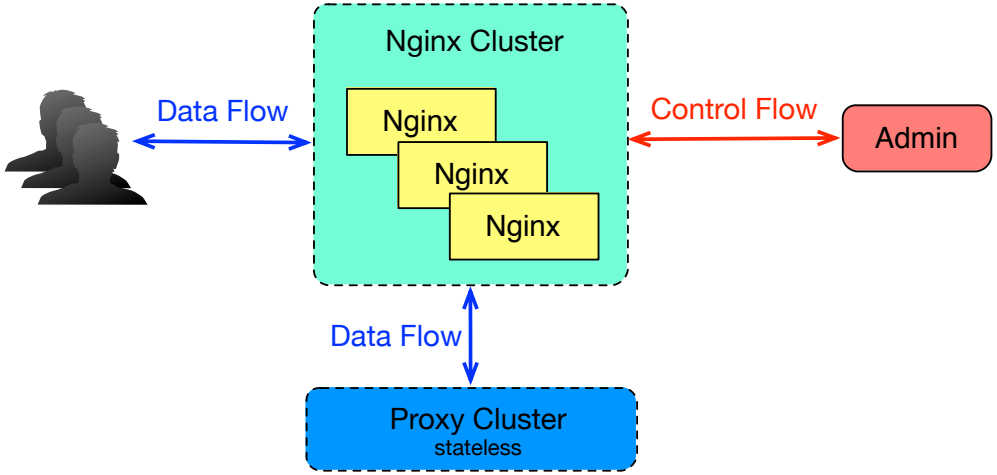
流量隔离-2

- 相关功能
  - 整桶限速
  - 整桶限并发连接
  - 整桶限TPS
  - 单连接限速
  - 实时统计监控
  - 智能流控
  - 全局限速
- 核心模块：Nginx Lua
  - limit_rate：设置下载速度
  - shared_dict：共享内存字典
  - **access_by_lua：请求开始时被调用**
  - **body_filter_by_lua：每读128K被调用**
  - **log_by_lua：请求结束时被调用**

流量隔离-3

- 桶连接数限制
  - access_by_lua被调用时，相关桶连接数加一；
  - log_by_lua被调用时，相关桶连接数减一；
- 桶TPS限制
  - 充值模式，每隔5秒充值；
  - access_by_lua被调用时，相关桶TPS余额减一；
- 桶流量限制
  - access_by_lua被调用时，limit_rate赋予初始流量限制；
  - body_filter_by_lua被调用时，根据采样情况动态调整limit_rate；
  - log_by_lua被调用时，回收当前流量limit_rate；

流量隔离-4



- 隔离示例

```
Limit music163 BUCKET_RATE  = 10M   # 限速10M
Limit music163 BUCKET_CONN = 1000  # 最多支持1000个并发连接
Limit music163 BUCKET_QPS   = 500   # 每秒Query数量最多500个
Limit music163 CONN_RATE    = 500K  # 单连接最大流量500K
```

其他

- 防攻击：类TMD
- 多版本：支持历史版本
- 对象去重：客户端和服务端去重
- 移动端优化：ing
- 存储层优化：ing
- 上传和下载加速：ing

Q&A

THANKS

SequeMedia 盛拓传媒 | ChinaUnix | ITPUB

weibo@dtrees

laidongmin@corp.netease.com