TEG技术发言人

存储世界，不止如此 - EB级存储引擎背后的技术

stephenzou(邹方明)

CNUTCon 2017
全球运维技术大会

上海·光大会展中心大酒店 | 2017.9.10–11

智能时代的新运维

大数据运维
DevOps
安全
SRE
Kubernetes
Serverless
游戏运维
AIOps
智能化运维
基础架构
监控
互联网金融

主办方 Geekbang. 极客邦科技 InfoQ

StuQueue
斯达克学院

实践驱动的IT教育

斯达克学院(StuQ)， 极客邦旗下实践驱动的IT教育平台。通过线下和线上多种形式的综合学习解决方案 ， 帮助IT从业者和研发团队提升技能水平。

人工智能 | 大数据 | 前端开发 | 后端开发 | 架构设计

移动开发 | 运维设计 | 产品测试 | 产品经理 | 技术管理

10大职业技术领域课程

http://www.stuq.org

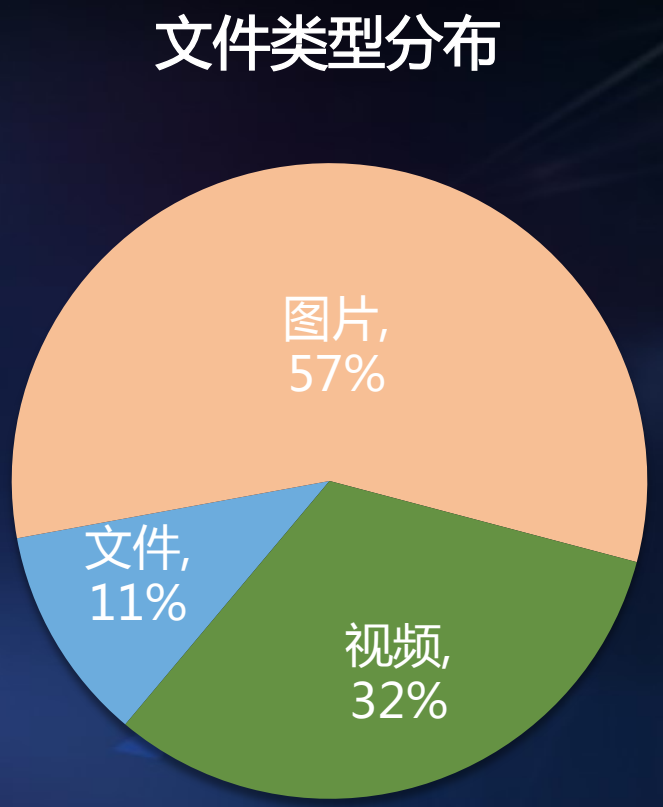
摘要

主要分享从2006开始，腾讯内部从无存储平台到存储量达EB级别的TFS2.0存储平台这一过程中所经历的技术问题。在社交图片和视频盛行的时代，存储系统的设计和运营如何进行适配，揭秘微信c2c图片和视频如何提升体验、降低成本，以及在云时代，腾讯是如何开放内部的存储技术的。

内容

- 1. EB级存储系统的发展过程
- 2. 图片时代存储设计以及在运营上的思考
- 3. 视频时代的存储设计以及在运营上的思考
- 4. 云时代的开放之路

存储数据增长趋势及内容分布



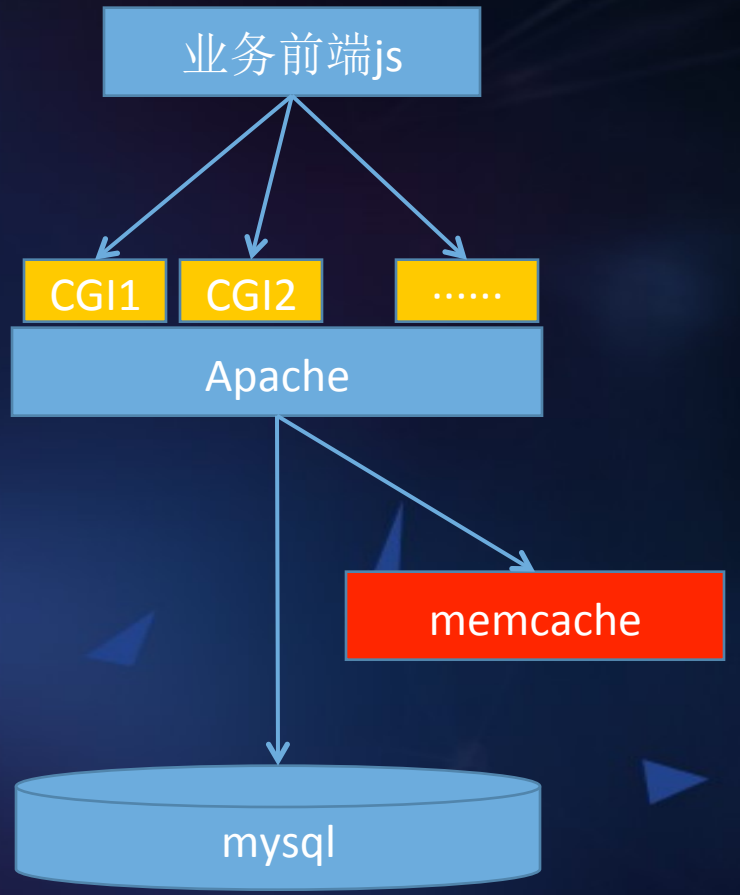
混乱时代

- 业务众多，各自存储
- 方案多样，运营复杂
- 性能低，成本高
- 容灾不完善

典型业务

相册

早期存储架构

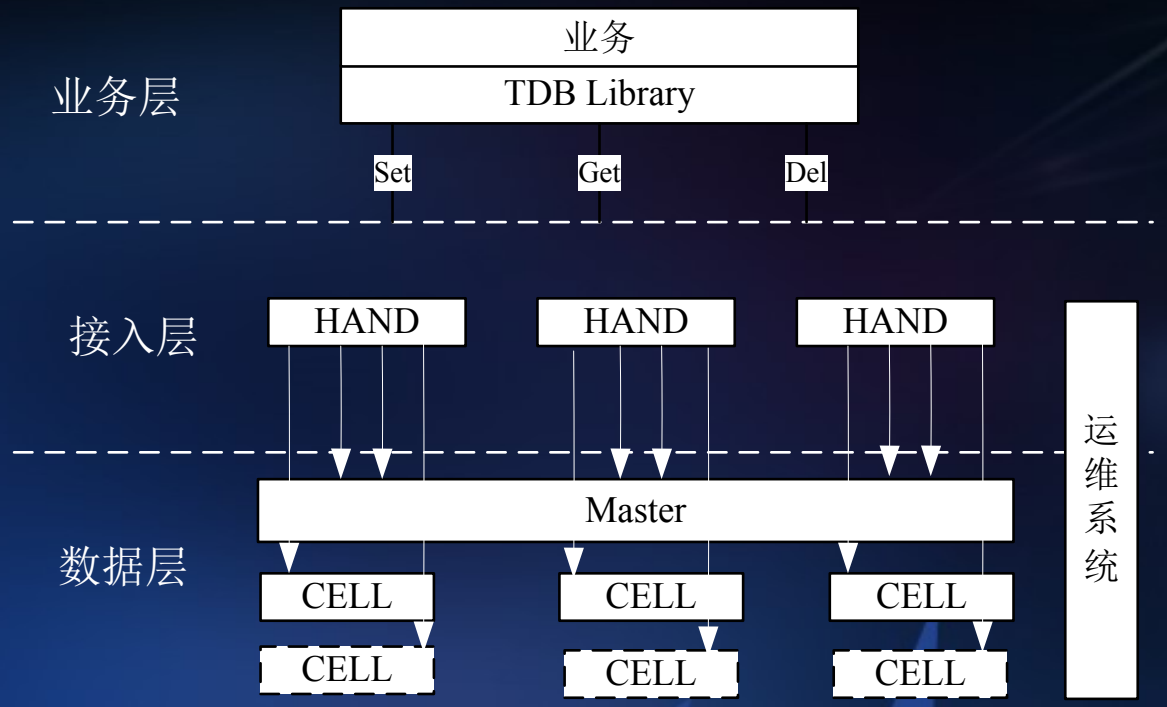


存储的第一代需求

- UGC业务爆发，IO扛不住
- 频繁扩容，运维伤不起
- 存储成本高

TFS第一代之KV存储系统

- 经典三层架构
- 接口简单，快速接入
- 一致性哈希分布
- 易扩展，高性能
- 半自动化运营



通用协议

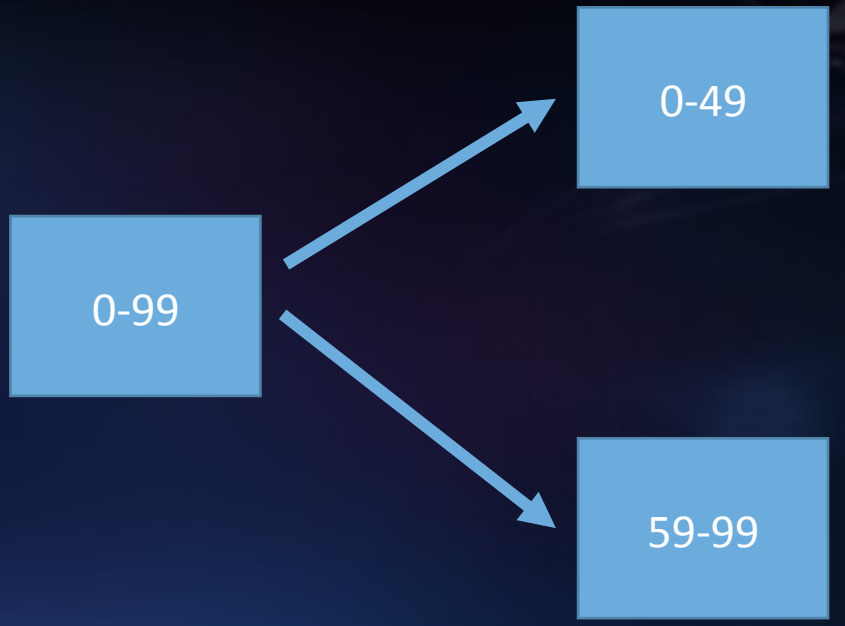
- GET—SET—DELETE　　接口通用
- bid—tid—key（string）　　KEY通用
- Value（string）　　返回通用
- KV系统，异构升级的基础

扩展性设计—自动分裂

[0 - 99]	**tablet**
[100 - 199]	**tablet**
……	……
[999900 - 999999]	**tablet**

路由项

10000个小表



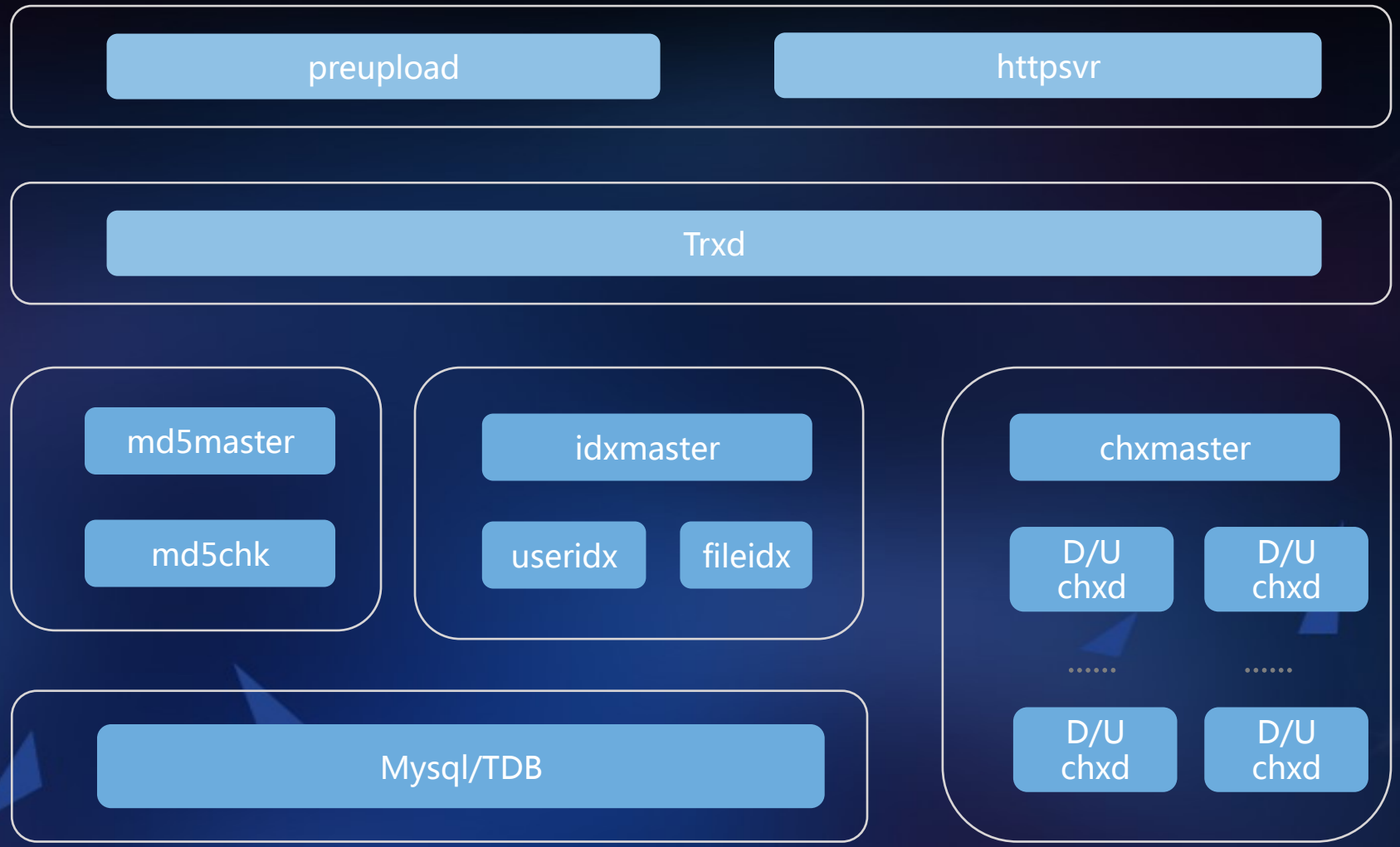
10000个小表组成的初始路由表，最大支撑2.56T

高性能保证

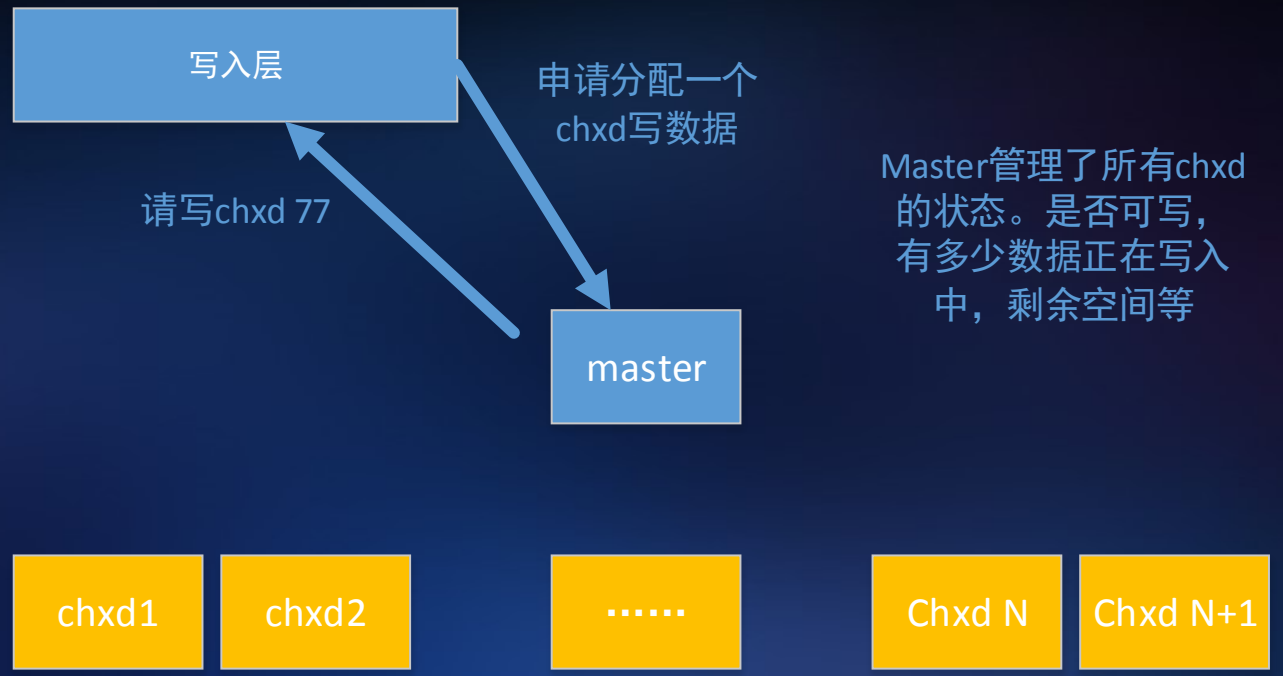
- 磁盘数据访问
  - -12个基本页中的基本块看成是一个平面的Hash表
  - -12个基本页公用4个溢出页
  - -落在同一个基本块的数据用TLV的方式打包存储
  - -超出基本块大小的数据使用连续的溢出块存储

- 效果
  - 无庞大的内存索引开销
  - 80%的记录一次IO
  - 按page的方式快速搬迁数据

TFS第一代—文件系统

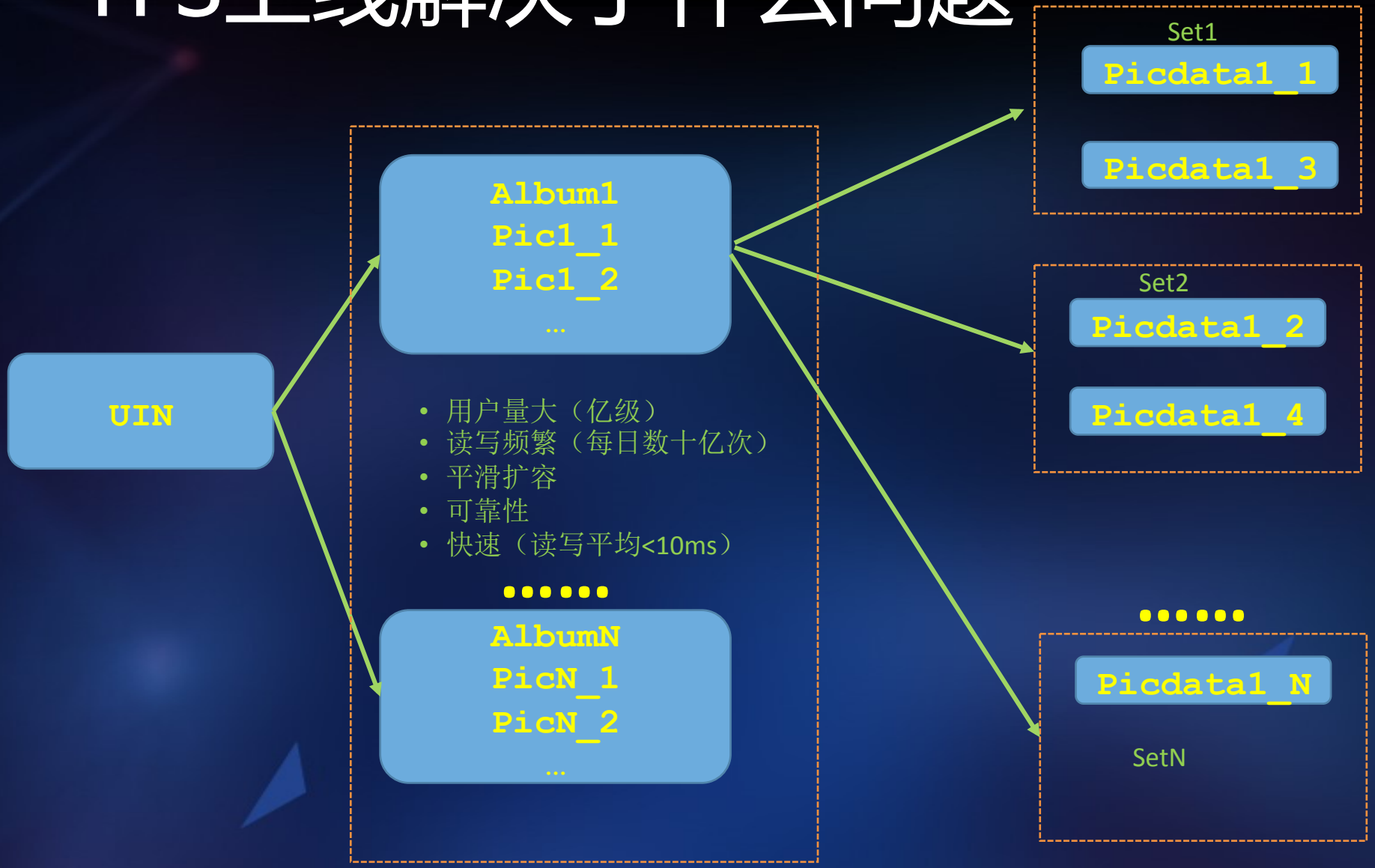

preupload
httpsvr
Trxd
md5master
md5chk
idxmaster
useridx
fileidx
chxmaster
D/U
chxd
D/U
chxd
.....
.....
D/U
chxd
D/U
chxd
Mysql/TDB

IO调度


写入层
申请分配一个chxd写数据
请写chxd 77
master
Master管理了所有chxd的状态。是否可写，有多少数据正在写入中，剩余空间等
chxd1
chxd2
……
Chxd N
Chxd N+1

空间管理

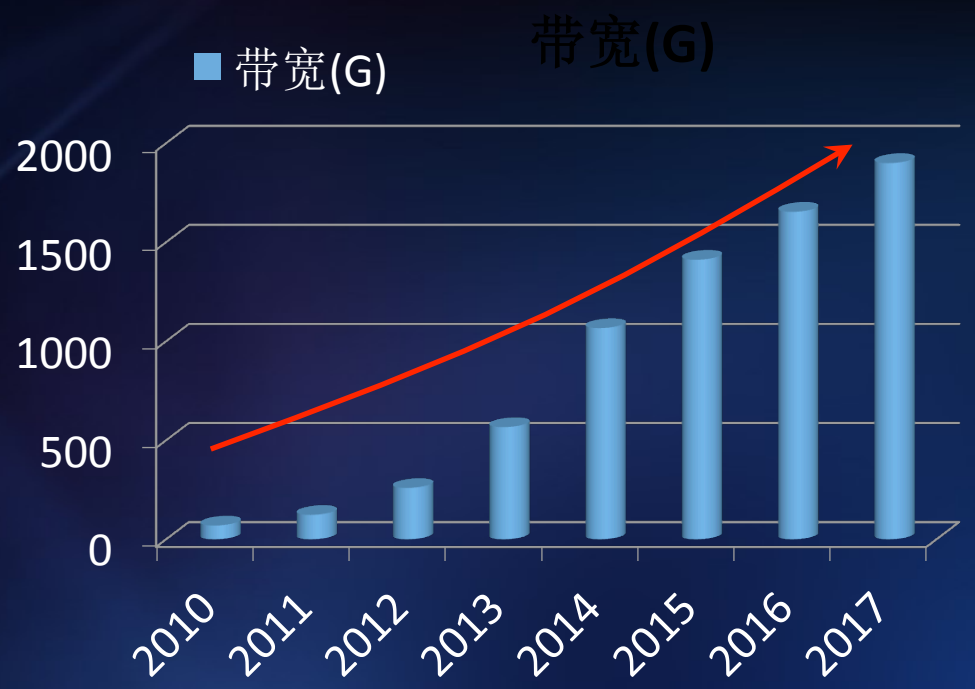
TFS上线解决了什么问题



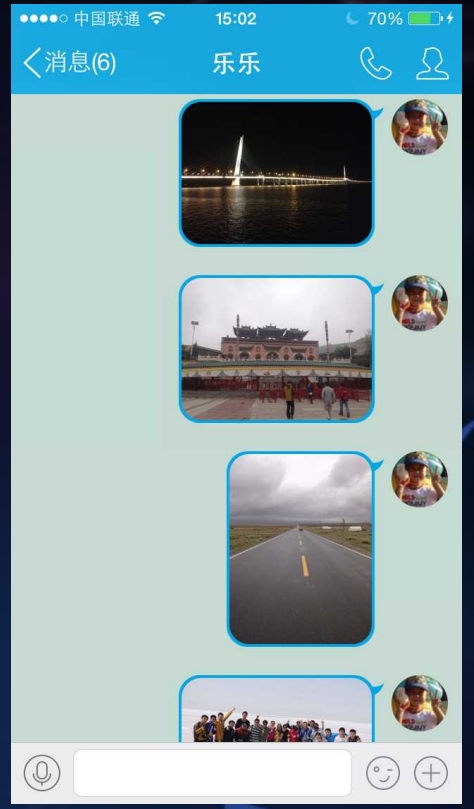
- 用户量大（亿级）
- 读写频繁（每日数十亿次）
- 平滑扩容
- 可靠性
- 快速（读写平均<10ms）

- 数据可靠性
- 成本节省（千万RMB）
- 读写数据量大（日增数十TB，日下载流量数百G）
- 平滑扩容
- 高效运维(坏盘探测与更换、set化运维)

图片时代的来临



数百PB存储、T级下载带宽。
日上传数十亿次、日下载数千亿次。



图片时代存储系统遇到的挑战

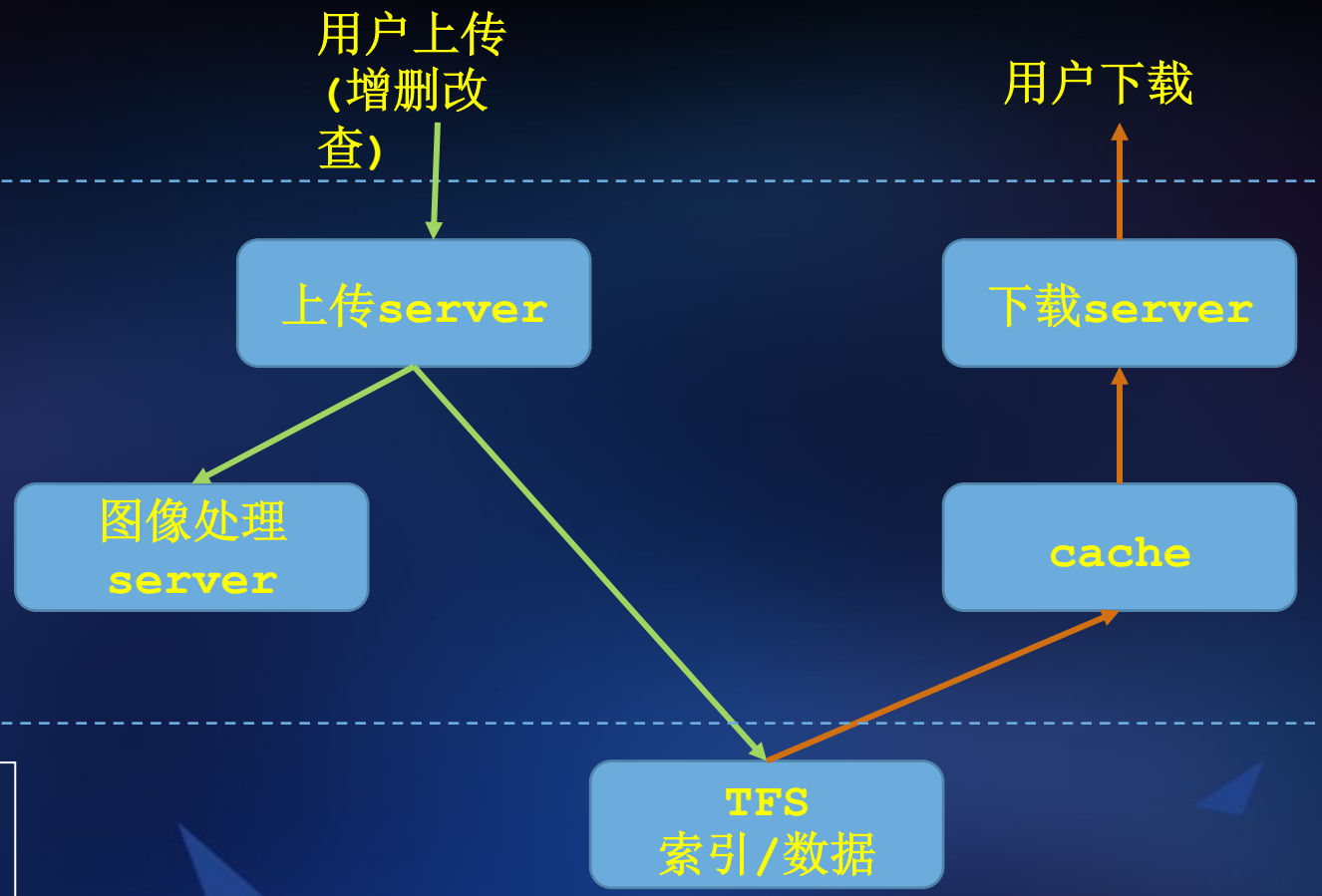
- 文件小，50%删除
- 吞吐极大

图片定制存储系统

上传加速
- 提供SDK和标准http协议
- SDK支持单机故障容灾
- 支持就近上传，IDC容灾

优化的图像处理
- 支持自定义多种尺寸
- 有限支持下载压缩
- 自定义水印
- 裁剪、锐化、旋转等
- 支持webp/jpeg压缩
- 质量相同，压缩比例更高

安全审核
- 最大社交平台十多年恶意库积累；
- 相似度指纹识别；
- 节省99%以上人工审核成本

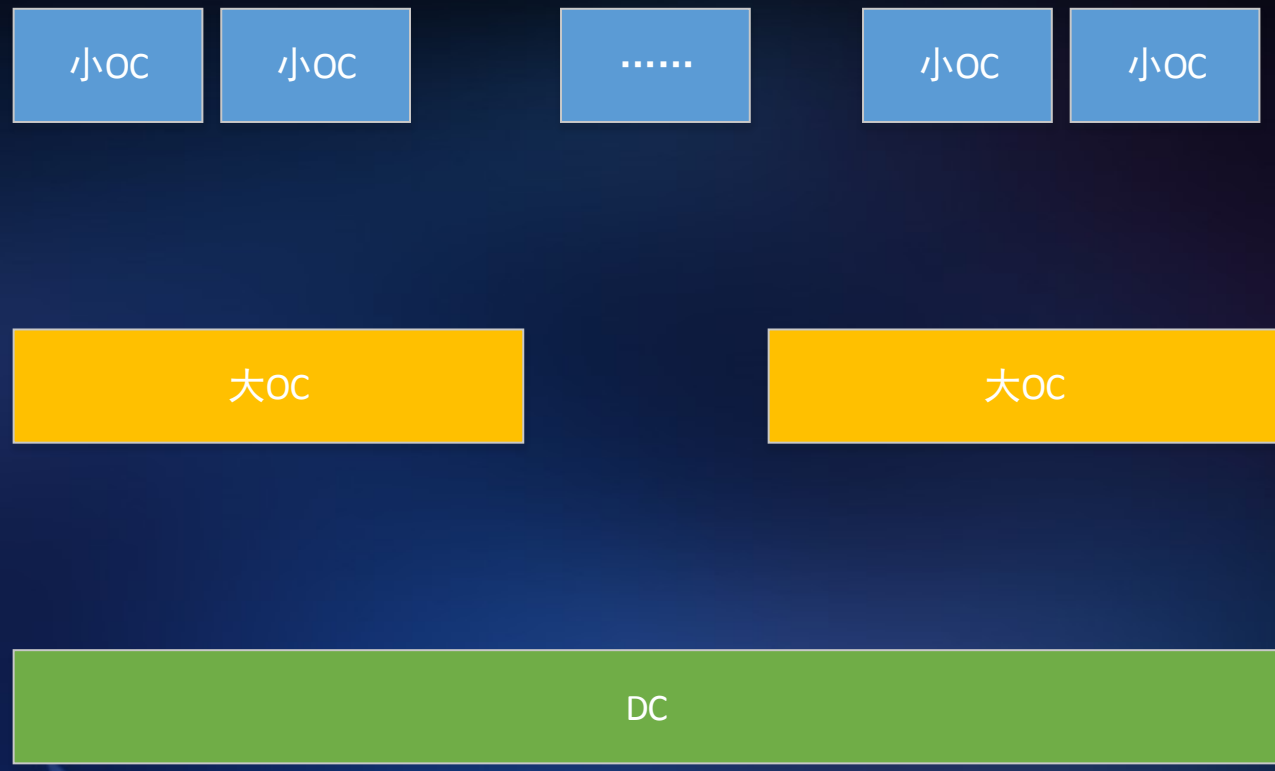


下载加速
- 腾讯高速CDN加速
- 支持防盗链
- 支持多种Cache策略
- 支持Webp/Jpeg等多种格式下载

立体监控
- 服务器端监控
- 用户体验评测
- 纬度丰富的检测扫描系统

高速回收

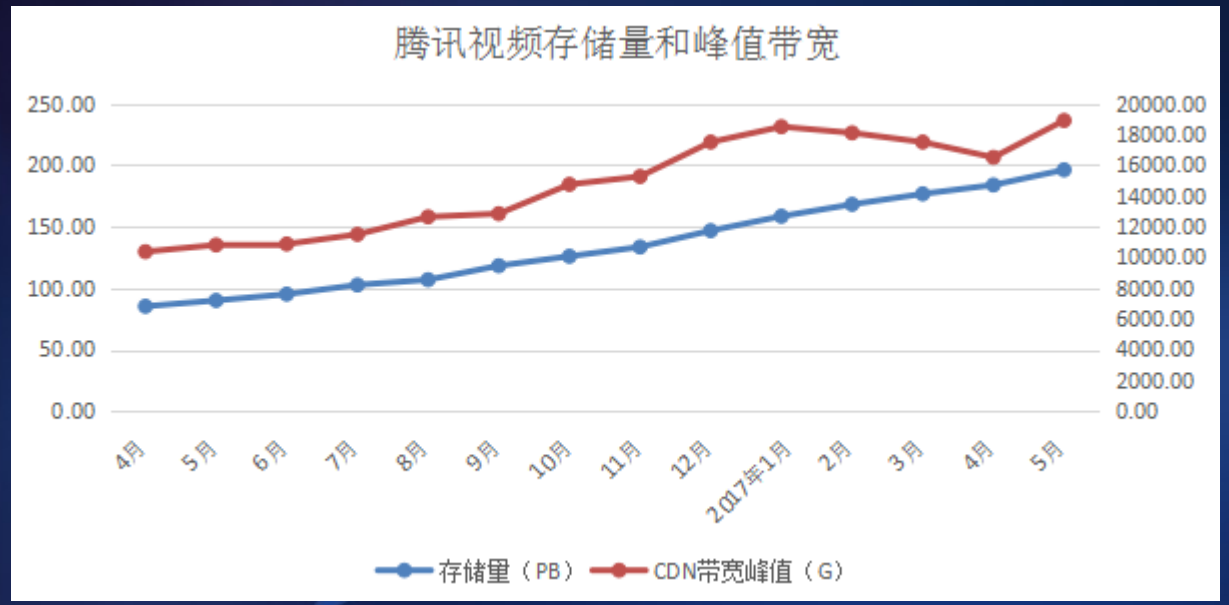
Cache设计



视频时代的来临

随着电影IP的兴起，各大视频应用呈现井喷式发展

付费VIP用户+广告收入增长迅速，以腾讯视频为例，16年流水达到百亿级别



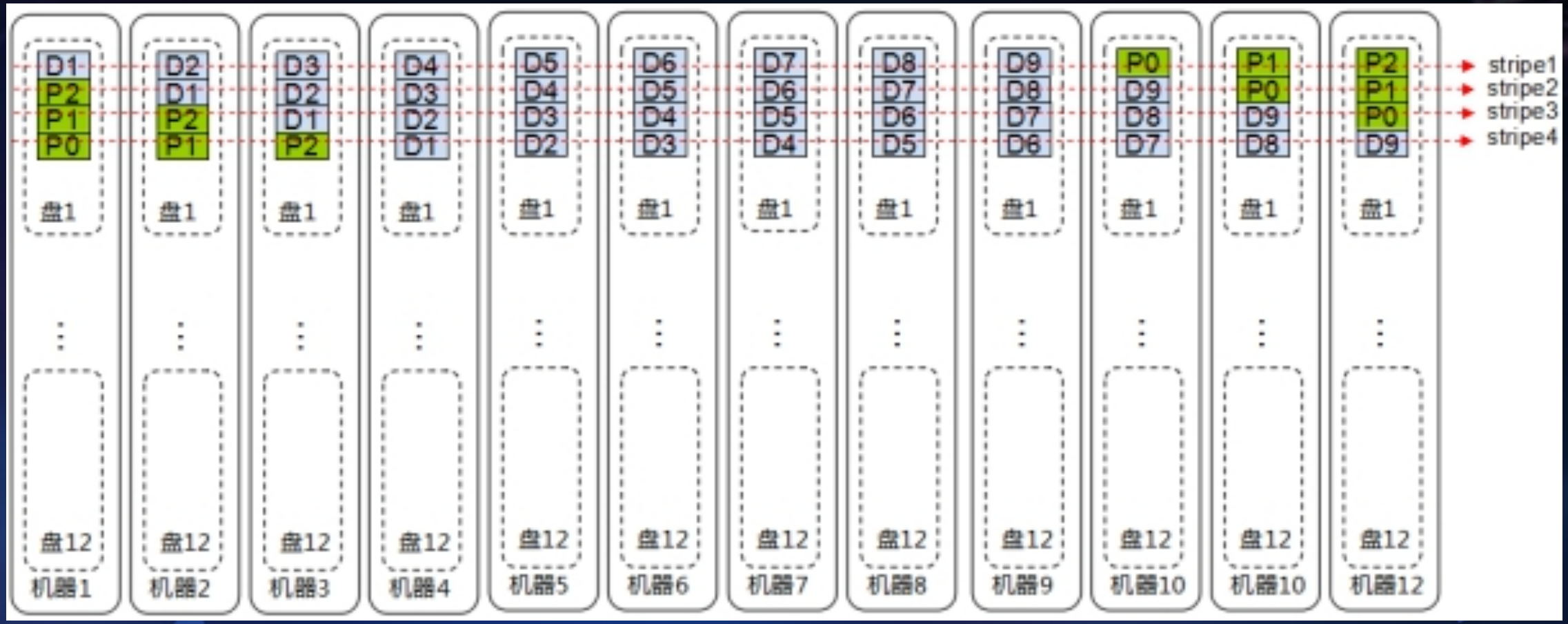
视频时代的存储挑战

- 超大文件（50T）
- 极冷文件的合理安置

超大视频


50T
一个50T的文件
5G
块1
5G
块2
5G
块10000
1M
分片1
1M
分片2
1M
分片n
1M
分片1
1M
分片2
1M
分片n
1M
分片1
1M
分片2
1M
分片n

极冷视频


D1
P2
P1
P0
D2
D1
P2
P1
D3
D2
D1
P2
D4
D3
D2
D1
D5
D4
D3
D2
D6
D5
D4
D3
D7
D6
D5
D4
D8
D7
D6
D5
D9
D8
D7
D6
P0
D9
D8
D7
P1
P0
D9
D8
P2
P1
P0
D9
盘1
盘12
机器1
机器2
机器3
机器4
机器5
机器6
机器7
机器8
机器9
机器10
机器10
机器12
stripe1
stripe2
stripe3
stripe4

云存储时代

我们准备开放什么

腾讯公司的存储技术

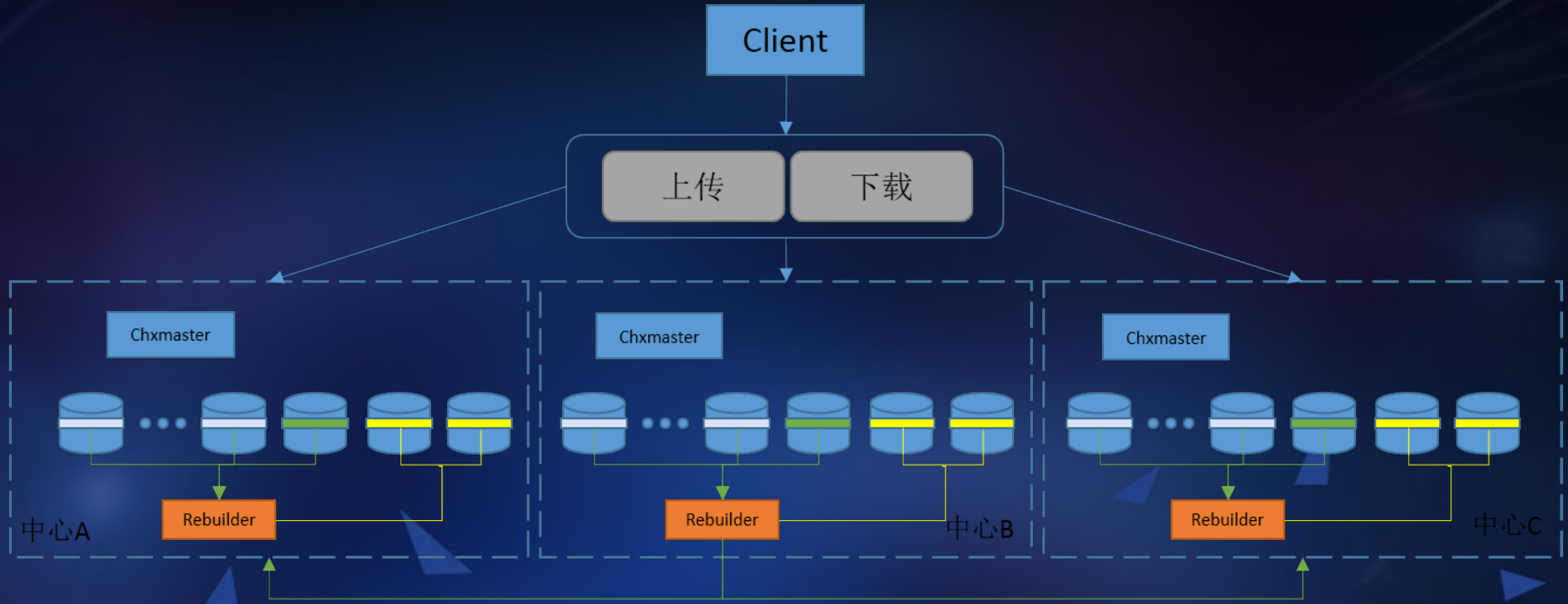
对数据深度理解的定制能力

对场景深度理解的业务能力

手机备份需求

- 多IDC（间隔80公里）
- 一地IDC灾难，数据保证完整可恢复

三地多中心解决方案


Client
上传
下载
Chxmaster
Chxmaster
Chxmaster
Rebuilder
Rebuilder
Rebuilder
中心A
中心B
中心C

归档需求

- 我们只需要价格低
  - 一年只需要读写入数据的万分之一
  - 可以延迟读

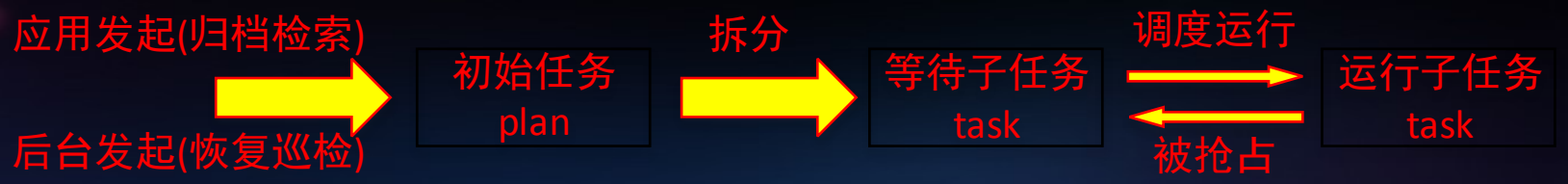
类glacier方案

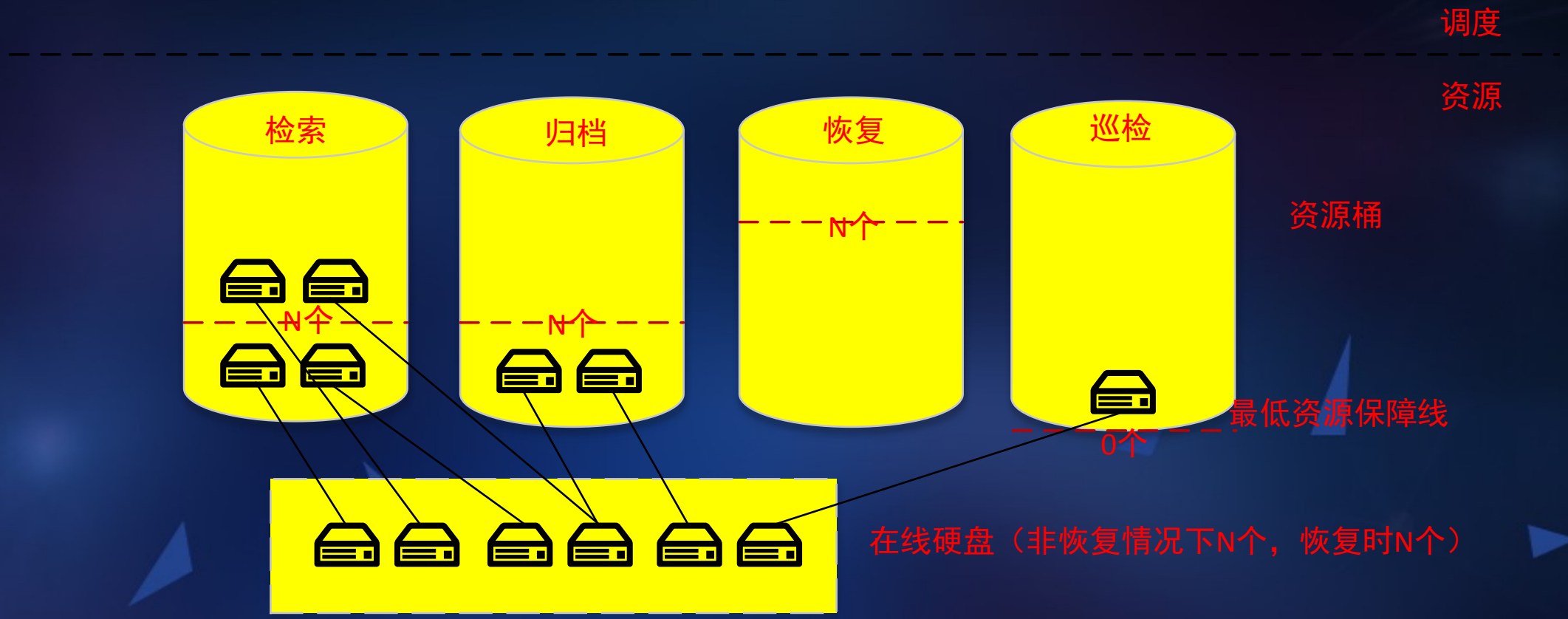
- 通过牺牲可用性（允许存在恢复时间）以及利用业务特性（批量导入，大文件），将成本发挥到极致的空间管理方法。

整机柜技术


JBOD
JBOD
JBOD
2个JBOD
一个故障域
每个硬盘切分NGB的chunk
挑选N个chunk组成stripe
分片1
分片2
按照选定粒度
切分

磁盘休眠


应用发起(归档检索)
后台发起(恢复巡检)
初始任务
plan
拆分
等待子任务
task
调度运行
被抢占
运行子任务
task


调度
资源
检索
归档
恢复
巡检
N个
N个
N个
资源桶
最低资源保障线
0个
在线硬盘（非恢复情况下N个，恢复时N个）

更多内容

谢谢

让创新技术推动社会进步

HELP TO BUILD A BETTER SOCIETY WITH
INNOVATIVE TECHNOLOGIES

Geekbang.
极客邦科技

InfoQ

专注中高端技术人员的技术媒体

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

高端技术人员学习型社交平台

StuQ 斯达克学院

实践驱动的 IT 教育平台

地址：北京市朝阳区洛娃大厦C座8层1801室　　网址：www.geekbang.org