

# 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  
安全  
SRE  
DevOps  
Kubernetes  
Serverless  
游戏运维  
AI Ops  
智能化运维  
基础架构  
监控  
互联网金融



# StuQ

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。



10大职业技术领域课程

<http://www.stuq.org>

# 摘要

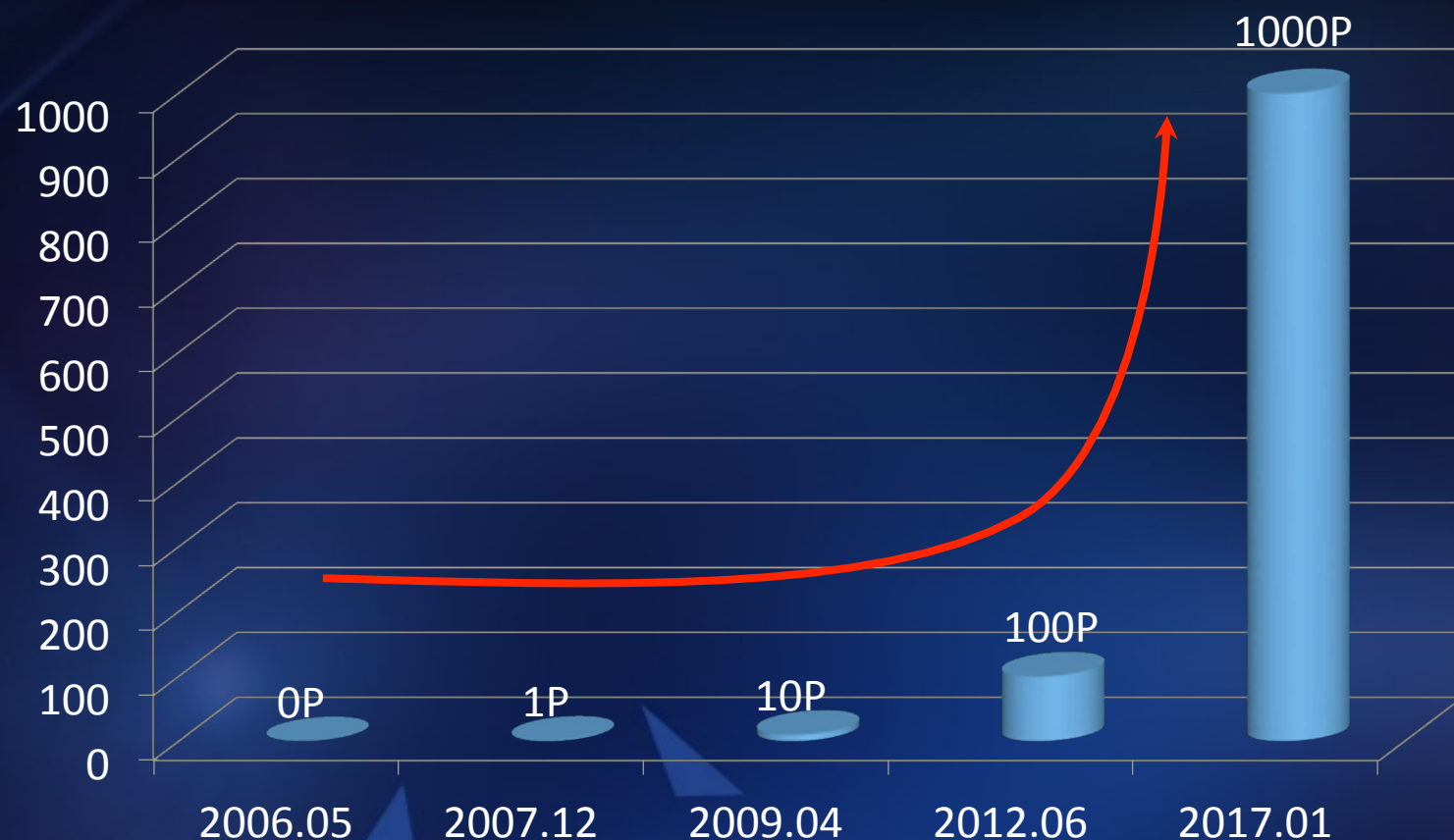
主要分享从2006年开始，腾讯内部从无存储平台到存储量达EB级别的TFS2.0存储平台这一过程中所经历的技术问题。在社交图片和视频盛行的时代，存储系统的设计和运营如何进行适配，揭秘微信c2c图片和视频如何提升体验、降低成本，以及在云时代，腾讯是如何开放内部的存储技术的。

# 内容

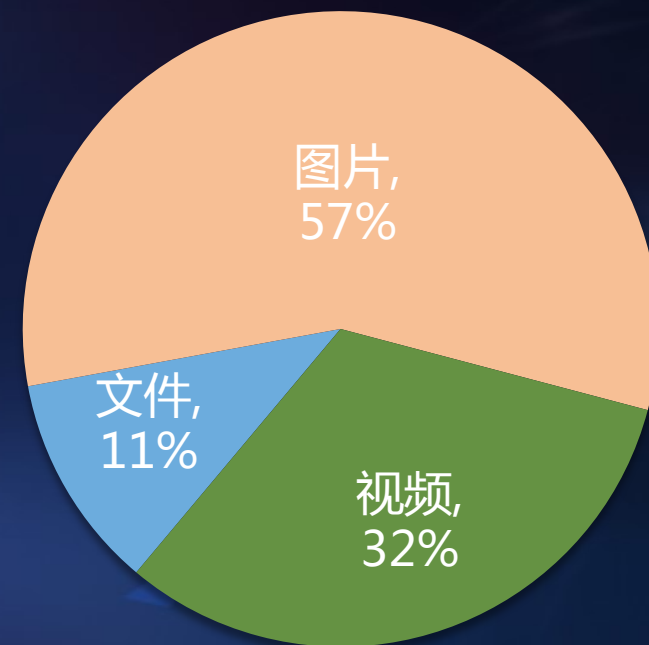
- 1. EB级存储系统的发展过程
- 2. 图片时代存储设计以及在运营上的思考
- 3. 视频时代的存储设计以及在运营上的思考
- 4. 云时代的开放之路

# 存储数据增长趋势及内容分布

## 存储量增长趋势



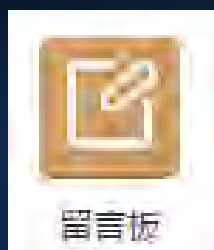
## 文件类型分布



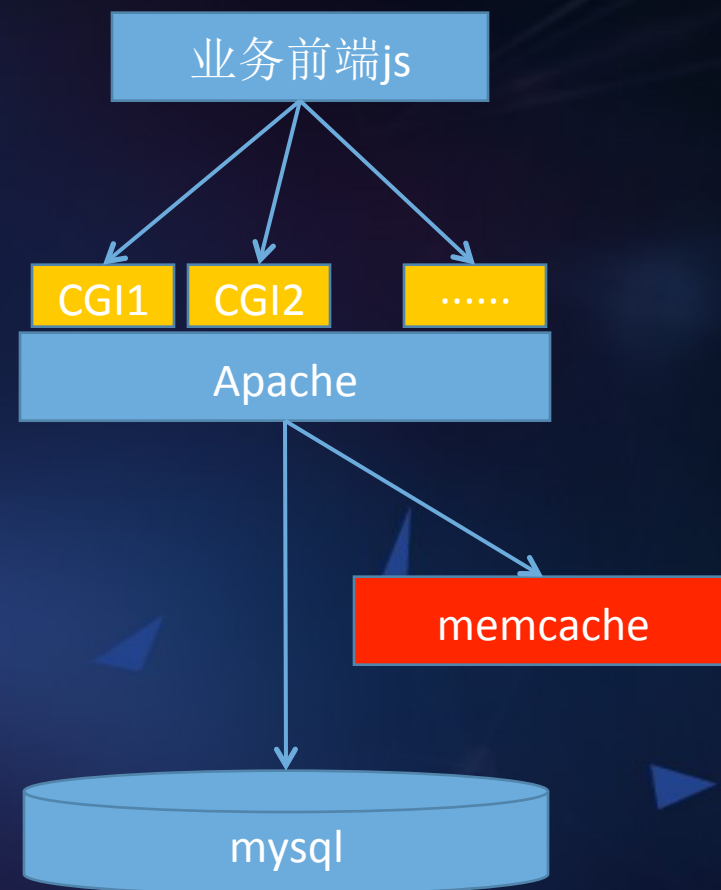
# 混乱时代

- 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



早期存储架构



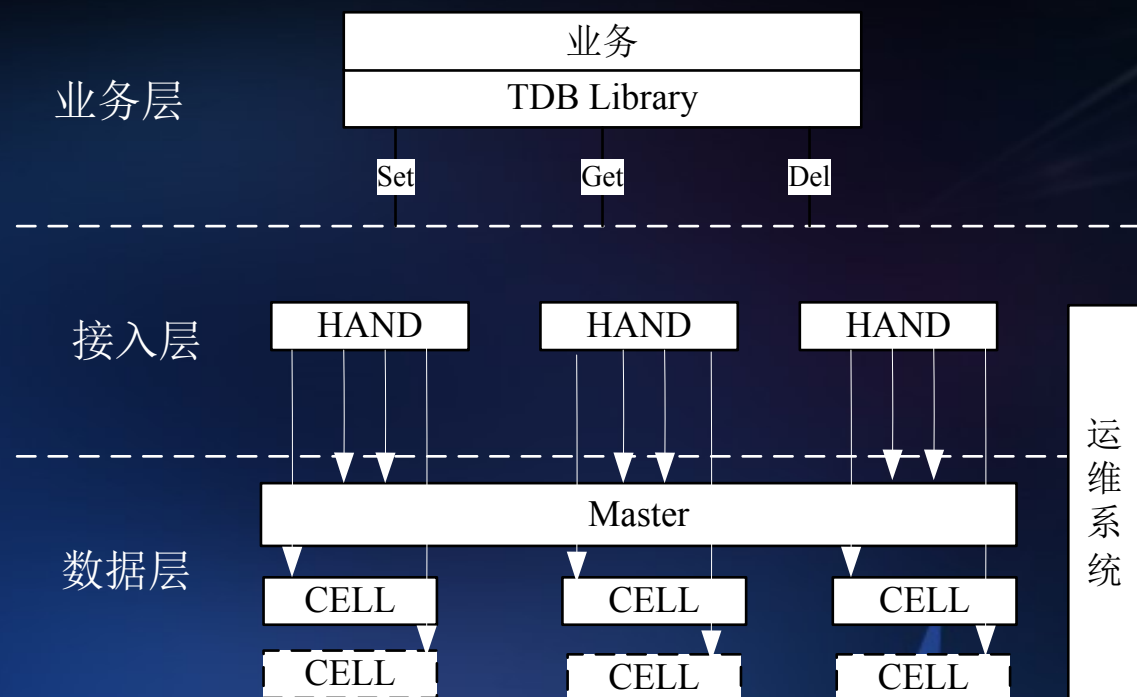
# 存储的第一代需求

- 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

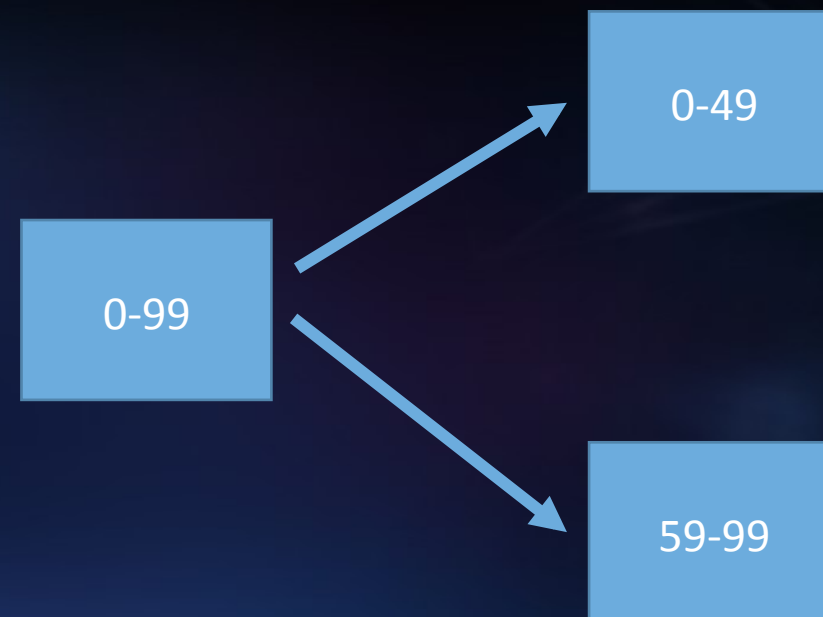
# 扩展性设计—自动分裂

[ 0 - 99]	tablet
[ 100 - 199]	tablet
.....	.....
[ 999900 - 999999]	tablet

路由项

.....

↓  
10000个小表  
↓

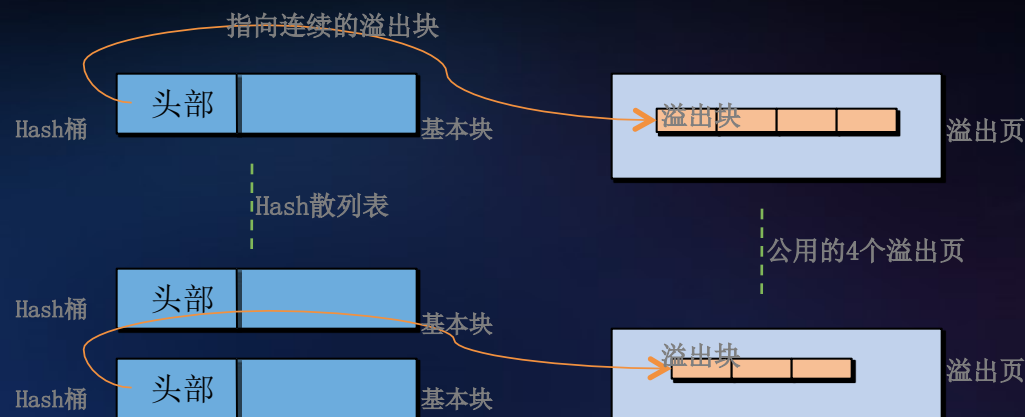


10000个小表组成的初始路由表，最大支撑  
2.56T

# 高性能保证

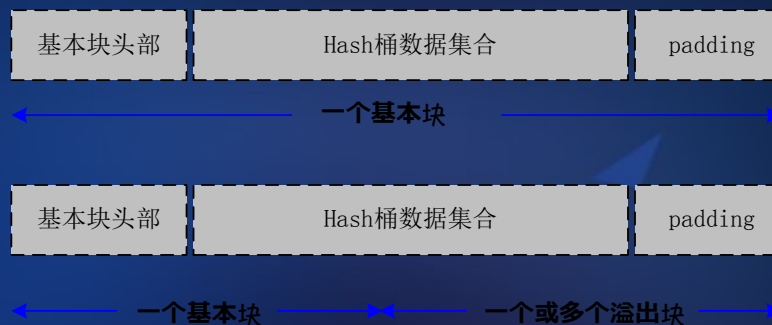
## ● 磁盘数据访问

- 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

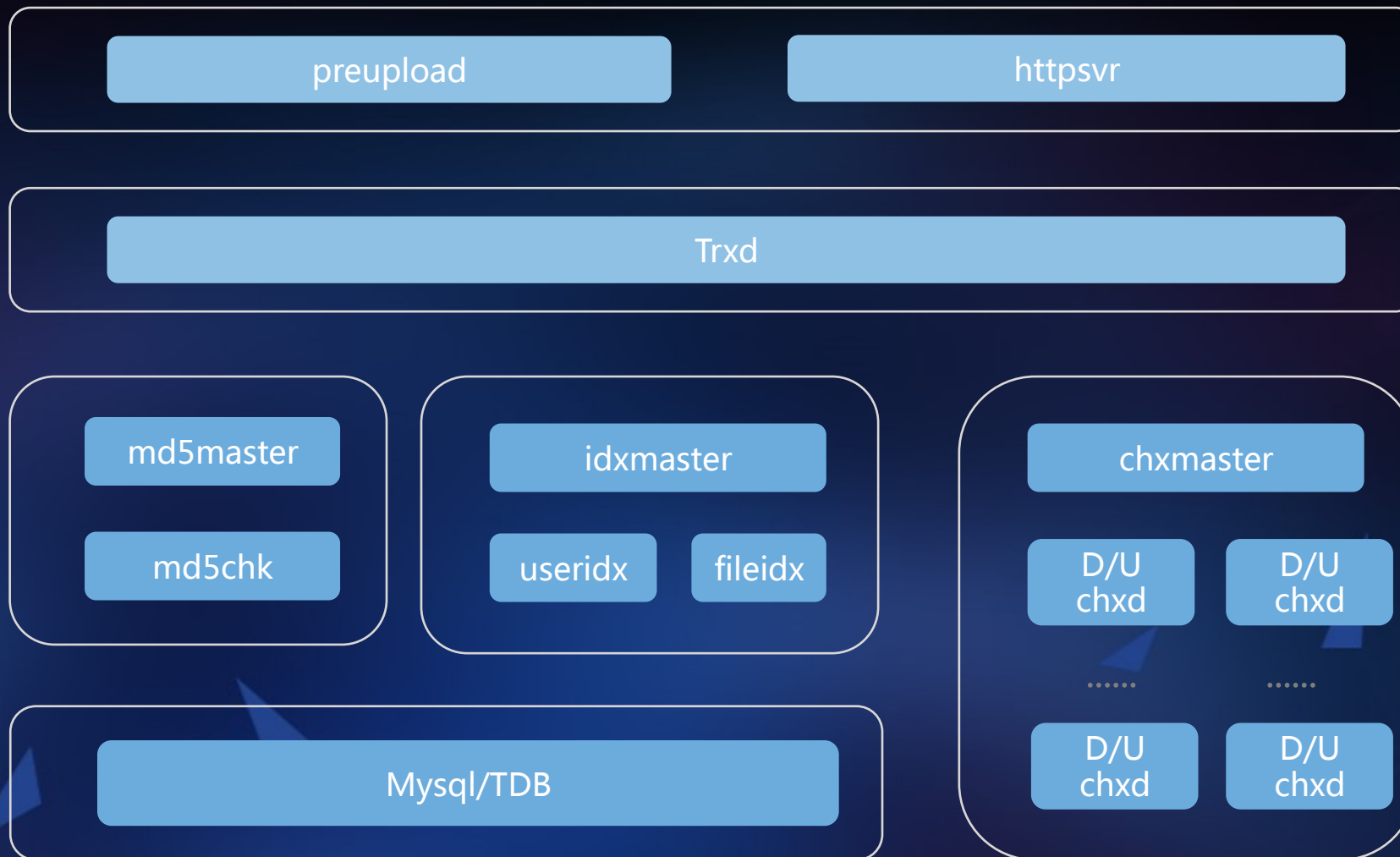


## ● 效果

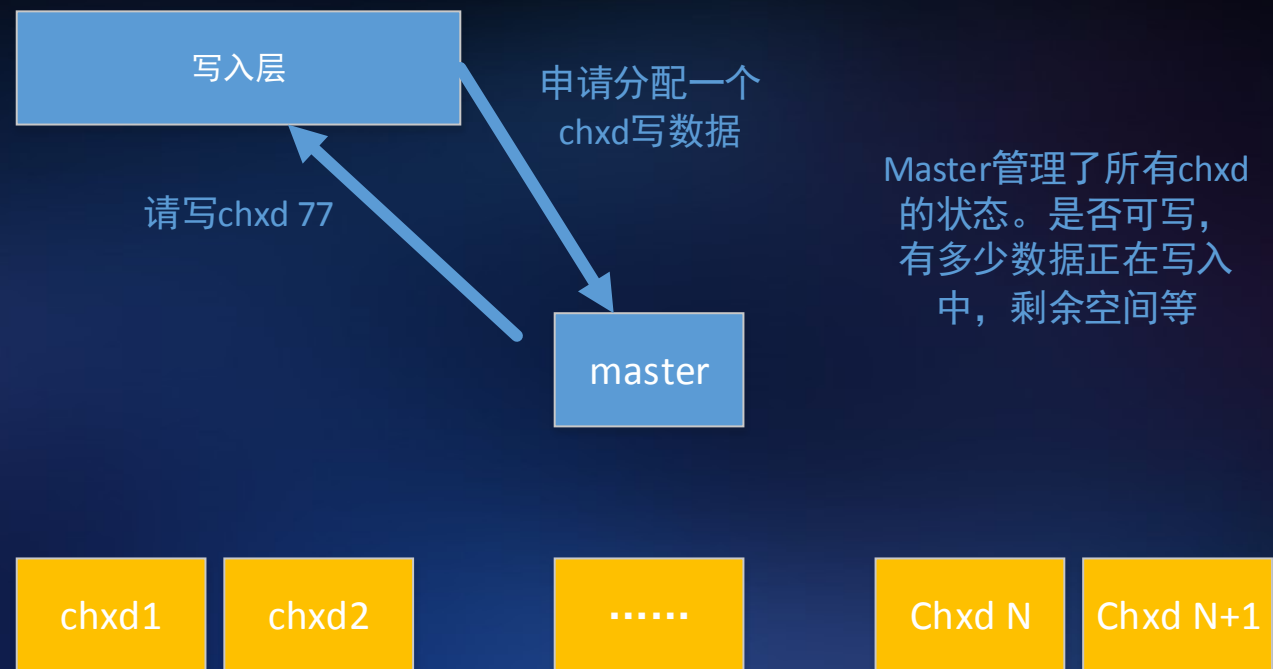
- 无庞大的内存索引开销
- **80%的记录一次IO**
- 按page的方式快速搬迁数据



# TFS第一代—文件系统



# IO调度



# 空间管理

数据写入



数据回收前



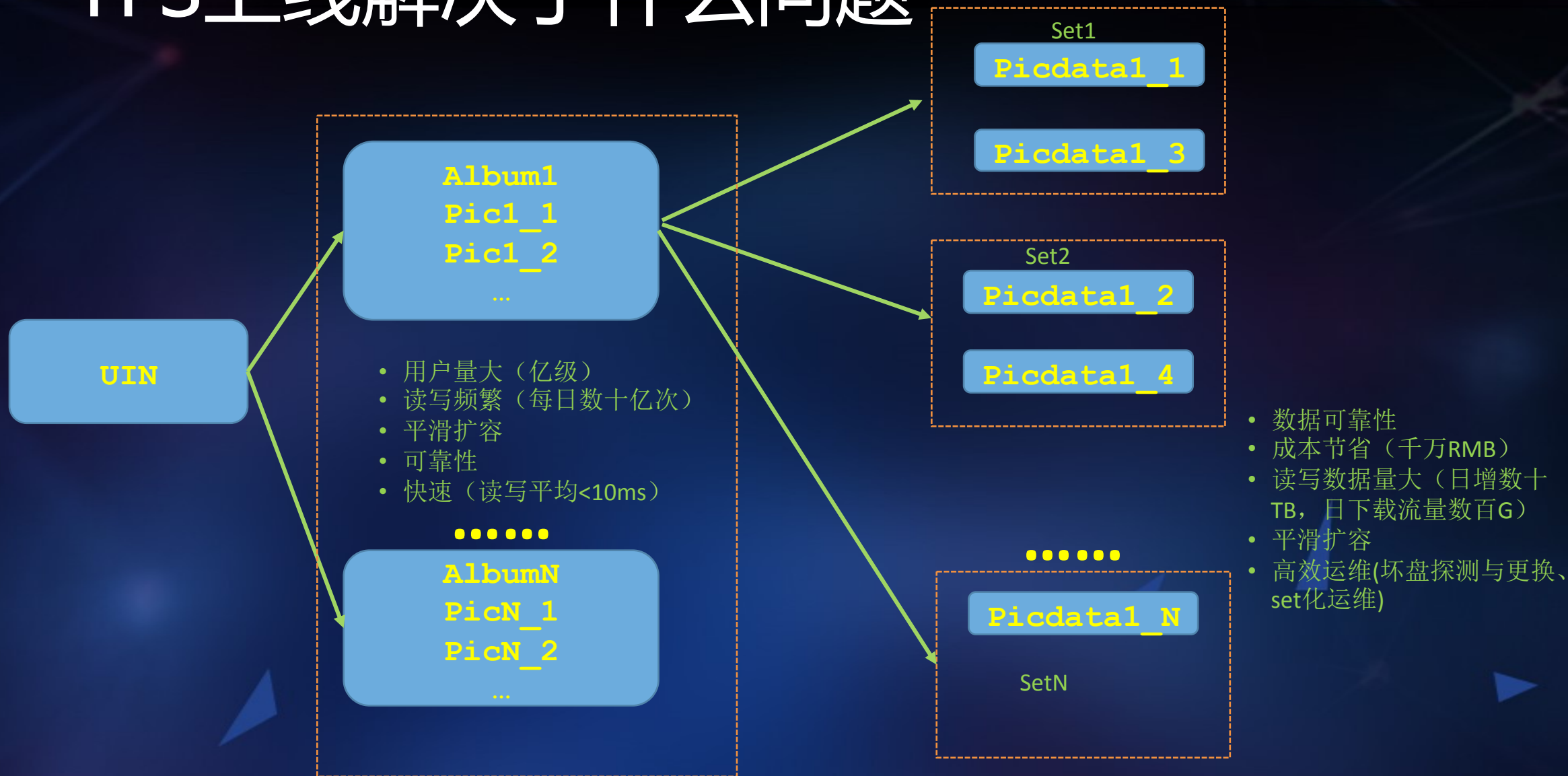
数据回收后



数据回收后写入

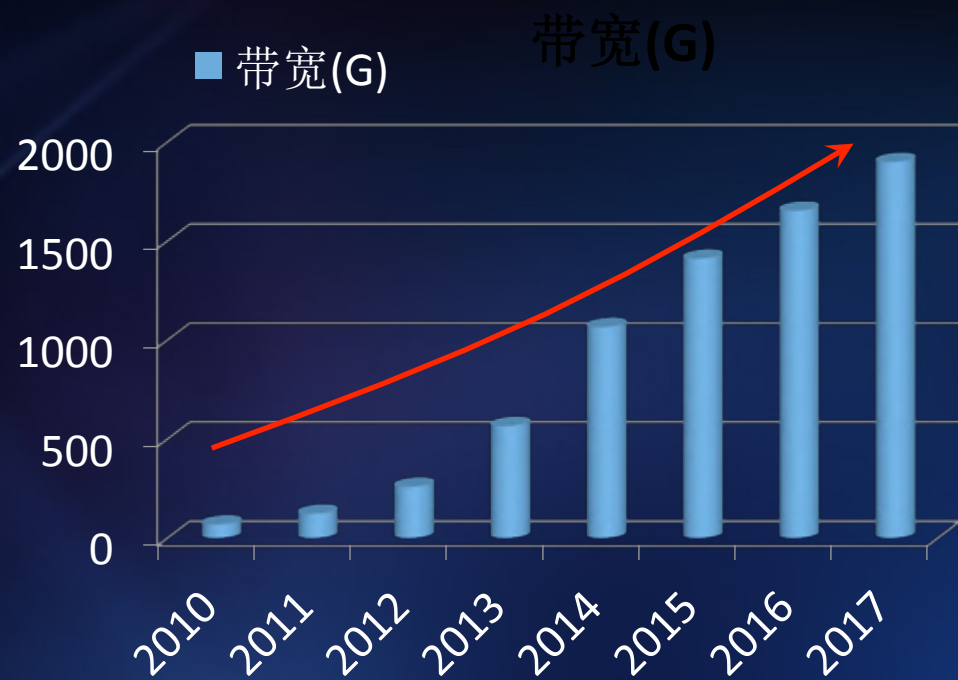


# TFS上线解决了什么问题

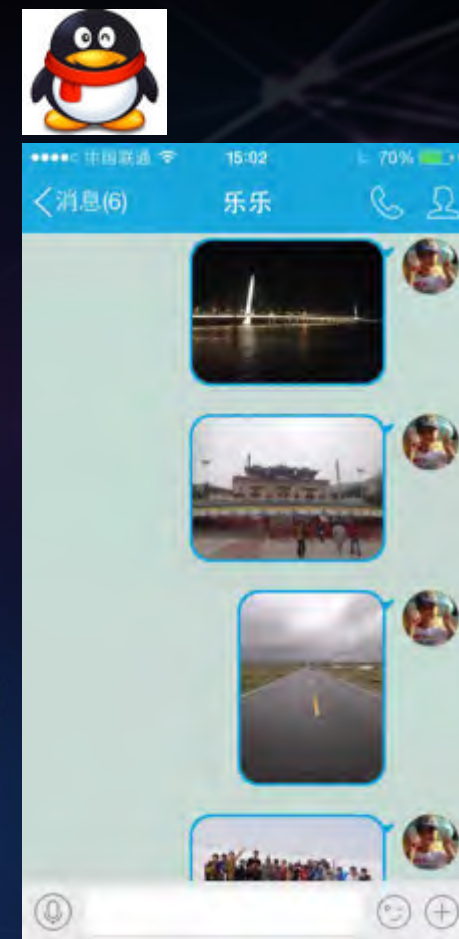




# 图片时代的来临



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

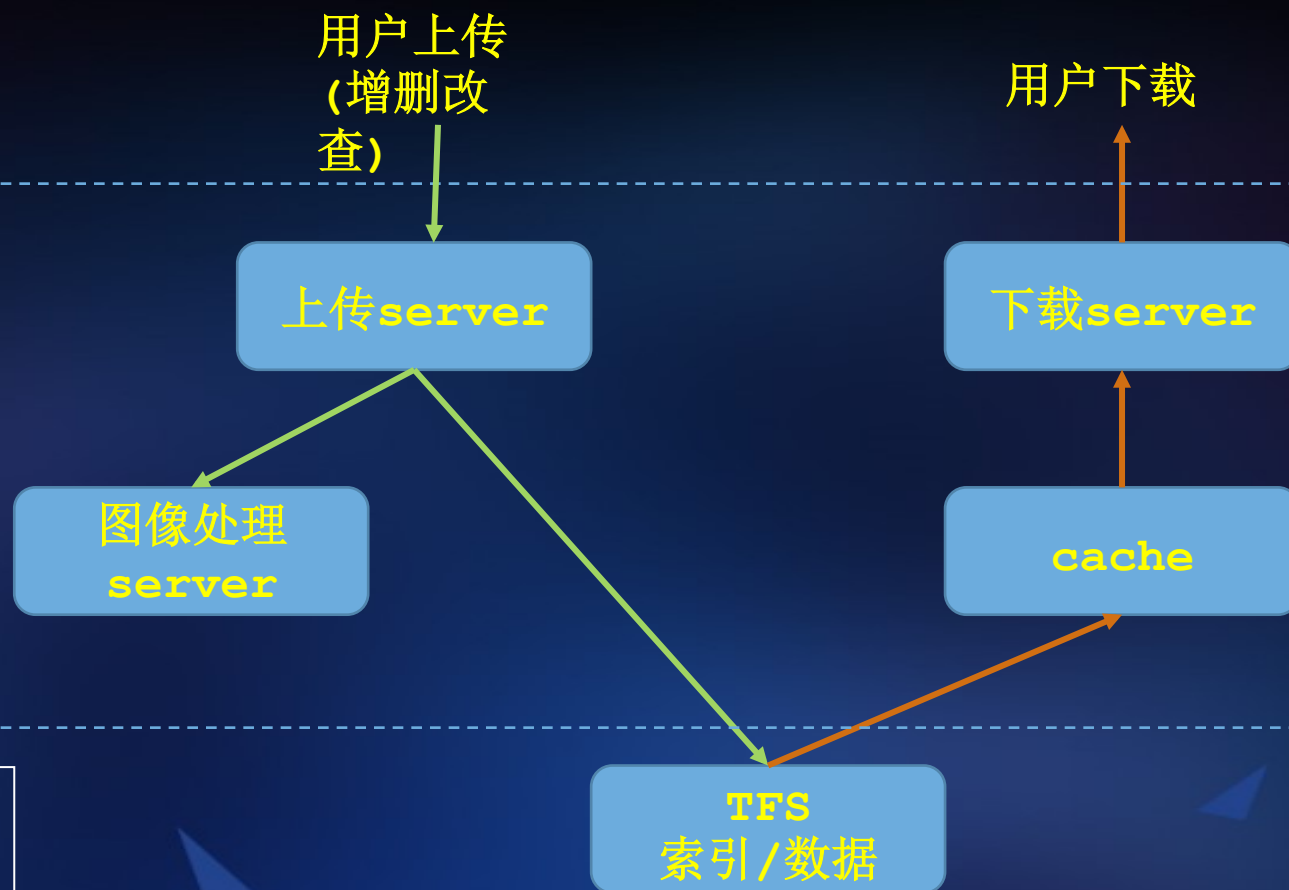
- 提供SDK和标准http协议
- SDK支持单机故障容灾
- 支持就近上传，IDC容灾

## 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



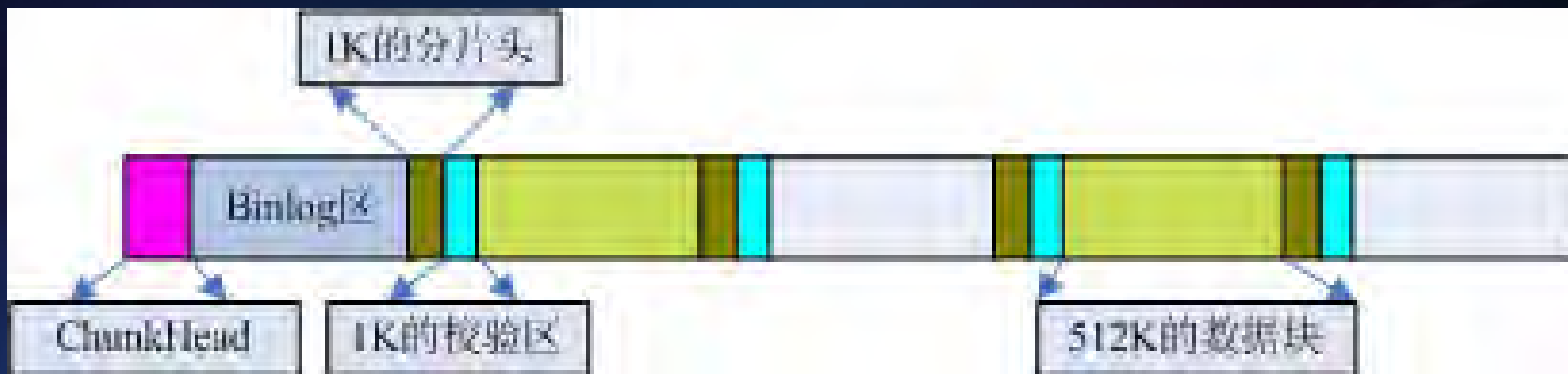
## 下载加速

- 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

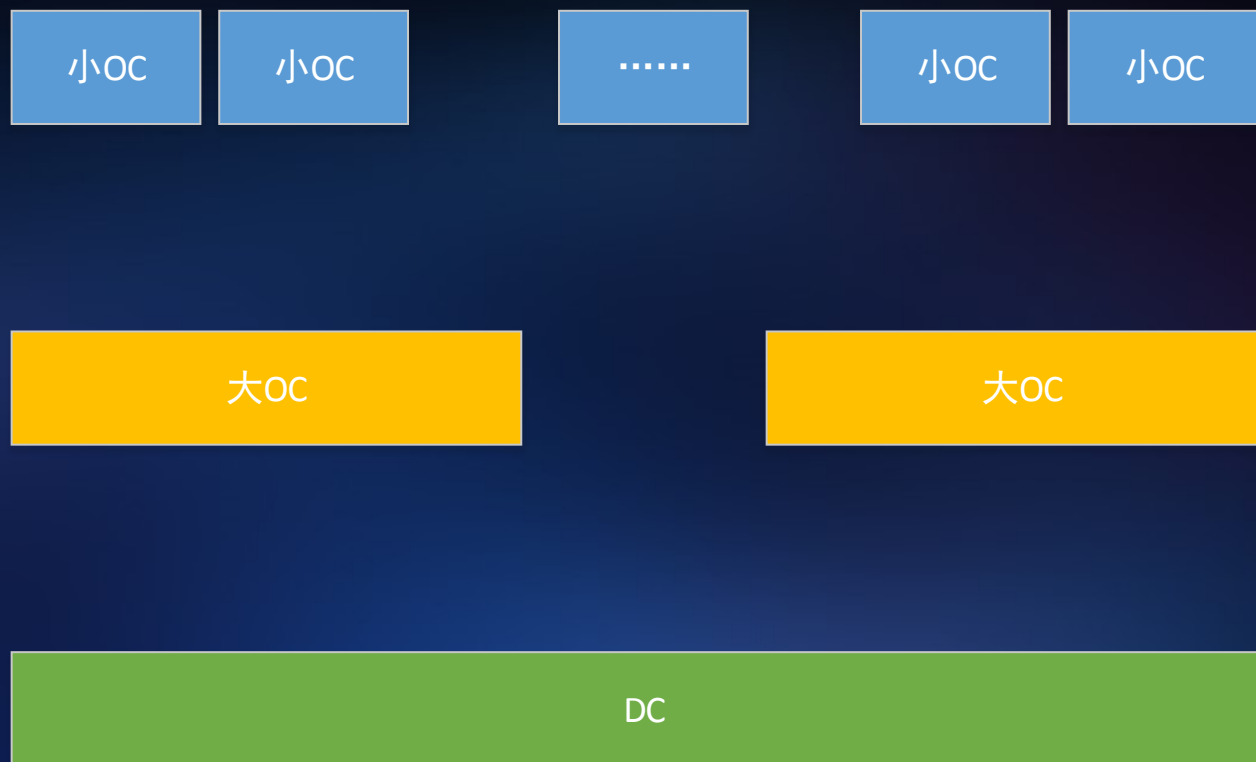
## 立体监控

- 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



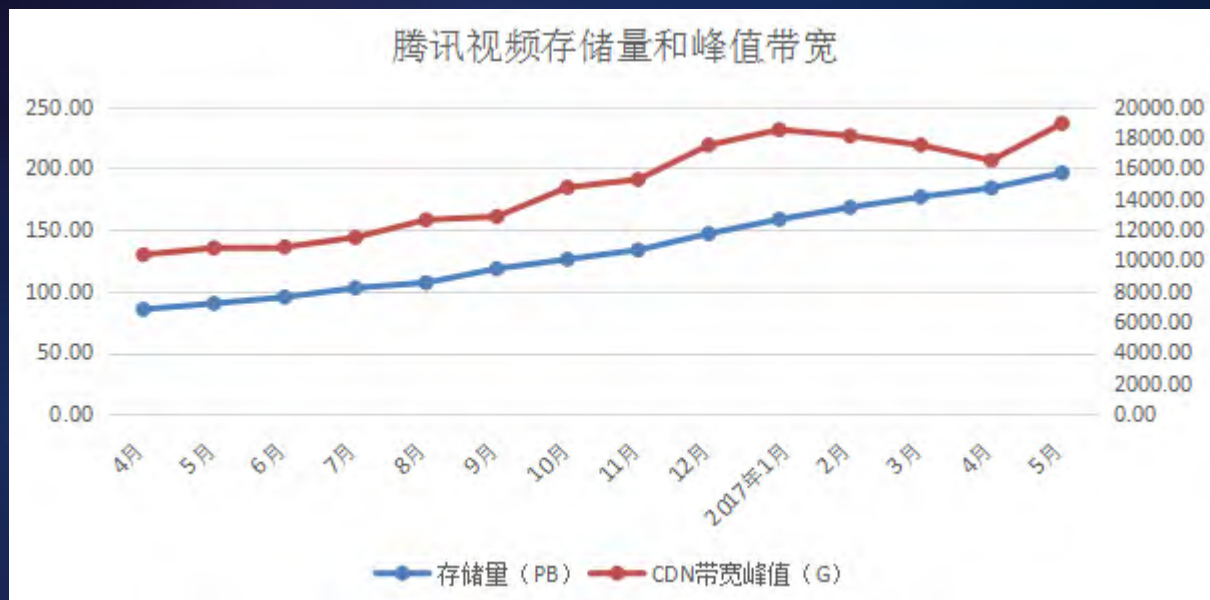
# Cache设计



# 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

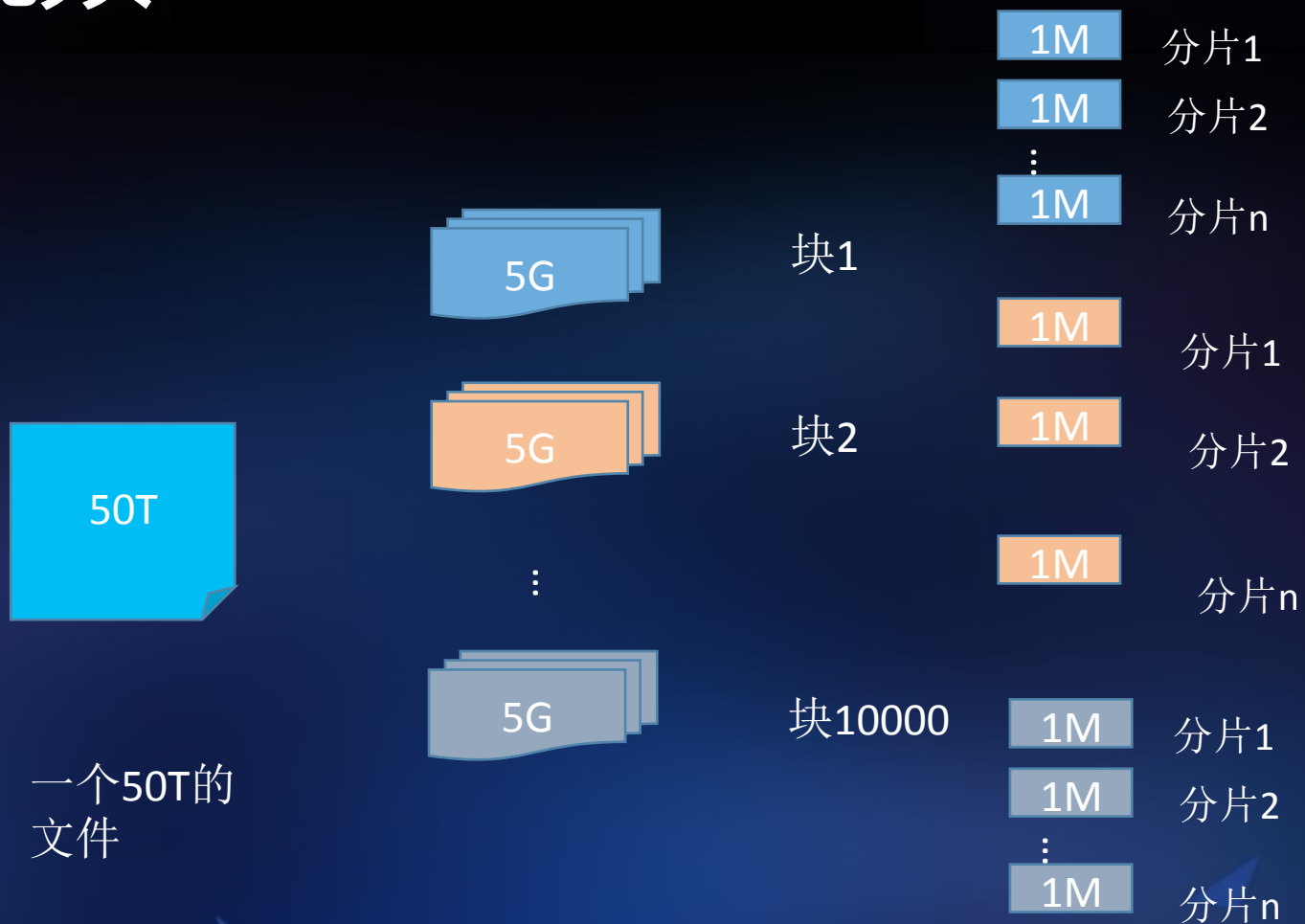
付费VIP用户+广告收入增长迅速，以腾讯视频为例，16年流水达到百亿级别



# 视频时代的存储挑战

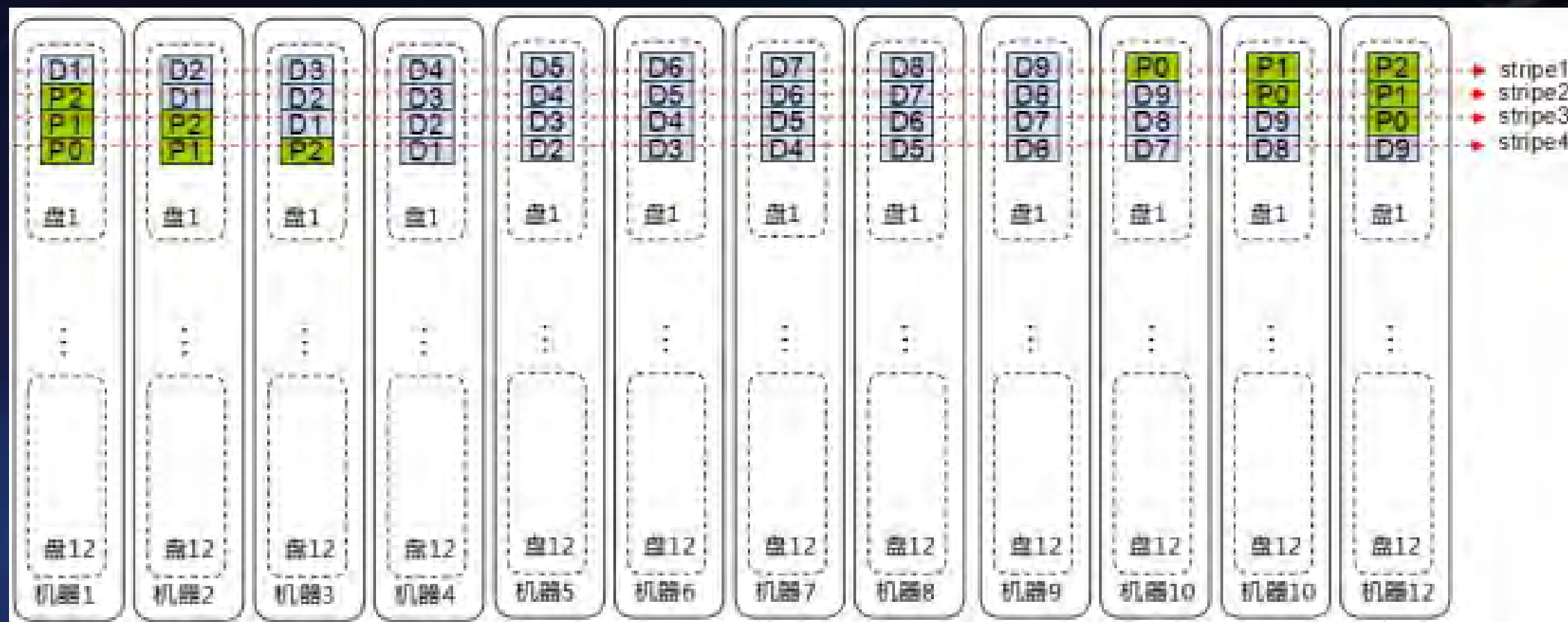
- 超大文件（50T）
- 极冷文件的合理安置

# 超大视频





# 极冷视频



# 云存储时代



# 我们准备开放什么

腾讯公司的存储技术

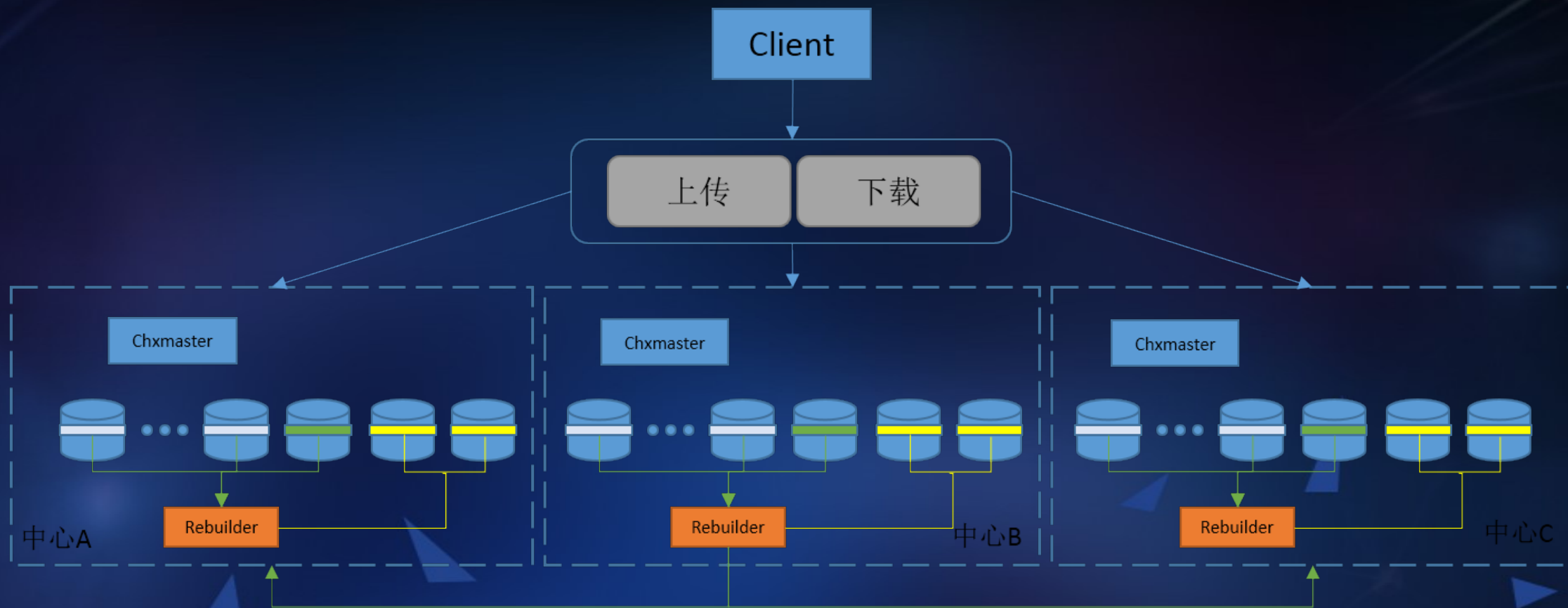
对数据深度理解的定制能力

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



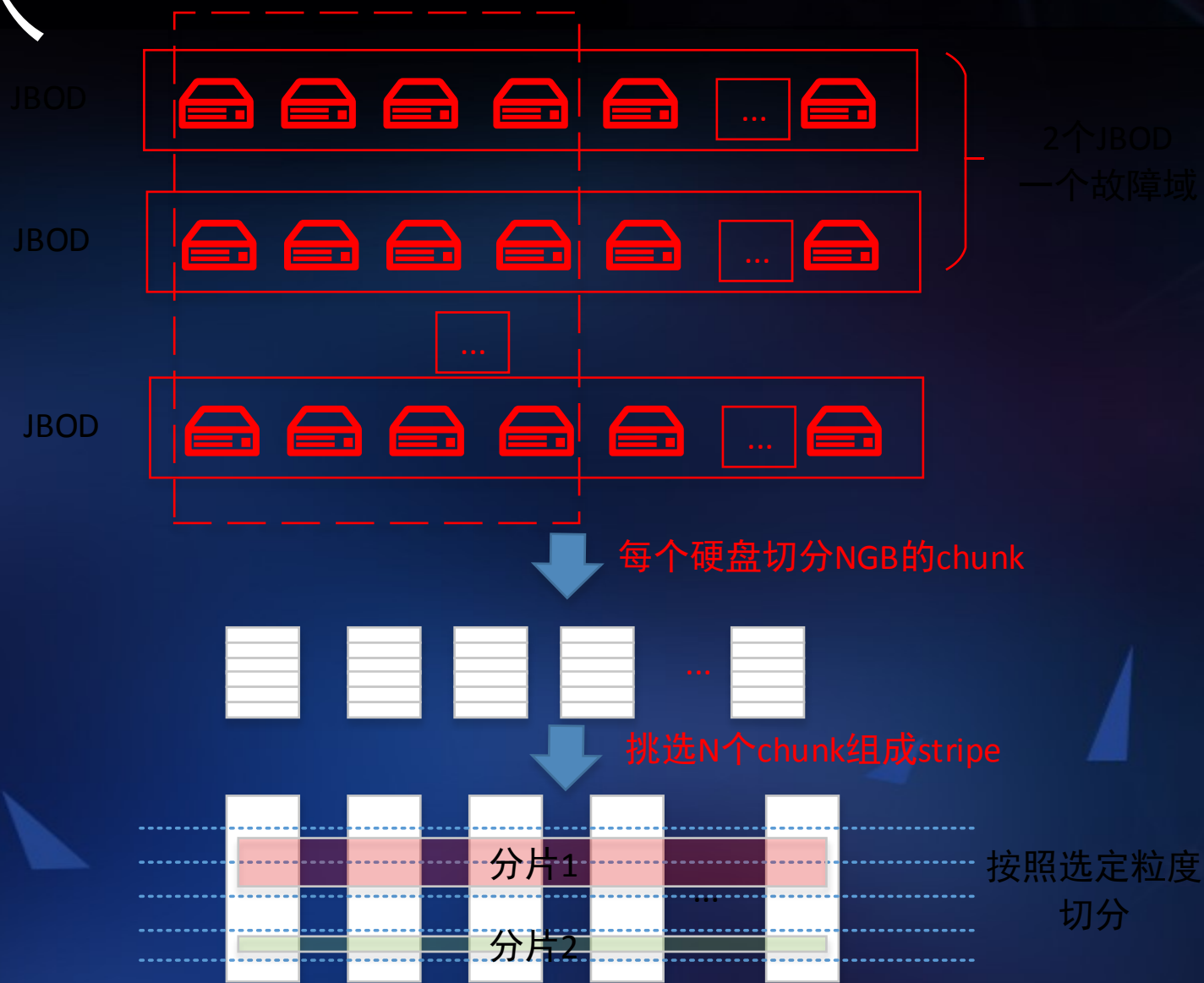
# 归档需求

- 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

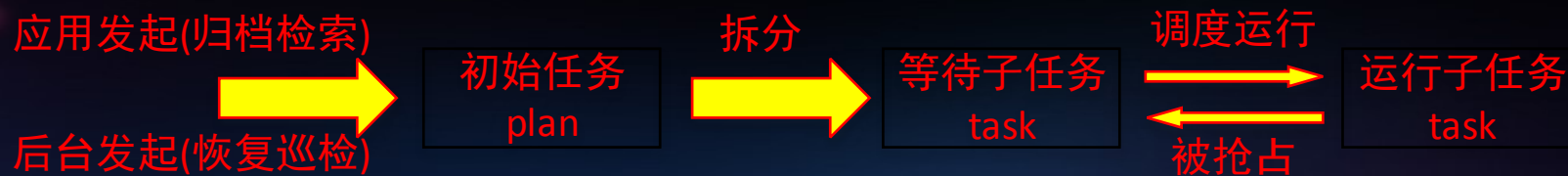
- 通过**牺牲可用性**（允许存在恢复时间）以及利用**业务特性**（批量导入，大文件），将**成本**发挥到极致的空间管理方法。

# 整机柜技术



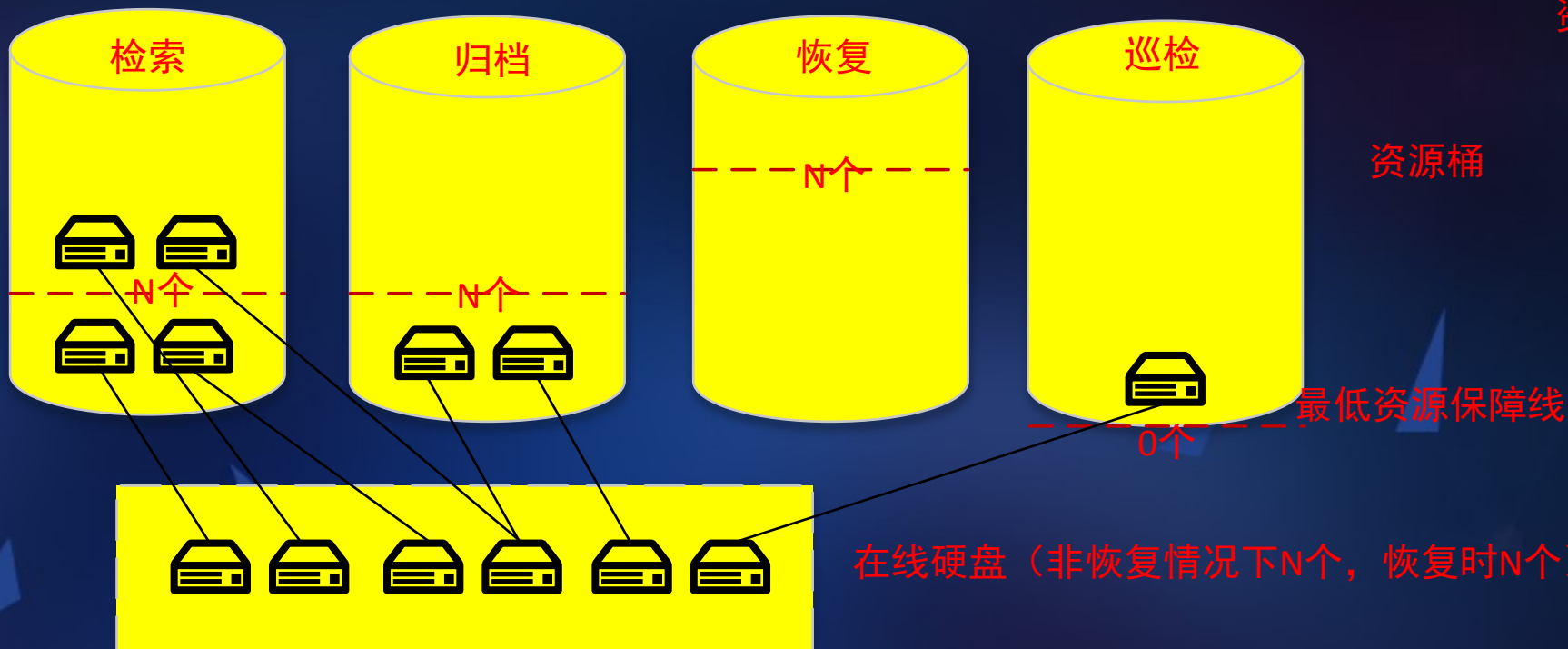


# 磁盘休眠



调度

资源



# 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**  
ueue

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**  
ueue

斯达克学院

实践驱动的 IT 教育平台

