淘宝网
Taobao.com

Taobao
海量图片存储与CDN系统

章文嵩 （正明）
淘宝核心系统部

淘宝网 Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

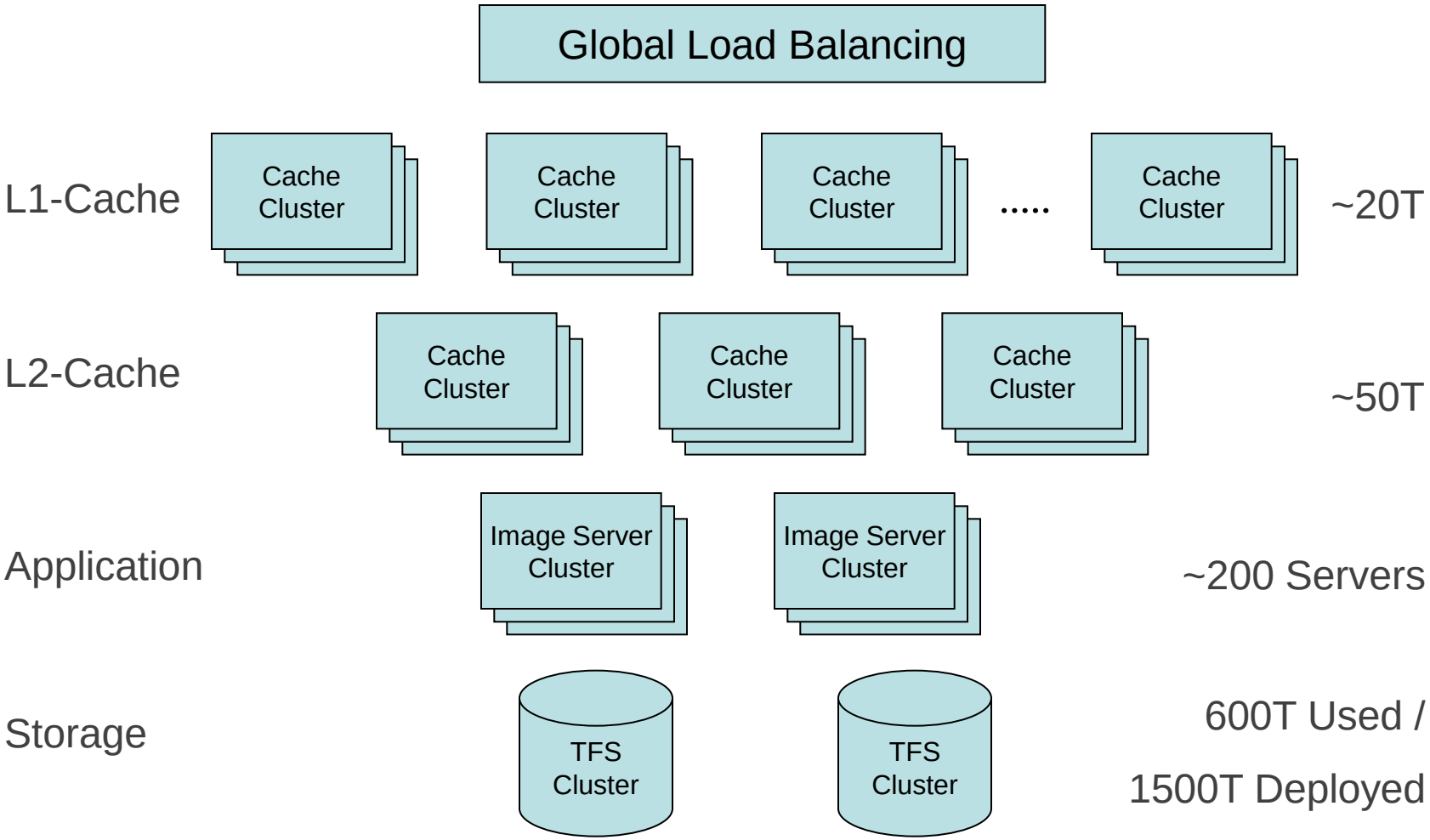
三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

图片存储与CDN系统全貌



淘宝网
Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

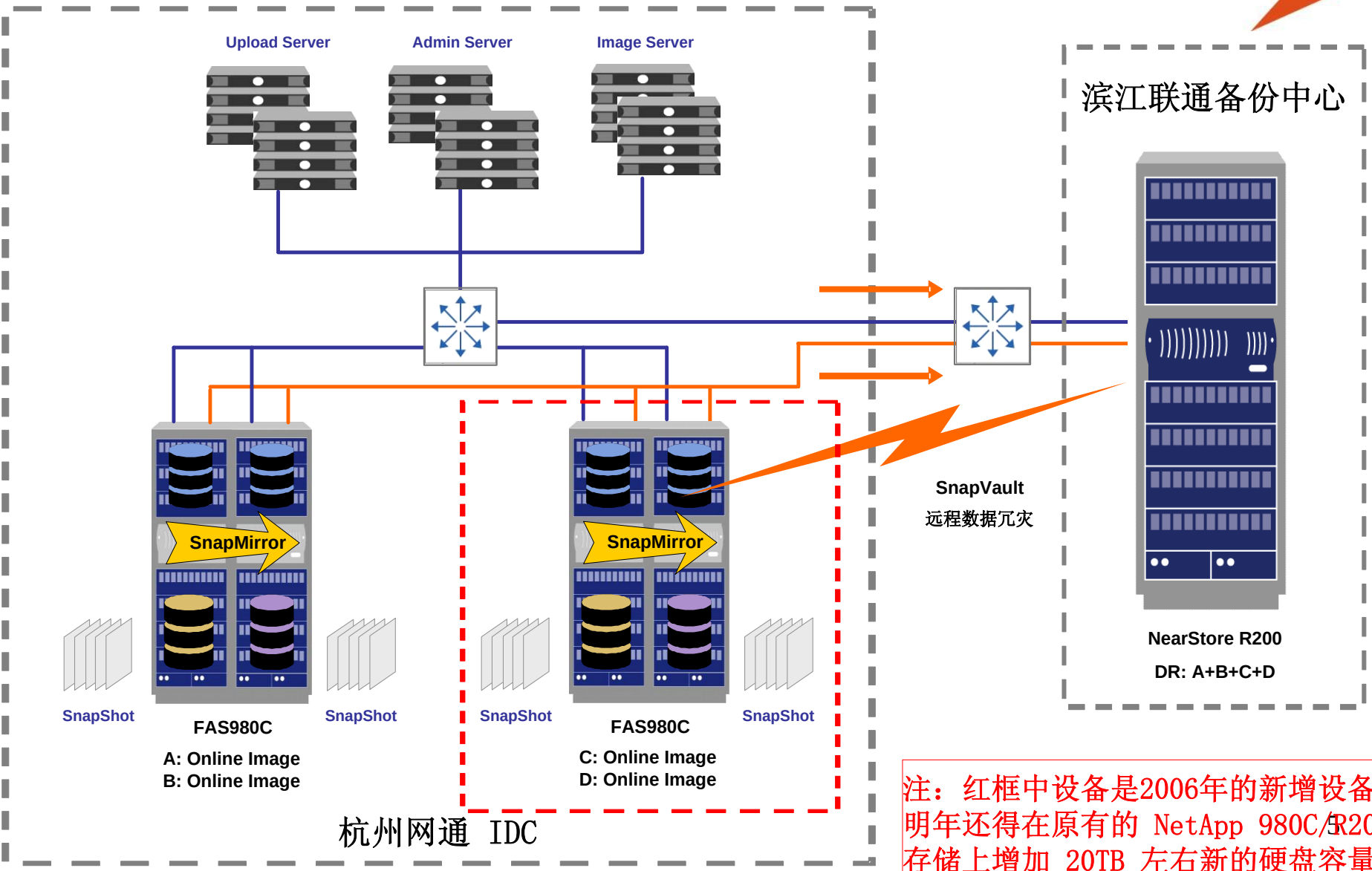
三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

2007年之前的图片存储系统



注：红框中设备是2006年的新增设备，明年还得在原有的 NetApp 980C/R200 存储上增加 20TB 左右新的硬盘容量。

淘宝网 Taobao.com

商用产品无法满足需求

- 系统需求
  - 淘宝的影响越来越大，数据的安全也更加重要
  - 数据存储量以每年二倍的速度增长（即原来的三倍）
- 商用存储产品
  - 对小文件的存储无法优化
  - 文件数量大，网络存储设备无法支撑
  - 连接的服务器越来越多，网络连接数已经到达了网络存储设备的极限
  - 扩容成本高，10T的存储容量需要几百万￥
  - 单点，容灾和安全性无法得到很好的保证

淘宝网 Taobao.com

TFS 1.0

- 2007年6月

  淘宝自主开发的分布式的文件系统

  TFS（Taobao File System）1.0上线运行

  主要解决海量小文件的分布式存储

  集群规模：200台PC Server(146G*6 SAS 15K Raid5)

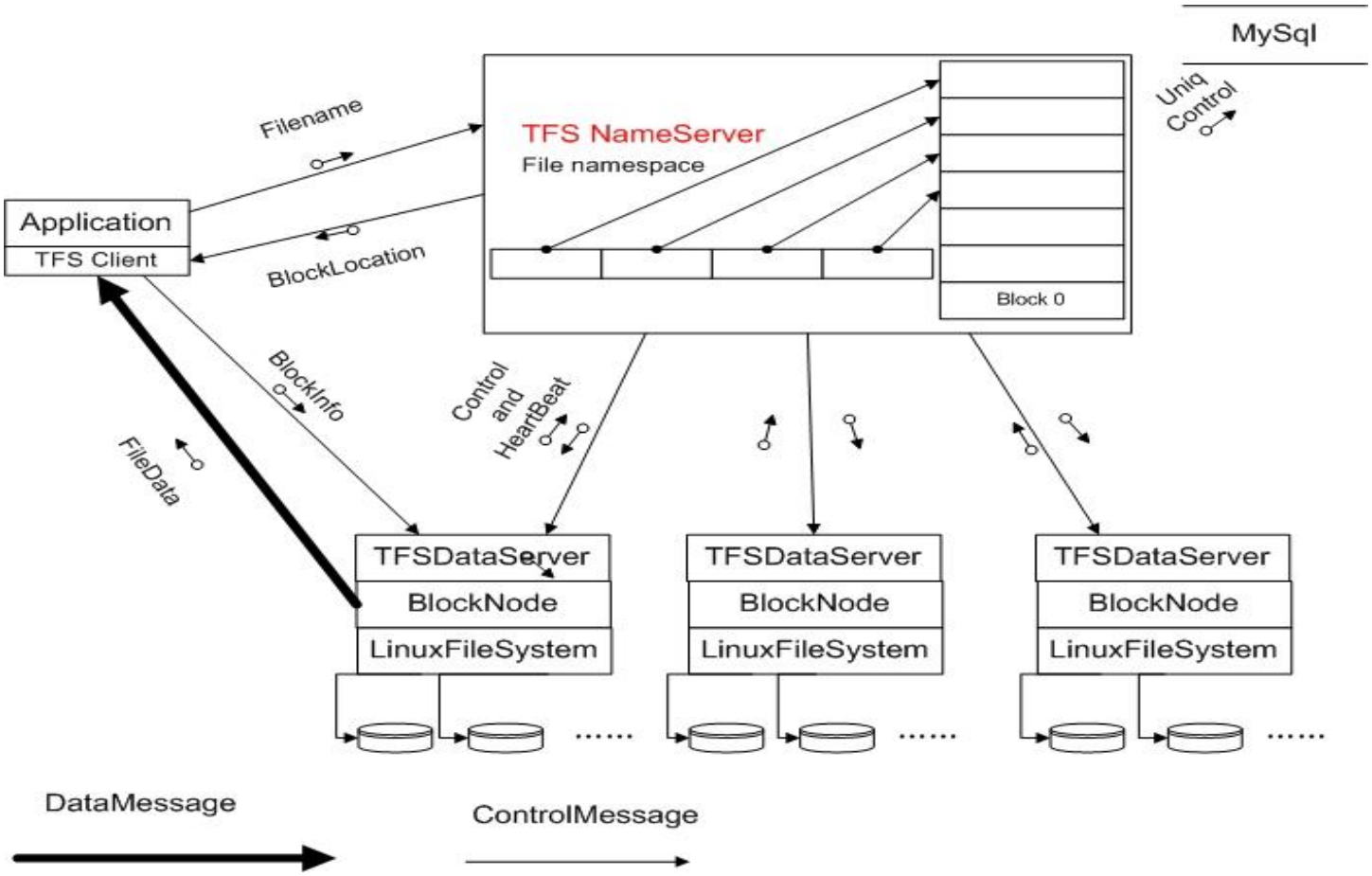
  文件数量: 亿级别

  系统部署存储容量: 140 TB

  实际使用存储容量: 50 TB

  单台支持随机IOPS 200+，流量3MBps

TFS 1.0的逻辑结构


MySql
Uniq Control
TFS NameServer
File namespace
Block 0
Filename
BlockLocation
Application
TFS Client
BlockInfo
FileData
Control and HeartBeat
TFSDataServer
BlockNode
LinuxFileSystem
TFSDataServer
BlockNode
LinuxFileSystem
TFSDataServer
BlockNode
LinuxFileSystem
DataMessage
ControlMessage

淘宝网 Taobao.com

TFS1.0的特性

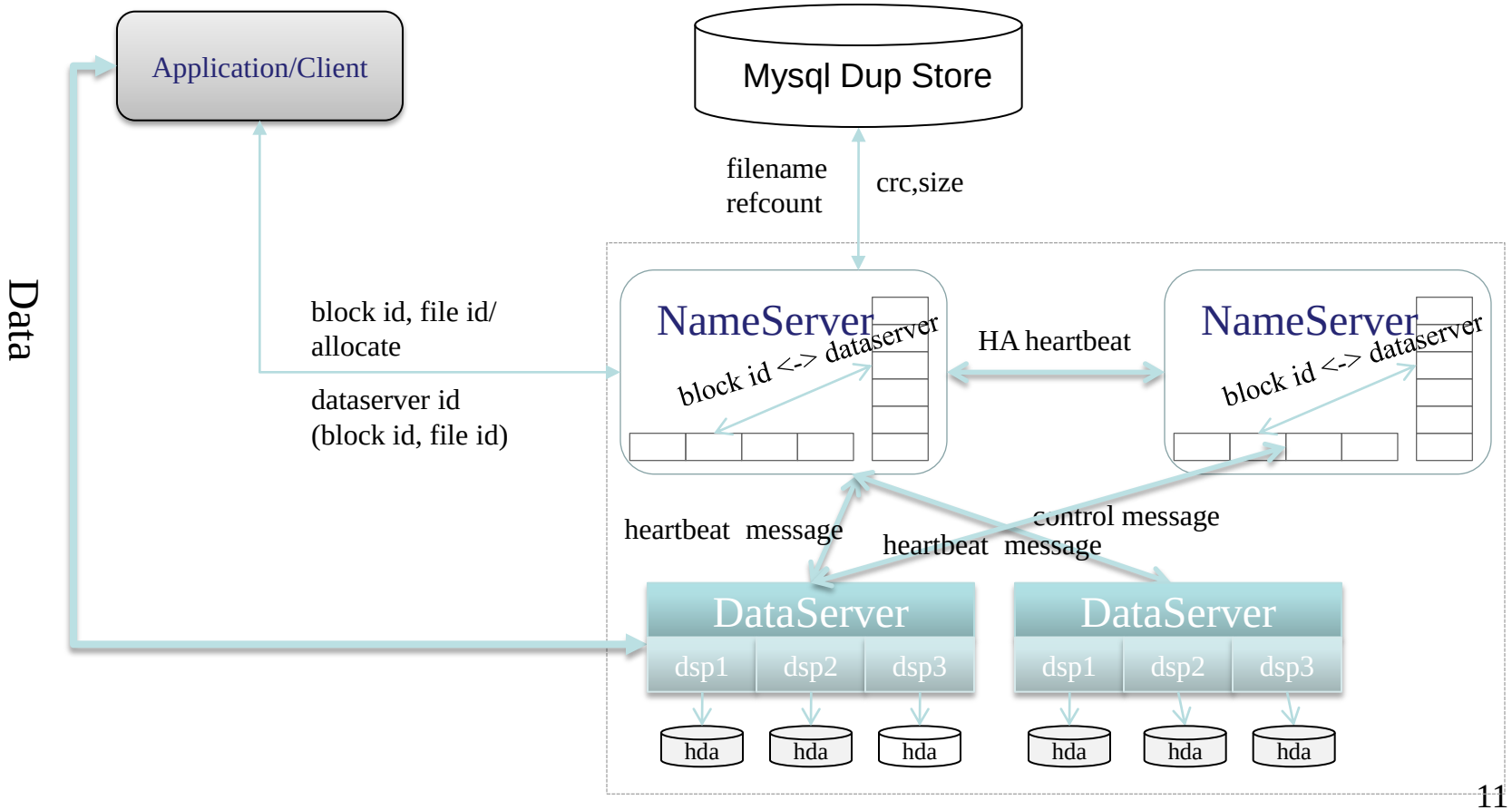
- 集群由一对Name Server和多台Data Server构成
- 每个Data Server运行在一台普通的Linux主机上
- 以block文件的形式存放数据文件(一般64M一个block)
- block存多份保证数据安全
- 利用ext3文件系统存放数据文件
- 磁盘raid5做数据冗余
- 文件名内置元数据信息，用户自己保存TFS文件名与实际文件的对照关系

TFS 1.3

- 2009年6月

  TFS（Taobao File System）1.3上线运行
- 集群规模
  - 440台PC Server (300G*12 SAS 15K RPM)
  - 文件数量： 百亿级别
  - 系统部署存储容量： 1580 TB
  - 当前实际存储容量： 600TB
  - 单台Data Server支持随机IOPS 900+，流量15MB+

TFS1.3的逻辑结构


Application/Client
Mysql Dup Store
filename
refcount
crc,size
Data
block id, file id/
allocate
dataserver id
(block id, file id)
NameServer
block id <-> dataserver
HA heartbeat
NameServer
block id <-> dataserver
heartbeat message
control message
heartbeat message
DataServer
dsp1
dsp2
dsp3
DataServer
dsp1
dsp2
dsp3
hda
hda
hda
hda
hda
hda

淘宝网 Taobao.com

TFS1.3的特性

- TFS1.3提供了一些重要的功能特性
  - 所有的元数据全部都内存化
  - 清理磁盘空洞
  - 容量和负载的均衡策略
  - 平滑的扩容
  - 数据安全性的冗余保证
  - 几秒内完成Name Server故障自动切换
  - 容灾策略
  - 性能大幅提升

淘宝网 Taobao.com

TFS的发展

- TFS2.0正在开发中
  - 目录功能的支持
  - 用户权限的校验
  - 元数据细化到文件级别，支持用户命名文件
  - 大小文件的共存，大文件的分片存储
  - 分级存储机制的建立，针对访问特性的文件迁移

  ……
- TFS即将开源，希望更多人来使用和改进TFS

淘宝网 Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

淘宝网 Taobao.com

图片处理与缓存

- 现状
  - 有200多台服务器Image Server，在Apache上实现的，从TFS取原图生产相应的缩略图
- 改进目的
  - 图片访问的热点一定存在，在Image Server实现Cache，提高响应速度，也减轻对后端TFS的压力
- Image Server上处理方式
  - 若请求图片在Cache中，直接发送
  - 没命中，若本地有原图，则根据原图做处理并缓存
  - 没命中，从TFS读取原图并添加到缓存，处理并缓存

淘宝网 Taobao.com

系统实现

- 将图片处理与缓存编写成基于Nginx的模块
  - Nginx是目前性能最高的HTTP服务器（用户空间）
  - 代码清晰
  - 模块化非常好
- 使用GraphicsMagick进行图片处理
  - 比ImageMagick性能更好
- 面向小对象的缓存文件系统
- 前端有LVS+Haproxy将原图和其所有缩略图请求都调度到同一台Image Server

淘宝网 Taobao.com

图片处理

- 从TFS存储中读取文件
- 将文件根据需要的尺寸进行缩放
- 可根据需要将缩略图按一定质量压缩保存（75%～94%），通过配置文件设定
  - 可有效地降低缩略图的体积（30%~70%）
  - 节约传输的带宽

淘宝网 Taobao.com

本地缓存对象文件系统

- 文件定位
  - 内存hash做索引
  - 最多一次读盘
- 写盘方式
  - Append方式写
  - 淘汰策略FIFO，主要考虑降低硬盘的写操作，没有必要进一步提高Cache命中率，因为Image Server和TFS在同一个数据中心

淘宝网 Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

淘宝网 Taobao.com

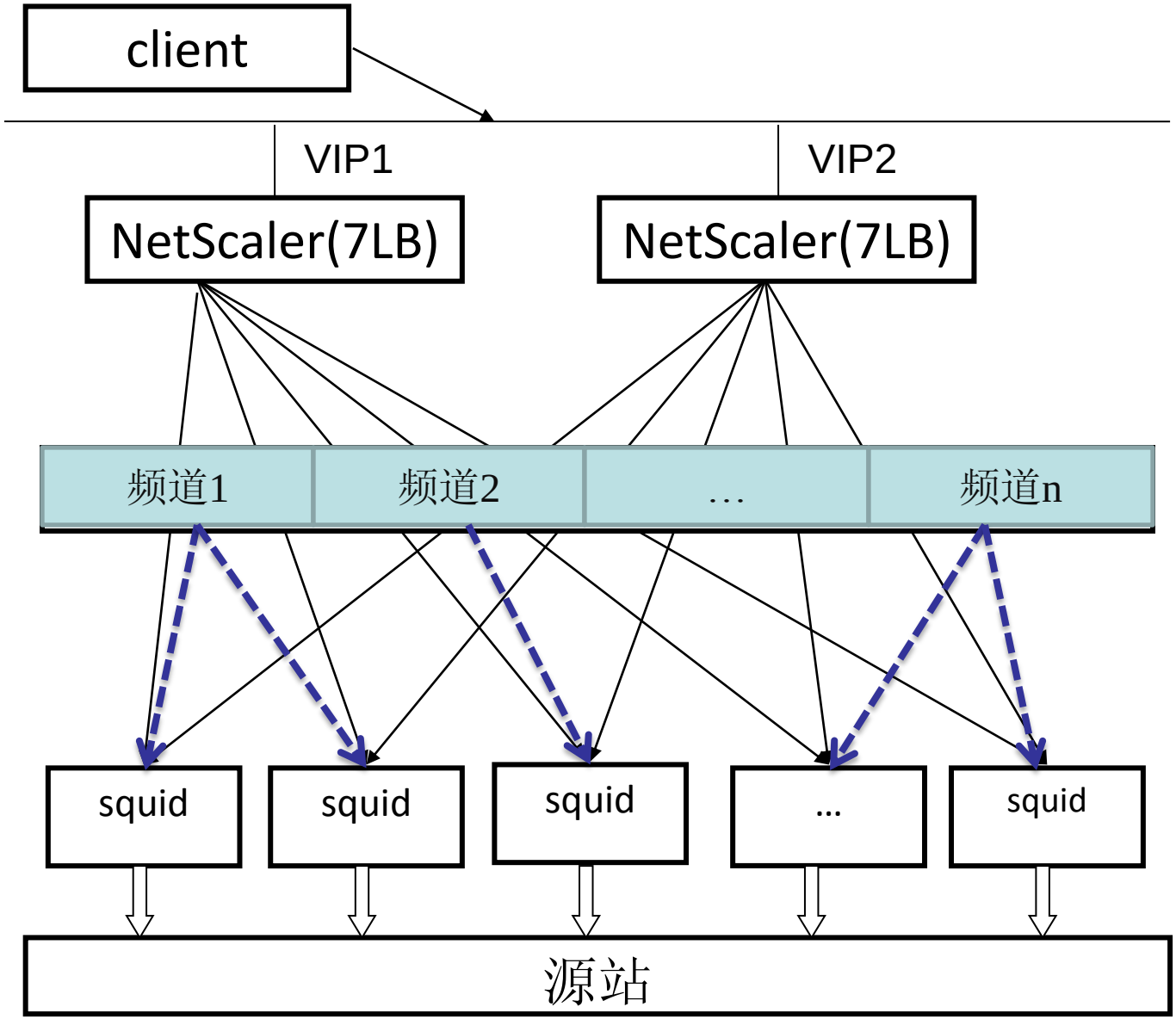
淘宝CDN系统

- CDN服务的图片规模
  - 150T容量的原图 + 150T容量的缩略图
  - 200亿左右的图片数，平均图片大小是15K
  - 8K以下图片占图片数总量的53%，占存储容量的15%
- CDN部署规模
  - 20个节点，部署在网民相当密集的主要中心城市
  - 每个节点目前处理能力在4~10G
  - CDN部署的总处理能力已超过150G
  - 目前承载淘宝流量高峰时119G，和一些集团子公司的流量

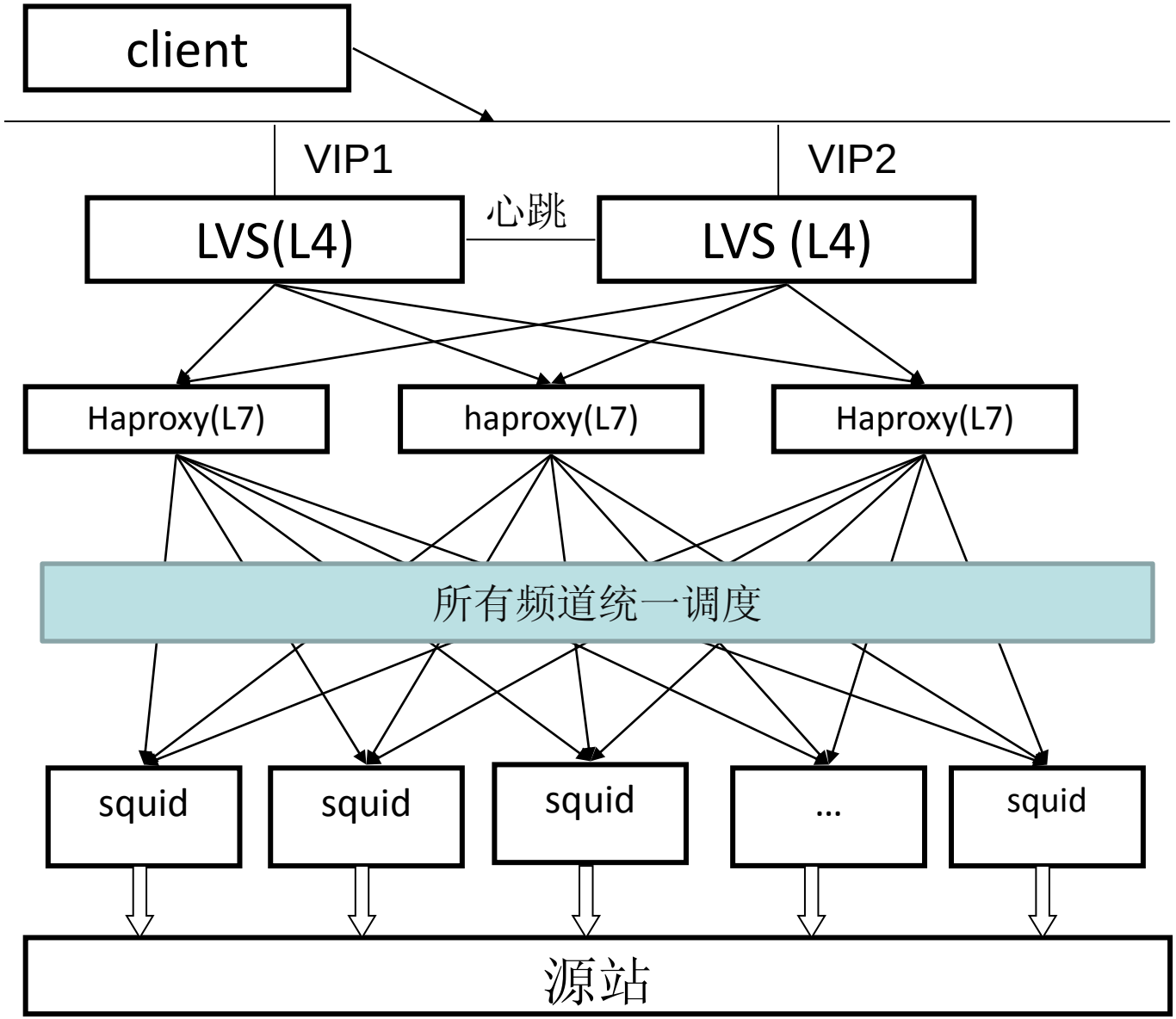
淘宝CDN系统的研发

- 主要解决现有的问题
  - 商用产品的性能瓶颈、功能欠缺，以及不稳定性
  - 整个系统的规模、性能、可用性和可管理性
- 开发完全自主的CDN系统
  - CDN节点的新架构和优化
  - CDN监控平台
  - 全局流量调度系统支持基于节点负载状态调度和基于链路状态调度
  - CDN实时图片删除
  - CDN访问日志过滤系统
  - 配置管理平台

CDN节点的架构对比-老架构



CDN节点的架构对比-新架构



淘宝网 Taobao.com

CDN节点的架构对比

对比项 \ 节点	新架构	老架构
流量分布均匀性	☆☆☆☆☆	☆☆☆
可维护性	☆☆☆	☆☆☆
抗攻击能力	☆☆☆☆	☆☆☆☆
自主控制能力	☆☆☆☆☆	☆☆☆
价格	☆☆☆☆☆	☆☆☆
扩展能力	☆☆☆☆☆	☆☆
灵活性	☆☆☆☆☆	☆☆

- 流量分布均匀性：所有的频道统一调度到128台squid，而不是将squid按频道分组，可提高命中率2%以上
- 扩展能力：在一个VIP上新架构可以扩展到近100G的流量（当然要用万兆网卡）
- 灵活性：一致性Hash调度方法使得增加和删除服务器非常方便，只有1/(n+1)的对象需要迁移

淘宝网 Taobao.com

Squid改进和优化

- 改进后的Squid可支持1T大小的COSS文件（blocksize为512 Bytes）
- Squid内存优化，一台Squid服务器若有一千万对象，大约节省250M内存，更多的内存可以用作Squid Memory Cache
- 改进Squid的对象淘汰策略，在不牺牲对象命中率的前提下，降低硬盘的写操作
- 用sendfile来发送缓存在硬盘上的对象
- 在Squid服务器上使用SSD作缓存，对数据量比较小的图片非常有效

淘宝网 Taobao.com

节点运行情况

- 节点规模：32台 DELL R710服务器
- 逻辑结构：2 LVS + 32 Haproxy + 128 Squid
- 时间：12月21日上线运行
- 当前最大服务流量：10.58 Gbps
- 理论最大负载能力：15Gbps以上
- 单台R710服务器可到500Mbps以上的吞吐率
- 单squid最大object数目：1000万
- Cache请求命中率：97%
- Cache字节命中率：97%
- **最重要的是命中率提高，大大改善用户的访问体验**

淘宝网 Taobao.com

CDN系统的发展

- CDN系统的研发与运维
  - 持续提高节点性能和稳定性
  - 优化GTM全局调度系统
  - 持续提高CDN系统可运维性
  - 不断完善CDN内容管理系统
- CDN系统的建设
  - 思路正在转向“部署更多的小节点，尽可能离用户近一些”
  - 定制化和快速部署

淘宝网 Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

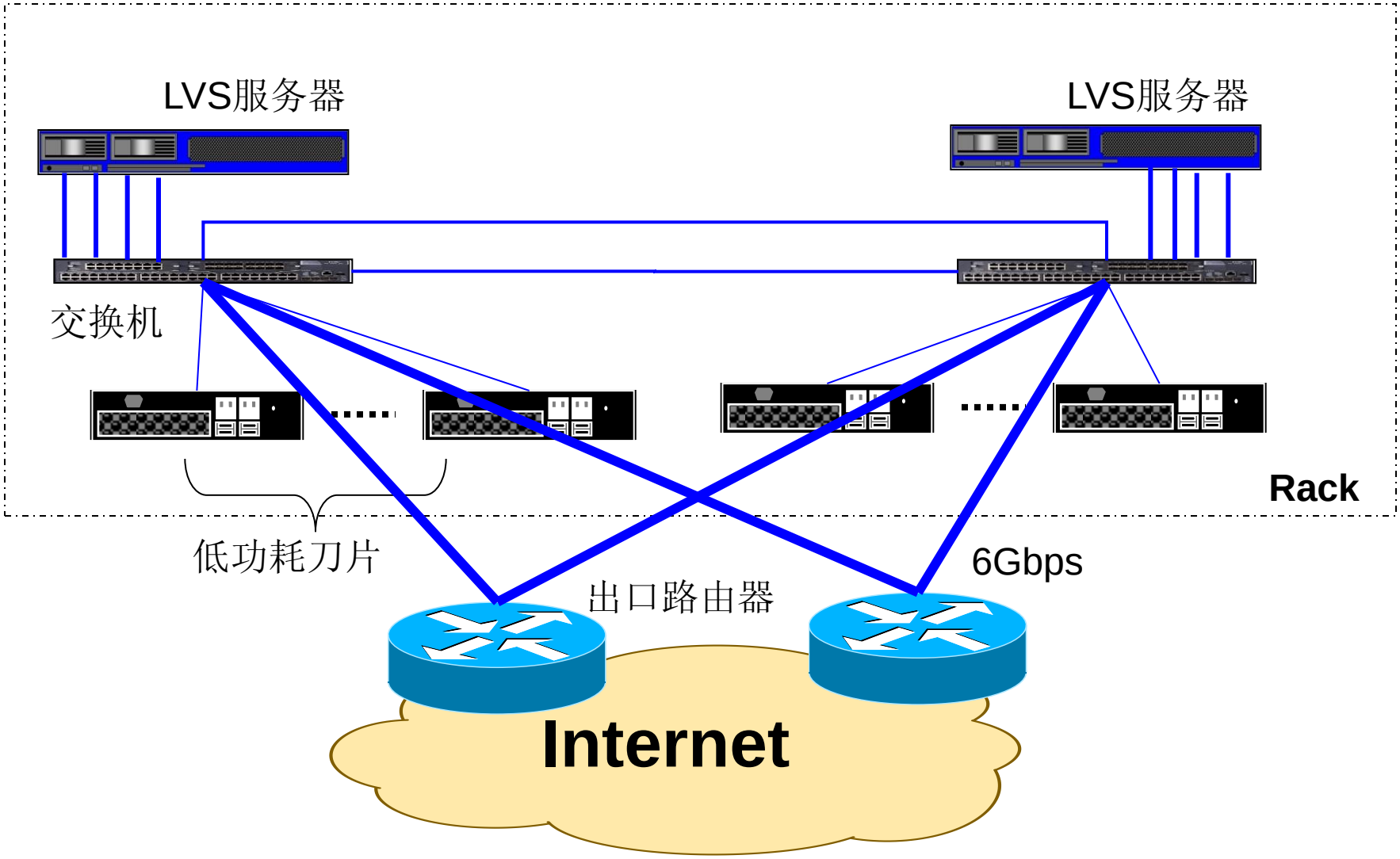
淘宝网 Taobao.com

低功耗硬件平台

- 低功耗硬件平台
  - 低功耗的CPU，如Intel ATOM, VIA Nano等
  - 低功耗的Chipset；SSD或低功耗的SATA硬盘
  - 关闭GPU和USB Controller等
- 适用不需要太多CPU计算的I/O类型应用
  - 例如CDN Cache Server、memory cache、存储节点、静态文件Web Server等
- 好处（大大降低成本）：
  - 降低电力消耗，减少碳排放
  - 单位空间(机柜)下有更高的I/O吞吐率
  - 降低硬件购置成本和运营成本

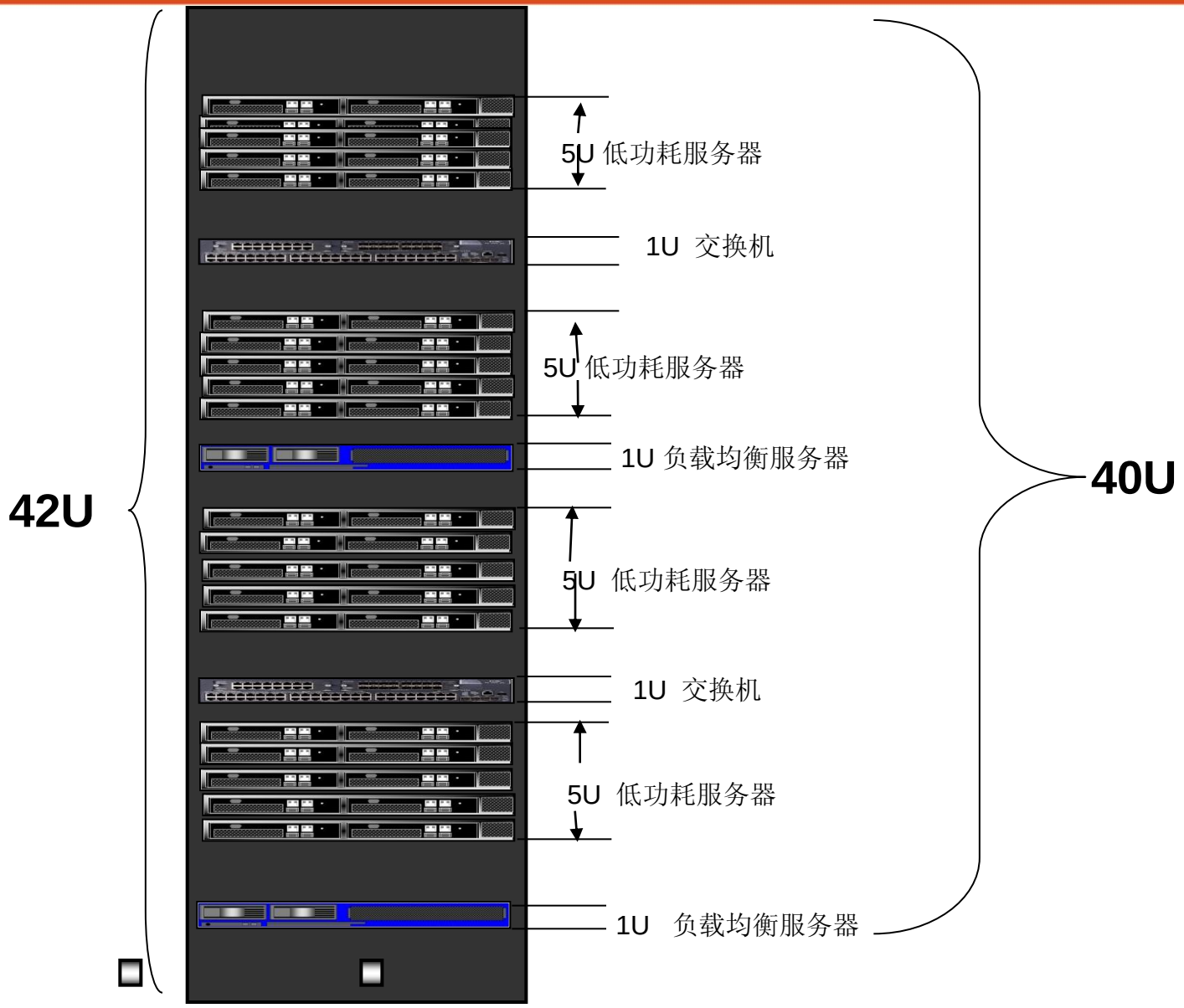
淘宝网 Taobao.com

单机柜方案:节点网络结构



淘宝网 Taobao.com

单机柜(6Gbps)方案:节点机架布局



淘宝网
Taobao.com

议程

一、系统全貌

二、Taobao图片存储系统--TFS

三、Image Server与Cache

四、CDN系统

五、低功耗服务器平台

六、经验

淘宝网 Taobao.com

经验

- 采用开源软件与自主开发相结合
- 规模效应，研发投入都是值得的
- 可以在软件和硬件多个层次优化
- 优化是长期持续的过程

讨论

Q & A

谢谢!