淘宝网
Taobao.com

淘宝软件基础设施实践

章文嵩（正明）

拥抱开源-企业IT自主之路

2013.6

淘宝网 Taobao.com

自我介绍

- 章文嵩（正明） 博士
- 淘宝高级研究员、核心系统负责人
- LVS开源项目的创始人与主要作者
- 曾为TelTel的首席科学家与联合创始人，国防科技大学副教授、ChinaCluster的联合创始人、Red Hat Kernel Developer

淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

三、事例：图片存储、CDN与DB

四、淘宝开源策略

五、小结

淘宝网 Taobao.com

淘宝网简介

- 2011年网购交易额约8019亿元，2012年到达13205亿元，同比增长64.7%。淘宝天猫到11600亿元，占8成以上
  - 源：中国电子商务研究中心的《2012年度中国网络零售市场数据监测报告》 http://www.100ec.cn/zt/2010bgdz/
- 网络流量排名（Alexa统计）
  - 国际：11 （10～18）
  - 国内：3
- 2012年双11大促活动的一些数据
  - 双11购物狂欢节支付宝总销售额191亿
  - 第1分钟超过1000万人涌入，1分钟成交19.2万笔
  - 全天有2.13亿独立访客，占中国网民总数4成
  - CDN最高峰值流量到达2100Gbps

淘宝网 Taobao.com

淘宝网的下一个十年

- 淘宝网的下一个十年将是基于大数据的十年
  - 改变商业规则的C2B
  - 挑战传统贷款规则的阿里金融
  - 淘宝SNS化、社会化电商
  - 移动电子商务
  - 骨干物流网：大数据驱动的智能物流系统

淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

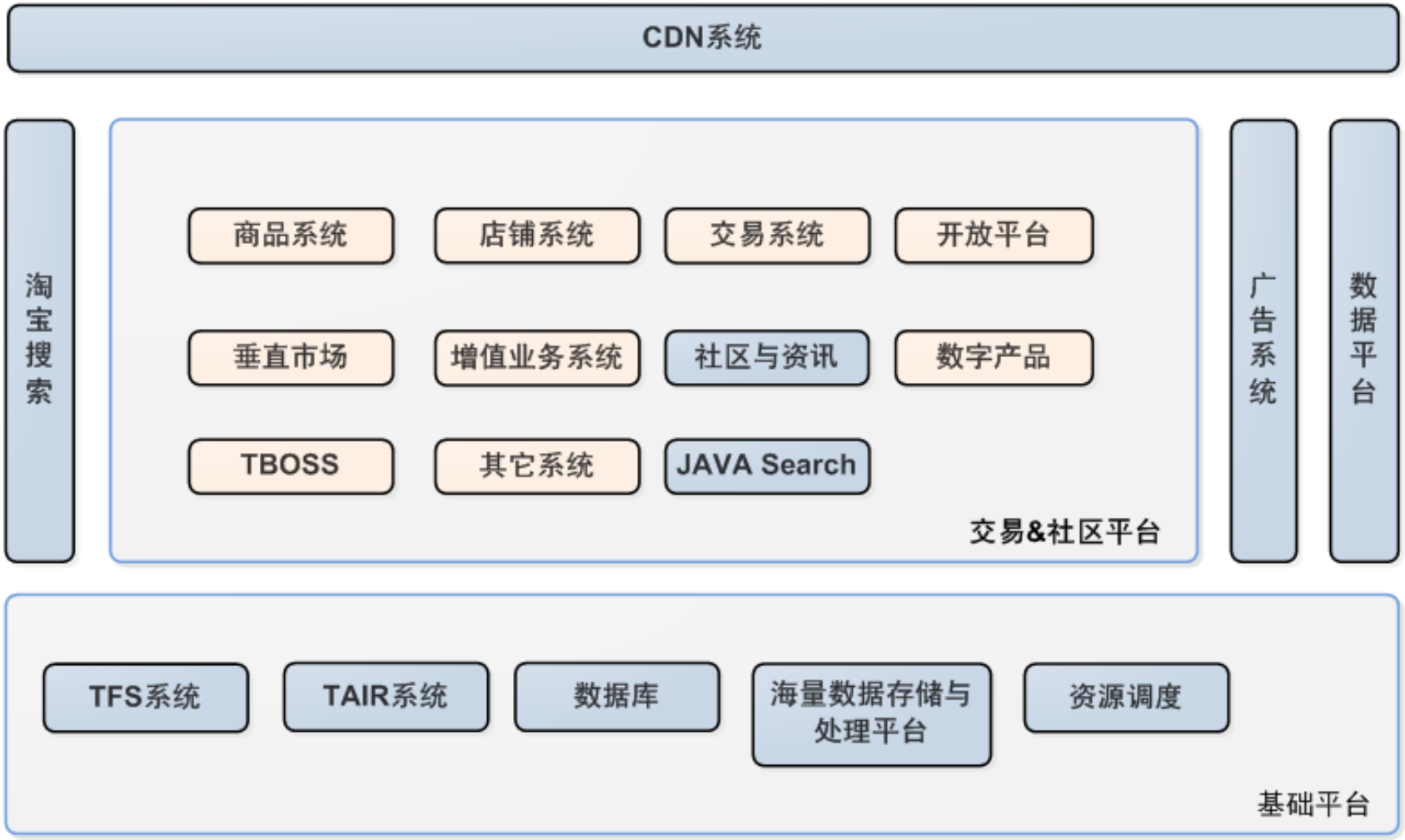
三、事例：图片存储、CDN与DB

四、淘宝开源策略

五、小结

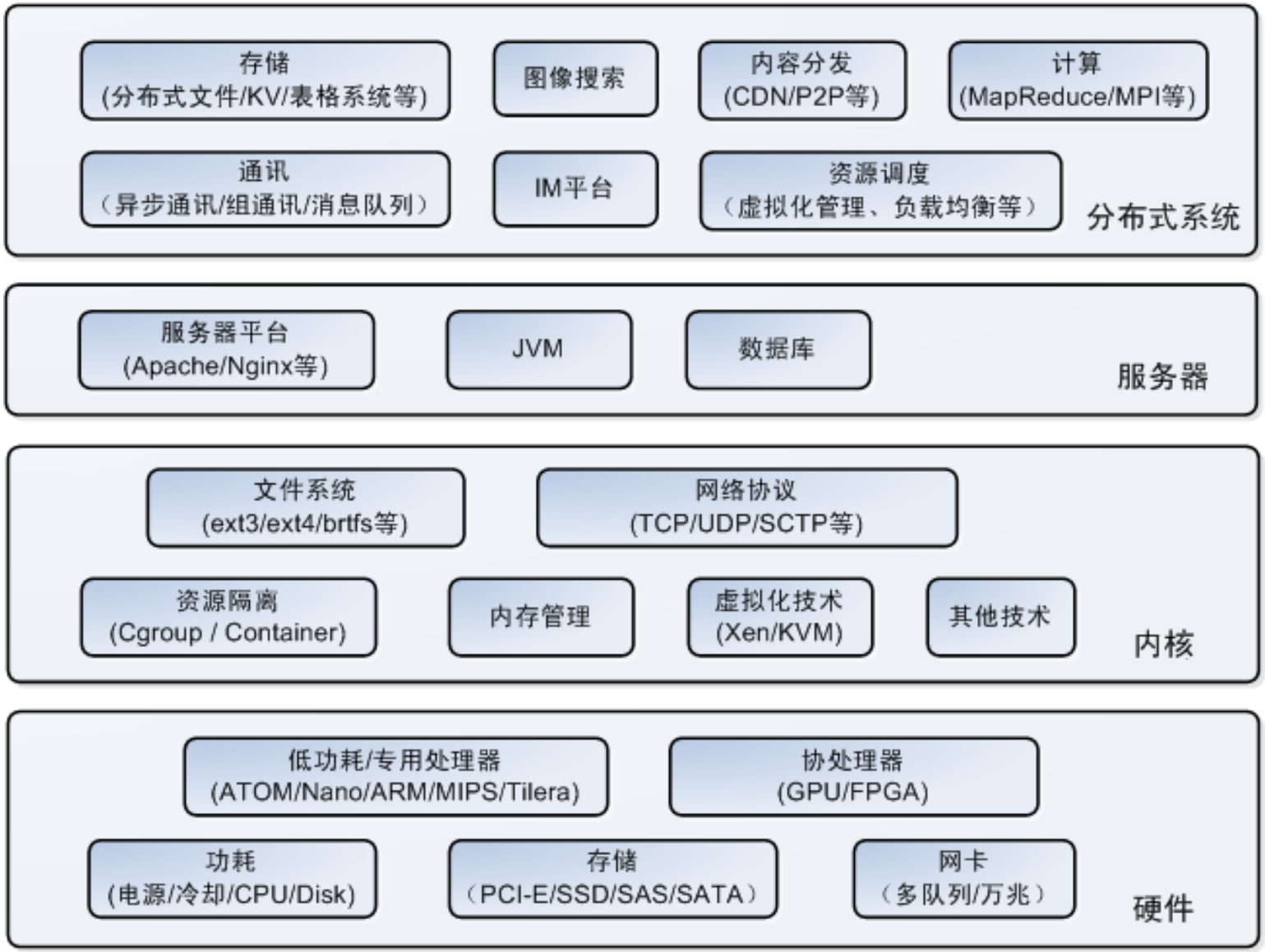
淘宝网
Taobao.com

淘宝网平台系统框架图



淘宝网 Taobao.com

软件基础设施的规划



淘宝网 Taobao.com

主要所采用的软件

- CDN：世界上流量最大的、面向图片的CDN系统
  - 基于开源软件LVS+Haproxy+Squid/TS+Bind上开发的CDN系统
  - 现有140个节点，能承载2400Gbps流量的能力
- TFS：自主开发的分布式对象存储系统
  - 可存储容量13.36P，实际使用10.28P容量
  - 图片空间每GB每年存储与运维成本从7.2元降到3.1元
- TAIR：淘宝的分布式缓存和K/V存储
  - 集成了开源的Redis和LevelDB存储引擎
  - 提供跨机房容灾的解决方案
- OceanBase：淘宝的分布式数据库系统
  - 支持千亿条记录级别的数据库、支持事务

淘宝网 Taobao.com

主要所采用的软件（2）

- 海量数据：采用开源的Hadoop平台
  - 现在单一集群到4200台服务器的规模
  - 系统可存储容量为86.7 PB，已使用69.1 PB，历史数据为压缩存储
  - 现在每天存储净增量约为259TB（使用量）
  - 每天运行的作业数约为15～18万
- 核心数据库：采用开源的MySQL，加上高速的非易失存储，以及多层级的系统优化
- 旺旺平台：自主开发；最高在线1231万，全年可用99.99%，机器数约1800台，服务了淘宝和B2B
- 服务器平台：Nginx 部署130个应用，约2000台机器，占有率20%；完成TMD等重要基础模块，Tengine项目开源

淘宝网 Taobao.com

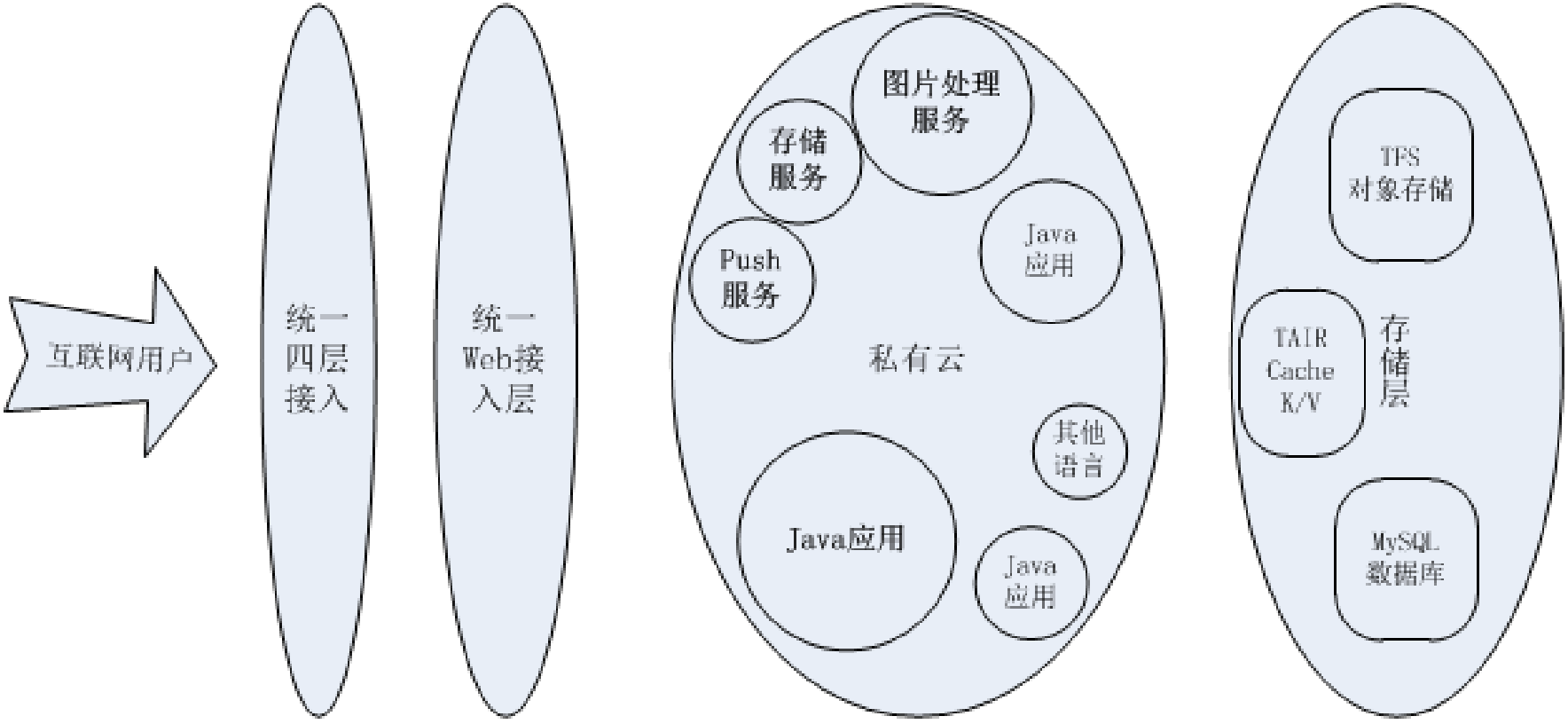
主要模块所采用的软件（3）

- 底层的支撑软件：
  - 在OpenJDK基础上开发和维护Taobao JVM
  - 在Red Hat基础上维护自己的Linux内核
  - 基于cgroup的轻量级弹性计算平台T4
  - 在Sheepdog上实现了KVM的虚拟化弹性计算平台
  - 在LVS基础上实现负载均衡解决方案
  - 用开源软件实现了高流量的网络镜像项目

- 可以说淘宝网平台建立在开源软件和自主开发的基础上。

淘宝网 Taobao.com

秒级自动扩展的在线服务平台



LVS
FullNAT
四层安全

Nginx+Cache
埋点Cookie
7层安全等

T4：私有弹性计算平台

存储层

TAIR作Memory Cache和K/V存储

淘宝网 Taobao.com

议程

一、淘宝网的简介

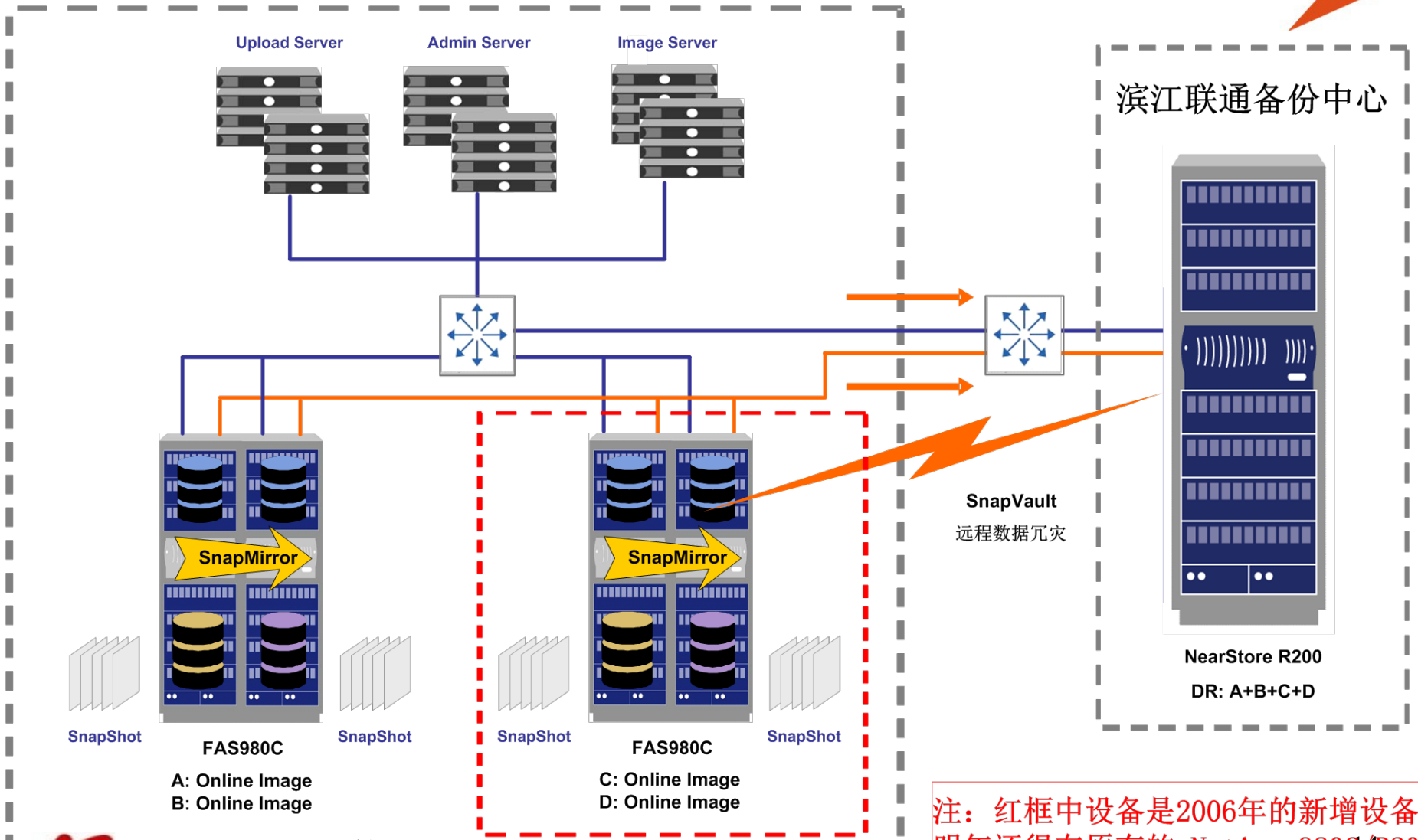
二、淘宝网电子商务平台

三、事例：图片存储、CDN与DB

四、淘宝开源策略

五、小结

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

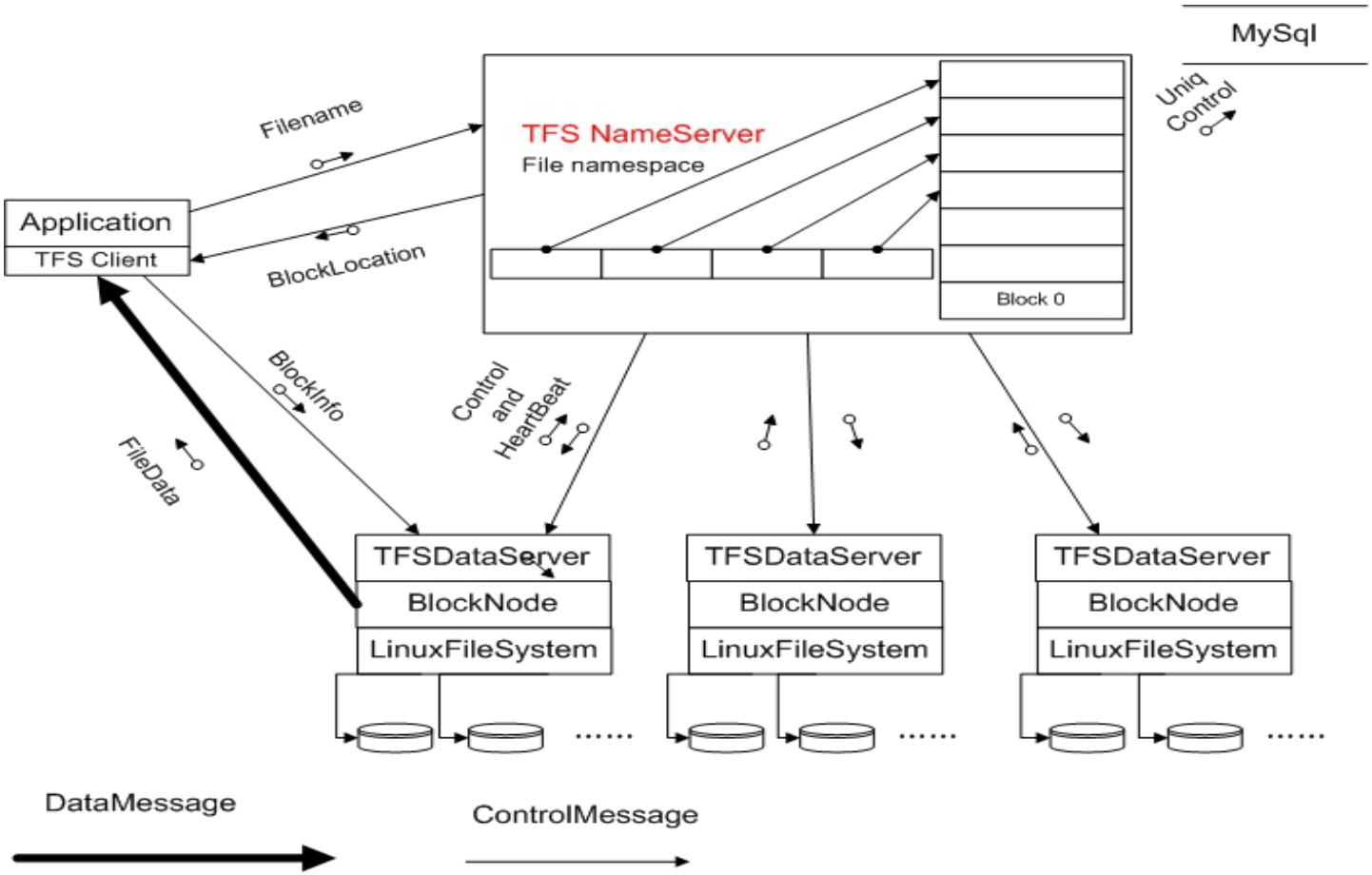
  系统部署存储容量：140 TB

  实际使用存储容量：50 TB

  单台支持随机IOPS 200+，流量3MBps

淘宝网 Taobao.com

TFS 1.0的逻辑结构



淘宝网 Taobao.com

TFS1.0的特性

- 集群由主备Name Server和多台Data Server构成
- Data Server运行在挂很多硬盘的Linux主机上
- 以block文件的形式存放数据文件(一般64M一个block)
- **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** – 使得元数据量特别小
  - 如T2auNFXXBaXXXXXXXX_!!140680281.jpg，名字中含有逻辑的block_no和object_no等
- 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

淘宝网 Taobao.com

TFS 1.3

- 2009年6月

  TFS（Taobao File System）1.3上线运行
- 集群规模（2010.8.22）
  - 440台PC Server （300G*12 SAS 15K RPM） + 30台PC Server （600G*12 SAS 15K RPM）
  - 文件数量： 百亿级别
  - 系统部署存储容量： 1800 TB
  - 当前实际存储容量： 995TB
  - 单台Data Server支持随机IOPS 900+，流量15MB+
  - 目前Name Server运行的物理内存是217MB（服务器使用千兆网卡）

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

- TFS在2010年9月开源，希望更多人来使用和改进
- TFS 2.0已经在生产系统中使用
  - 支持大文件存储
  - 在外围通过MySQL集群向应用提供目录支持
  - 加入资源中心，控制集群级别的权限
- 后续开发
  - 优化性能，提高扩展性，降低存储成本
  - RAID（Erasure Coding）
  - 分级存储机制（SSD/SATA），动态文件迁移等

淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

三、事例：图片存储、CDN与DB

四、淘宝开源策略

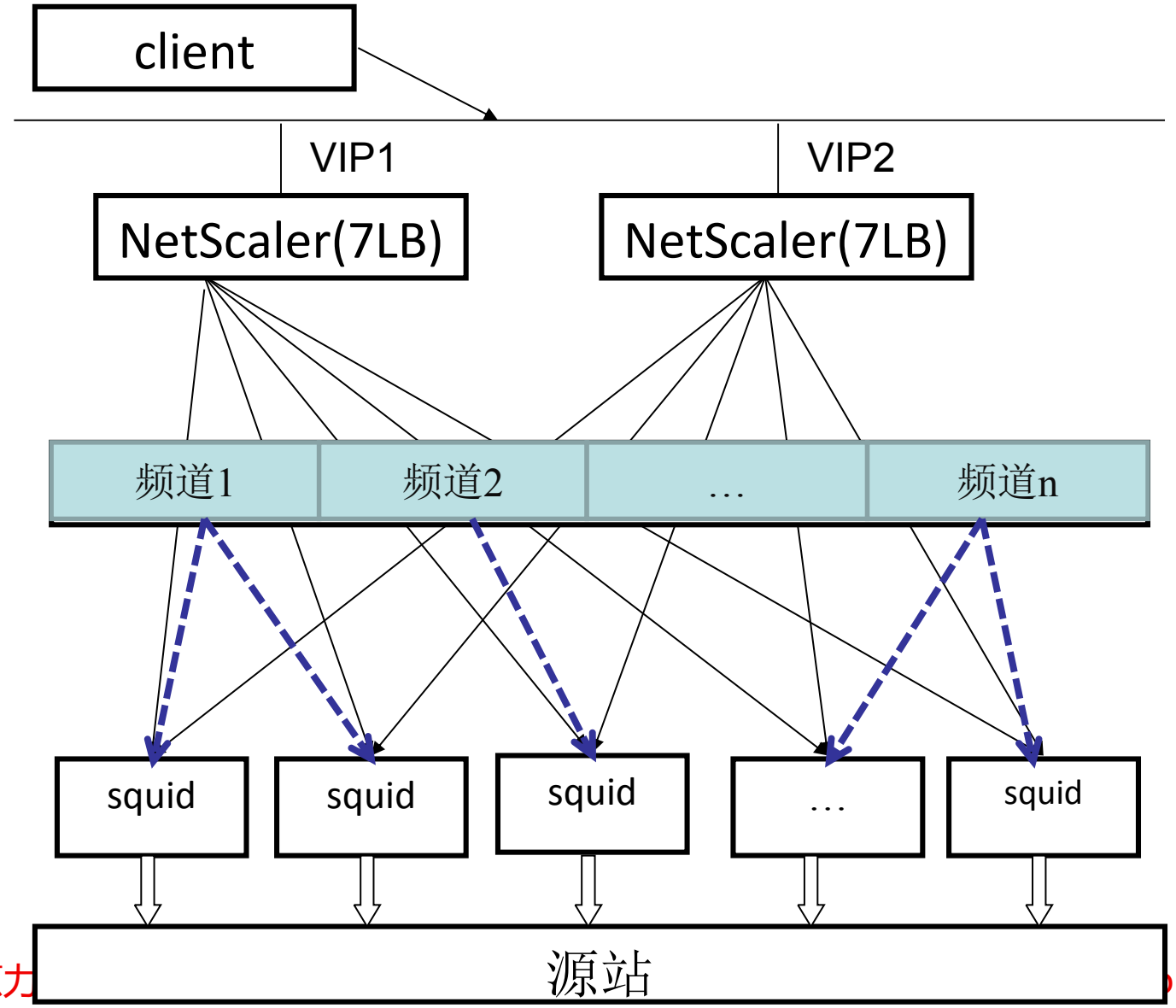
五、小结

淘宝网 Taobao.com

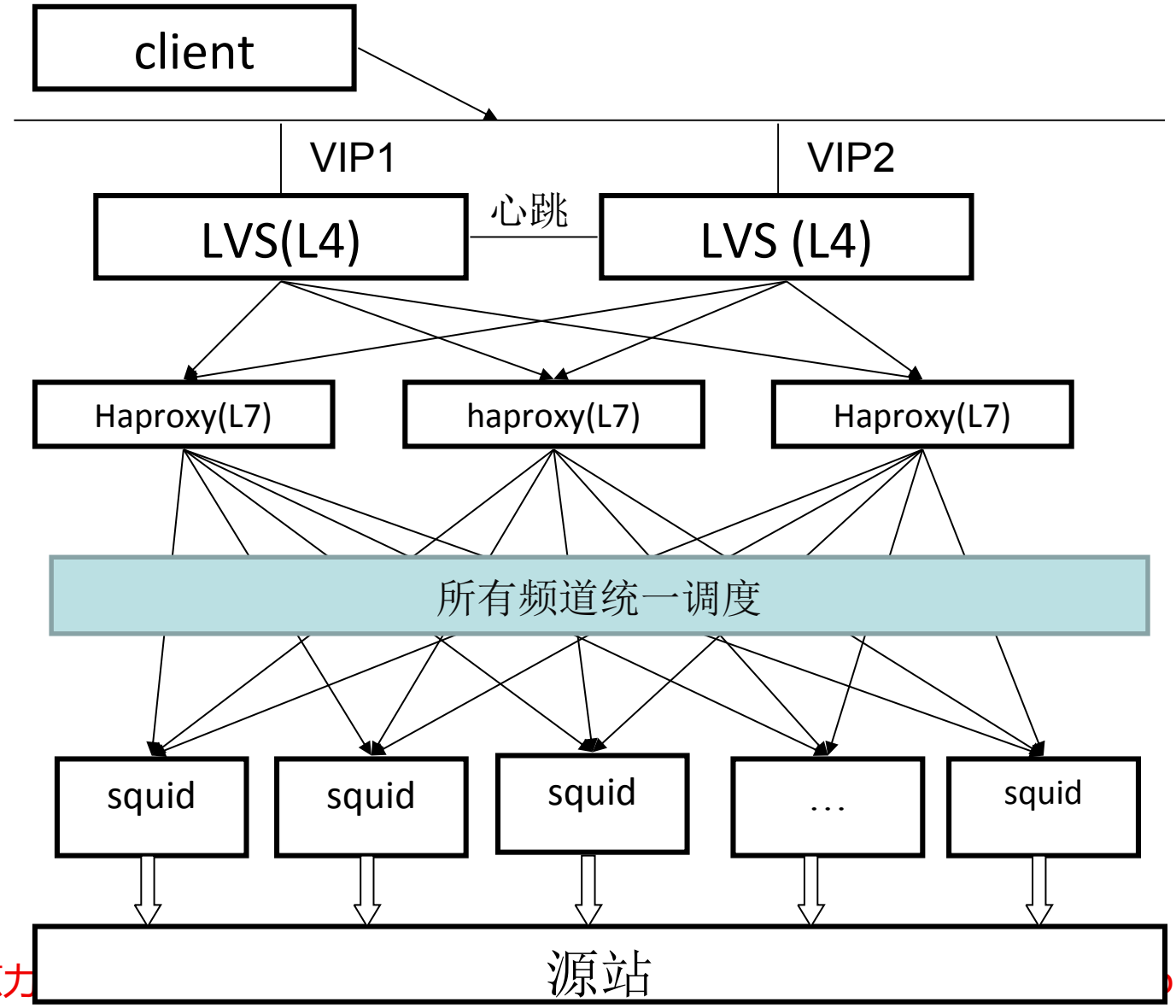
淘宝CDN系统的研发

- 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

CDN系统的历程

- 2010：LVS + Haproxy + Squid + GTM
  - 利用DELL 2950和混合存储（1SSD + 4*SAS + 1SATA）
  - Squid优化，ext2文件系统
  - 32个节点，320Gbps以上的能力
- 2011：LVS + Haproxy + Squid + GTM
  - 低功耗服务器和混合存储（1SSD + 3*SATA）
  - Squid优化，ext4+nojournal
  - 103个节点，1000Gbps以上的能力
- 2012：LVS + Haproxy + TS + Pharos为主
  - 40Gbps节点，Xeon L处理器+万兆网卡+6*SSD
  - 104个节点，320Gbps以上的能力
  - 开发了轻量级高性能的Cache软件Swift

淘宝网 Taobao.com

Squid改进和优化

- 在COSS存储系统基础上实现了TCOSS，FIFO加上按一定比例保留热点对象，支持1T大小的文件
- Squid内存优化，一台Squid服务器若有一千万对象，大约节省1250M内存，更多的内存可以用作memory cache
- 用sendfile来发送缓存在硬盘上的对象，加上page cache，充分利用操作系统的特性
- 针对SSD硬盘，可以采用DIRECT_IO方式访问，将内存省给SAS/SATA硬盘做page cache
- IO优化到平均一个请求需要做约0.9个IO操作
- 在Squid服务器上使用SSD+SAS+SATA混合存储，实现了类似GDSF算法，图片随着热点变化而迁移

$$migration\_weight * \frac{frequency}{size^{migration\_power}} \; ; \quad migration\_power \in (0,1]$$

热点迁移的数据对比（1）

- 简单按对象大小划分：小的进SSD，中的放SAS，大的存SATA
- SSD + 4 * SAS + SATA上的访问负载如下：

```
[root@cache161 ~]# iostat -x -k 60 | egrep -v -e "sd.[1-9]"
...
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           3.15    0.00    5.63   11.35    0.00   79.87

Device:         rrqm/s   wrqm/s   r/s   w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await
svctm  %util
sda              15.40     1.17 50.66  2.63  2673.22   124.85   105.01     0.55
10.39   6.27  33.41
sdb               0.07     0.03 447.29  1.02  4359.01   191.90    20.30     0.32
0.71   0.27  12.13
sdc               5.73     1.53 114.93  8.42  1264.86   100.58    22.14     1.05
8.48   3.56  43.94
sdd               5.57     2.07 121.83  9.57  1319.45   104.12    21.67     1.19
9.02   3.63  47.72
sde               5.53     1.45 111.45  8.52  1246.53   101.92    22.48     0.95
8.26   3.41  40.68
sdf               5.45     2.02 118.93  8.00  1281.92   106.25    21.87     1.19
9.27   3.74  47.44
```

其中：黑色为SATA，绿色为SSD，红色为SAS

4块SAS硬盘上的访问量和超过SSD硬盘上的访问量

热点迁移的数据对比（2）

- 按对象访问热点进行迁移：最热的进SSD，中等热度的放SAS，轻热度的存SATA
- SSD + 4 * SAS + SATA上的访问负载如下：

```
[root@cache161 ~]# iostat -x -k 60 | egrep -v -e "sd.[1-9]"
…
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           3.15    0.00    5.63   11.35    0.00   79.87

Device:         rrqm/s   wrqm/s   r/s   w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await
svctm  %util
sda               5.08     1.65 18.55  2.52  1210.07   119.00   126.18     0.14
6.50   5.46  11.51
sdb               1.68     0.05 610.53  1.75  6962.29   413.47    24.09     0.28
0.46   0.23  14.25
sdc               0.22     0.03 28.87  0.97  1172.93   189.13    91.31     0.16
5.28   4.40  13.13
sdd               0.23     0.02 29.70  0.77  1133.47   122.53    82.45     0.15
4.99   4.59  13.57
sde               [illegible]  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
5.00   4.24  12.40
sdf               0.10     0.02 28.42  0.55  1090.27   115.00    83.22     0.15
5.04   4.44  12.86
```

其中：黑色为SATA，绿色为SSD，红色为SAS

SSD硬盘上的访问量是4块SAS硬盘上访问量之和的5倍以上，SAS和SATA的硬盘利用率低了很多

淘宝网 Taobao.com

CDN系统的未来发展

- CDN系统的研发与运维
  - 持续提高节点性能（应用软件、操作系统等）
  - 精细化和自动化全局调度系统
  - 优化视频支持（P2P结合）、移动网络支持
  - 持续提高CDN系统可运维性，服务质量监测
  - 面向音视频通讯的中转支持
- CDN系统的建设
  - 系统进一步整合，优化不同规模节点的硬件配置，建设中型和大型节点
  - 定制化和快速部署

淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

三、事例：淘宝核心数据库

四、淘宝开源策略

五、小结

淘宝网 Taobao.com

商业解决方案

- IOE＝ IBM + Oracle + EMC
- 优点：
  - 稳定，功能非常强大
  - 支持完善，方便运维
- 缺点：
  - License贵，软硬件成本高
  - 集中式架构，不利扩展
  - 软件黑盒子

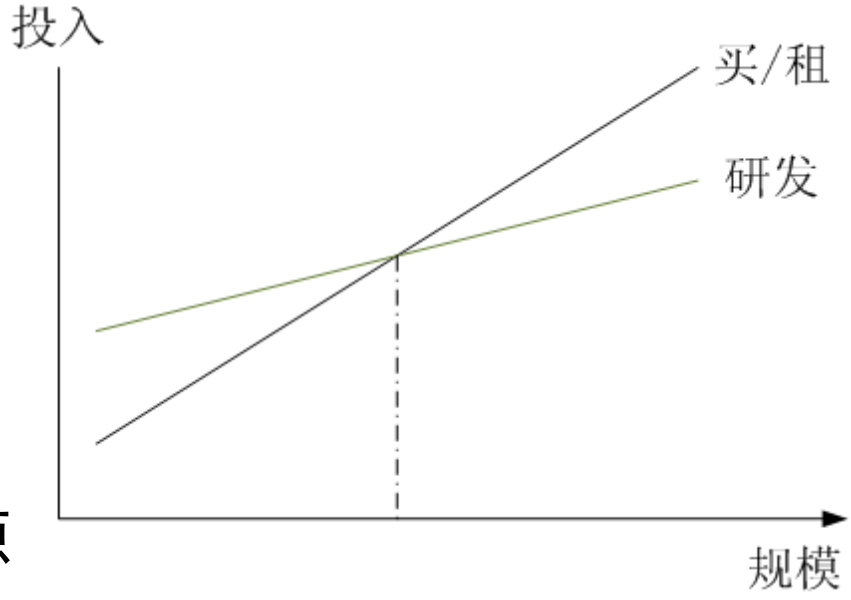
淘宝网 Taobao.com

MySQL发展

- 2008年开始周边的数据库使用MySQL
- 2010年 组建数据库开发团队
  - 结合高速的非易失存储设备，多层优化MySQL
  - 实现分库分表的TDDL中间件成熟
  - 核心业务开始迁移到MySQL
- 2011年 核心数据库迁移到基于MySQL解决方案
  - 商品库：MySQL 16*2
  - 交易库：MySQL 16*2，IOE成本是2千万，现在成本是近400万，TPS从9000提升到12.8万

经验

- 商用软件不能满足大规模系统的需求
- 采用开源软件与自主开发相结合，有更好的可控性，更高的可扩展性
- 规模效应，研发投入都是值得的
- 在软件和硬件多个层次优化，优化是长期持续的过程
- 先平台后业务 vs 先业务后平台，后者更顺
  - 发挥边际效应，提高资源利用率



淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

三、事例：图片存储、CDN与DB

四、淘宝开源策略

五、小结

淘宝网 Taobao.com

淘宝开源策略

- 淘宝系统中使用大量的开源软件，并在开源软件的基础上进行改进和定制，并把工作成果回馈给上流的开源社区
  - 淘宝维护了自己的Linux内核树并不断地向Linux社区贡献patch，目前淘宝在对Linux内核贡献补丁数的公司排名为104
  - 签署OCA，向Oracle回馈了JVM的补丁（接受了18个）和MySQL的补丁
  - 向Apache回馈hadoop、hbase和Traffic Server的补丁
  - 向Nginx回馈补丁，并开源了基于Nginx的Tengine服务器等

淘宝网 Taobao.com

Linux Kernel

- http://kernel.taobao.org
- Ext4文件系统核心开发团队之一
- 被Linux官方接受235+个patch
- 全球最活跃Linux开发团队排104名
  - 自2006年起统计
- 开源虚拟化项目sheepdog
  - 主要代码贡献者和维护者

No.102	LG Electonics	242(0.07%)
No.102	Calxeda	242(0.07%)
No.104	Tao Bao	235(0.07%)
No.105	Miracle Linux	232(0.07%)
No.106	P.A. Semi	231(0.07%)
No.107	OpenedHand	226(0.07%)
No.108	rPath	220(0.07%)
No.108	Embedded Alley Solutions	220(0.07%)
No.110	RisingTide Systems	219(0.06%)
No.111	Bull SAS	215(0.06%)
No.112	Myricom	214(0.06%)
No.113	LWN	208(0.06%)
No.114	Hansen Partnership	198(0.06%)
No.114	STRATO	198(0.06%)
No.116	M&N Solutions	192(0.06%)
No.116	Avionic Design Development GmbH	192(0.06%)
No.116	igalia	192(0.06%)
No.119	Hauppauge	185(0.05%)
No.120	EXAR	184(0.05%)
No.121	Voltaire	180(0.05%)
No.122	CSR	179(0.05%)
No.123	SANPeople	178(0.05%)
No.123	Collabora Multimedia	178(0.05%)
No.125	Toshiba	174(0.05%)
No.126	Tuxera	173(0.05%)
No.127	Philosys Software	172(0.05%)
No.128	Bitmer	170(0.05%)
No.129	tieto	169(0.05%)
No.130	HuaWei	164(0.05%)
No.131	MathEmbedded Consulting	163(0.05%)
No.132	OMICRON electronics	162(0.05%)
No.133	secunet Security Networks AG	160(0.05%)
No.134	Real-Time Remedies	156(0.05%)
No.135	Open Nandra	154(0.05%)
No.136	Barco	142(0.04%)
No.137	US National Security Agency	141(0.04%)
No.138	Realtek	137(0.04%)
No.138	Synopsys	137(0.04%)
No.140	Imagination Technologies	136(0.04%)
No.141	Apple	132(0.04%)
No.142	Candela Tech.	130(0.04%)
No.143	Eukrea Electromatique	129(0.04%)
No.144	Etersoft	124(0.04%)

Source: http://www.remword.com/kps_result/all_whole.html

淘宝网 Taobao.com

淘宝开源策略

- 淘宝建设了淘蝌蚪开源平台，开源多年开发主力的基础软件（共100余个）
  - 2010.6开源了分布式缓存和K/V系统TAIR
  - 2010.9开源了分布式存储系统TFS
  - 2011.8开源了海量数据库OceanBase
  - 阿里Web框架WebX
  - 分布式数据库中间件 TDDL
  - 异构数据源数据交换工具 DataX
  - 性能分析工具 TProfiler
  - 手机自动化测试框架Athrun
  - 分布式任务调度引擎TBSchedule等等。

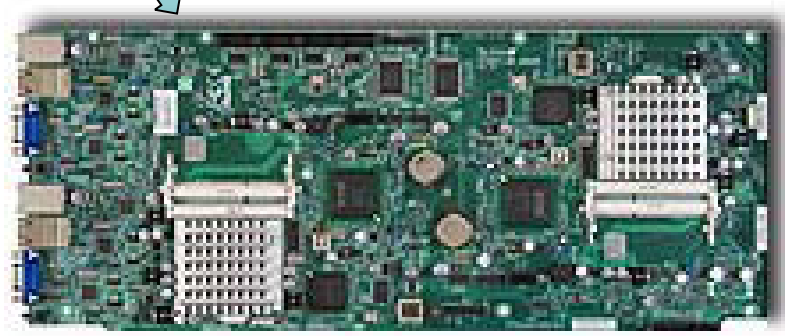
淘宝网 Taobao.com

淘宝开源平台

淘宝网 Taobao.com

定制的低功耗服务器

（背面）



一个板卡集成两个服务器系统

（正面，24个可插拔硬盘）

- 每个热插拔模块2个nodes
- 每个NODES 3块硬盘
- 支持 24 x 2.5” SATA/SSD
- 选择2U 8 nodes 的原因：
  - SuperServer: 2U 8 nodes 支持热插拔设计。
  - 降低功耗（2U 8nodes 共享4个系统FAN）
  - 成本更低（8nodes共享1个机箱，1u 4nodes 共享1个机箱）
  - 2U TWIN机箱和所使用的主板是成熟产品
- 单服务器配置：
  - Intel® Atom™ D525 with 2 cores
  - Intel® ICH9R Chipset
  - 4GB memory DDR23 800MHZ SO-DIMM w/o ECC
  - LAN: Intel 82574L 2*1GB
  - HDD:
    - 1* SSD 80G，
    - 2* 2.5” SATA 500GB

淘宝网 Taobao.com

低功耗服务器开源

- 开源网站 http://www.greencompute.org/

开源绿色计算

English Version

首页 | 项目介绍 | 设计规范 | 合作赞助 | 论坛讨论 | 联系方式 | 新闻公告

设计规范名称	采用CPU型号	版本	发布时间	下载地址	面向应用
主板设计规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用
机箱和电源设计规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用
服务器测试规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用

淘宝网 Taobao.com

GreenCompute开源项目

- 目标是推动互联网整体硬件基础设施（包括服务器、网络设备、IDC机房、机架和电源等）的节能环保；
- 组织方式是采用多方合作的机制吸纳业内同行共同参与该项目，
- 运转方式是根据不同的设施类型分成不同的子项目，分别有特定的参与方负责推动在该方向上“绿色”设备的定制化、产品化和规模化；
- 成果将以开源的方式发布到项目网站上供业内的人士参考。

淘宝网 Taobao.com

议程

一、淘宝网的简介

二、淘宝网电子商务平台

三、事例：图片存储、CDN与DB

四、淘宝开源策略

五、小结

淘宝网 Taobao.com

小结

- 淘宝是开源系统的受益者，并积极参与开源生态系统的建设，促进开源生态系统的发展，积累更好的口碑，凝聚人才，迎接未来更大的技术挑战。
- 淘宝公司希望以更开放的方式与业界一起进行技术创新

淘宝网 Taobao.com

讨论

- Q&A
- 谢谢！