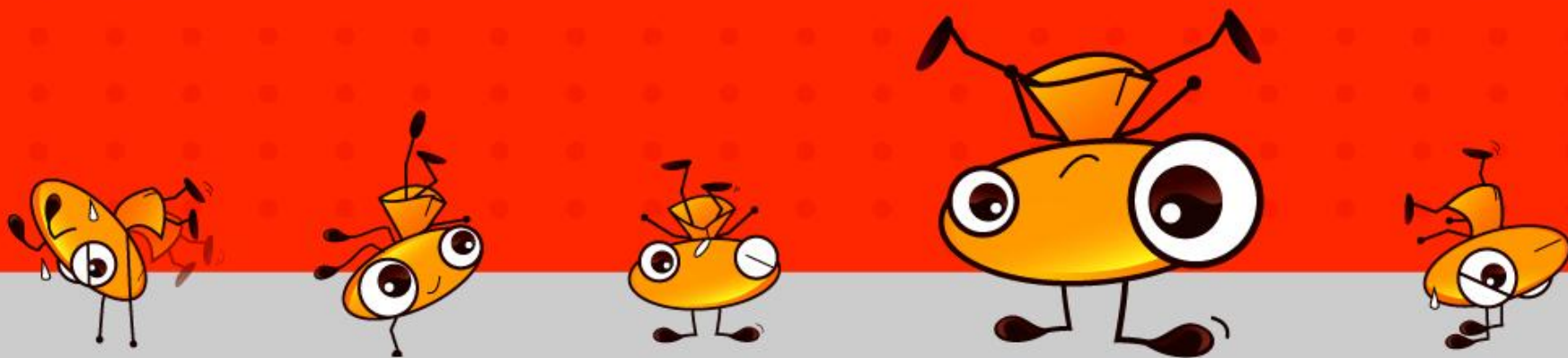
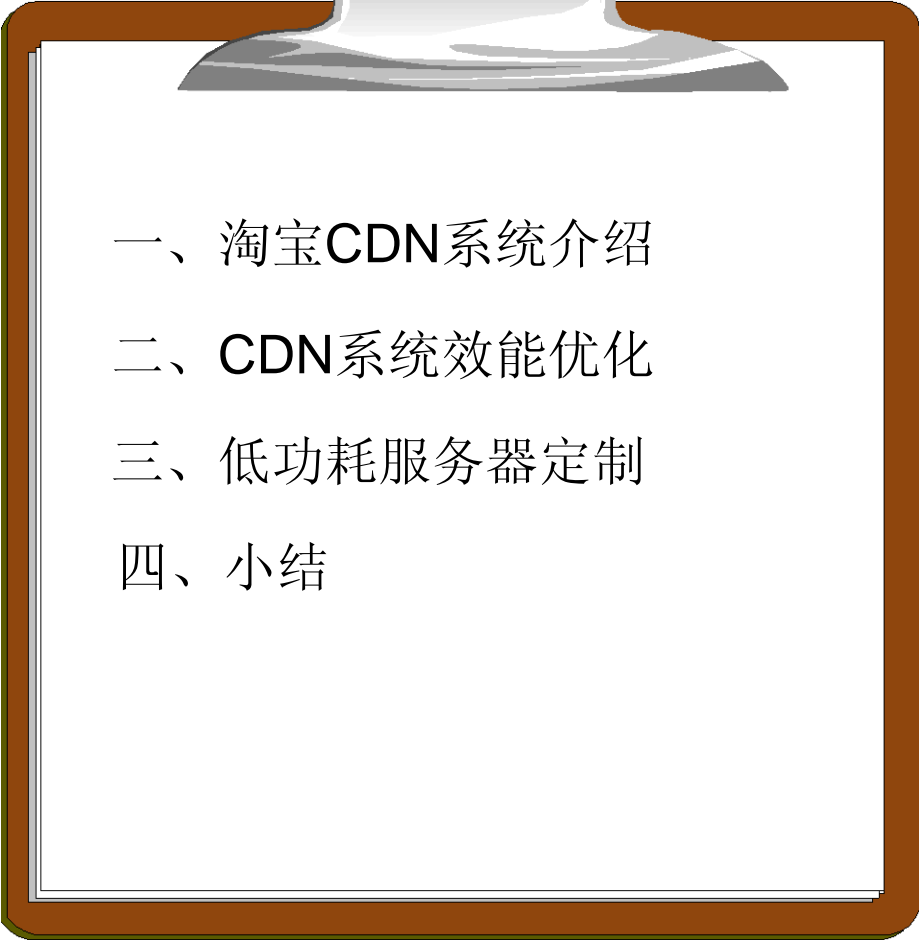


打造高效能的CDN系统

任卿 (易统)
淘宝核心系统研发部





- 
- 一、淘宝CDN系统介绍
 - 二、CDN系统效能优化
 - 三、低功耗服务器定制
 - 四、小结



什么是CDN

- **CDN(Content Delivery Network)**内容分发网络，简单的说就是不同地点缓存内容，然后通过负载均衡等技术将用户请求定向到最合适的缓存服务器上获取内容，提高用户访问网站的响应速度。
- 通过**CDN**服务提高网站的访问性能及稳定性，保障网站服务品质。

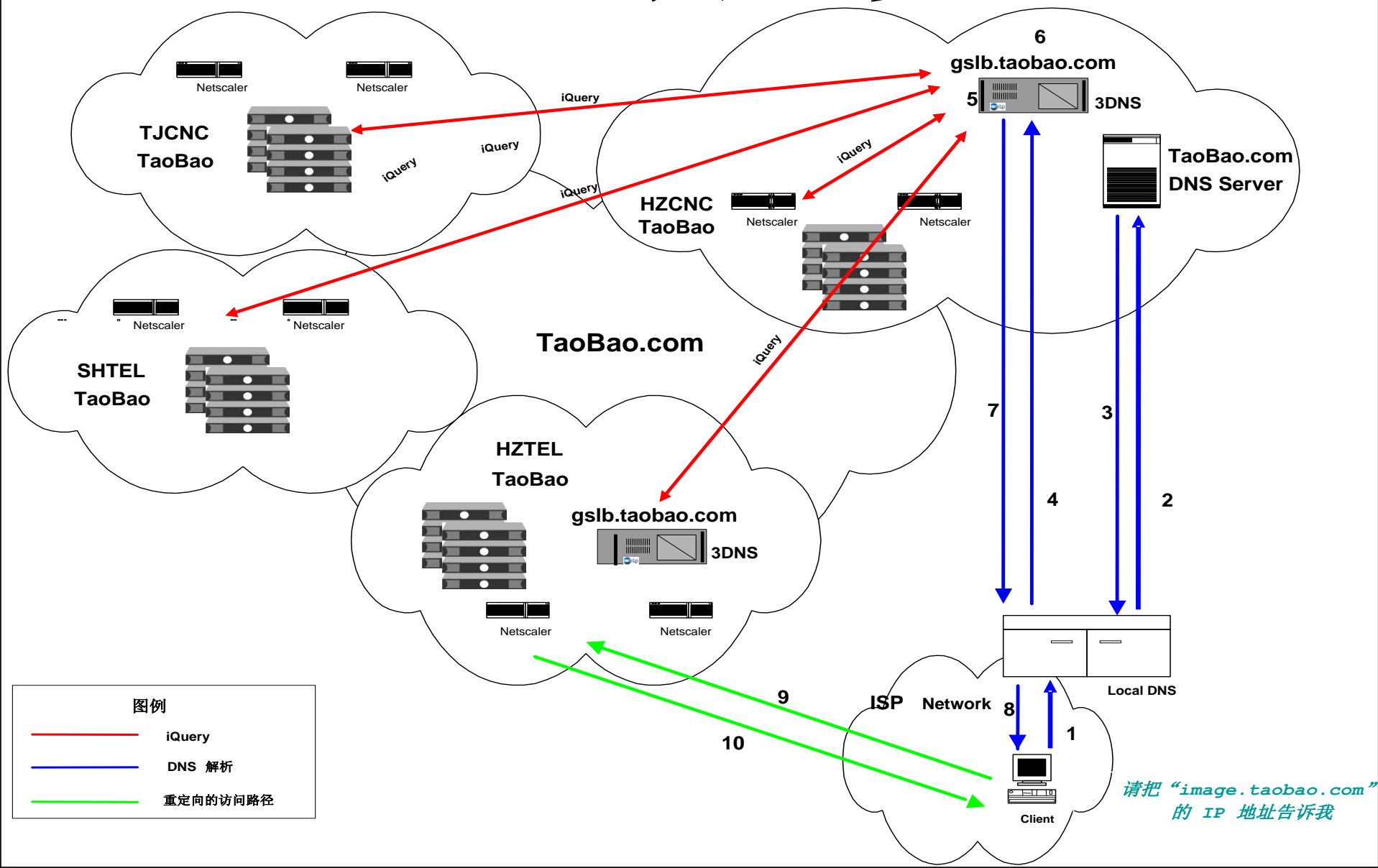


淘宝CDN的一些数字

- CDN系统的规模
 - 500T容量的原图 + 500T容量的缩略图
 - 约700亿左右的缓存图片数，平均图片大小约20KB
 - 18KB以内的对象数量占总数量的80%
- CDN部署的规模
 - 近100个节点，部署在网民相对密集的主要中心城市
 - 每个节点目前处理能力在10G左右
 - CDN部署的总处理能力800G左右
 - 目前承载淘宝流量高峰时近400G流量



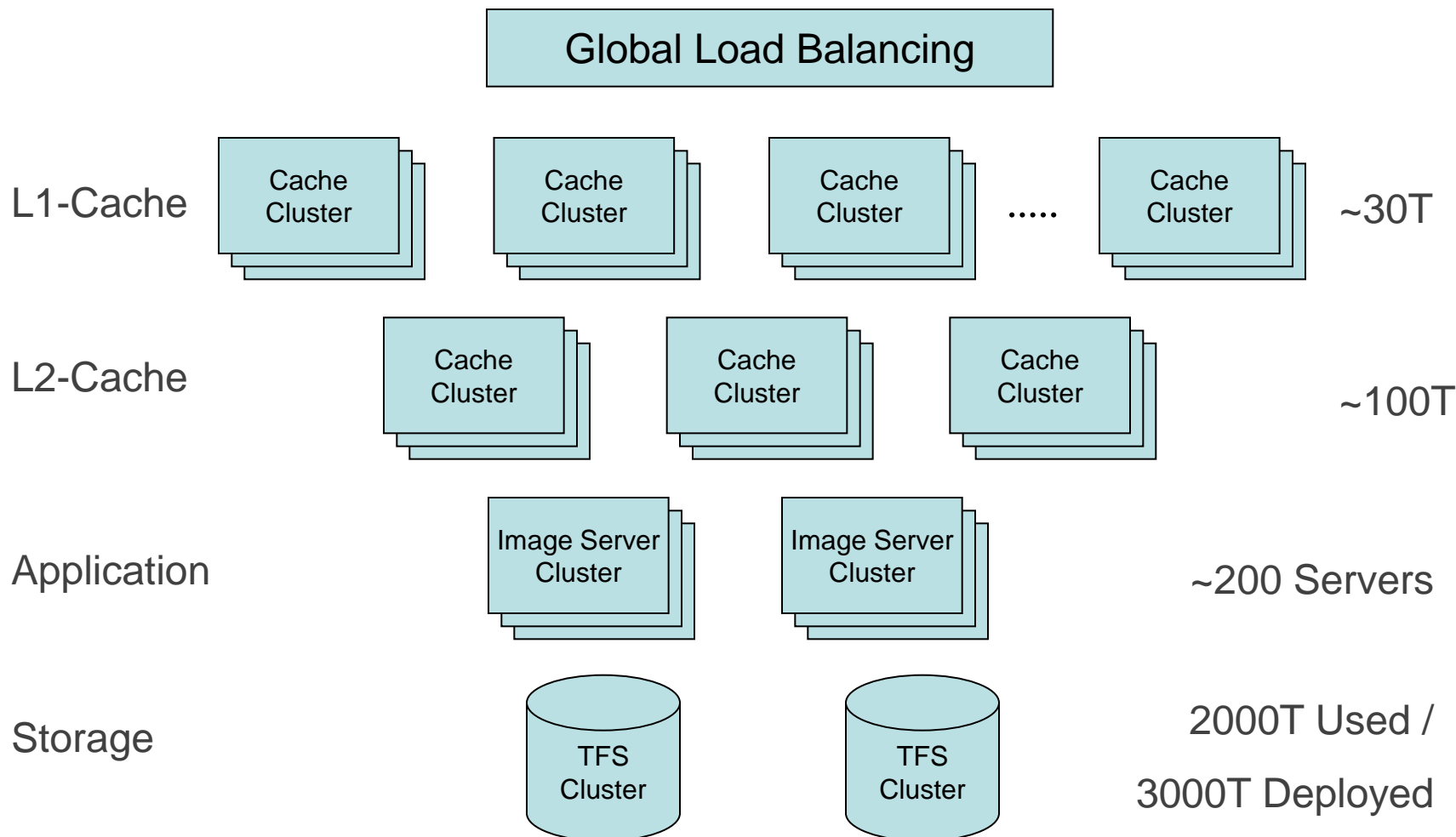
CDN系统总览



请把“image.taobao.com”
的 IP 地址告诉我

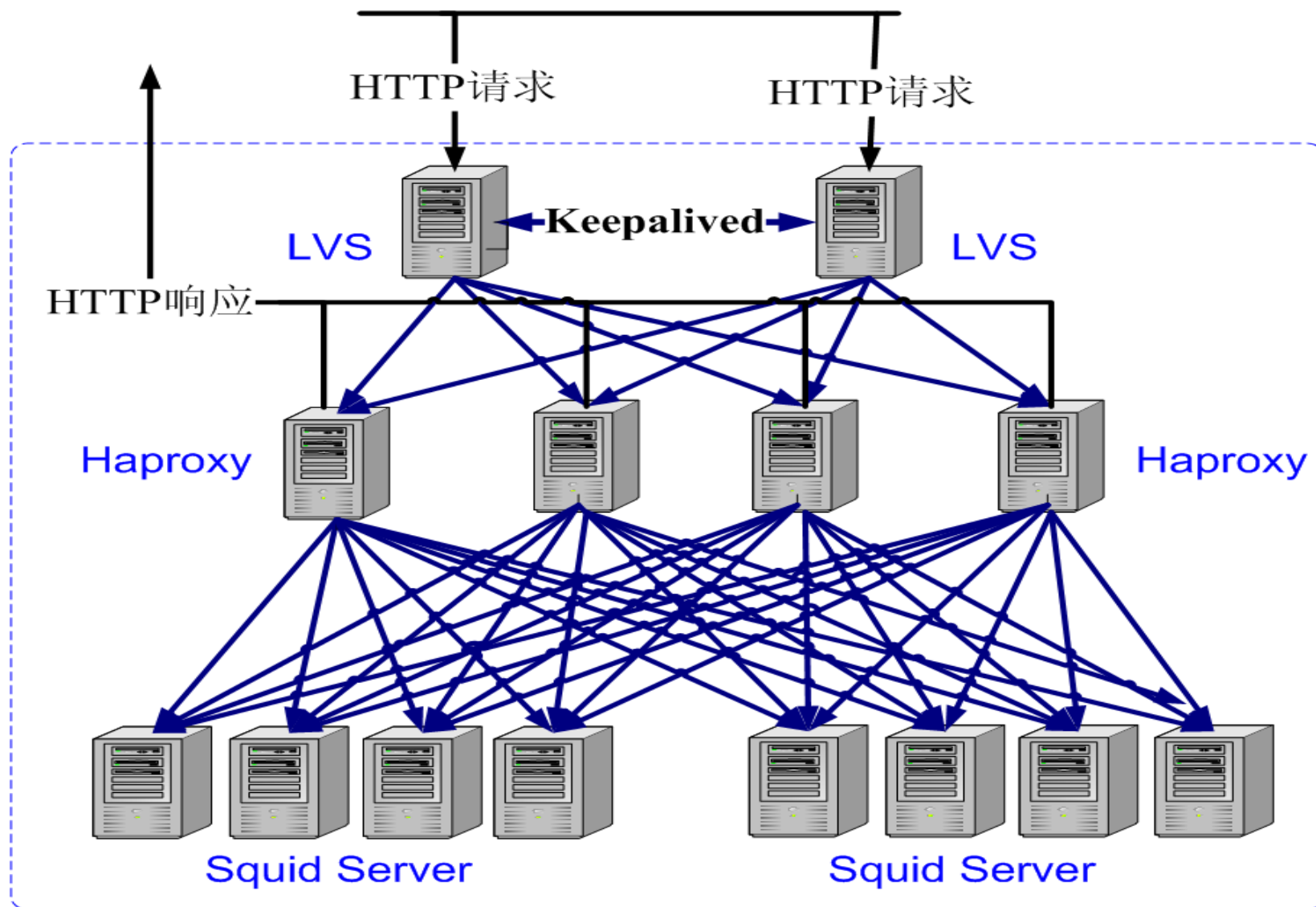


淘宝CDN系统体系结构

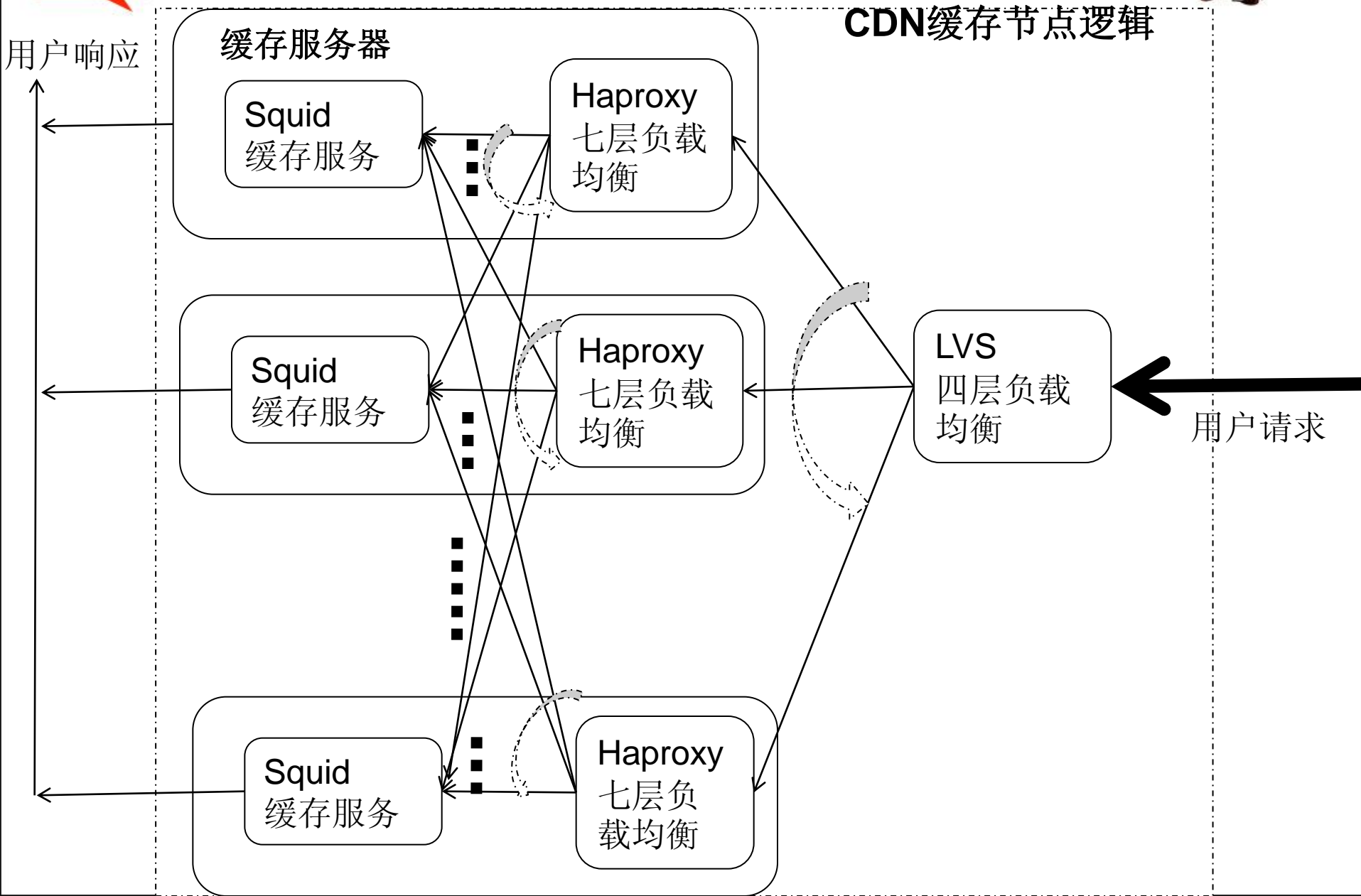




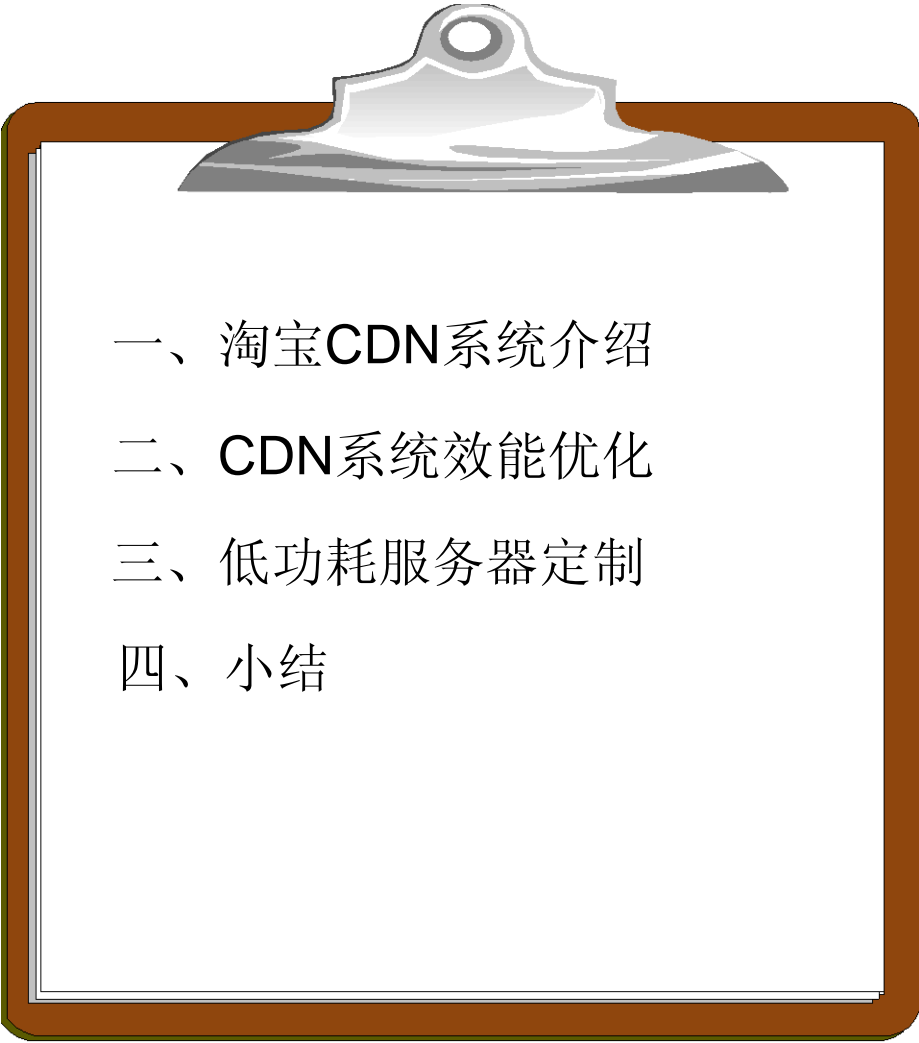
软件负载均衡节点部署架构



CDN节点逻辑架构





- 
- A large, brown-bordered clipboard with a silver clip at the top, containing the agenda list.
- 一、淘宝CDN系统介绍
 - 二、CDN系统效能优化
 - 三、低功耗服务器定制
 - 四、小结



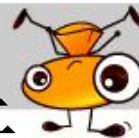
CDN系统效能优化

- 负载均衡优化
- 网络层优化
- 存储优化

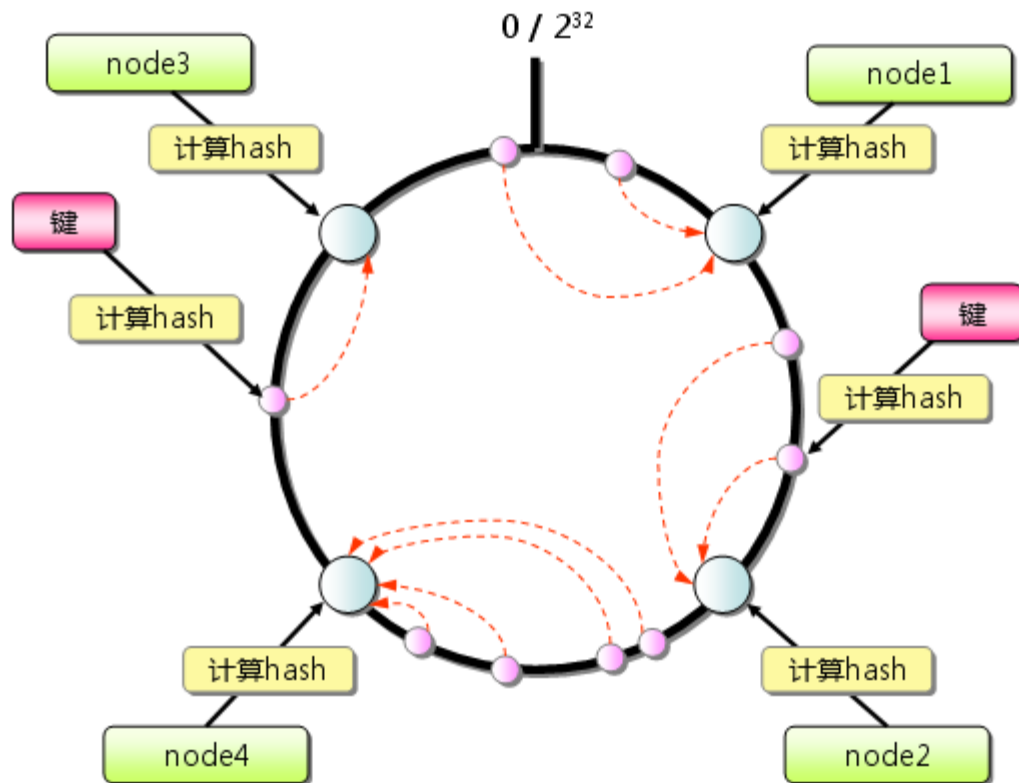


Haproxy软件优化

- hash优化，提高效率、消除短板
 - servers=128 nodes=128 stdvar=3.010755e+06 stdvar/avg=0.08973
 - servers=128 nodes=256 stdvar=1.974319e+06 stdvar/avg=0.05884
- 精确的调度和数据清理
 - 基于一致性哈希调度请求
 - 基于调度历史做精确清理，避免全量清理操作
- 支持Cache功能
 - 将最热的内容缓存在haproxy中
 - 改善性能，应对访问的热点



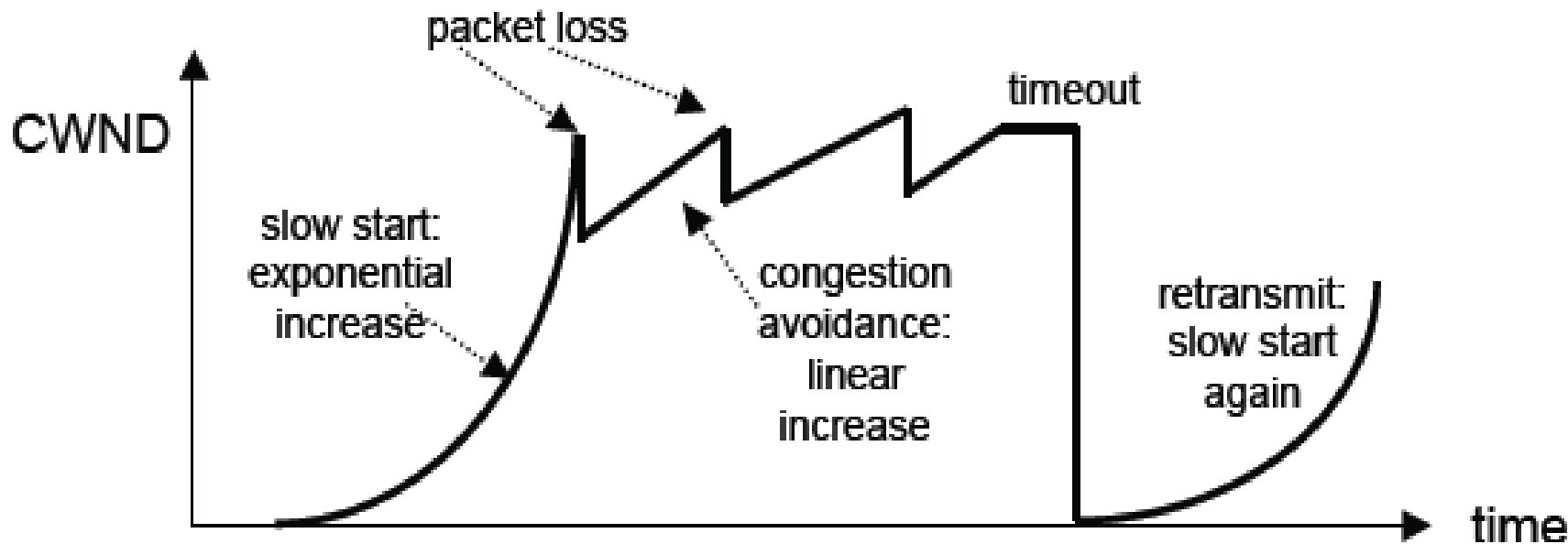
基于一致性哈希的调度算法





Haproxy长链接支持

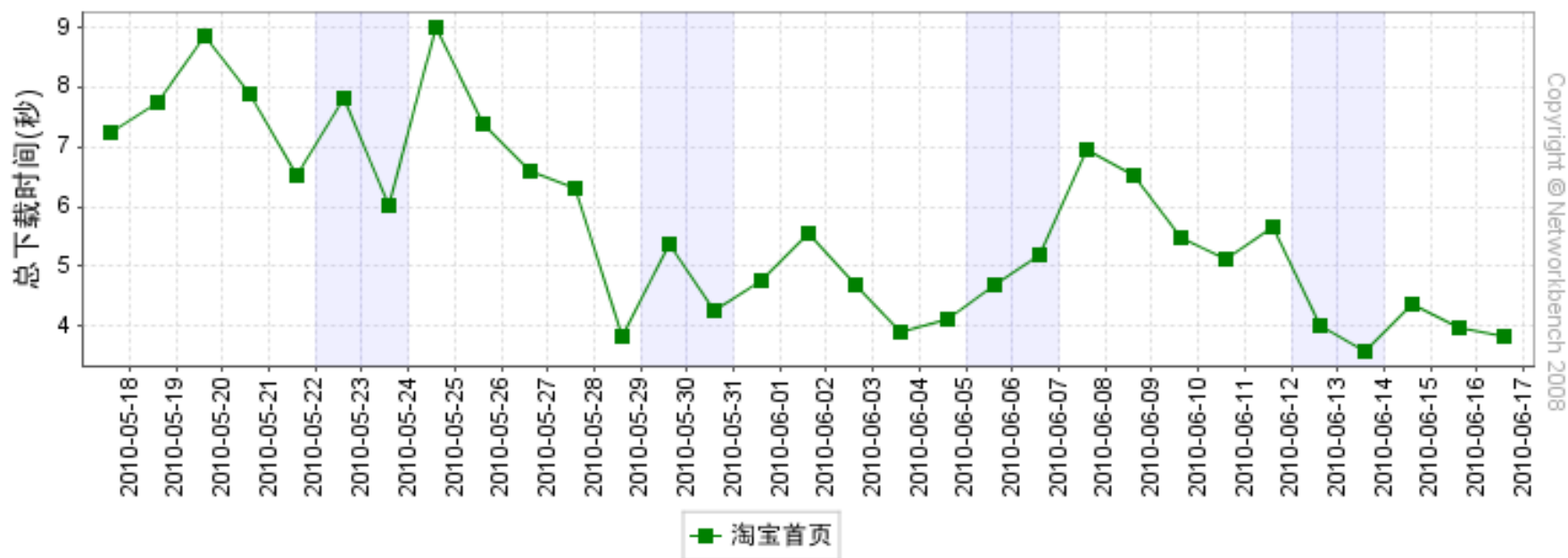
- 长链接的作用
 - 对客户端的keepalive, 提高用户响应速度
 - 对服务端的keepalive, 提高服务器处理能力
- TCP拥塞控制





Haproxy长链接效果

- 挖掘淘宝访问的业务特点，平衡系统开销和加速效果
- 提升用户体验，响应时间最多提升50%+





动态内容加速

- 针对不能被缓存的动态内容做加速
- 基于**TCP**协议原理，优化网络通讯
- 内核协议栈调优
- 充分利用**CDN**节点和中心站点之间的“高速公路”



动态内容加速效果

- 性能提升**15%**左右
- 目前已经有两个应用上线测试

before:detail_orig=v_ok_real after:detail_haproxy=v_ok_real

Compare Files:/tmp/b,/tmp/a

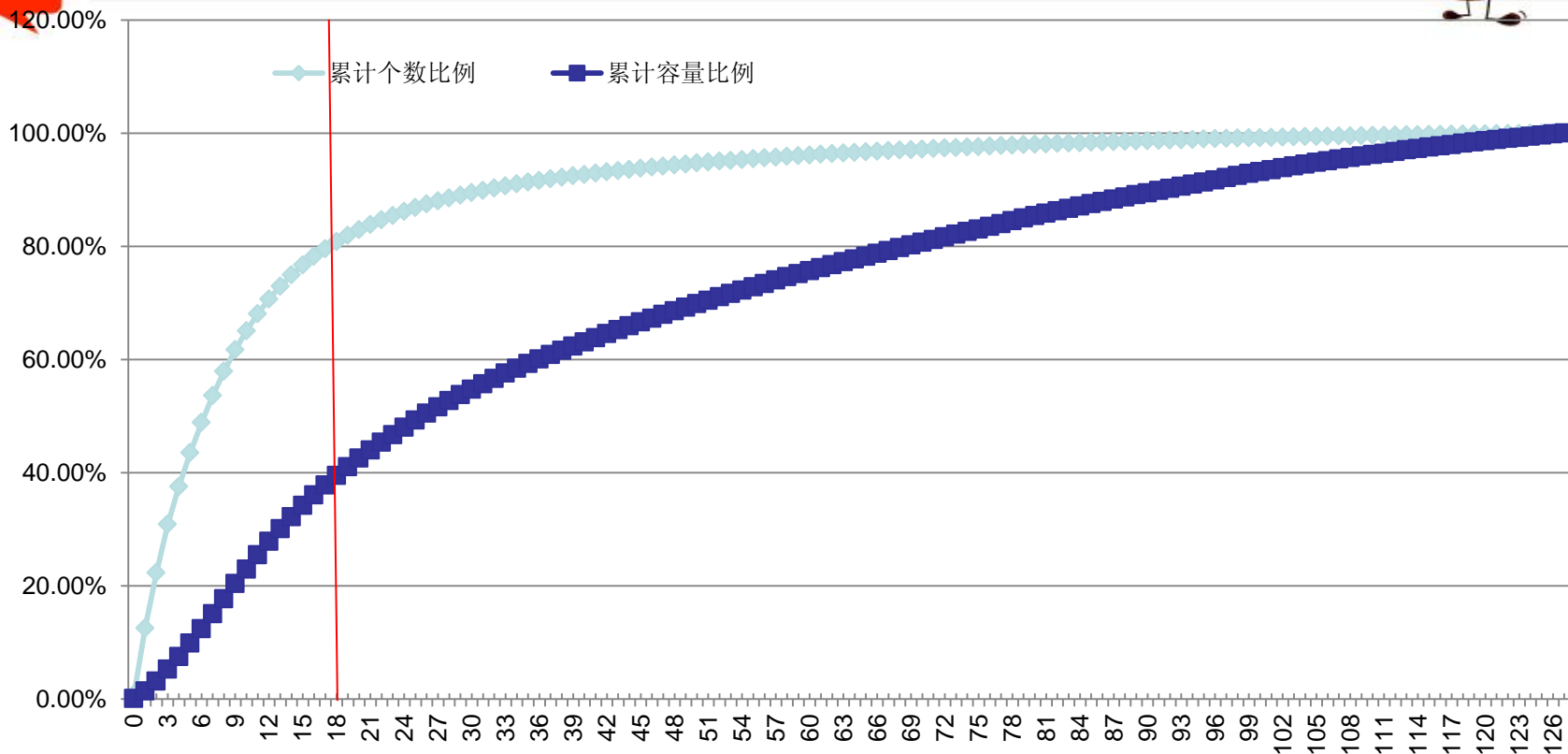
quantiles(%)	original(ms)	optimized(ms)	diff(ms)	diff(%)	org_mean(ms)	opt_mean(ms)	dif_mean(ms)	dif_mean(%)
5.0	593.0	507.0	-86.0	-14.5	458.1	373.3	-84.8	-18.5
10.0	750.0	656.0	-94.0	-12.5	565.6	477.8	-87.8	-15.5
20.0	1000.0	906.0	-94.0	-9.4	724.7	627.9	-96.8	-13.4
30.0	1266.0	1125.0	-141.0	-11.1	859.3	756.2	-103.1	-12.0
40.0	1594.0	1422.0	-172.0	-10.8	1001.3	885.1	-116.2	-11.6
50.0	2000.0	1823.0	-177.0	-8.8	1159.8	1029.9	-129.9	-11.2
60.0	2547.0	2344.0	-203.0	-8.0	1341.3	1202.8	-138.5	-10.3
70.0	3407.0	3094.0	-313.0	-9.2	1571.2	1417.7	-153.5	-9.8
80.0	4657.0	4422.0	-235.0	-5.0	1872.1	1699.6	-172.5	-9.2
90.0	7469.0	6938.0	-531.0	-7.1	2309.4	2127.1	-182.3	-7.9
95.0	11547.0	9404.0	-2143.0	-18.6	2672.5	2430.0	-242.5	-9.1
99.0	27819.0	23375.0	-4444.0	-16.0	3234.6	2877.1	-357.5	-11.1
99.9	98453.0	68094.0	-30359.0	-30.8	3590.4	3158.8	-431.6	-12.0
99.99	463593.0	200248.0	-263345.0	-56.8	3750.5	3263.8	-486.7	-13.0
99.999	1161078.0	215717.0	-945361.0	-81.4	3788.7	3278.1	-510.6	-13.5
100	1161078.0	215717.0	-945361.0	-81.4	3830.9	3285.6	-545.3	-14.2



CDN节点存储优化

- 充分了解缓存内容特点
- 充分了解存储介质特点
- 资源合理组合配置
- 优化缓存处理逻辑

CDN缓存对象的特性



- 0~18KB的对象数量占总数量的80%，而存储量只有不到40%
- 80%被访问到的对象，其存储占用只有不到20%
- 访问的局部性，决定分层次的对象存储



	内存	Sata固态硬盘	Sata机械硬盘
IO能力	6.4Gbps	<10000 iops	<160iops
存储容量	4~16GB	80 ~160GB	500~1000GB
单价（每G的成本）	150	20	2



存储系统优化思路

- 充分利用访问局部性
- 通过控制将热点内容存储在内存和SSD，降低对Sata机械盘的访问
- 服务器IO的瓶颈在Sata机械盘
- 增加Sata机械硬盘提高存储能力，降低存储成本



存储系统优化实践

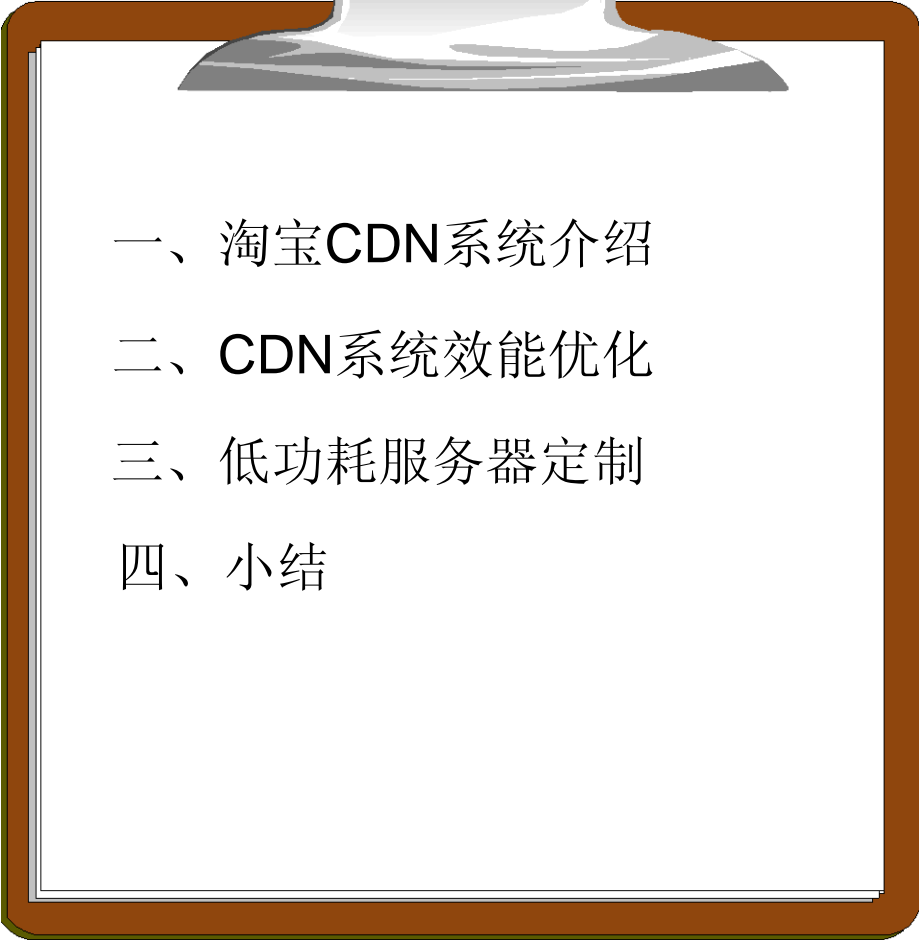
- 改进Squid的COSS文件
- 根据对象大小和访问特点切分，分级存储
- 支持热点迁移的**TCOSS**文件系统
- 用sendfile来发送缓存在硬盘上的对象
- Squid内存优化， 一台Squid服务器若有一千万对象，大约节省400M内存，更多的内存可以用作Squid Memory Cache



存储系统优化效果

- 缓存字节命中率：97%以上
- 缓存请求命中率：97%以上
- 缓存响应时间：10ms以内
- 单台服务器缓存对象数：6000万以上



- 
- 一、淘宝CDN系统介绍
 - 二、CDN系统效能优化
 - 三、低功耗服务器定制
 - 四、小结



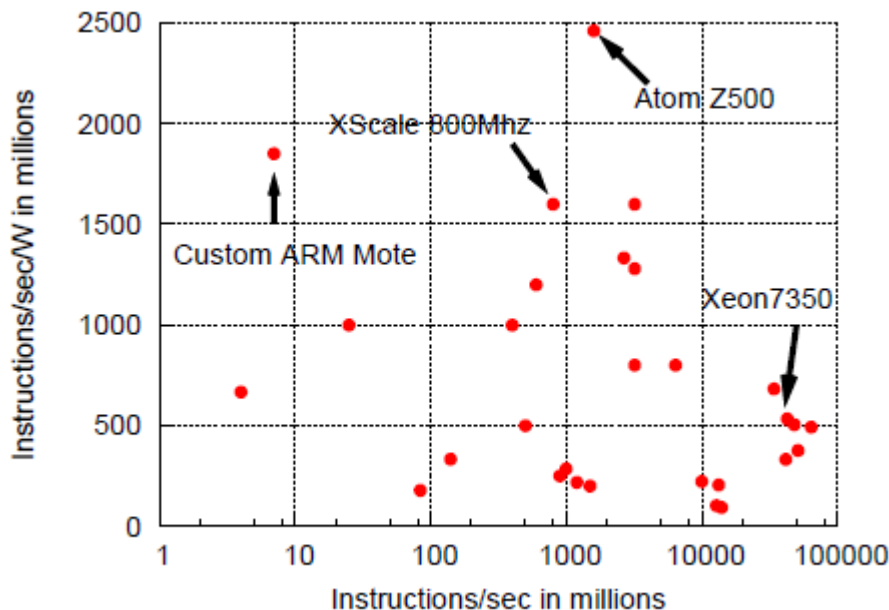
- 请求-响应的服务模型
 - 请求的处理
 - 基本上是通过hash(URL)对响应内容进行查询
 - 逻辑简单，消耗CPU较少
 - 响应的处理
 - 命中--》磁盘--》内存缓冲--》网卡--》网络--》用户
 - 不命中--》网卡--》网络--》网卡--》内存缓冲--》网卡--》网络--》用户
- I/O密集型服务
 - CPU处理少
 - 绝大部分是IO操作
 - 容量越大，命中率越高，目标：98%

为什么考虑低功耗服务器



- 不断增大的CPU与IO之间的差距
 - 对于IO密集型服务，硬盘、网卡是瓶颈
 - 在消耗<30%CPU时，硬盘IO已满

■ CPU功耗的增加快于处理速度的增长



来源：FAWN - A Fast Array of Wimpy Nodes

- 降低CPU的峰值功率比动态调整功率使用更能减低系统能耗
 - 芯片的泄露电流随频率变化很小
 - 性能强的CPU，外围组件（如网卡、总线等）能力都较强，因此耗电也相对高
- 传统CPU峰值功耗高限制了IDC服务器的密度
 - IDC中每个机柜都有额定的功率
 - 虽然空间上可以放置几十台服务器，但是由于传统服务器的功耗高，使得整个机架只能放置几台服务器，造成空间浪费
- 在成本和功耗不增加的情况下，将原有单台高性能服务器承载的流量分摊到多台低功耗服务器上，降低单台服务器故障带来的影响
- 更高密度的存储能力



- 低功耗
 - 单位服务器满载功耗控制在20~30Watts
- 高密度
 - 单位空间放置尽可能多的服务器和存储
- 兼容性
 - 通用的硬件方案
- 高性价比
 - 单位服务能力的成本及功耗为衡量标准
 - 成本包括：投入成本和运营成本
- 可运维性

ATOM低功耗服务器



1. 2U的机箱，集成8个单独的服务器硬件系统，每两个服务器系统集成在一个板卡上；每个服务器间只共享电源
2. 每个服务器系统配置：
 - Dual Core Intel® Atom™ D525(1.8GHz 13W)processor + Intel® ICH9R Chipset
 - 4GB **Non-ECC** DDR3 1333MHz SO-DIMMs (per node) support
 - 3x 2.5" Hot-swap SATA HDD (RAID 0, 1)
 - 2GE with Intel 82574L
3. 可热插拔硬盘与主板分离，共提供24个盘位支持
4. 内置BMC支持IPMI
5. 720W 冗余高效金牌电源，支持PMbus



一个板卡集成两个服务器系统



(正面，24个可插拔硬盘)

方案对比：存储IO配置对比



• 服务器

	Atom低功耗	Xeon偏低功耗	Xeon服务器
CPU	Atom D525 -1*2 cores - 1.80Ghz - 1MB cache	Intel L3406 -1*2cores -2.26Ghz -4MB cache	Intel E5620 -1*4Cores -2.66GHz -12MB cache
内存	2*2GB	4*4GB	3*4GB
SSD	1*80GB	1*160GB	2*160GB
SAS	NA	NA	6*600GB
Sata	2*500GB rpm7200 HyBrid	3*500GB rpm7200 EN	NA

■ 机械硬盘

机械硬盘	容量 (G)	单盘IOPS
Seagate SATA混合盘	500	120
SAS硬盘	600	180
SATA企业盘	500	130

■ 节点存储与IO

	单机SSD数	单机SATA数	单机SAS数	Cache服务器数目	机械盘总IOPS	节点SSD总容量 (G)	节点硬盘总容量 (G)	节点总容量 (G)
Xeon偏低功耗	1	3		22	8580	3520	33000	36520
Atom低功耗	1	2		64	15360	5120	64000	69120
Xeon服务器	2		6	10	10800	3200	36000	39200



服务器功耗估算对比



		单位功耗（瓦）	数量	功耗小计(瓦)
Atom低功耗	ATOM D525	13	1	13
	西数混合盘	2.2	2	4.4
	SSD	2	1	2
	网卡	1.9	1	1.9
	内存	2	2	4
	合计			25
Xeon低功耗	Intel L3406	30	1	30
	西数企业盘RPM7200	3.2	3	9.6
	SSD	2	1	2
	网卡	1.9	1	1.9
	内存	4	4	16
	合计			60
Xeon服务器	Intel E5620	80	1	80
	SAS	9	6	54
	SSD	2	2	4
	网卡	1.9	1	1.9
	内存	10	4	40
	合计			180



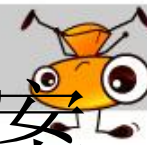
	机械盘总 IOPS	机械盘最大 利用率	内存和SSD 命中率	机械盘 COSS命 中率	单位请求消 耗机械盘 IOPS数	估算 QPS	平均访问对 象大小 (KB)	节点服务 能力 (Gps)
Atom低功耗	15360	80%	~92%	5.5%	2.14	104401	18	15.5
Xeon偏低功耗	8580	80%	~91%	5.0%	2.14	64150	18	9.5
Xeon服务器	10800	80%	~90.8%	5.2%	2.14	77642	18	11.5

	缓存服务器 功耗	cache数量	LVS服务器 功耗	LVS数量	交换机功耗	交换机数 量	总功耗 (瓦)
Atom低功耗	25	64	150	2	80	2	2060
Xeon偏低功耗	60	22	58	2	80	1	1516
Xeon服务器	180	10	150	2	80	1	2180

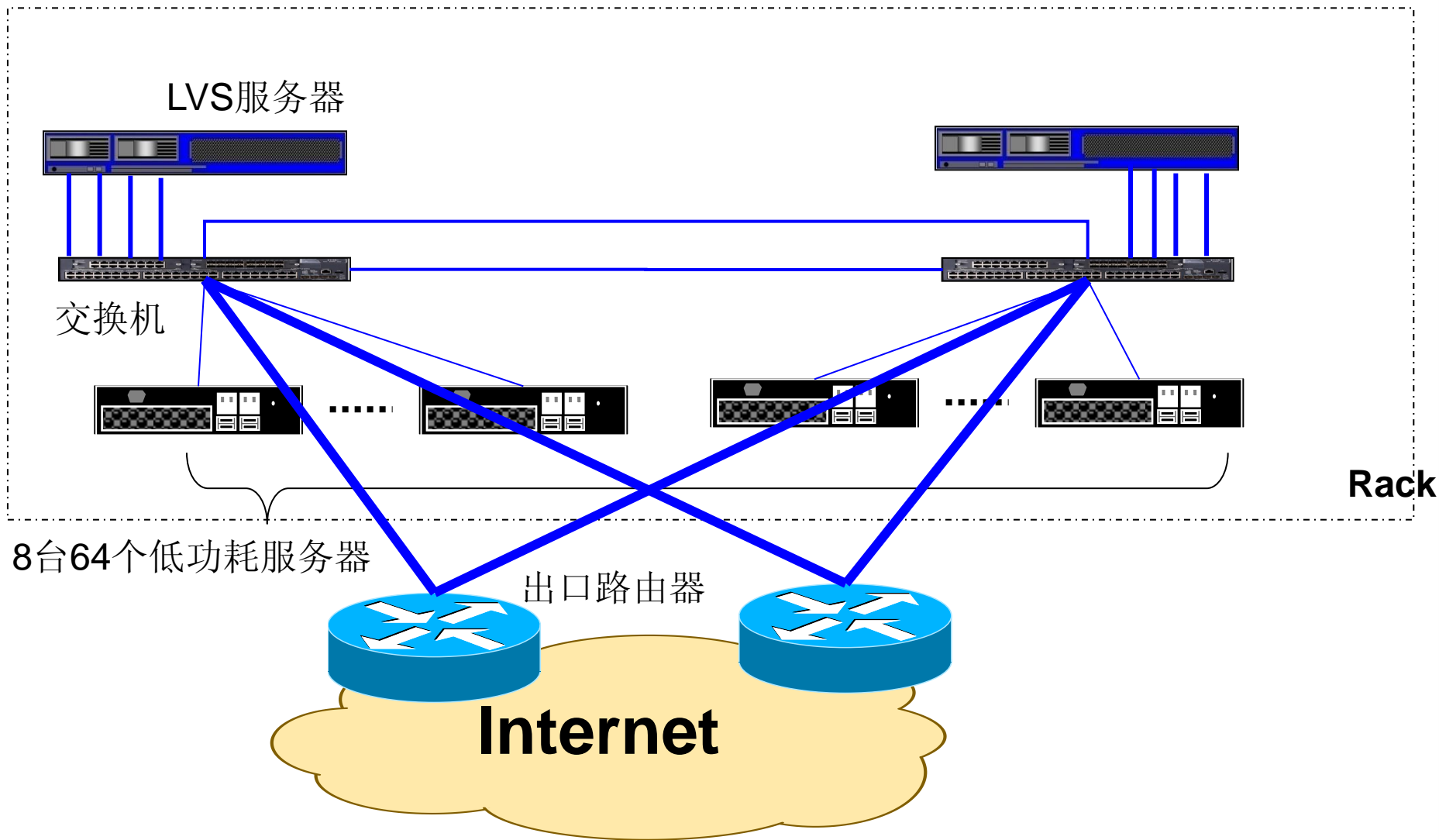
节点性价比与功耗比预估



	服务能力性价比 (kbps/元)	存储性价比 (MB/元)	服务能力性耗比 (Mbps/瓦)
Atom低功耗	1.72	2	7.7
Xeon偏低功耗	1 (基准值)	1 (基准值)	6.42
Xeon服务器	1.3	1.15	5.41

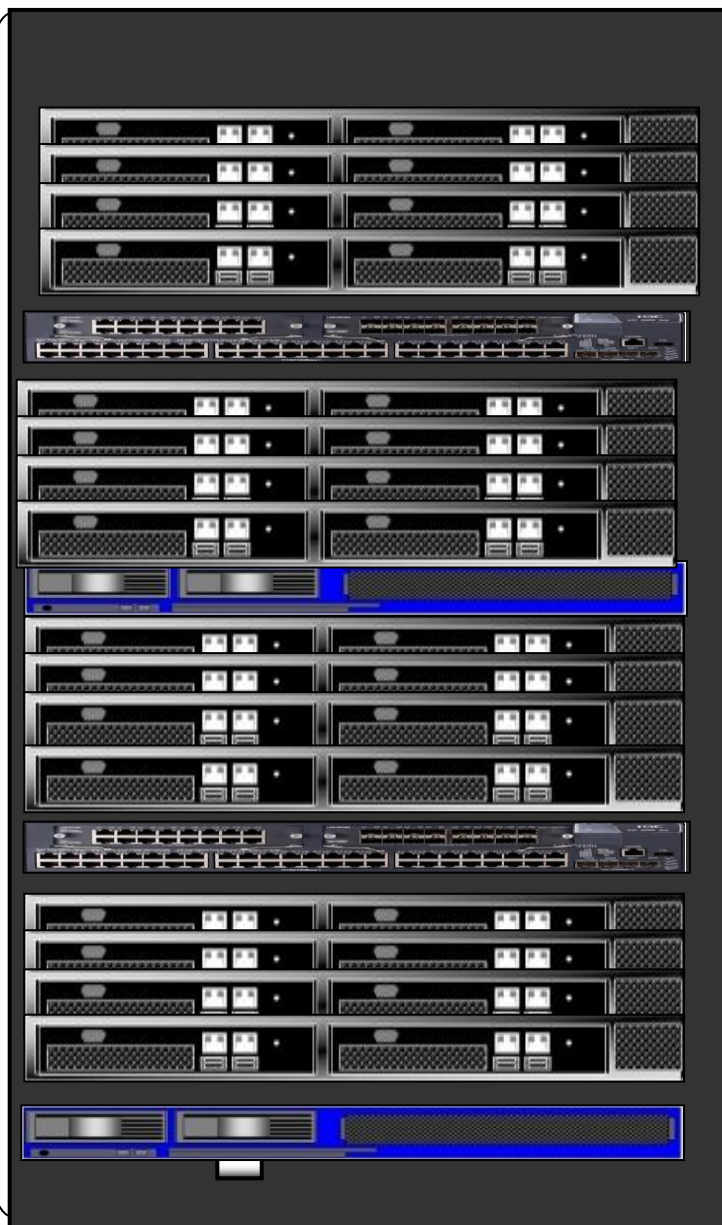


低功耗CDN一级缓存节点方案





机架



2U*2 低功耗服务器



1U 交换机



2U*2低功耗服务器



1U 负载均衡服务器



2U*2 低功耗服务器



1U 交换机



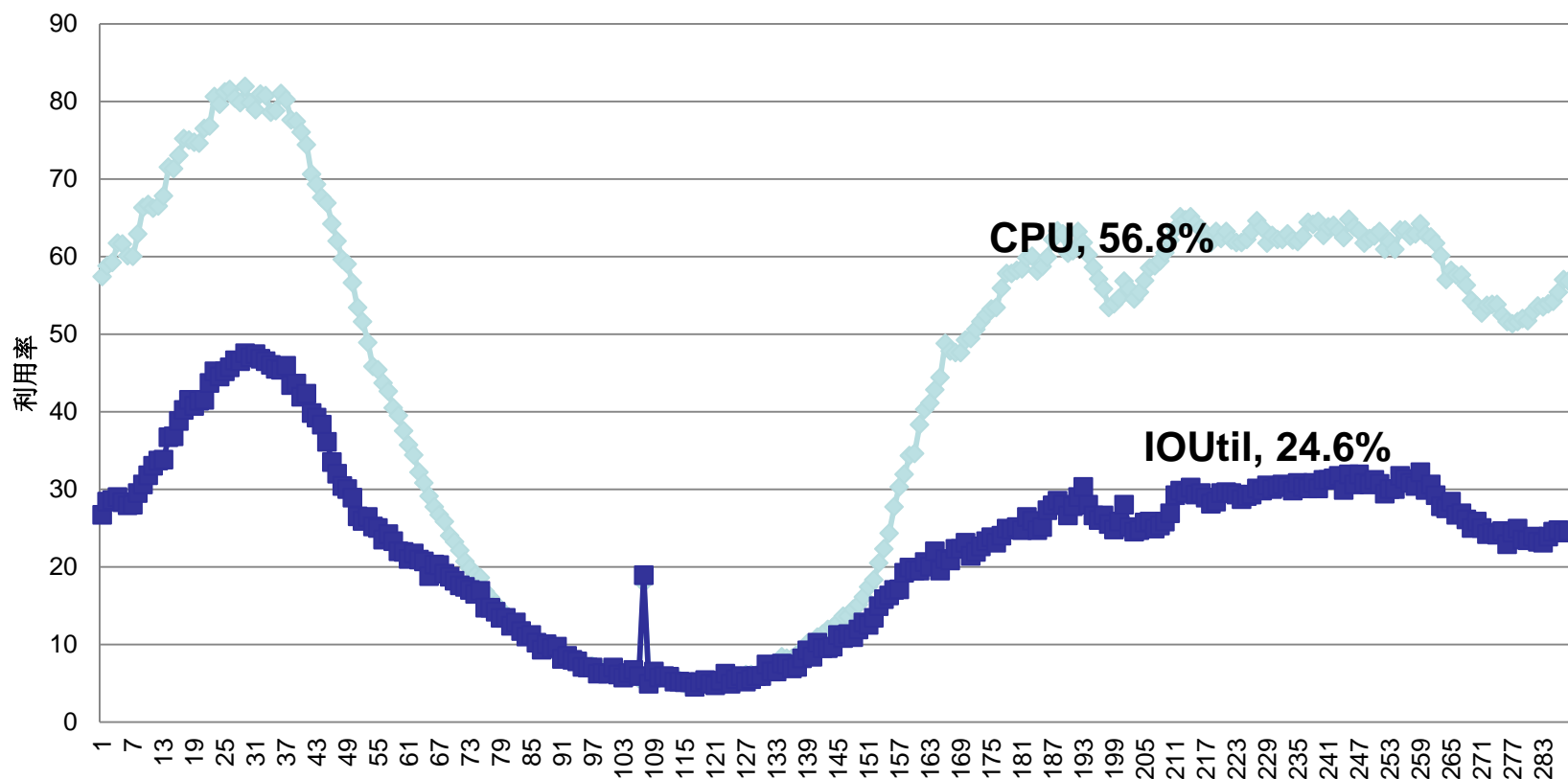
2U*2 低功耗服务器

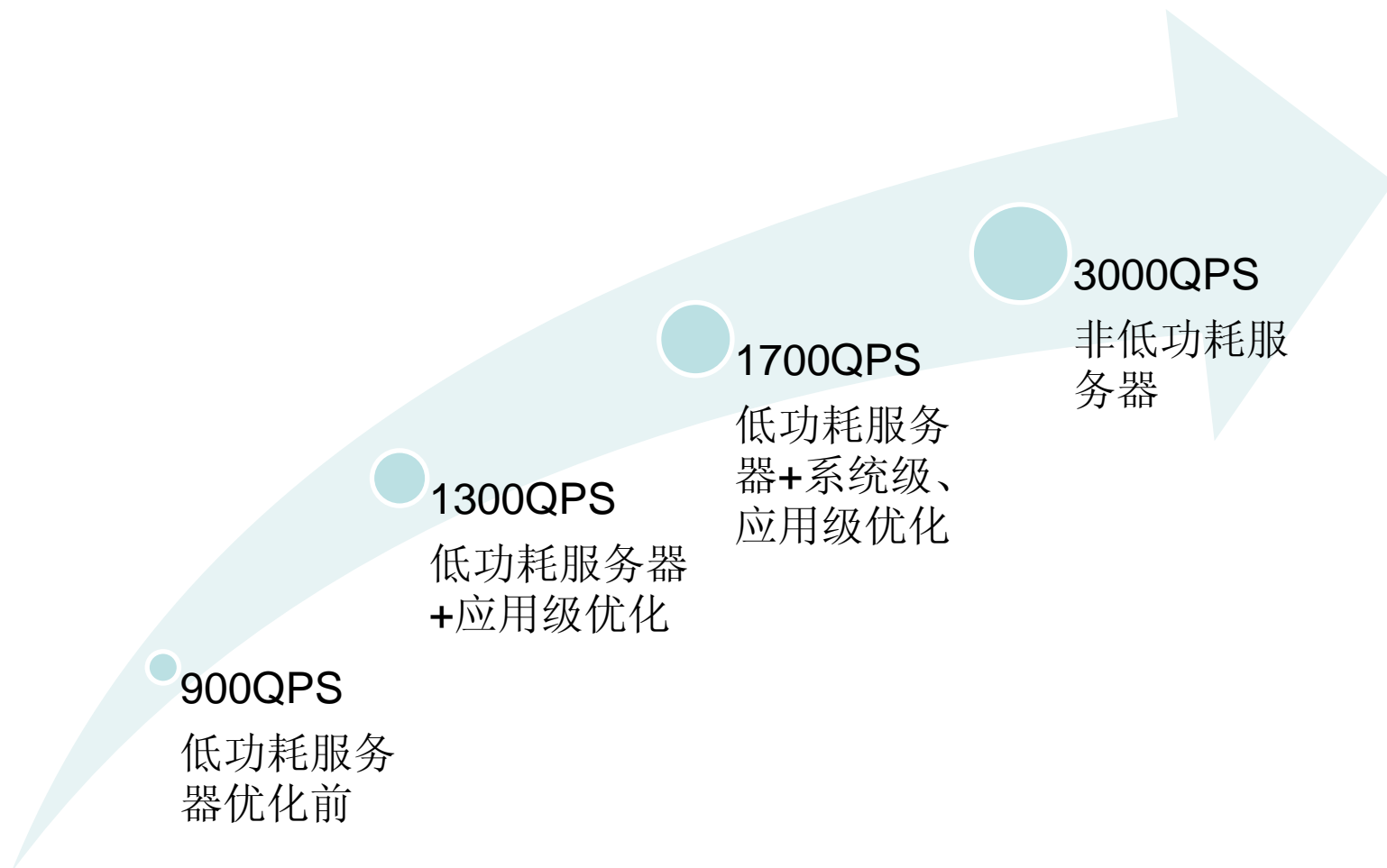


1U 负载均衡服务器



低功耗服务器



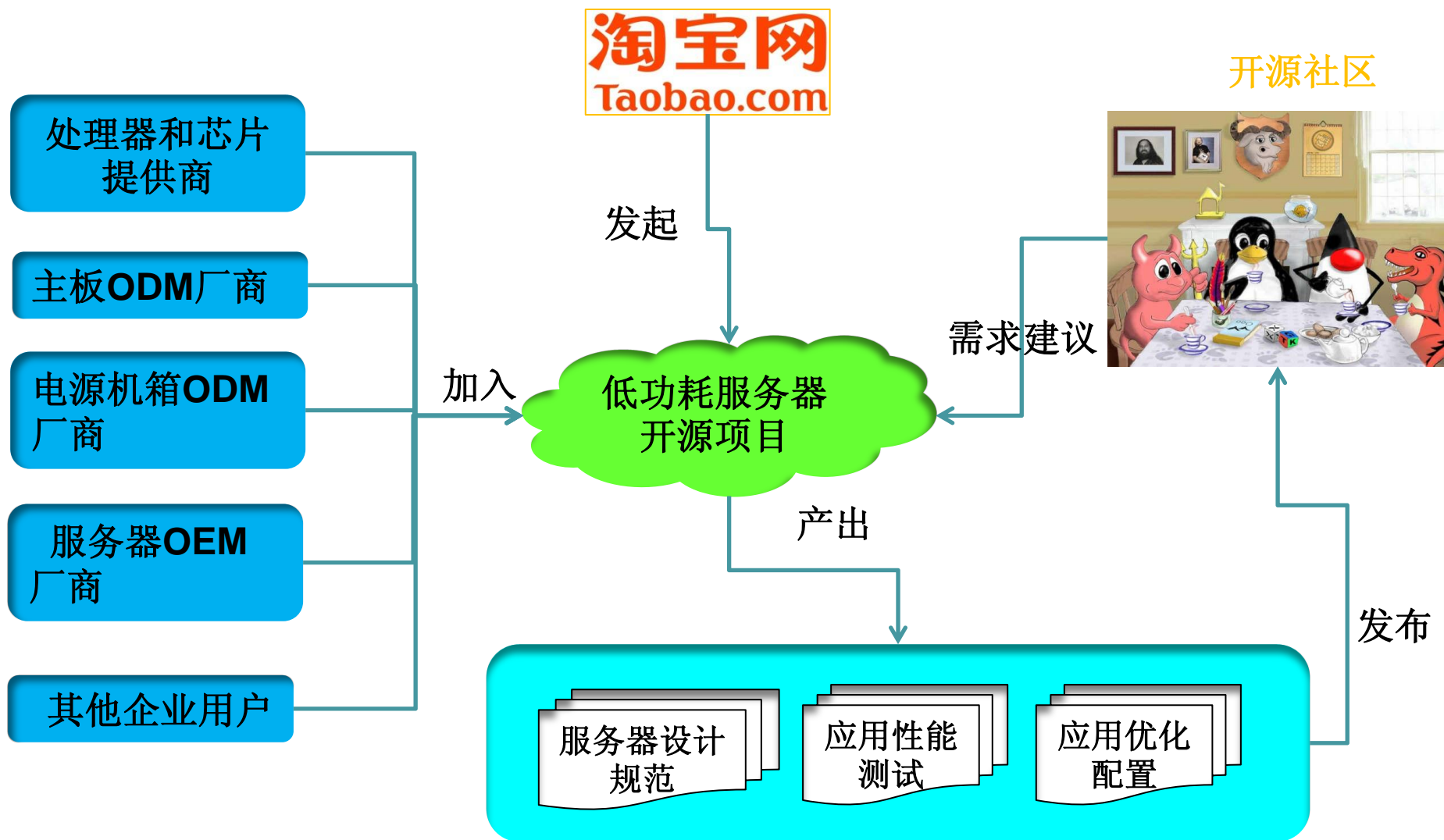




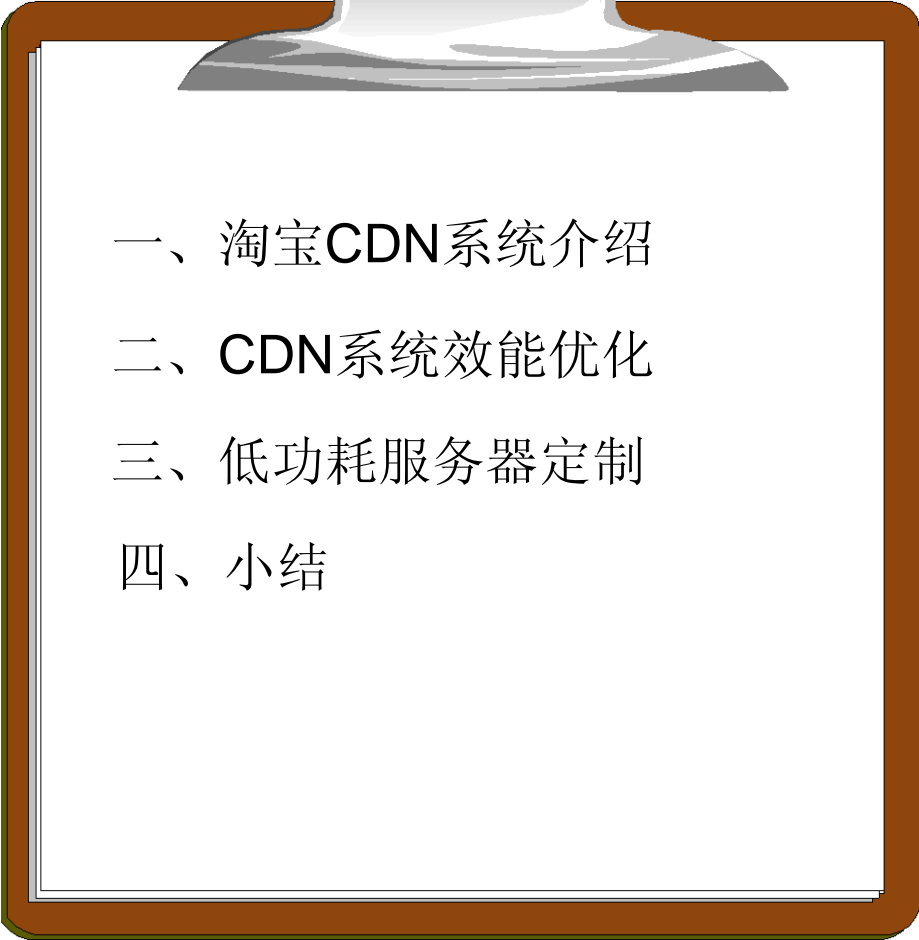
- 功耗优化
- 成本优化
- 性能优化
- 定制方案开源



低功耗项目的开源策略





- 
- 一、淘宝CDN系统介绍
 - 二、CDN系统效能优化
 - 三、低功耗服务器定制
 - 四、小结





小结

- 速度是网站的根本，**CDN**是优化网站速度的利器
- 系统优化是多层次的，软硬件结合
- 从关注性能到关注效能



谢谢!