

低功耗服务器定制与绿色计算

章文嵩（正明）

阿里集团核心系统部

2011.12.6

Velocity China 2011

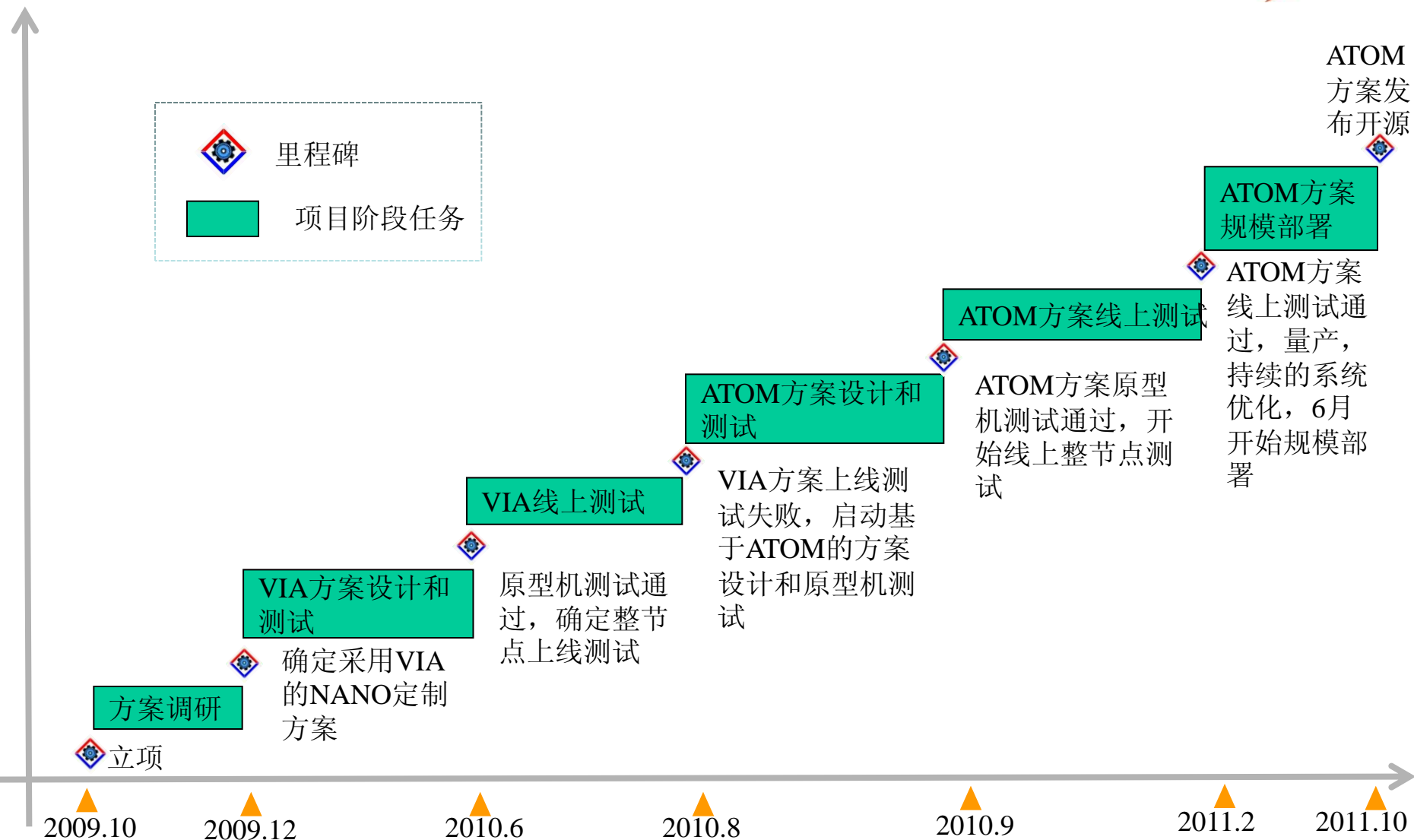


- 一、低功耗服务器项目
- 二、淘宝CDN系统
- 三、低功耗服务器应用
- 四、绿色计算项目
- 五、小结

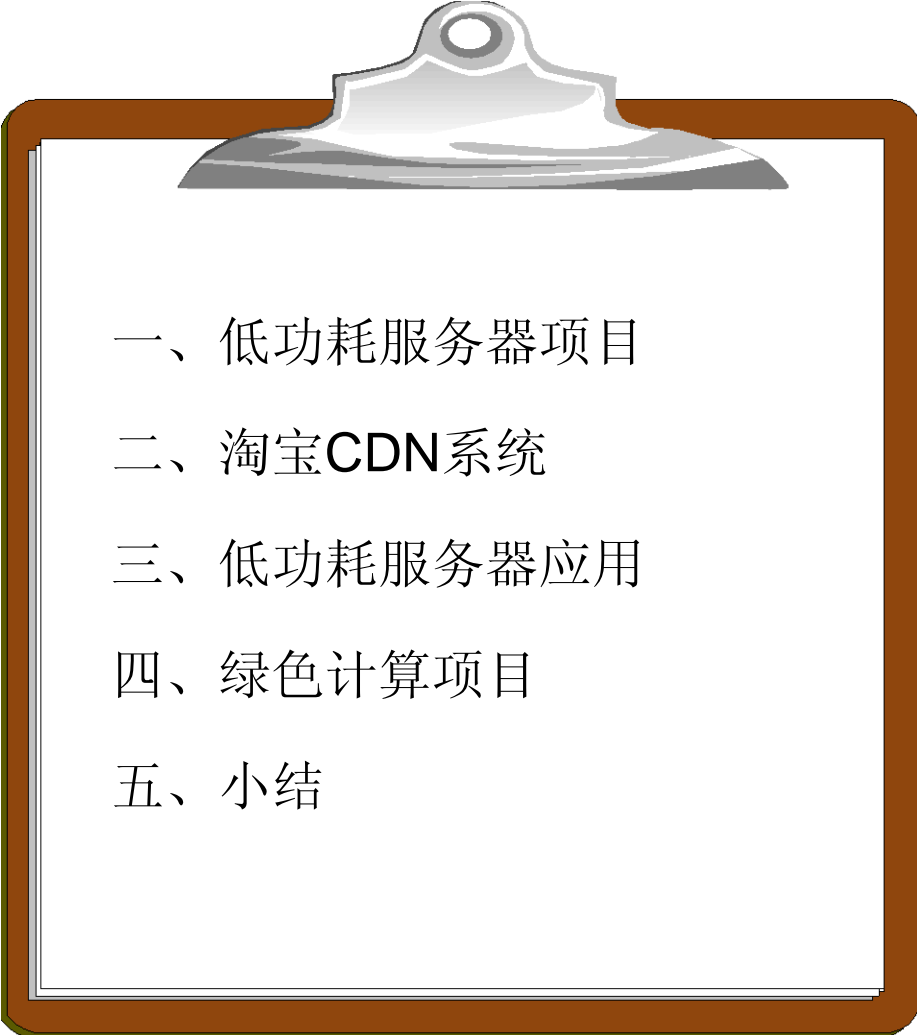
- 低功耗的CPU来做云是非常有前景的，尤其对于数据密集型的应用
 - 2008年基于ARM下载盒，空载时功耗大概为1W，当CPU 100%跑着，外接USB硬盘也在读写时，功耗大概为9W。
 - FAWN: fundamentally Power-efficient Clusters; HOTOS, May 2009
 - Gordon: Using Flash Memory to Build Fast, Power-efficient Clusters for Data-intensive Applications; March 2009
- 2009年底启动了定制低功耗服务器项目，考虑到迁移的成本，先选择IA架构处理器



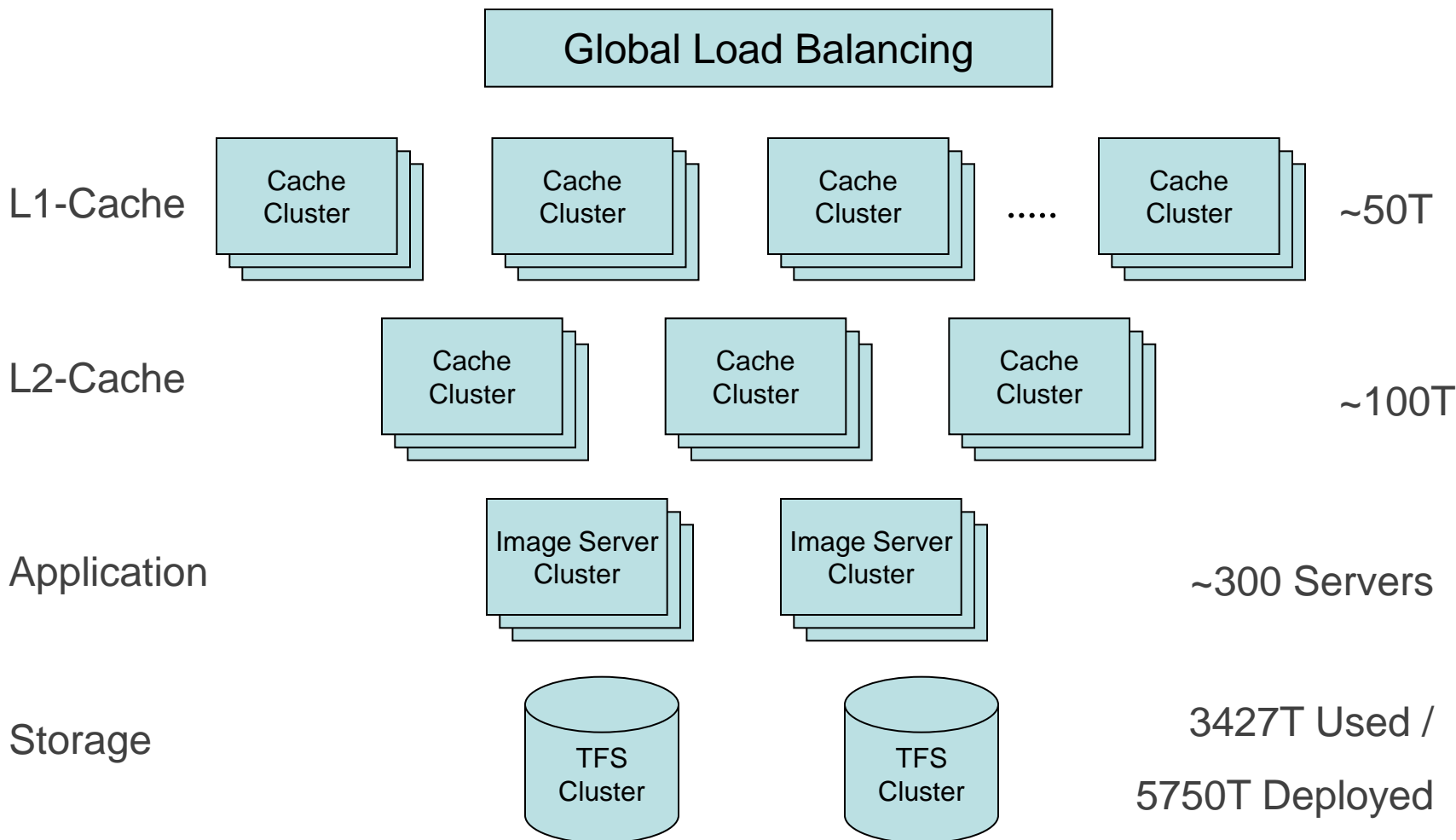
低功耗服务器项目的发展过程



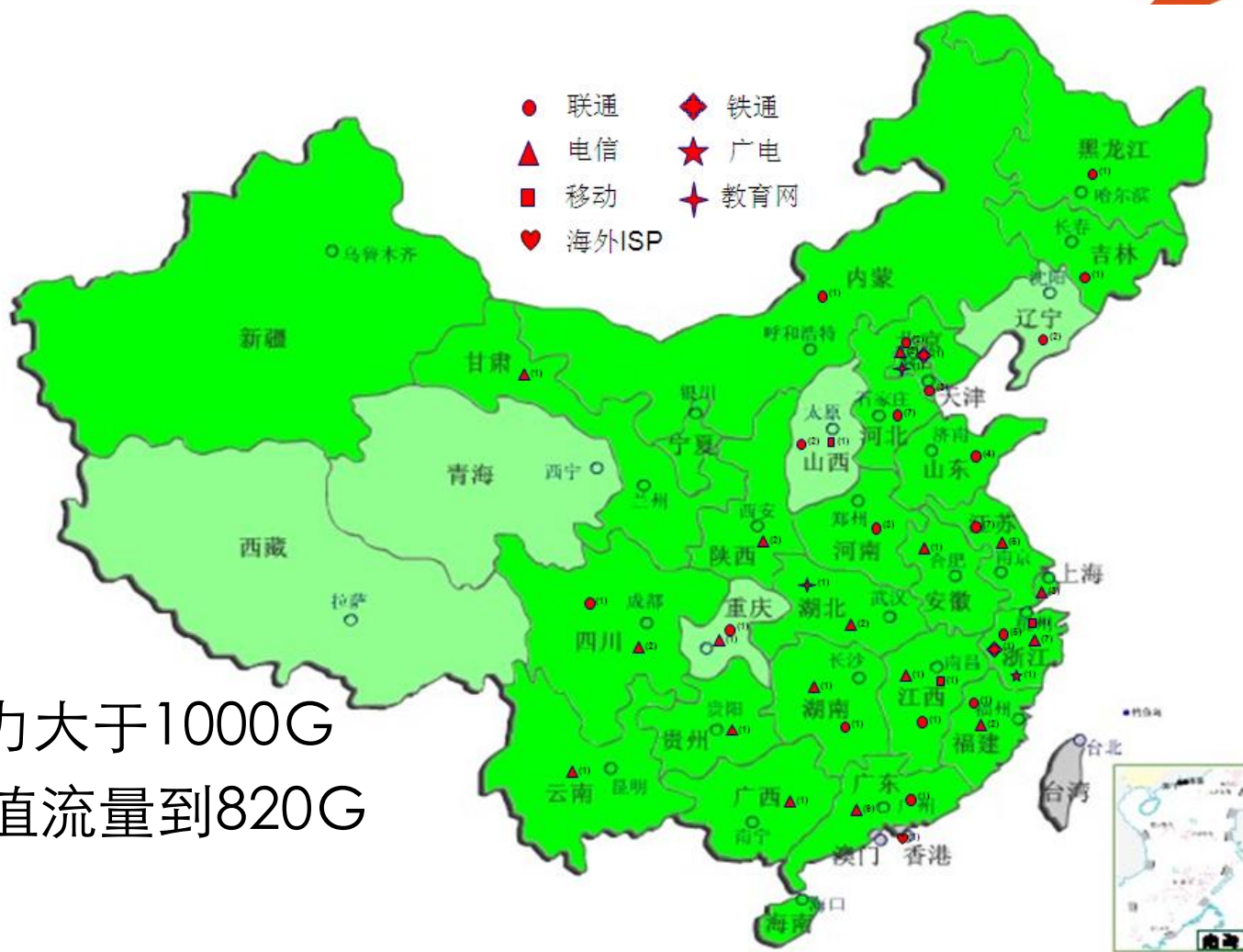


- 
- 一、低功耗服务器项目
 - 二、淘宝CDN系统
 - 三、低功耗服务器应用
 - 四、绿色计算项目
 - 五、小结

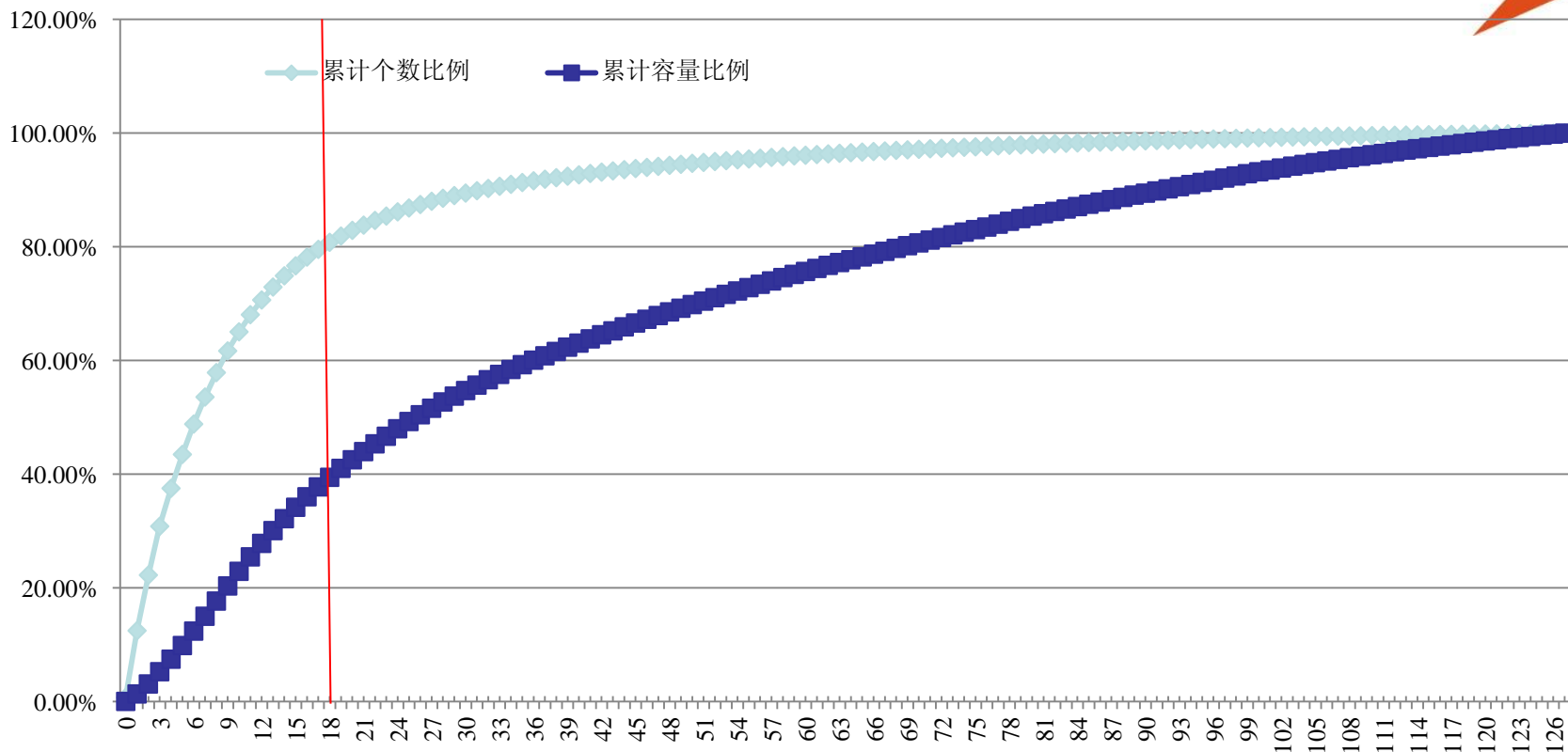
- 淘宝网：网络购物
- 淘宝的愿景：提供电子商务的基础设施服务，打造电子商务的生态圈
- 网络流量排名（Alexa统计）
 - 国际：12~15
 - 国内：3
- 现在每天6000万以上的UV
- 淘宝网站上约有600个应用
- 90%以上的流量用于图片传送



- 覆盖国内所有主流运营商
- 103个节点
- 单节点服务能力 >10Gbps
- 重服务能力大于1000G
- 双11时峰值流量到820G

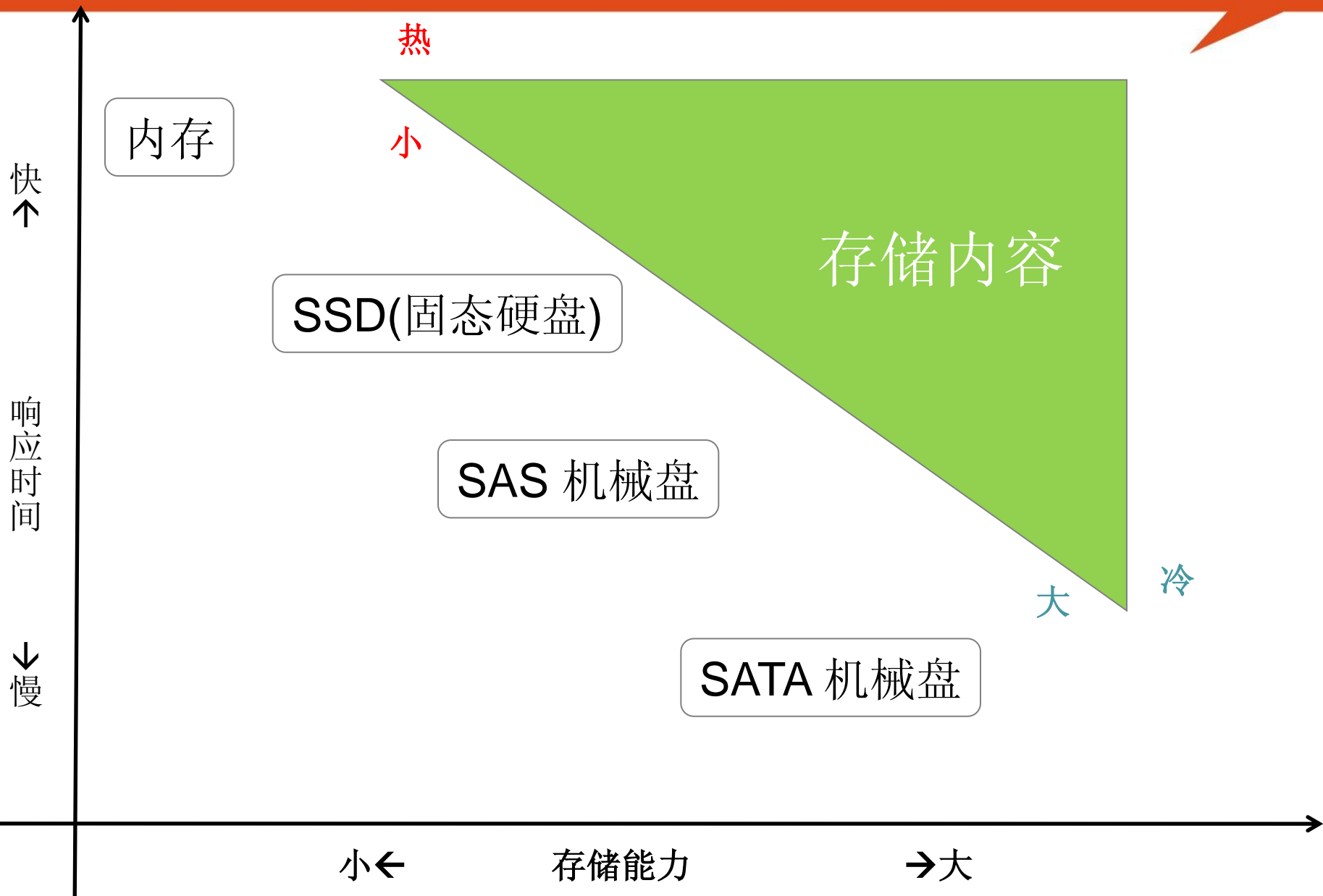


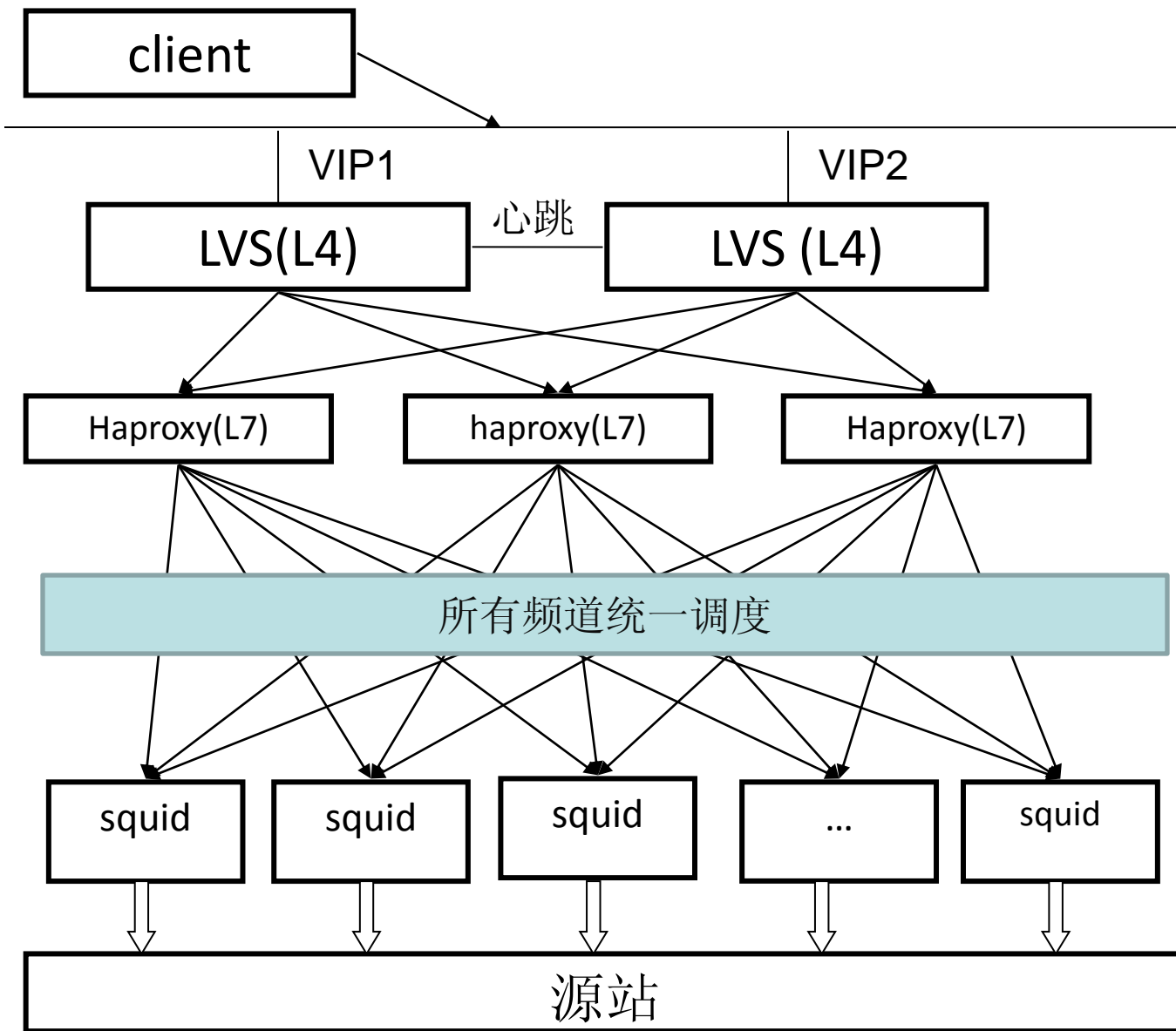
- CDN服务的图片规模
 - 约800T容量的原图 + 800T容量的缩略图
 - 约880亿左右的图片数，平均图片大小约为18K
- CDN功能
 - 支持图片、视频和静态数据；
 - 支持动态页面加速
 - L1-Cache字节命中率为96~98%，L2-Cache的为70~80%，大大改善用户的访问体验，节约回源带宽
 - 针对教育网的CDN部署优化：BGP解决LDNS乱设问题



- 18KB以下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

分层存储机制





- 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

$$migration_weight * \frac{frequency}{size^{migration_power}} ; migration_power \in (0, 1]$$

- 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	7.88	3.42	41.06
sdf	5.45	2.02	118.93	8.00	1281.92	106.25	21.87	1.19	9.37	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

- 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]"
```

...

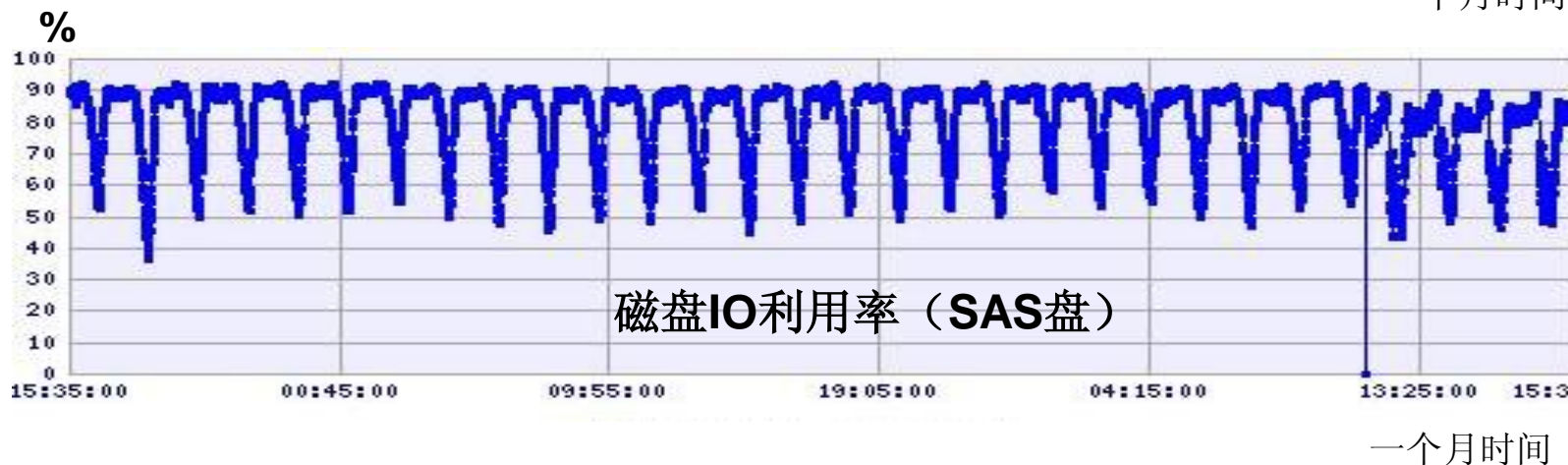
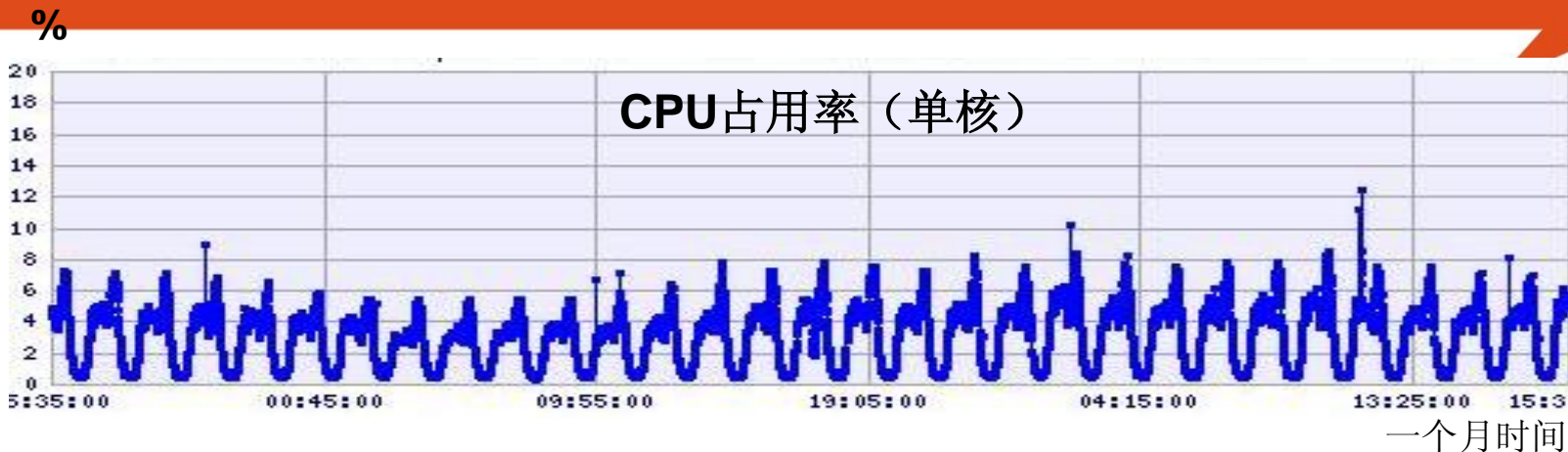
```
avg-cpu:  %user  %nice %system %iowait  %steal   %idle
           3.15   0.00   5.63  11.35   0.00  79.87
```

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	4.39	13.37
sde	0.18	0.03	28.23	1.03	1078.73	206.27	87.81	0.15	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

CDN缓存服务是IO密集型应用

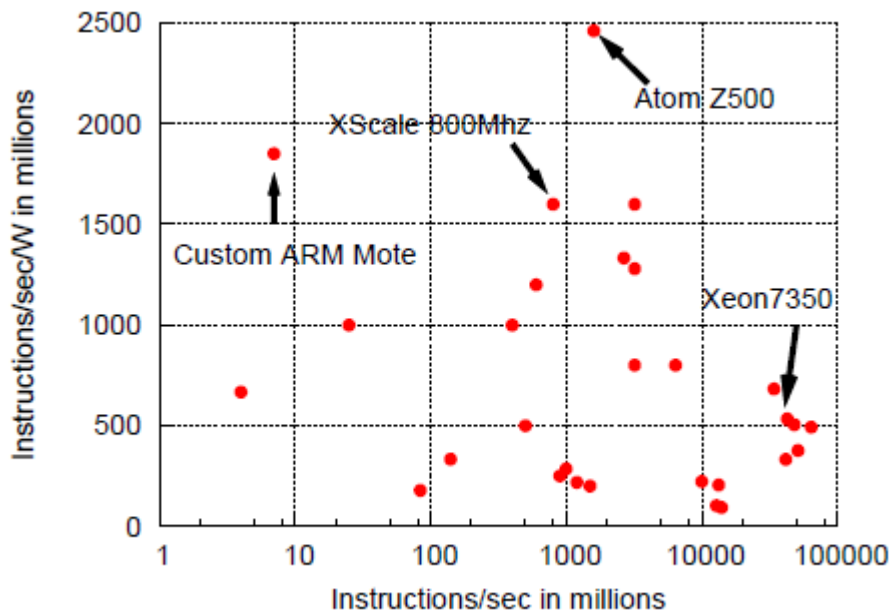


- 缓存应用在一台高性能服务器上测试一个月的数据
- CPU的占用率最大不超过10%，而IO已经接近饱和

- 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

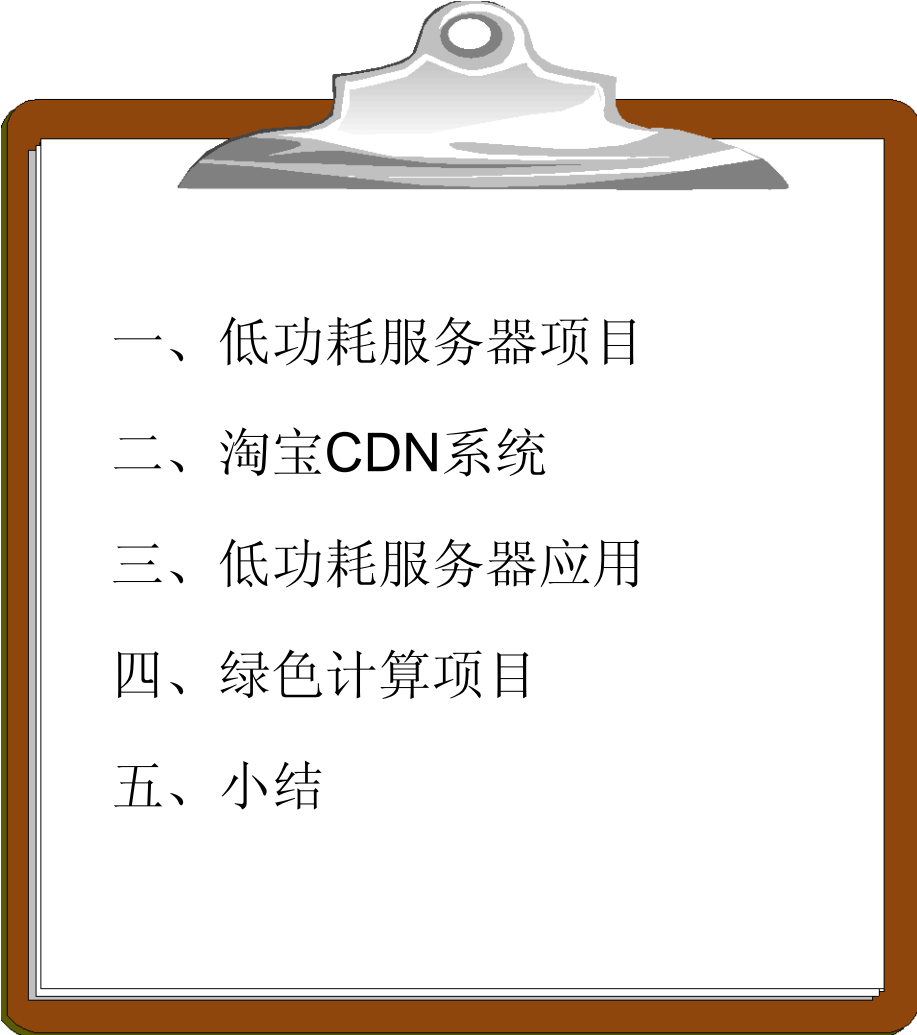
- CPU功耗的增加快于处理速度的增长



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
- 传统CPU峰值功耗高限制了IDC服务器的密度
- 服务器数量大，降低单台服务器故障带来的影响
- 更高密度的存储能力



- 
- 一、低功耗服务器项目
 - 二、淘宝CDN系统
 - 三、低功耗服务器应用
 - 四、绿色计算项目
 - 五、小结

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Node 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)
- 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
 - HDD:
 - SSD 80G * 1,
 - 2.5" SATA 500GB * 2

25W

• 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

	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)
超微2u8方案	15360	80%	~92%	5.5%	2.14	104401	17	14.6
Dell 3U8方案	8580	80%	~91%	5.0%	2.14	64150	17	9.0
HP D320G6方案	10800	80%	~90.8%	5.2%	2.14	77642	17	10.9

	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)
超微2u8方案	25	64	150	2	80	2	2060
Dell 3U8方案	60	22	58	2	80	1	1516
HP D320G6方案	180	10	150	2	80	1	2180

	节点服务能力 (Gps)	总功耗 (瓦)	性价比 (kbps/元)	性耗比 (Mbps/瓦)
超微2u8方案	14.6	2060	30.52	7.27
Dell 3U8方案	9.0	1516	17.78	6.07
HP D320G6方案	10.9	2180	23.09	5.11

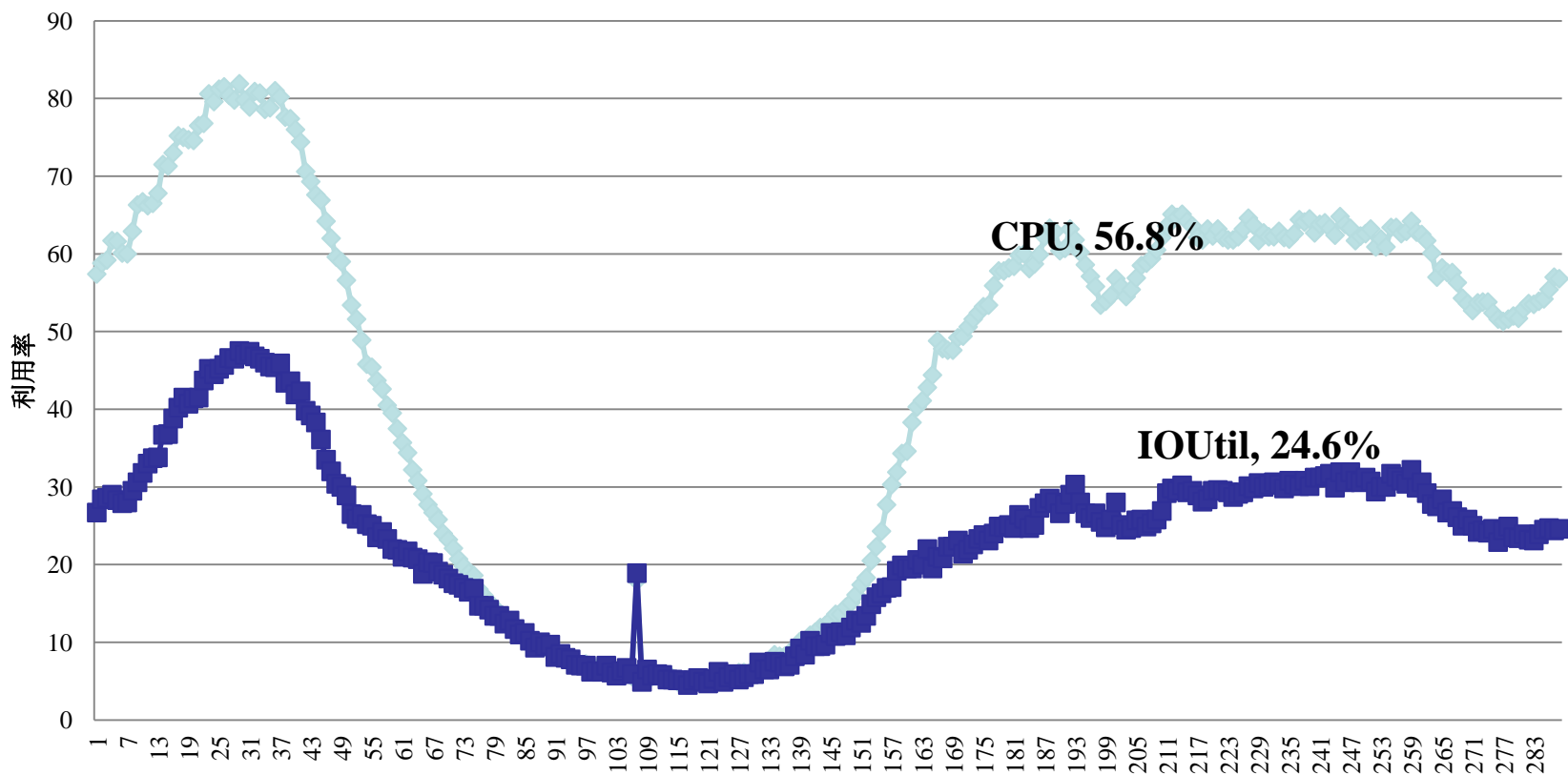
这一款低功耗服务器
针对CDN应用的优化



- 从64台低功耗服务器精简到48台低功耗服务器，即8个2U机框变成6个2U机框
- 交换机精简到一台
- 最大性能约在11.7G

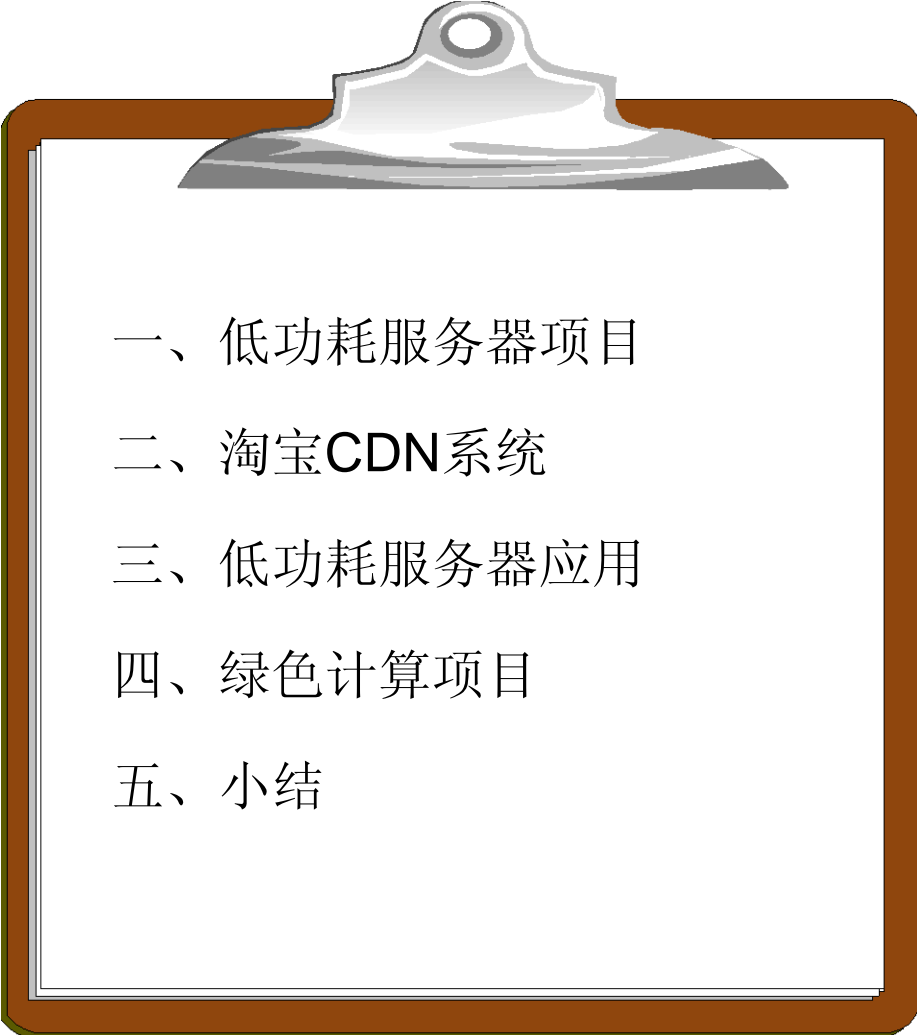
	节点服务能力 (Gps)	总功耗 (瓦)	性价比 (kbps/元)	性耗比 (Mbps/瓦)
6框方案	11.7	1600	33.18	7.51
8框方案	14.6	2060	30.52	7.27

低功耗服务器



- 在CDN中部署了约800台ATOM低功耗服务器，分别为Level-1节点，Level-2节点。
- 经受住了淘宝双11的考验
 - 有些节点跑到9.7G
 - 整体的IO在50%左右
 - 整体squid的rt值都在20ms以下
- CDN应用软件和系统需要持续优化
 - 调整对象迁移的参数，寻找新配置下的平衡点
 - 简化软件，节约CPU消耗，提高性能
- 这款服务器的可运维性还有待提高



- 
- 一、低功耗服务器项目
 - 二、淘宝CDN系统
 - 三、低功耗服务器应用
 - 四、绿色计算项目
 - 五、小结

- 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

- 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。

开源社区



发起

加入

需求建议

开源绿色计算项目

<http://www.greencompute.org/>

产出

发布

处理器和芯片
提供商

主板ODM厂商

电源机箱ODM
厂商

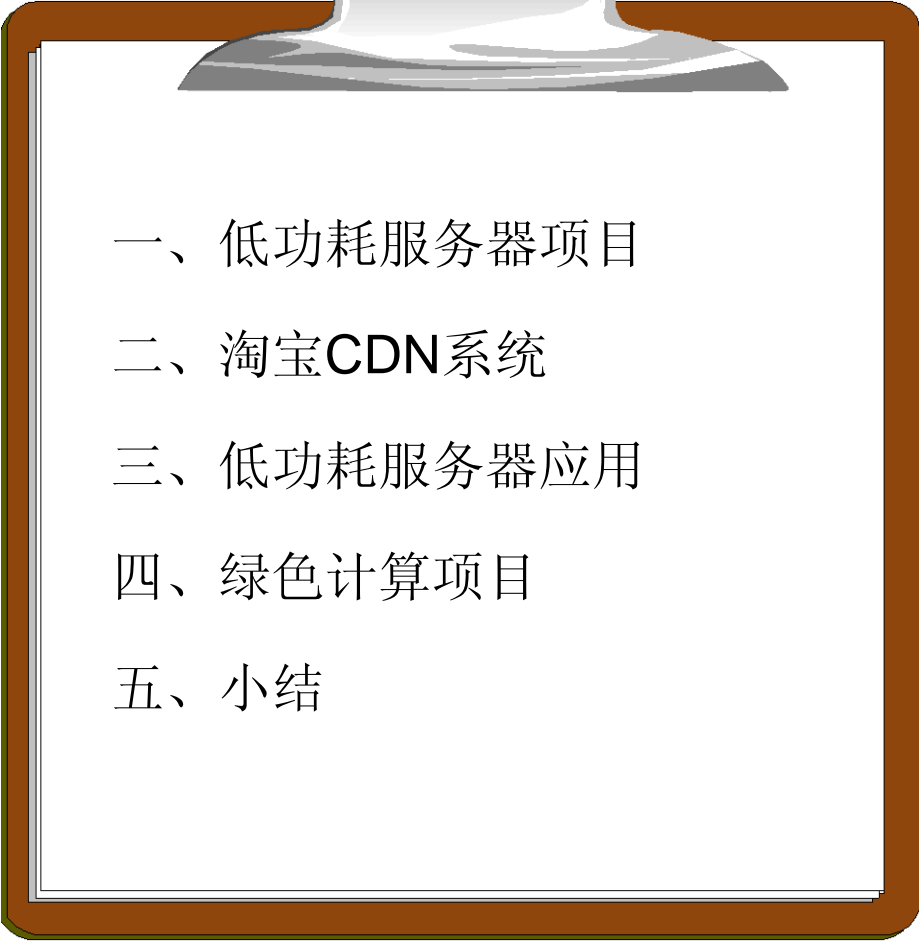
服务器OEM
厂商

其他企业用
户

服务器设计
规范

应用性能
测试

应用优化
配置

- 
- 一、低功耗服务器项目
 - 二、淘宝CDN系统
 - 三、低功耗服务器应用
 - 四、绿色计算项目
 - 五、小结



- 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

- 持续优化低功耗的CDN系统
- 关注低功耗处理器：ATOM ECC支持，低功耗 Sandy Bridge，ARM等
- 定制和评估新的低功耗服务器
- 拓展绿色CDN以外可能的数据密集型应用
 - 存储系统
 - Memory Cache
 - 海量数据处理等
- 欢迎大家加入到绿色计算的行列!

Q & A
谢谢!



<http://www.greencompute.org>