40G 的挑战

2012 年 cdn 架构

赵永明（永豪）

前 言

2011 年我们的 CDN 面临着 860G 的压力，我们坚持住了． 2012 年我们面临流量再翻番的挑战，如何能够支撑住 CDN 系统的 100+% 增幅？如何能够从软件硬件全方位进一步提升能力？ 10 台 cache 真能负载 40G 流量吗？软件硬件极限都在哪里？

提 纲

1. 我们的 CDN
2. 2011 我们怎么过的
3. 2012 更大的挑战
4. 硬件的发展，摩尔定律的指挥棒
5. 如何设计高性能 cache 系统，如何发挥全部硬件的能力？
6. 能力翻番，同时成本减半，能达到吗？
7. 高效运维，释放生产力
8. 展望 2013

去年我们的 CDN

- 100+ 节点，分布在全国 30 多个省市
- 去年 860Gbps 峰值流量，今年继续翻番涨
- 每节点 30TB 容量
- 节点的命中率 95%
- Lvs + Haproxy + Squid(TS)

今天我们的 CDN

- 30TB 能过保证 93% 命中， 50TB 才到 95%
- 1% 硬盘容量的内存仅能够负载 50% 请求量
- 平均文件大小 18->36kB

2011 我们怎么过的

- 节点数量扩容 50%
- 总负载能力翻番到 1TB
- 经过 6 个月时间准备，刚好撑过双 11 促销

2012 更大的挑战

- 流量继续翻番，我们需要建设 2Tbps 容量
  - 如果按照每个节点平均 10G 建点，我们需要 200 个节点
  - 总的命中率维持 93% ，会有 140G 左右流量回源

2012 更大的挑战

- 相比 10 年底，节点命中率下降 3 个点
- 如何找回这 3 个点？

  再增加一倍的存储？
- 新增节点 30%
- 新增流量容量 100%
- 总体成本下降 10%

2012 更大的挑战

成本

Capex = 节点成本 / 节点实际计费流量

Opex = 节点实际计费流量

- 通常的节点成本 << 一个月带宽 30G 带宽成本

Cache 的硬件系统

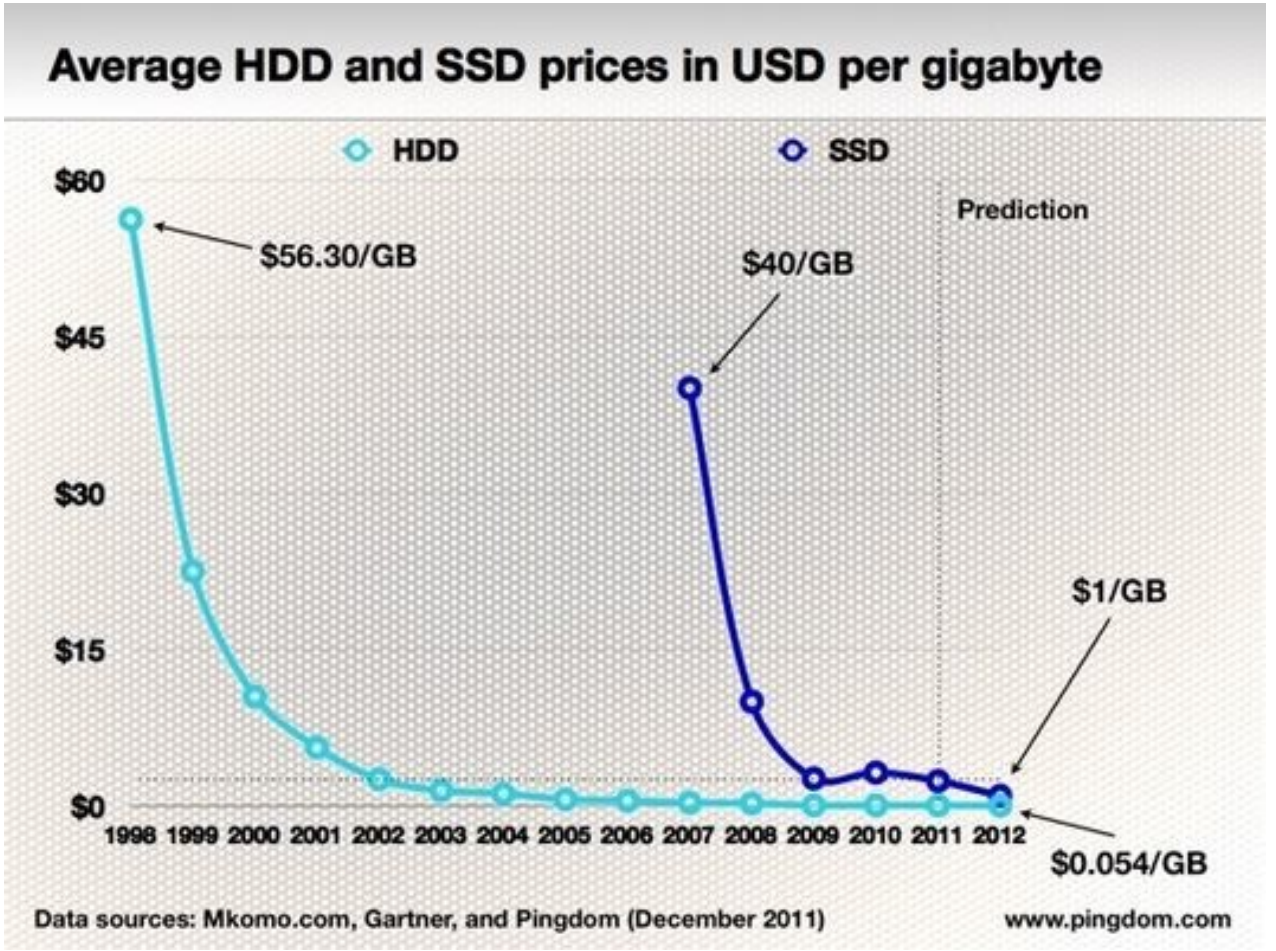
- CPU - 决定了负载流量的能力
- 内存 – 存放热点数据，能过极大的减少磁盘的负载压力
- 存储 - CDN 系统的关键，命中率的保证
- 网络 – 流量保证

我们选择用性价比最高的

硬件的发展，摩尔定律的指挥棒

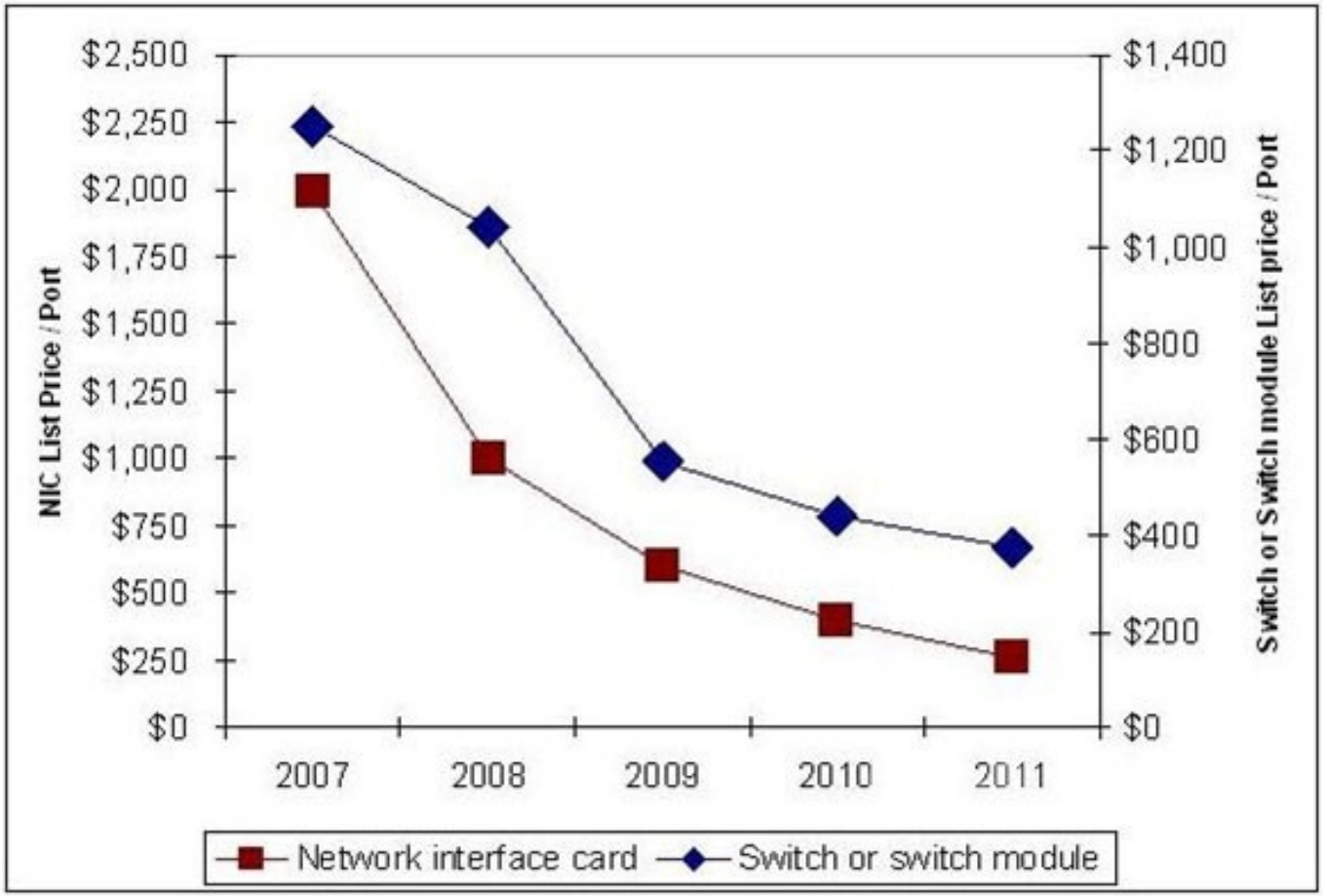
- SSD 价格将降低到 $1/GB
- 淘宝网上价格：

  256G 1500 元



硬件的发展，摩尔定律的指挥棒

- 计算能力飞升， Intel 推出新 E5 主流平台
- 万兆网价格逐步降到 $600 每端口



高性能 cache 系统

- 大容量存储　xxTB
- 高性能吞吐　xxGB
- 牛 X 的软件
- 高速 cpu ？很多核？
- 很大内存？

．．．

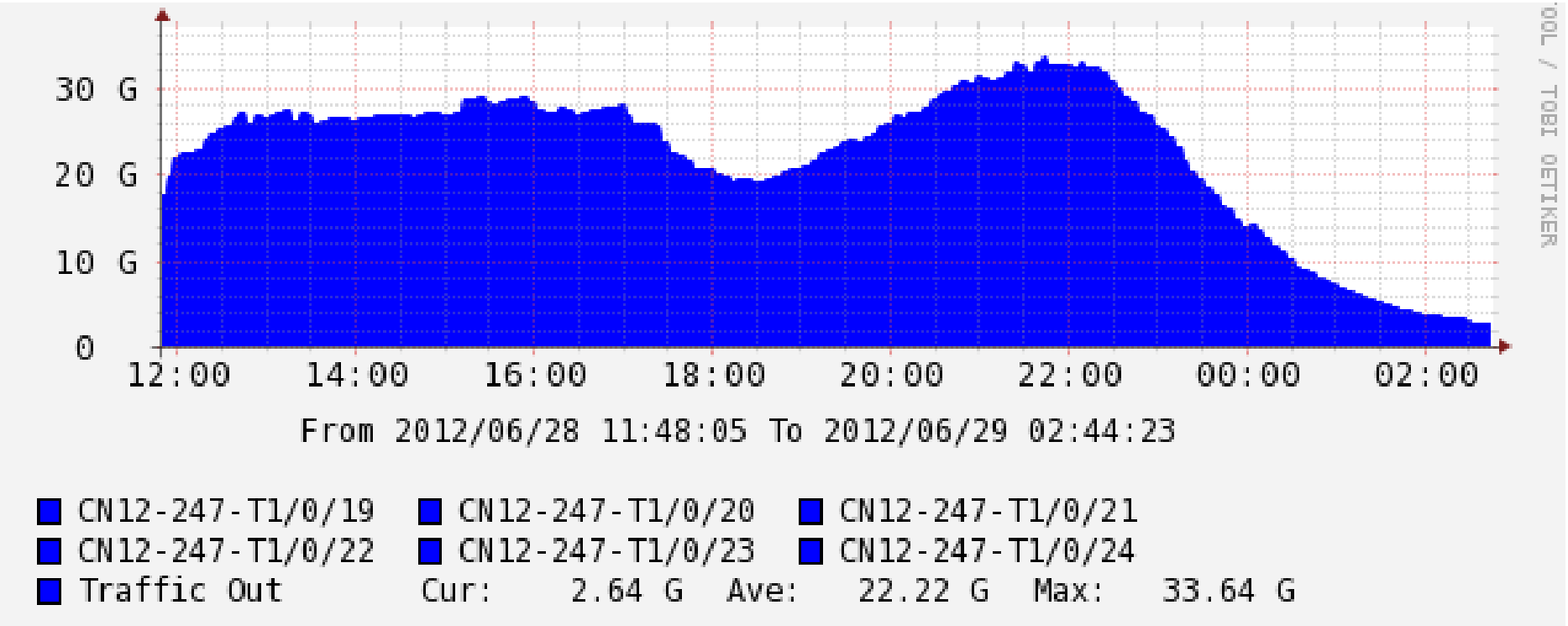
高性能 cache 系统

- 当前的硬件独立性能：
  - SSD 能力： 270GB/s 吞吐， >>6000 IOPS
  - CPU 计算： 1gbps/core
  - 网卡： 10gpbs （还可以 bonding ）
  - 内存，性能是问题吗？

高性能 cache 系统

- 我们配置 16core 8×600G 万兆网卡能跑出多少流量来？
- TS 能：
  - 实验室单机记录： 16gbps ， 112kqps
  - 线上试验记录： 7.7gbps （全硬盘命中）
  - 实际运行记录： 3.4gbps ， 17kqps（配 haproyx)

高性能 cache 系统


30 G
20 G
10 G
0
12:00
14:00
16:00
18:00
20:00
22:00
00:00
02:00
From 2012/06/28 11:48:05 To 2012/06/29 02:44:23
CN12-247-T1/0/19
CN12-247-T1/0/20
CN12-247-T1/0/21
CN12-247-T1/0/22
CN12-247-T1/0/23
CN12-247-T1/0/24
Traffic Out
Cur: 2.64 G
Ave: 22.22 G
Max: 33.64 G
TOBI OETIKER / TOOL

目标：能力翻番，同时成本减半

- 能力的翻番借助于：
  - 稍好的硬件投入
  - 性能瓶颈的释放
- 成本的控制：
  - 大流量节点的大包买
  - 硬件投入的增长 < 性能的提升速度

稳定高效才能释放生产力

- TS 的大规模部署安全性
  - 使用 Raw Disk 模式减少磁盘，坏盘不怕
  - 快速重启，快速服务能力， x 秒级即可提供服务
- TS 的稳定性
  - 自身健康监控（磁盘，内存，服务）
  - 出问题会记录 stack
  - 多种报警机制（可以脚本定制）

展望 2013

- 更大的集群， 10 台 - ＞ 30 台？
- 更多的硬盘， 8 块 4.8T - ＞ 24 块 48T ？
- 更大的流量， 4G / 台 - ＞ 8G / 台 ？

淘宝 CDN 架构组成员

- 永豪
- 阔台
- 陶锐

- 莫涵

欢迎加入这个充满挑战的团队！