

Richbox系统架构演化

作者：张勇(SuperMan)

什么是Richbox?

- A. 推广
- B. NBA Box
- C. 搜索结果
- D. 新闻BOX
- E. 搜索历史
- F. 广告
- G. 还在搜
- H. 其它..

网页 图片 视频 音乐 搜吧 问问 博客 新闻 更多>

SOSO搜搜 nba 搜搜

搜索到约15,700,000项结果, 用时0.001秒

> 网页结果

图片

视频

博客

新闻

> 全部时间

一天内

一周内

一月内

一年内

自定义日期范围

> 全部来源

搜搜

tom

搜狐

搜索历史 - 清除

nba

天气

北京天气

600580

600518

拍拍网2010热销NBA单品排行榜 腾讯搜索推广

腾讯拍拍网, 精选畅销NBA单品, 品种丰富, 质优价廉!

www.papai.com

NBA - 添加到个人中心 - QQ提醒

客队	比分	主队	比赛进程	文字直播	转播频道
奇才	83 - 112	魔术	已结束	暂无直播	
太阳	110 - 94	爵士	已结束	暂无直播	

[前线报道](#) [明日赛程](#) [球队排名](#) [球员统计](#)

NBA频道

NBA中国官方网站, 提供NBA每日赛事直播、丰富的视频集锦、海量的新闻快讯、权威的官方消息、丰富的赛事图片、最新的姚明火箭新闻、准确快速的NBA数据、专业的NBA知识等。

nba.tom.com/2010-10-29 - 网页快照 - 预览

.....

nba 搜搜新闻

[央视NBA直播关键时刻切信号 领导拍桌大骂](#) 大众网 1小时前

*北京时间27日, 火箭与湖人的首战最后2.4秒未在CCTV5中播出, 频道总监江和平接受采访时承认存在失误, 并表示会在央视网上登载一则道歉声明。...

[碧昂斯亮相NBA赛场 有意模仿迈克尔·杰克逊](#) 网易 2小时前

[因太脏洗衣机推荐 NBA直播安静看](#) 腾讯网 4小时前

NBA 2010-11赛季 网易体育

网易nba是报道赛事最快, 报道新闻最专业的网站, 除了掌握第一手nba资讯以外, 内容还囊括了大量姚明新闻、易建联新闻、nba赛事比分、nba赛事直播、nba视频、nba图片、...

sports.163.com/nba 2010-10-26 - 网页快照 - 预览

1 2 3 4 5 6 7 8 9 10 11 下一页>

相关搜索: [nba中文网](#) [nba直播](#) [nba官网](#) [nba总决赛第七场视频](#) [nba搜狐](#)

[虎扑nba](#) [nba赛程表](#) [nba视频](#) [www.nba.com](#) [nba视频直播](#)

腾讯搜索推广

[睡觉减肥 疯狂掉肉 值](#)

不打针不吃药 贴哪里哪里方便

尚品用掉肥油 秀出好身材

www.lago68.cn

[创业加盟, 选择一次](#)

观直播吧, 找火爆商机, 做一

免费留言咨询, 创业少走弯路

baos.com

网友还在搜

[火箭队](#)

[nba中文网](#)

[湖人](#)

[nba直播](#)

[姚明](#)

[科比](#)

[白羊座](#)

[深圳天气](#)

[湖南卫视](#)

[双鱼座配射手座](#)

刘德华



共2697万张

刘德华 - 搜狗百科

刘德华（1961年9月27日—），生于香港新界，著名演员和歌手，是“四大天王”之一。1981年以全优成绩毕业于TVB艺训班并签约出道，《猎鹰》大红，名列“无线五虎”；1983年主演《神雕侠侣》在香港创纪录。1985年首发专辑，1990年凭专...

刘德华的博客

刘德华的影视作品

电影 电视剧 资讯

共148部>>



风月
6.4分



寒战
7.4分



我知女
5.6分

刘德华的图片 共2697万张>>



QQ2010

SuperMan - [隐身]
又一哥们离开去创业了...

☆ 3 115 8 7

搜QQ好友或群，搜网页信息

我的好友 [6/92]
PHPer [9/67]
Tencent [3/26]

8月22日 农历七月十三

北京 **28°C** (实时天气)
多云 微风

QQ提醒您: 天气

今天	明天	后天
20°C ~ 29°C	22°C ~ 30°C	22°C ~ 30°C

报告错误地理信息 更多城市天气

香港星光大道表彰人员



梁朝伟
忧郁影帝



周润发
香港演技之神



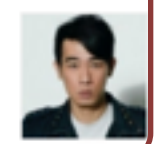
周星驰
无厘头文化领袖



成龙
国际功夫巨星



曾志伟
龙套帝到影帝



SuperMan LV5 个人中心 设置 管理

星座运势

运势首页 好友对碰 《搜搜月历》8月号：揭秘扑克牌里的中国导演生存

今日 | 明日 | 本周 | 本月 | 今年 写运势日志

白羊座 今日运势 ★★

爱情: ★★	健康: ★★★★★	速配: 白羊座
财运: ★★	工作: ★★★★★	幸运色: 绿

运势分析 来自腾讯星座

邓紫棋介入张杰谢娜婚姻 最会炒作的女人 ent.taiwan.cn/

靠绯闻上位的男星：李易峰恋李多海文章不离婚张杰谢... 烟台大众网

曾不被看好的明星夫妻：张杰谢娜没离婚孙俪生二胎 中国日报

13小时前
13小时前
14小时前
1天前

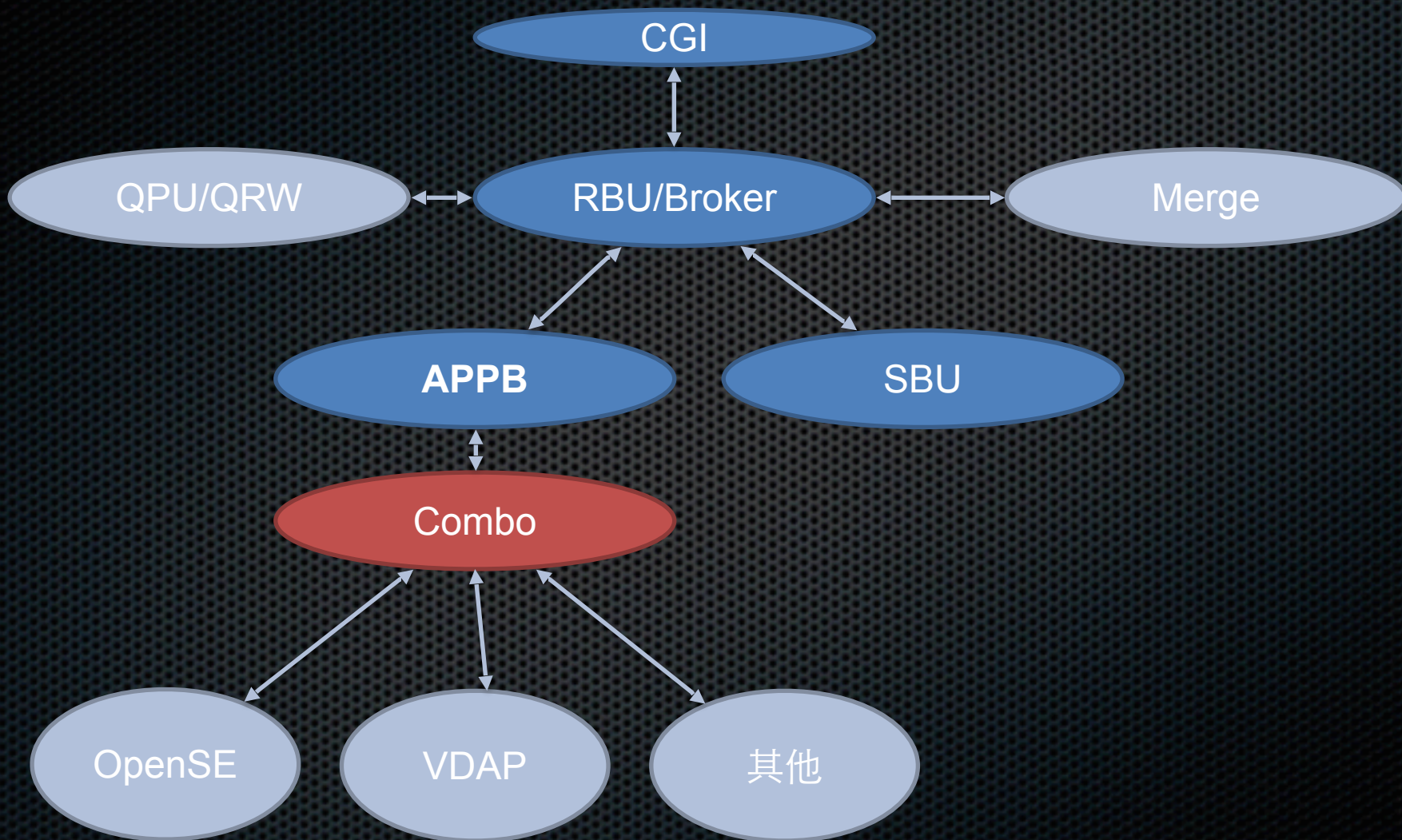
自然搜索结果和Richbox对比

	网页自然搜索	RichBox
后台数据	依照爬虫爬取的	通过如开放平台收录，垂直引擎收录，第三方数据收录的
用户体验	通用的标题，摘要，链接模板	富态化展现，结构化数据的表达，减少用户信息查询的深度。
数据结构	标准化数据（标题；摘要；时间， <i>anchor</i> 等）	多样的结构化数据。
排序方法	BM25 等文本相关性算法	基于数据挖掘的排序
查询词覆盖	覆盖所有查询词	依照重要程度，依次覆盖

如何做好它？

- 分需求分类做好
- 避免引入其他伤害（噪音）
- 扩大召回
- 多意图识别
- solution体系（知识图谱）
- 各种优化召回率、准确度、Ranking...

不过？



我们今天要讲的是？

- ▶ Web 架构的演化过程
- ▶ 数据接入平台到开放平台的演进
- ▶ 调查：
 - 1、Web开发
 - 2、后台开发
 - 3、Devops?

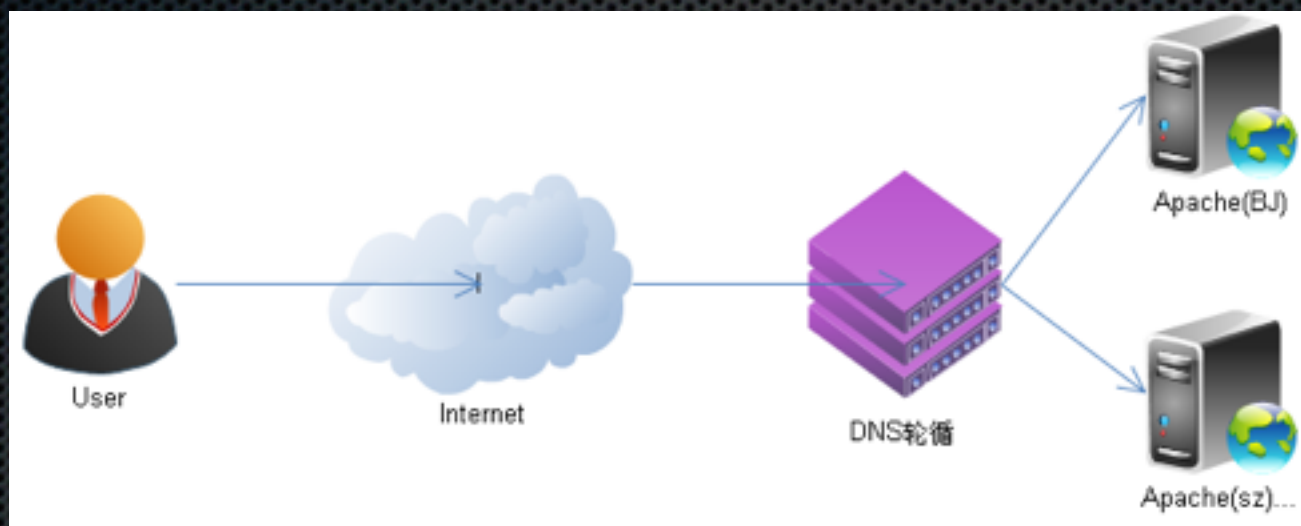
巨人的肩膀



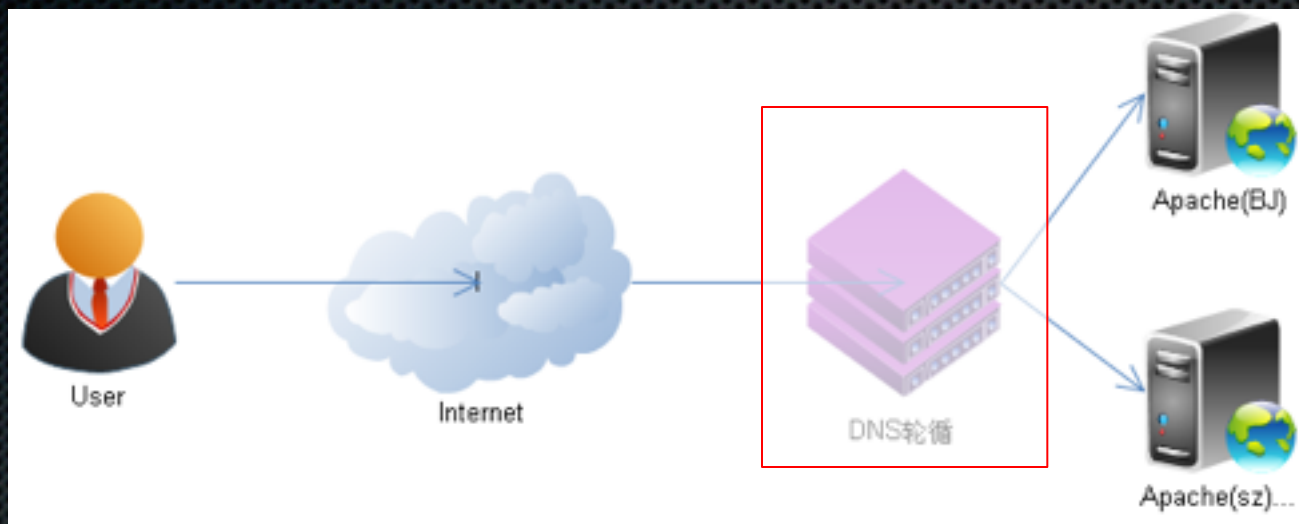
Memcached



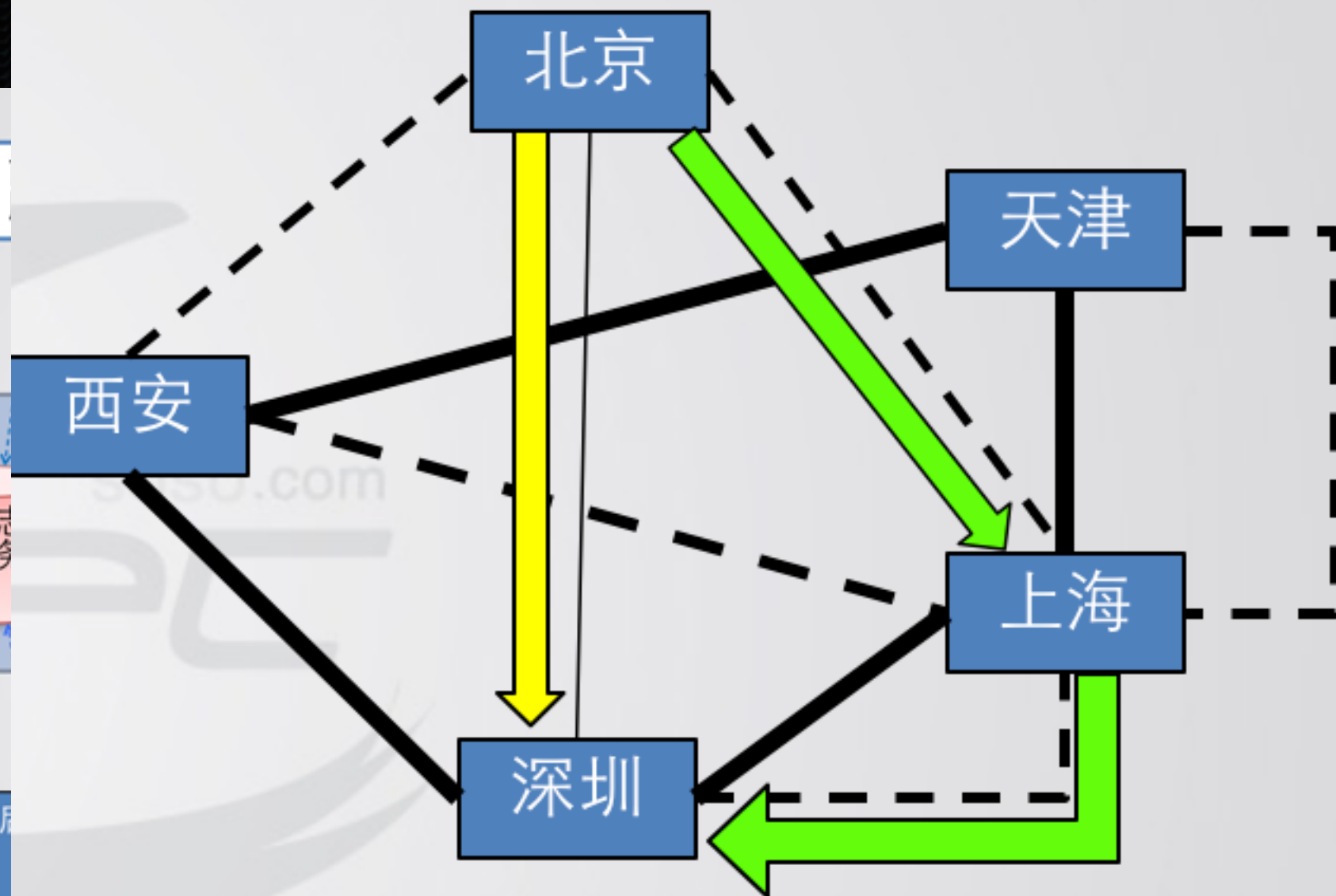
我们是从这里开始的



问题与挑战

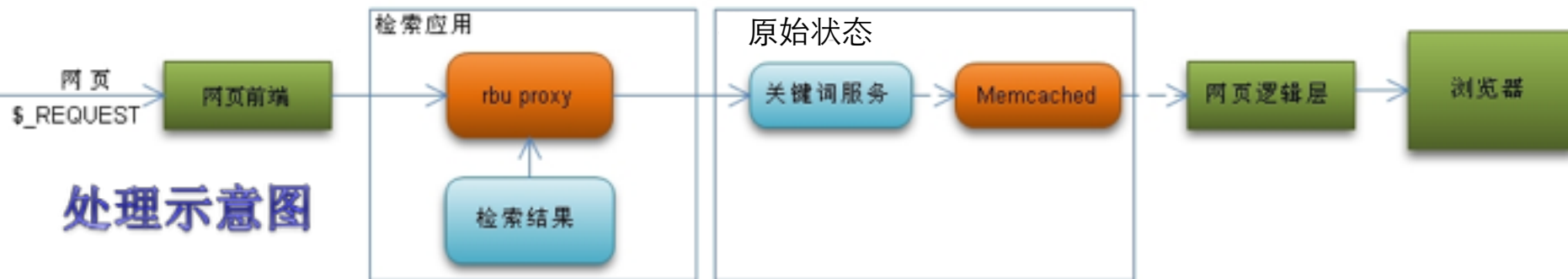


- 用户越来越多
- DNS不准?
- 用户反映访问速度慢
- 大家都用专线?

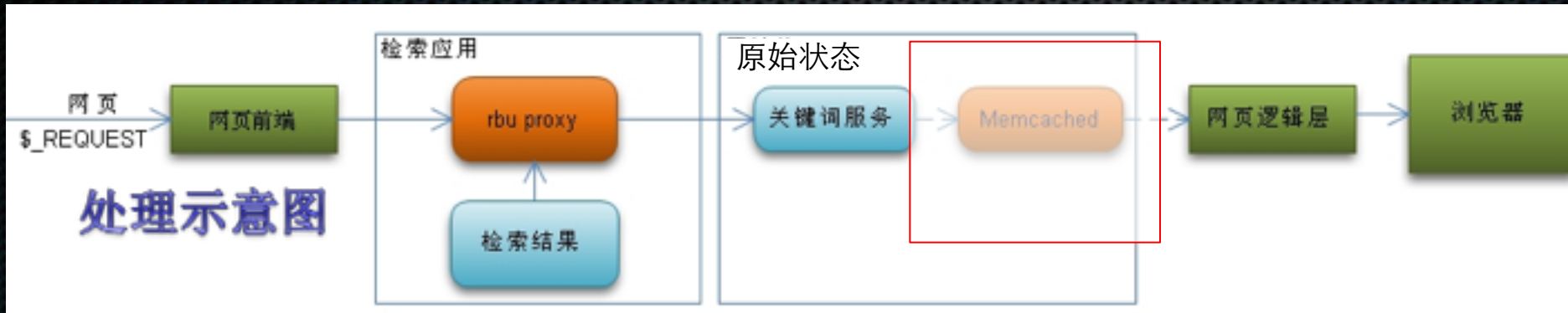


回到Web端

处理示意图

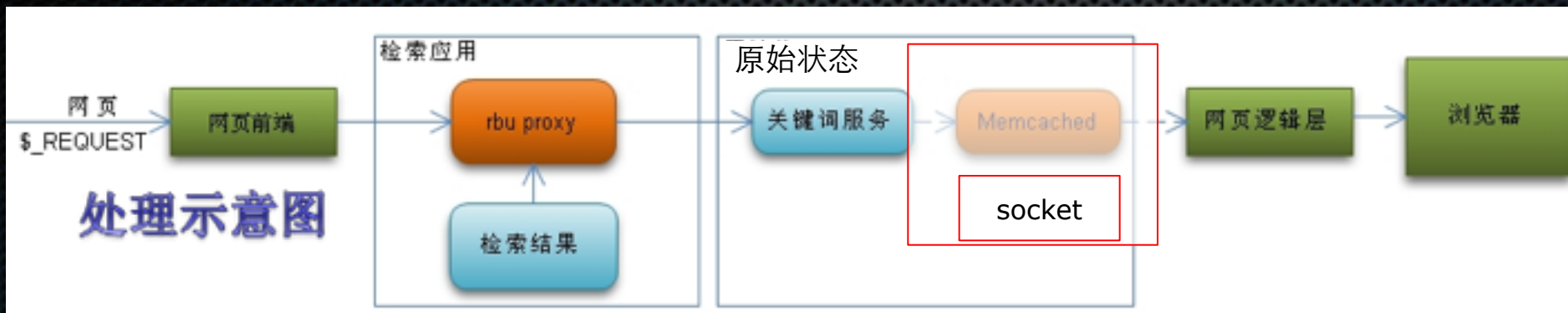


问题与挑战



- ▶ 响应时间经常超过1秒甚至2秒 (server端)
- ▶ Memcache timeout (1s)

解决办法



- ▶ 使用socket+Memecache协议读数据
- ▶ 控制timeout在秒级以内（50ms）
- ▶ Memcached

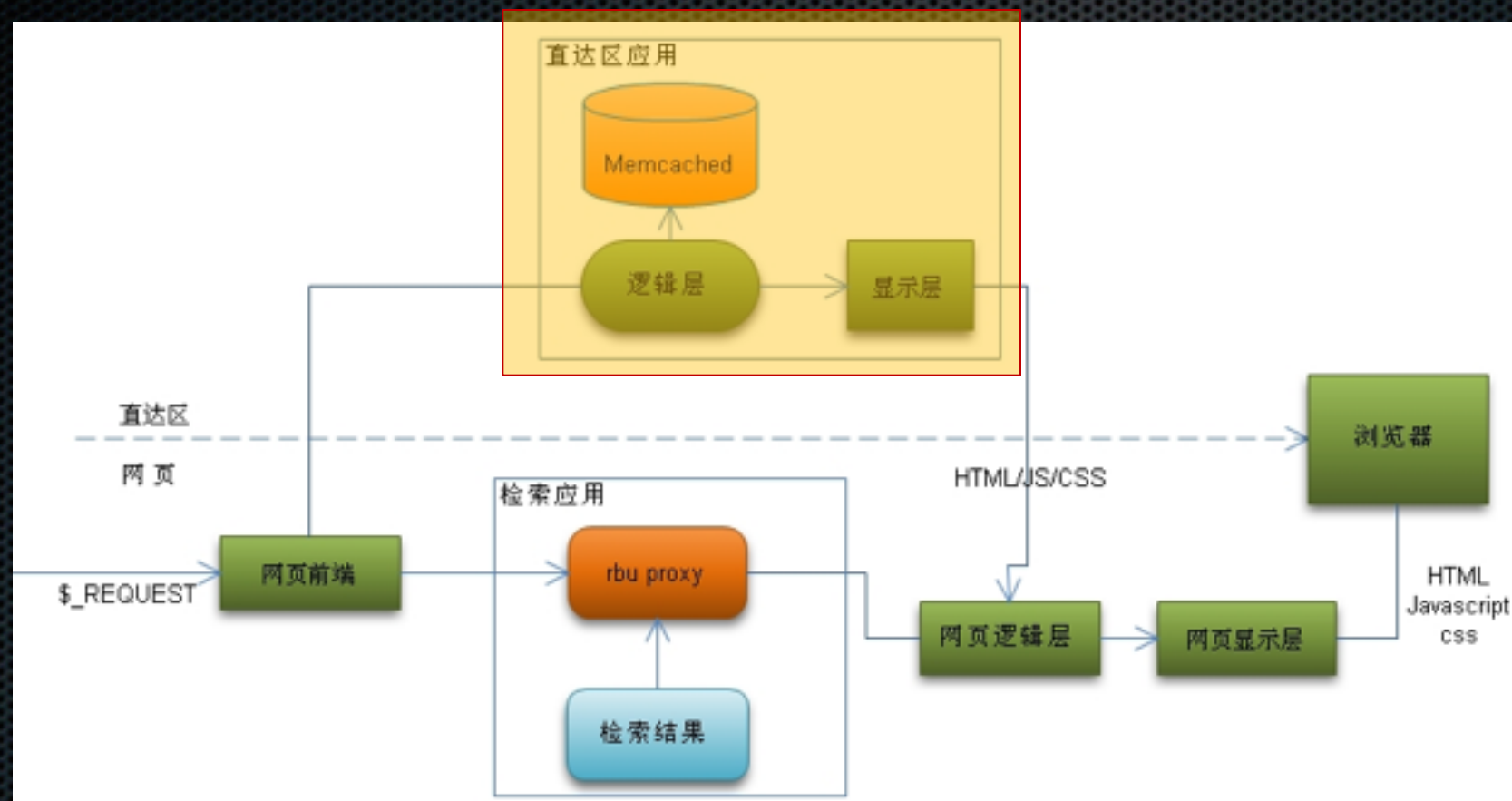
问题与挑战

- ▶ 应用越来越多：天气、股票、NBA。。。。
- ▶ 交互越来越重，越来越“rich”
- ▶ 它已经从Box，华丽丽滴变成了Richbox!!
- ▶ 而请求和渲染却越来越慢了

解决与应对

- ▶ 业务拆分
- ▶ 建立命中模块 (ComboHit)
- ▶ 并行检索与Box请求
- ▶ 合并输出
- ▶ 有损服务

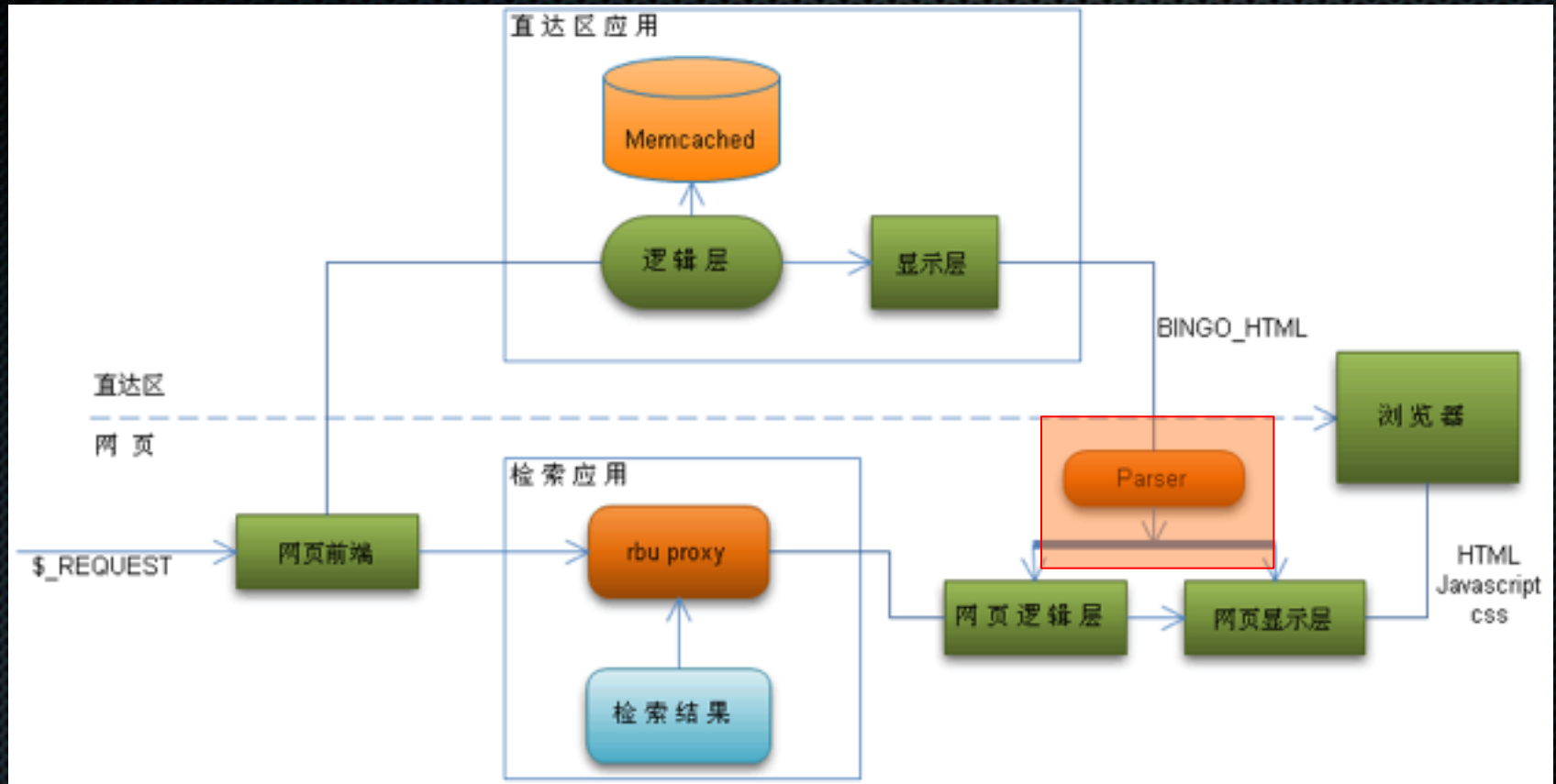
解决与应对



越来越不快乐

- 业务增长
 - 用户
 - Box数量
 - 数据越来越多，内存吃紧
- 团队合作成本高，编码规范缺失
- 脚本管理混乱，复用性低，维护成本高
- 代码冲突：box vs box vs websearch

规划 = 重构+优化



心得体会

- 用户至上
- 毫秒必争
- 不是聪明就可以，要有基础组件支持
- 简单、务实
- 用数据说话

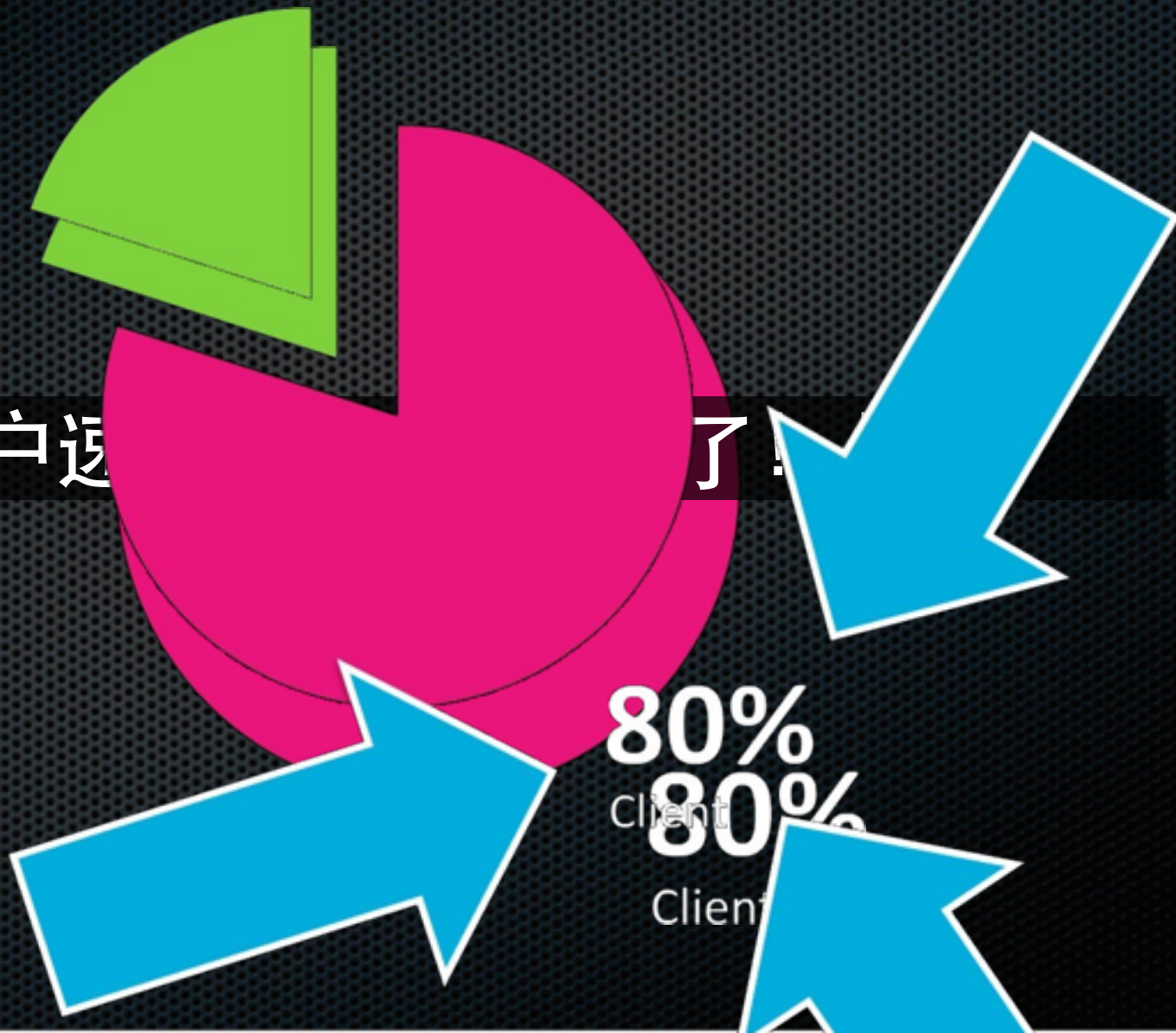
终于可以快乐的玩耍了!

20%
Server
Server

用户选

了!

80%
Client
80%
Client



还是不快乐？

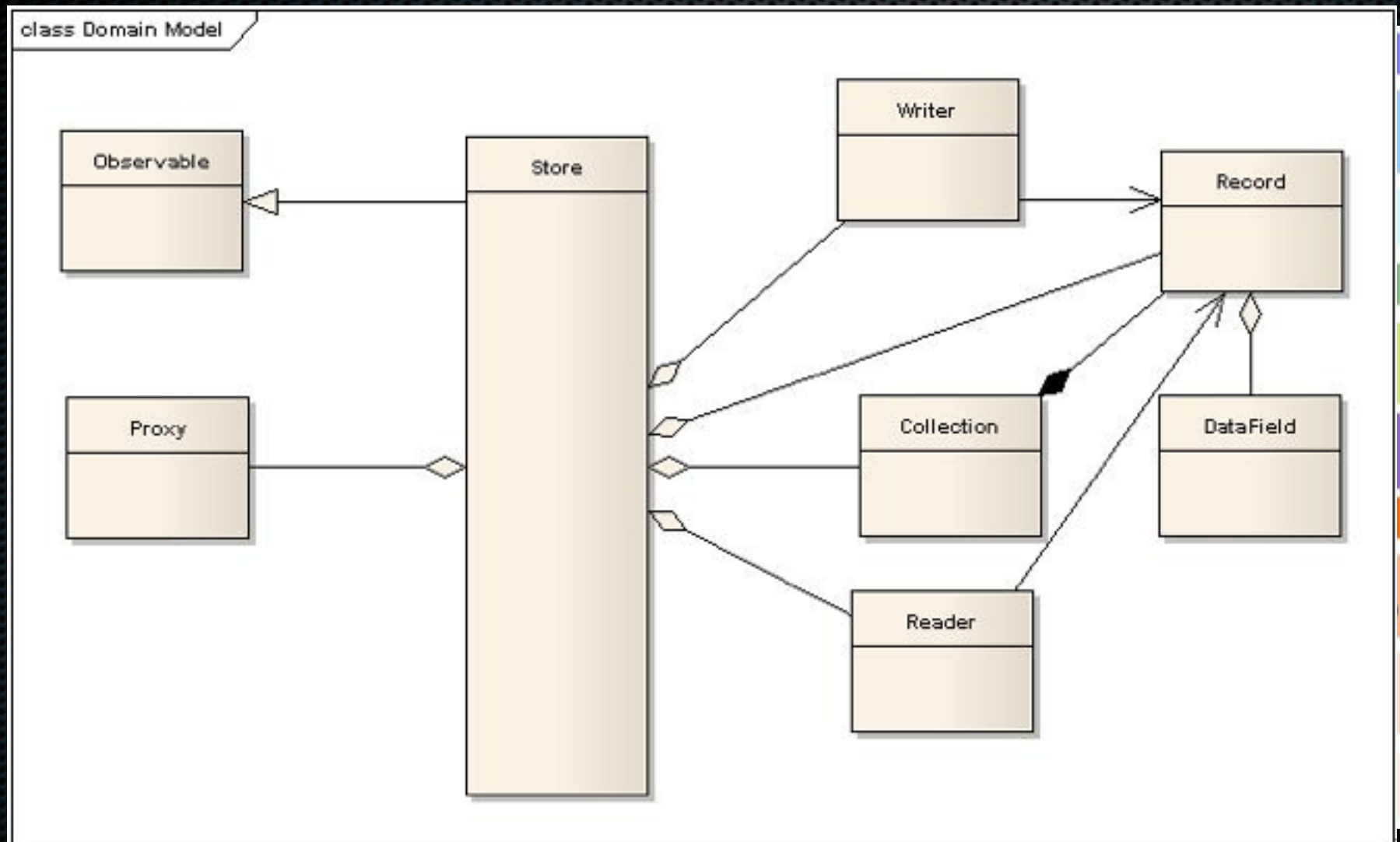


问题与挑战

- Box数量越来越多，数据源越来越多
- 每天陷入了更苦逼的深坑：

正则

接入平台诞生




```
2 $tConfig = array();
```

Array

```
5 * XML/(x)html 节点选择类
```

```
6 *
```

```
7 * @todo 3.2.7.5
```

```
8 * 1.(调整) 深度封装, 使其可以以类似jQuery的形式, 进行级联操作
```

```
9 * (See Updates 1)2.伪类支持, 所有选择器、属性、过滤器、伪类可以以任意顺序组合, 支持形如 div.foo:nth-child(odd)[@foo=bar].bar:first 这样的选择器
```

```
10 * 3.
```

```
11 * Updates:
```

```
12 * 1.完成todo:2 伪类支持, 所有选择器、属性、过滤器、伪类可以以任意顺序组合, 支持形如 div.foo:nth-child(odd)[@foo=bar].bar:first 这样的选择器
```

```
13 * 目前支持的选择器:
```

```
14 * 1.元素选择符
```

```
15 * *
```

```
16 * X
```

```
17 * X Y
```

```
18 * X > Y 或 X/Y
```

```
19 * X + Y
```

```
20 * X ~ Y
```

```
21 * 2.属性选择符
```

```
22 * 使用符号@, 例如div[@foo='bar'] => setWriter(new Flight());
```

```
23 * E[foo] 有属性"foo"
```

```
24 * E[foo=bar]
```

```
25 * E[foo^=bar]
```

```
26 * E[foo$=bar]
```

```
27 * E[foo*=bar]
```

```
28 * E[foo%=n]
```

```
29 * E[foo!=bar]
```

```
30 *
```

```
31 */
```

```
32 class S050_Base_Util_XMLQuery {
```

```
33 public $document;
```

```
34 private $options = array();
```

```
35 public $first;
```

```
36 public $last;
```

```
37 public $nextSibling;
```

```
38 public $matches = array();
```

```
39 *
```

```
40 public static $cache = array();
```

```
41 public $simpleCache = array();
```

```
42 public $valueCache = array();
```

```
43 // public $nonSpace = '\s';
```

```
44 public $trimRe = '/^\s+|\s+$/';
```

```
45 public $tplRe = "/^{\(\\d+\)}\}/U";
```

```
46 //public $modeRe = "/^(\s?[\>+-]\s?|\\s|$)/";
```

```
17 </discounts>
```

```
18 </flight>
```

任意节点

节点名为X的节点

X的所有节点名为Y的子孙节点

X的一级子节点, 节点名为Y

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

与X紧挨着的所有Y节点

[airline] => 中国南方航空公司[南方航空]

[model] => 空客321(中)

[depCity] => 成都

[depTime] => 19:35

[arrCity] => 北京

[arrAirport] => 北京首都国际机场

[arrTime] => 22:00

[discounts] => Array

[0] => Array

[1] => Array

[2] => Array

[3] => Array

[4] => Array

[5] => Array

[6] => Array

[7] => Array

[8] => Array

[9] => Array

[10] => Array

[11] => Array

[12] => Array

[13] => Array

[14] => Array

[15] => Array

[16] => Array

[17] => Array

[18] => Array

[19] => Array

[20] => Array

[21] => Array

[22] => Array

[23] => Array

[24] => Array

[25] => Array

[26] => Array

[27] => Array

[28] => Array

[29] => Array

[30] => Array

[31] => Array

[32] => Array

[33] => Array

[34] => Array

[35] => Array

[36] => Array

[37] => Array

[38] => Array

[39] => Array

[40] => Array

[41] => Array

[42] => Array

[43] => Array

[44] => Array

[45] => Array

[46] => Array

[47] => Array

[48] => Array

[49] => Array

[50] => Array

[51] => Array

[52] => Array

[53] => Array

[54] => Array

[55] => Array

[56] => Array

[57] => Array

[58] => Array

[59] => Array

[60] => Array

[61] => Array

[62] => Array

[63] => Array

[64] => Array

[65] => Array

[66] => Array

[67] => Array

[68] => Array

[69] => Array

[70] => Array

[71] => Array

[72] => Array

[73] => Array

[74] => Array

[75] => Array

[76] => Array

[77] => Array

[78] => Array

[79] => Array

[80] => Array

[81] => Array

[82] => Array

[83] => Array

[84] => Array

[85] => Array

[86] => Array

[87] => Array

[88] => Array

[89] => Array

[90] => Array

[91] => Array

[92] => Array

[93] => Array

[94] => Array

[95] => Array

[96] => Array

[97] => Array

[98] => Array

[99] => Array

[100] => Array

[101] => Array

[102] => Array

[103] => Array

[104] => Array

[105] => Array

[106] => Array

[107] => Array

[108] => Array

[109] => Array

[110] => Array

[111] => Array

[112] => Array

[113] => Array

[114] => Array

[115] => Array

[116] => Array

[117] => Array

[118] => Array

[119] => Array

[120] => Array

[121] => Array

[122] => Array

[123] => Array

[124] => Array

[125] => Array

[126] => Array

[127] => Array

[128] => Array

[129] => Array

[130] => Array

[131] => Array

[132] => Array

[133] => Array

[134] => Array

[135] => Array

[136] => Array

[137] => Array

[138] => Array

[139] => Array

[140] => Array

[141] => Array

[142] => Array

[143] => Array

[144] => Array

[145] => Array

[146] => Array

[147] => Array

[148] => Array

[149] => Array

[150] => Array

[151] => Array

[152] => Array

[153] => Array

[154] => Array

[155] => Array

[156] => Array

[157] => Array

[158] => Array

[159] => Array

[160] => Array

[161] => Array

[162] => Array

[163] => Array

[164] => Array

[165] => Array

[166] => Array

[167] => Array

[168] => Array

[169] => Array

[170] => Array

[171] => Array

[172] => Array

[173] => Array

[174] => Array

[175] => Array

[176] => Array

[177] => Array

[178] => Array

[179] => Array

[180] => Array

[181] => Array

[182] => Array

[183] => Array

[184] => Array

[185] => Array

[186] => Array

[187] => Array

[188] => Array

[189] => Array

[190] => Array

[191] => Array

[192] => Array

[193] => Array

[194] => Array

[195] => Array

[196] => Array

[197] => Array

[198] => Array

[199] => Array

[200] => Array

[201] => Array

[202] => Array

[203] => Array

[204] => Array

[205] => Array

你觉得祭出个大招就快乐了？

- 呵呵！

- 问题：

- 内存吃紧，进程退出
- 展现多样化(IM、Qzone、检索、微信...)
- 数据源成倍增长
 - 数据问题追踪杀死开发
 - 产品经理嗝死开发
 - 即使写Query，依然开发效率跟不上接入速度

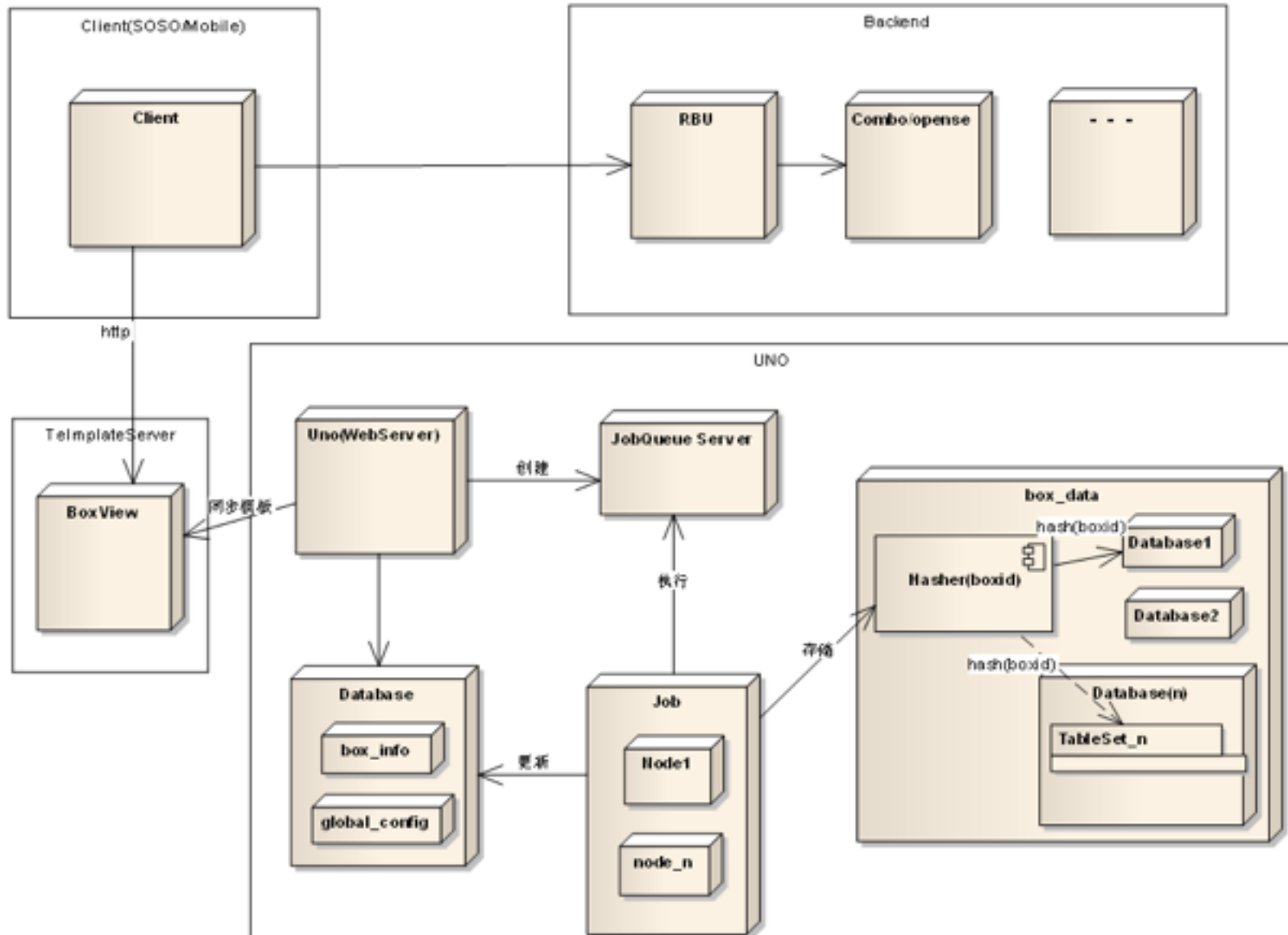
怎么破！

终极篇

- 流式解析
- 标签配置
- schema验证
- 自动报警
- 调度可控
- 完备日志信息
- 释放内存
- 释放开发
- 释放开发
- 释放开发
- 释放开发
- 释放开发

从猴子变成了人

deployment 基础设施层



心得体会

- 后台架构并不一味追求速度，人（开发）的因素是更重要的
- 面对快速迭代：优雅设计 + 龌龊实现
- 用已掌握的技术解决问题,权衡稳定与激情
- 避免过度设计（根据场景自然进化）
- 使用内存比使用磁盘来的爽的多
 - 但你要小心看护它
 - 需要做好规划，不做没“边界”的事
- 切分（水平、按功能）
 - 把工作负载分解成多个有能力驾驭的小单元，让每个单元都能维持良好的性价比
 - 依赖场景设计模块,而不是工具、中间件、哪个人
 - 使任何模块可替代

谢谢收看！

张勇 - SuperMan

QQ : 16732305

¥ 欢迎PHPers到北京发展 ¥