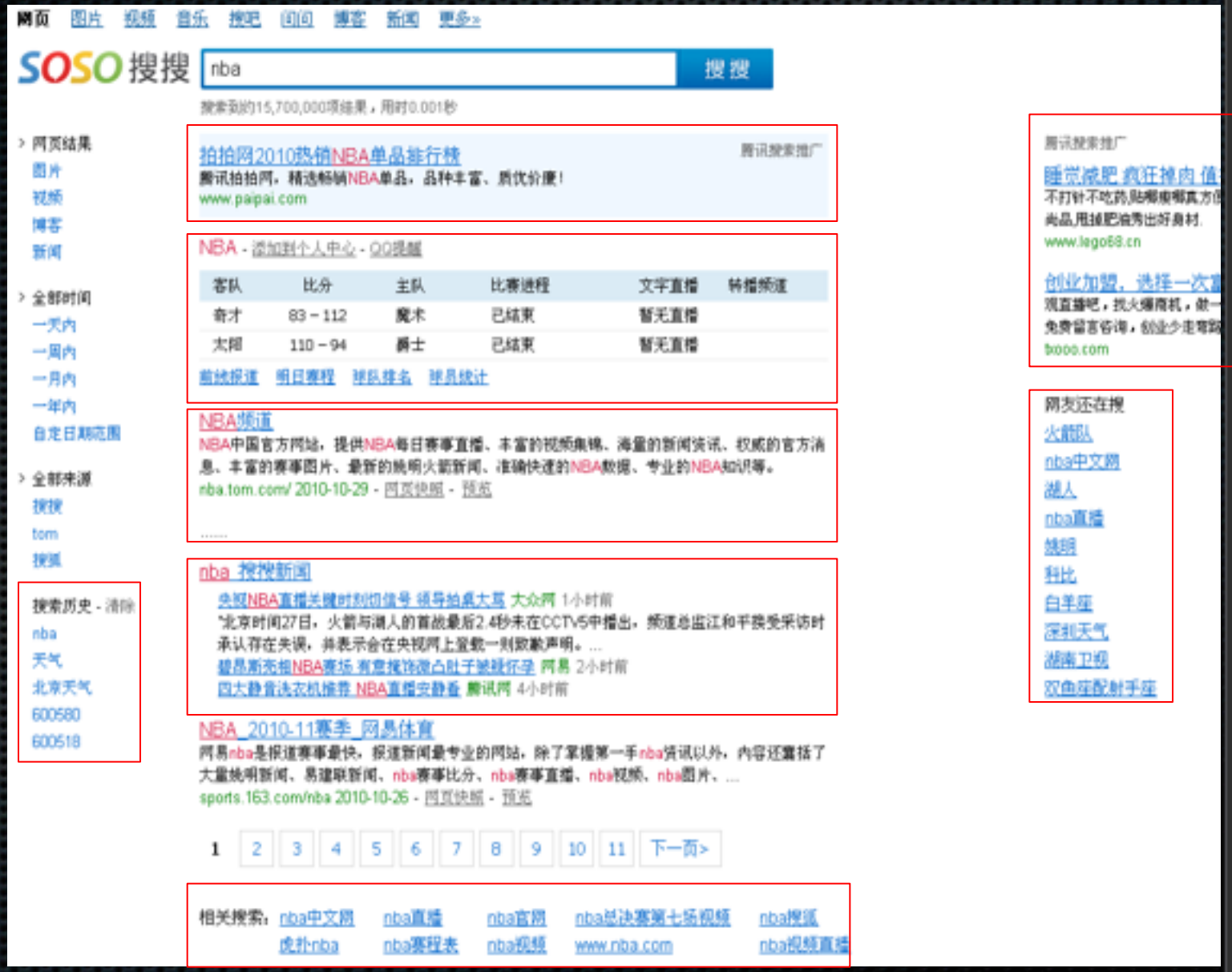
Richbox系统架构演化

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..



搜狗搜索 新闻 网页 微信 问问 图片 视频 音乐 地图 百科 更多>>

刘德华

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

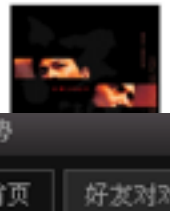
刘德华的影视作品

电影 电视剧 资讯 共148部>>


风暴 6.4分


寒战 7.4分


我知女 5.6分



刘德华的图片 共2697万张>>

香港星光大道表彰人员 展开


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

展开

邓紫棋介入张杰谢娜婚姻 最会炒作的女人 ent.taiwan.cn/ 13小时前

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 14小时前

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 1天前

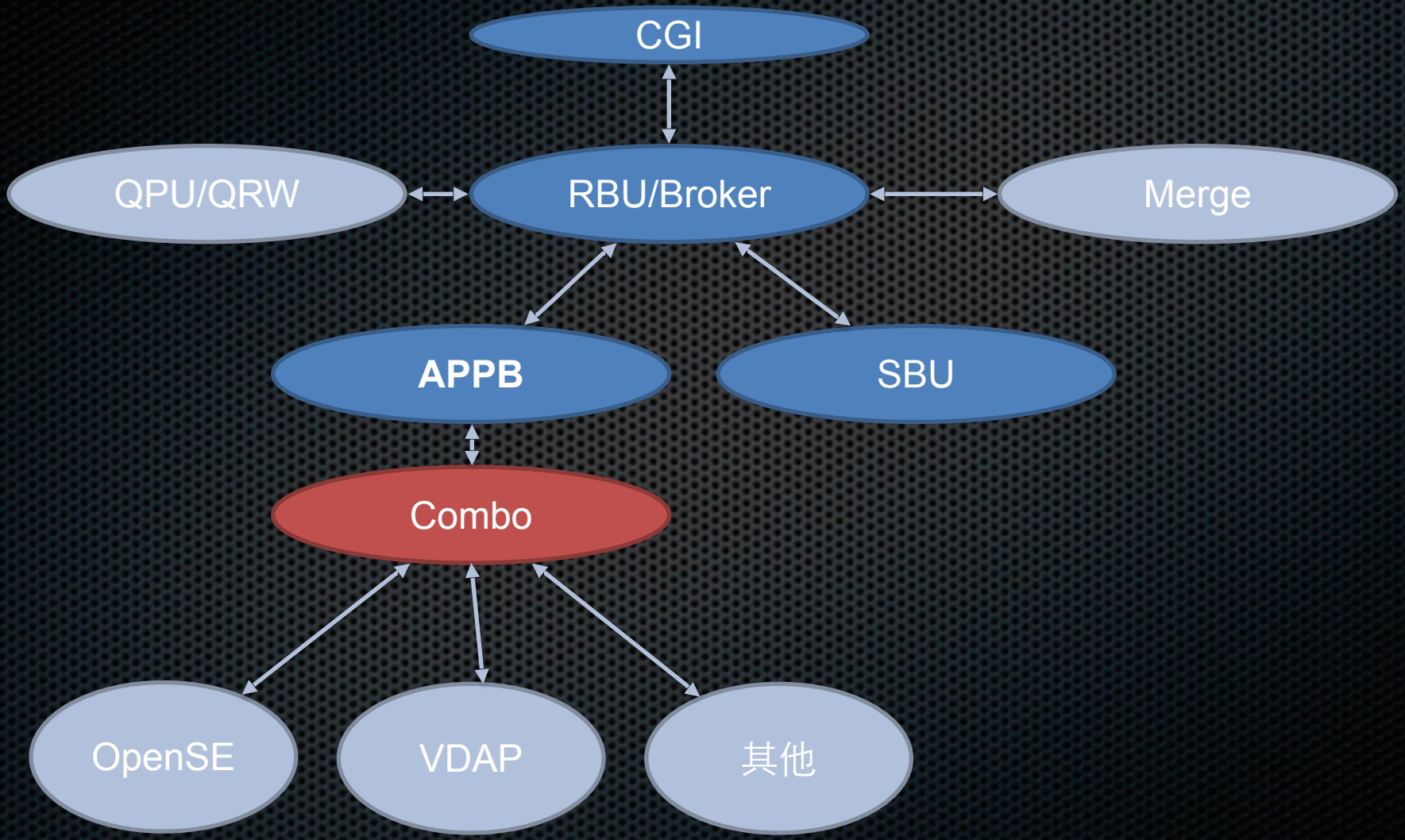
自然搜索结果和Richbox对比

	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，***anchor***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？



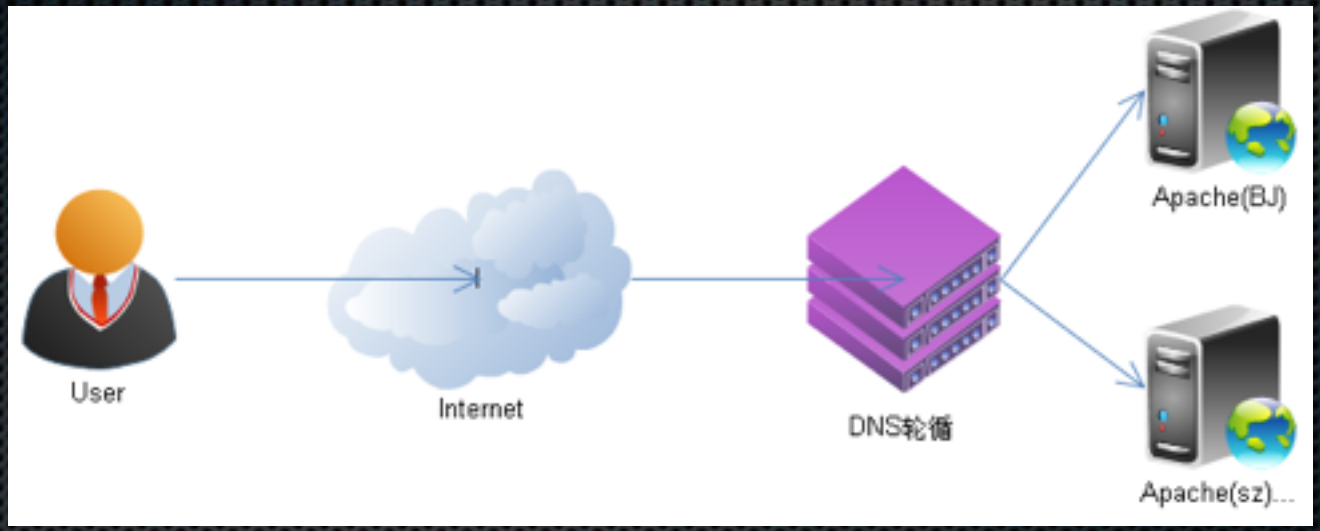
我们今天要讲的是?

- 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?

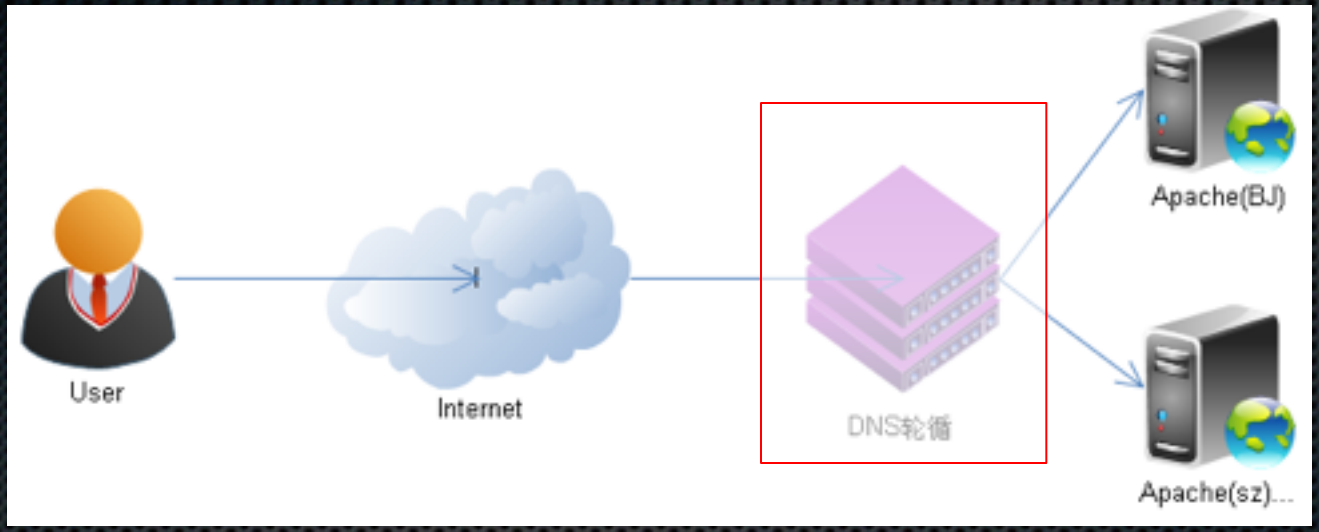
巨人的肩膀

Memcached

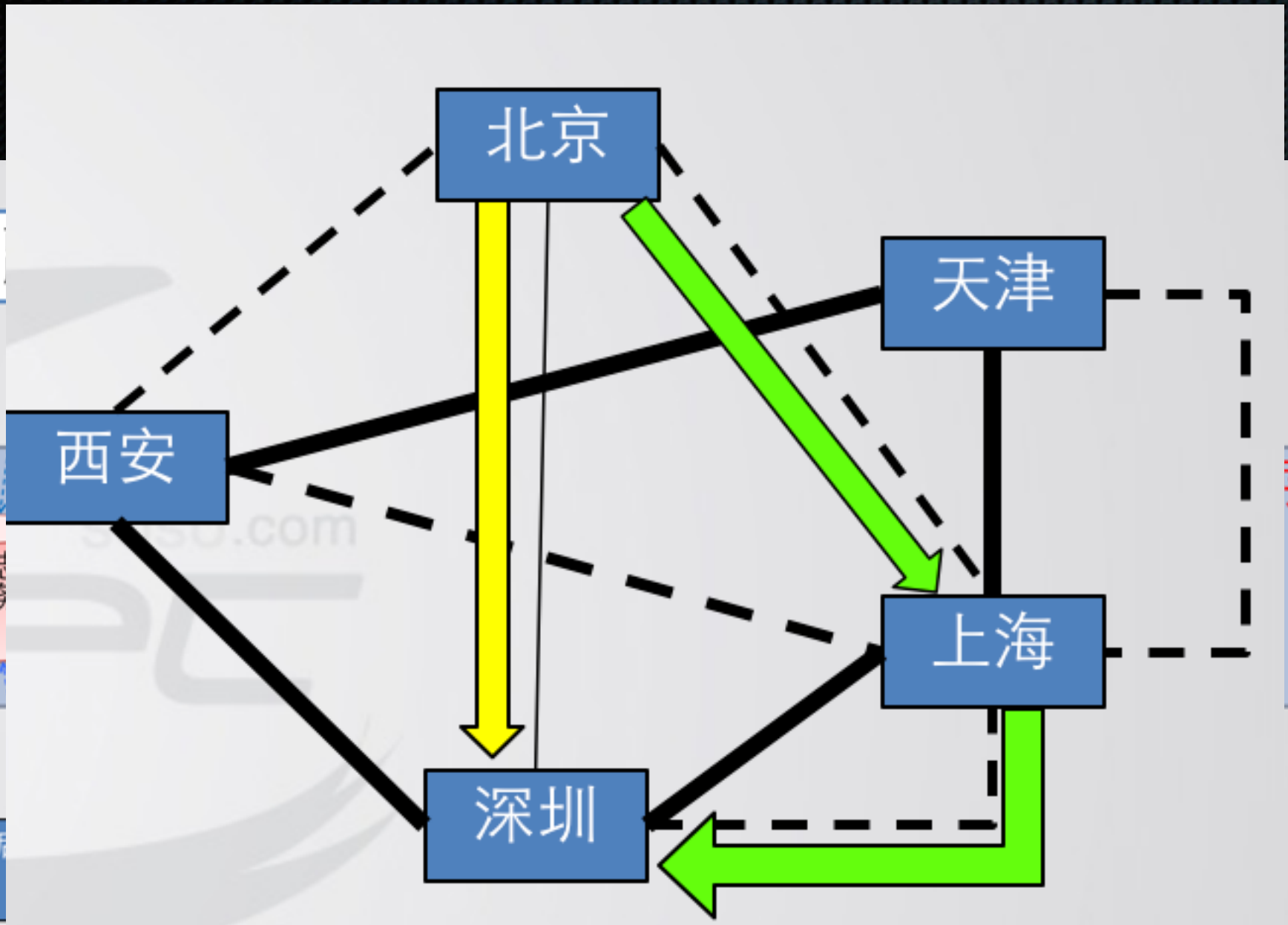
我们是从这里开始的



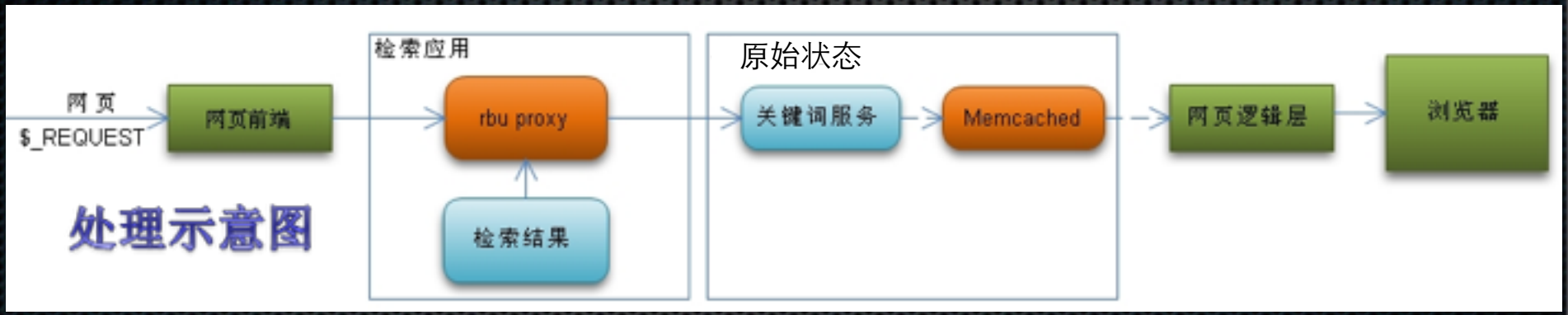
问题与挑战



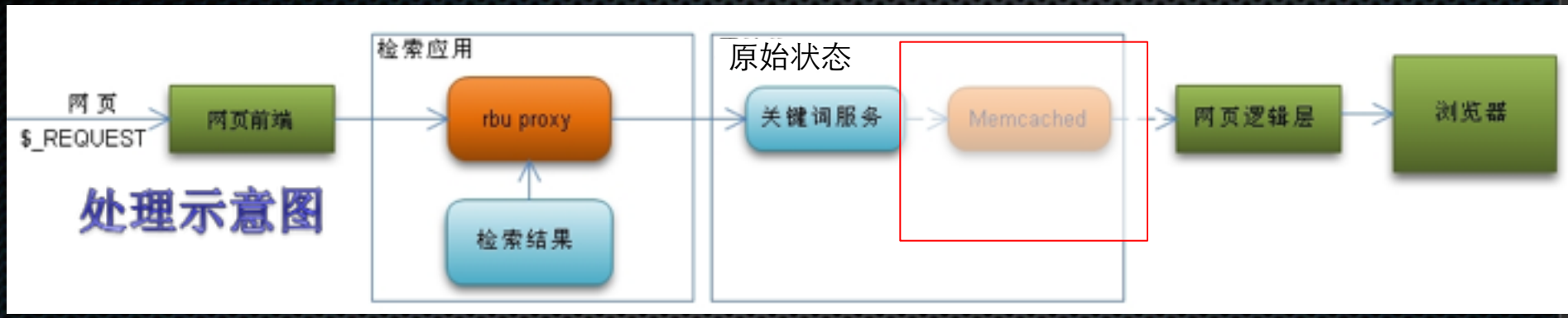
- 用户越来越多
- DNS不准?
- 用户反映访问速度慢
- 大家都用专线?


北京
天津
西安
上海
深圳

回到Web端

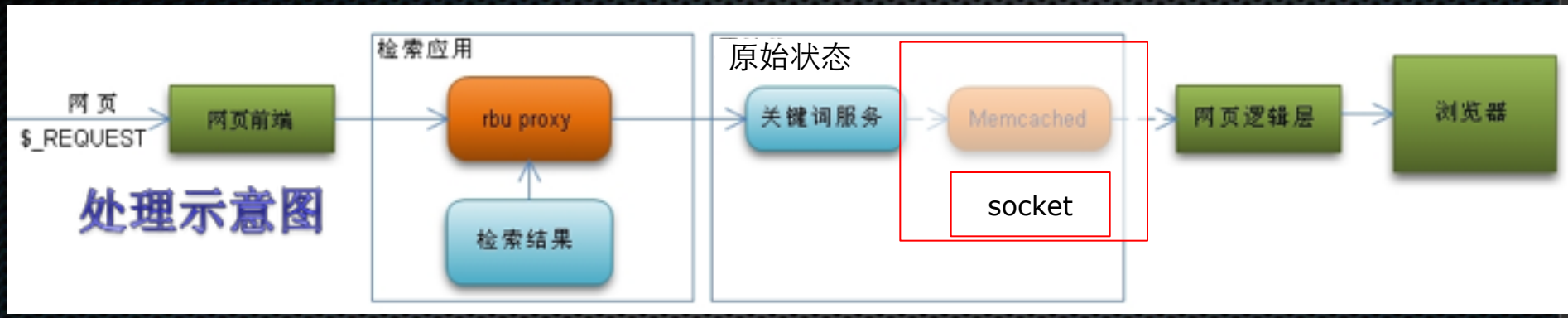

网 页
$_REQUEST
网页前端
检索应用
rbu proxy
检索结果
原始状态
关键词服务
Memcached
网页逻辑层
浏览器
处理示意图

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

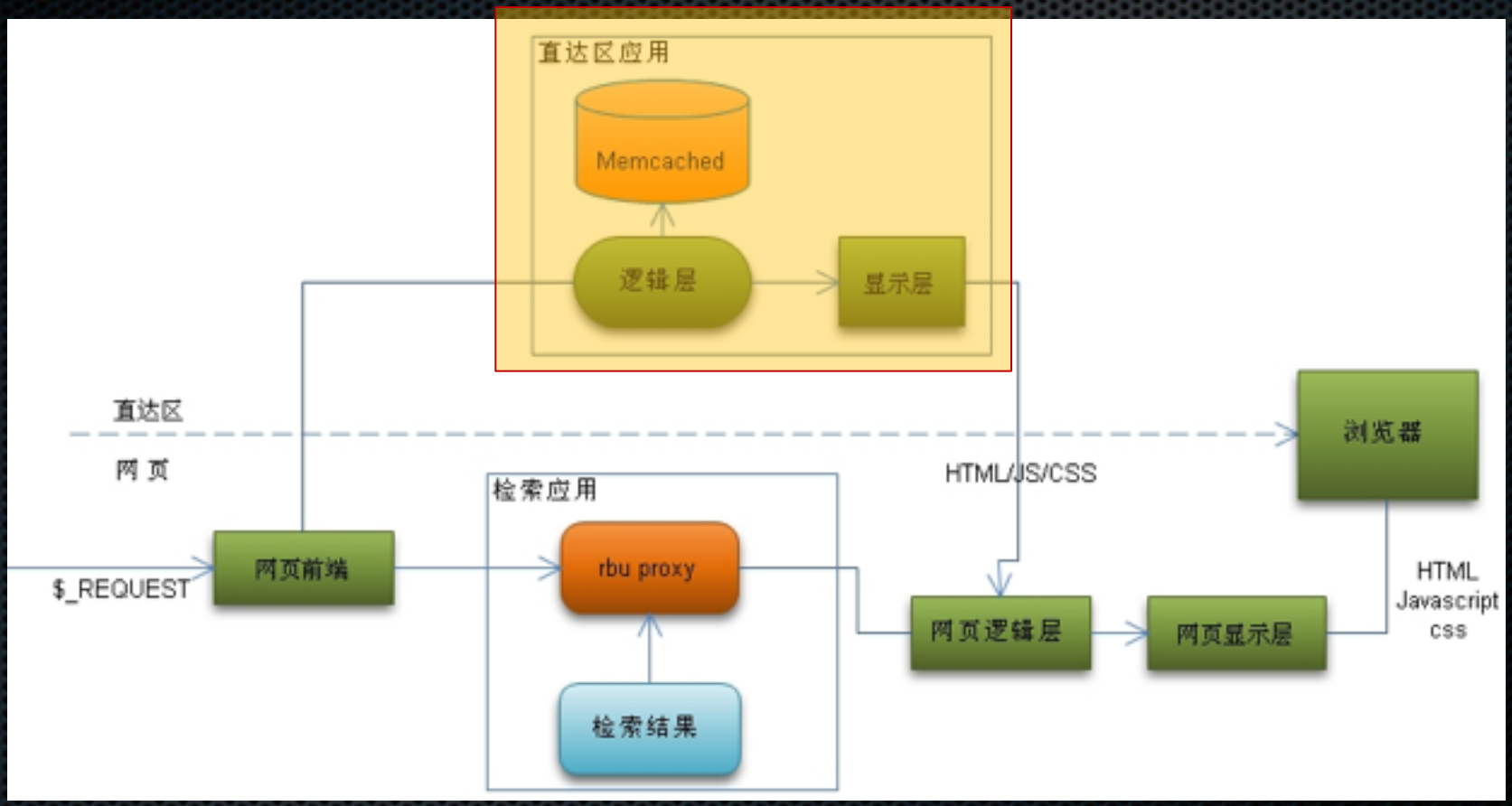
问题与挑战

- 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

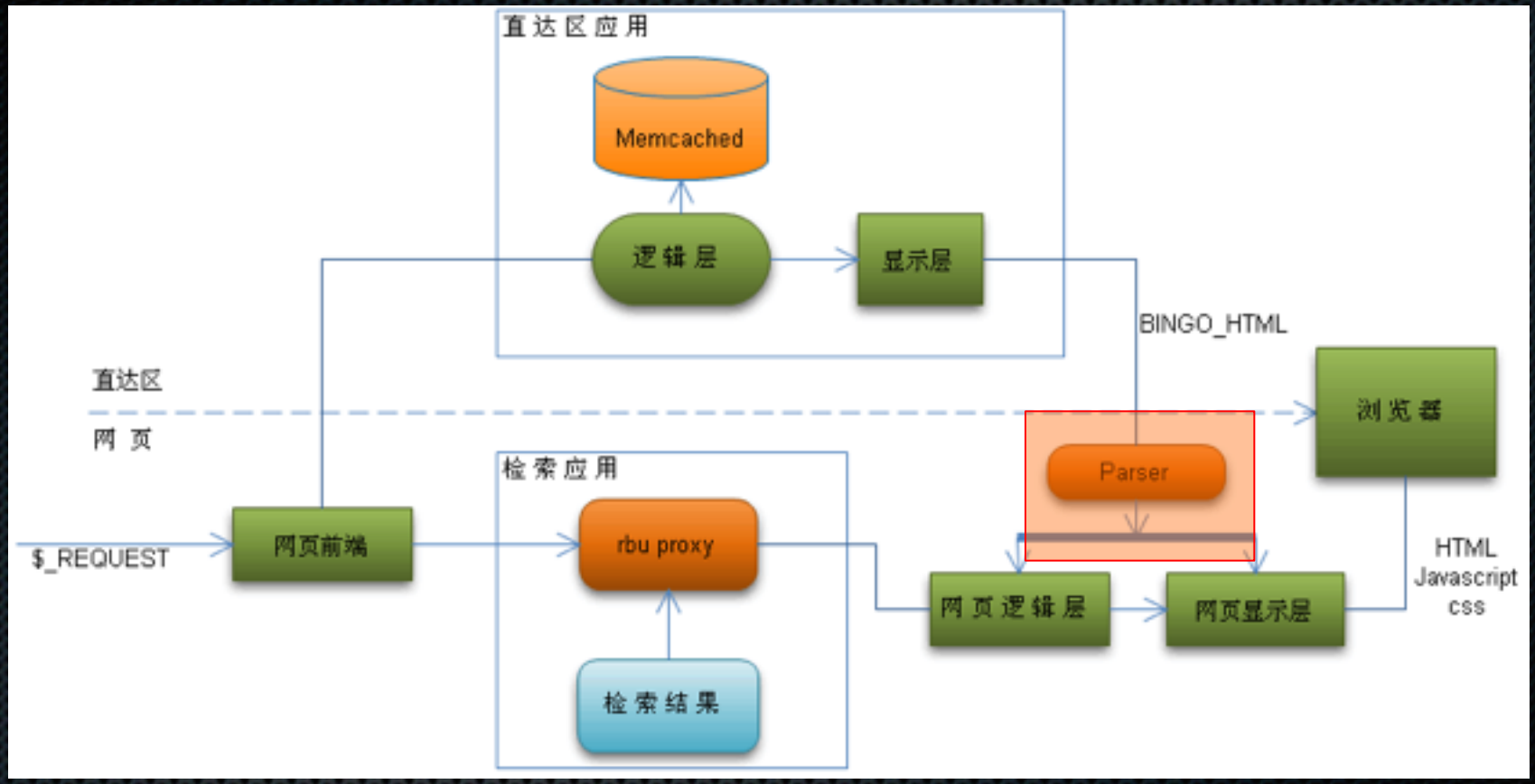
解决与应对


直达区应用
Memcached
逻辑层
显示层
直达区
网 页
浏览器
HTML/JS/CSS
检索应用
$_REQUEST
网页前端
rbu proxy
网页逻辑层
网页显示层
HTML
Javascript
css
检索结果

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

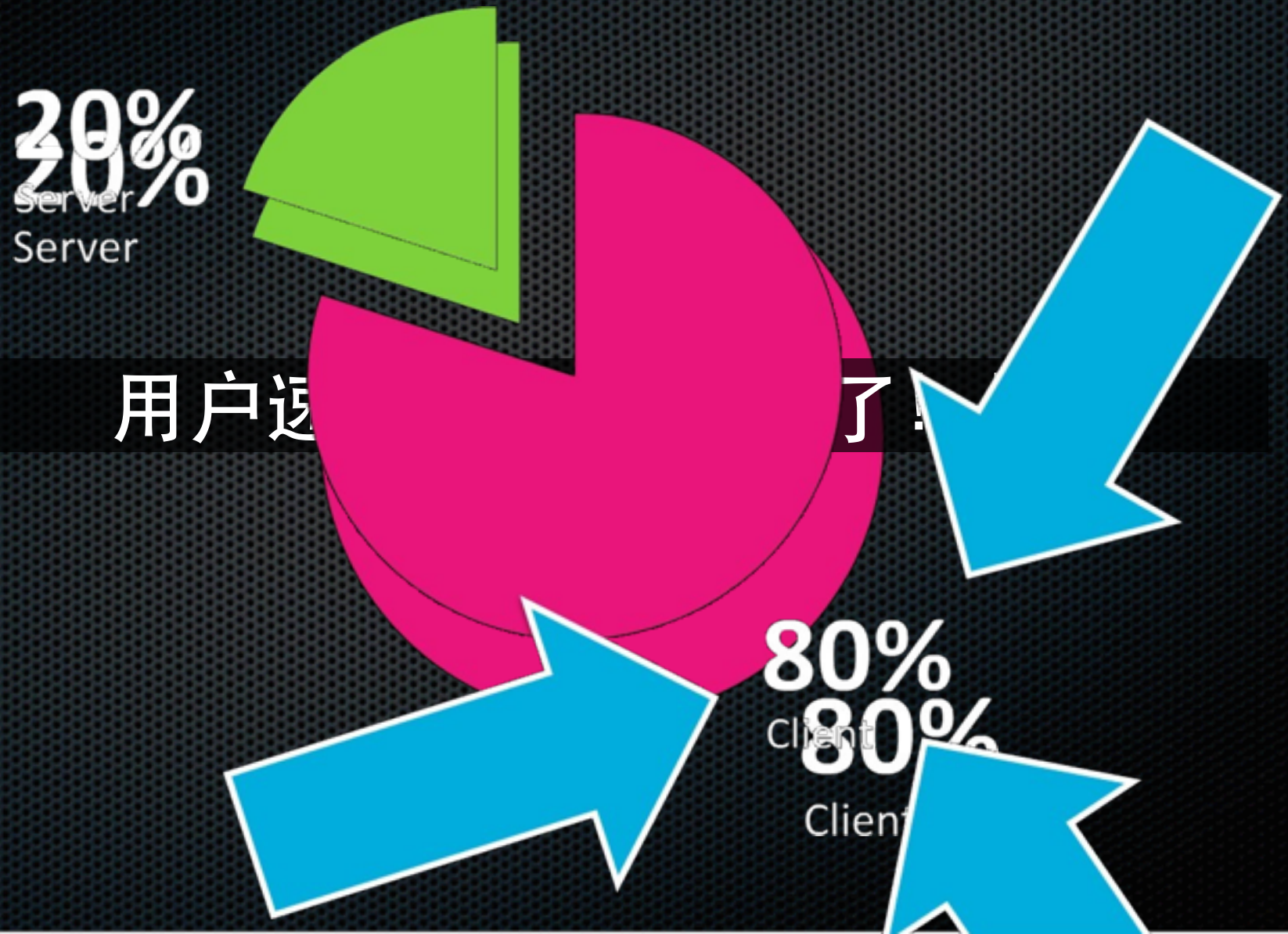
规划 = 重构+优化


直达区应用
Memcached
逻辑层
显示层
BINGO_HTML
直达区
网 页
浏览器
Parser
检索应用
$_REQUEST
网页前端
rbu proxy
网页逻辑层
网页显示层
HTML
Javascript
css
检索结果

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！



用户远 了！

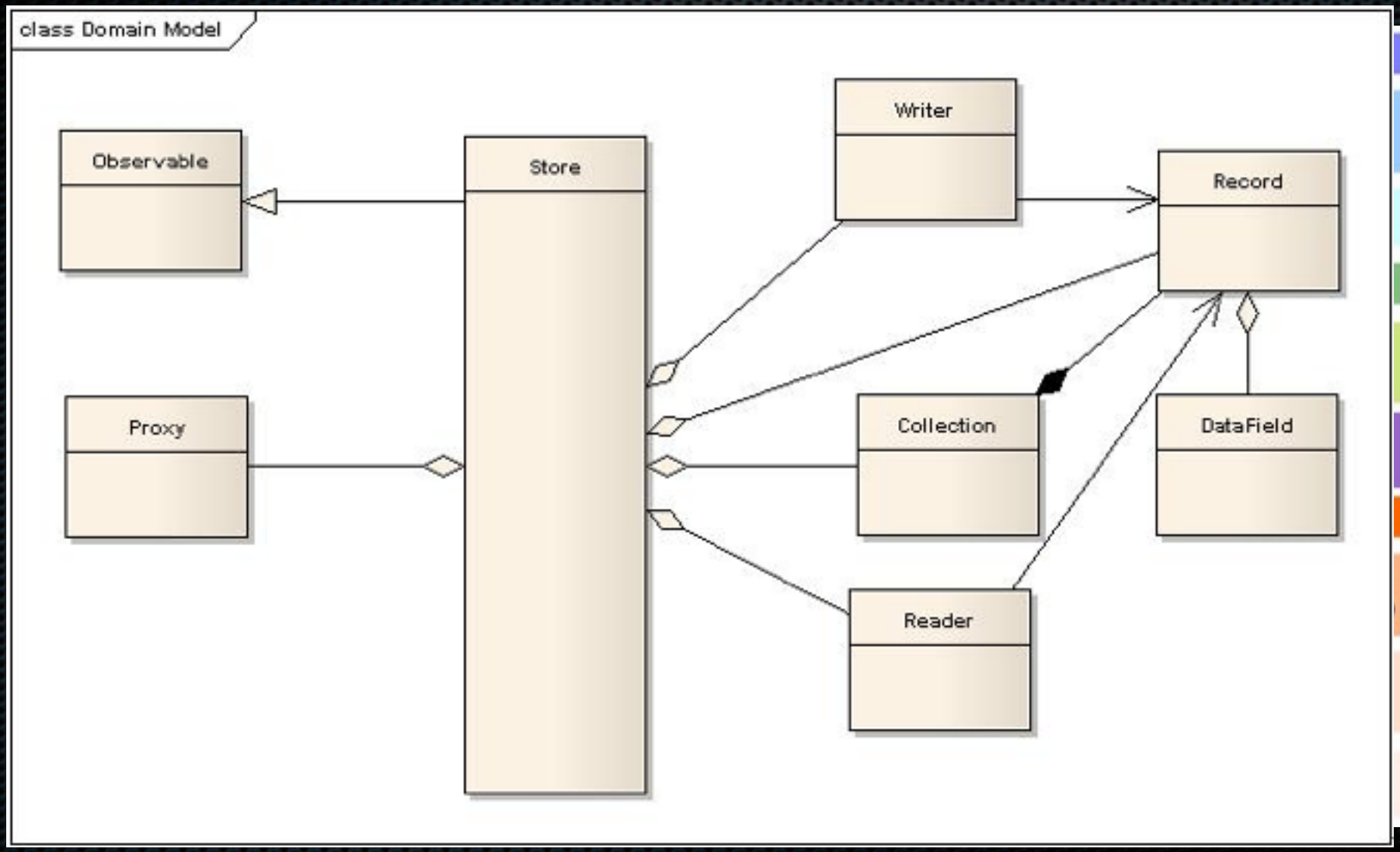
还是不快乐?

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


class Domain Model
Observable
Store
Writer
Record
Proxy
Collection
DataField
Reader

```
2  $tConfig = array();                                                    Array
5  * XML/(x)html 节点选择类
6  *
7  * @todo
8  *      1.(搁置) 深度封装，使其可以以类似jQuery的形式，进行级联操作
9  *  (See Updates 1)2.伪类支持 ,所有选择器、属性、过滤器、伪类可以以任意顺序组合，支持形如 div.foo:nth-child(odd)[@foo=bar].bar:first 这样的选择
10 *  3.
11 * Updates:
12 *  1.完成todo:2 伪类支持 ,所有选择器、属性、过滤器、伪类可以以任意顺序组合，支持形如 div.foo:nth-child(odd)[@foo=bar].bar:first 这样的选择器
13 * 目前支持的选择器：
14 *      1.元素选择符
15 *      *                       任意节点
16 *          X                   节点名为X的节点
17 *      X Y                 X的所有节点名为Y的子孙节点
18 *    X > Y 或 X/Y          X的一级子节点，节点名为Y
19 *      X + Y               与X紧挨着的所有Y节点
20 *          X ~ Y                   与X紧挨着的所有Y节点
21 *  2.属性选择符
22       使用符号@，例如div[@foo='bar']
23    E[foo] 有属性"foo"
24    E[foo=bar]            有属性"foo" 且等于 "bar"
25    E[foo^=bar]           有属性"foo" 且以"bar"开头
26    E[foo$=bar]           有属性"foo" 且以"bar"结尾
27    E[foo*=bar]           有属性"foo" 且含有"bar"作为一部分
28    E[foo%=n]             有属性"foo" 且被n取模值为 0 的节点
29    E[foo!=bar]           有属性"foo" 且不等于"bar"
30 *
31 */
32 class SOSO_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>
```

```
[url] => http://flight.qunar.com/site/one
[price] => 428
```

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
    - 产品经理嗲死开发
    - 即使写Query，依然开发效率跟不上接入速度

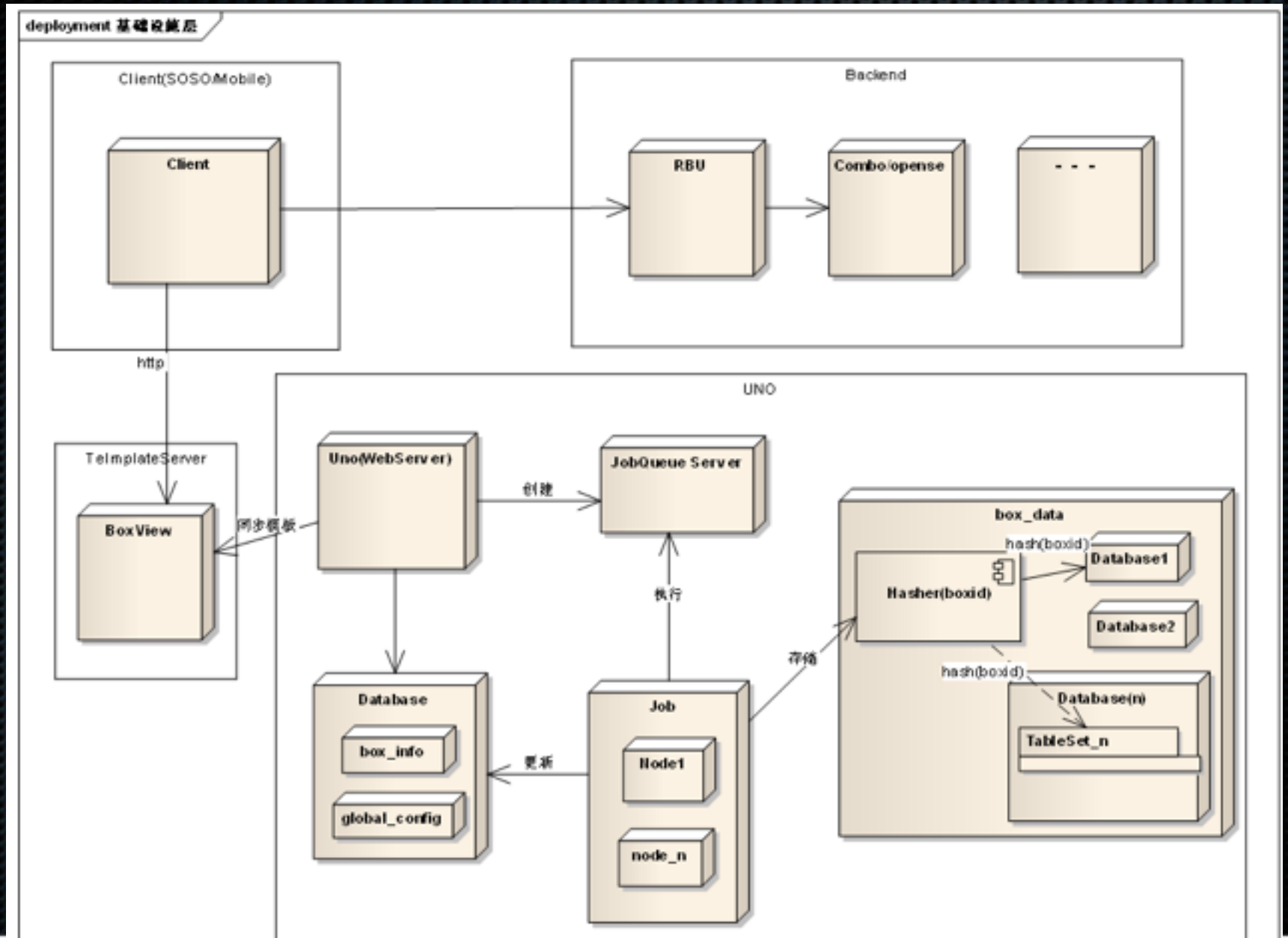
怎么破!

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

从猴子变成人了


deployment 基础设施层
Client(SOSO/Mobile)
Client
Backend
RBU
Combo/opense
- - -
http
UNO
TeImplateServer
BoxView
同步模板
Uno(WebServer)
创建
JobQueue Server
执行
Database
box_info
global_config
更新
Job
Node1
node_n
存储
box_data
hash(boxid)
Database1
Hasher(boxid)
Database2
hash(boxid)
Database(n)
TableSet_n

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
￥ 谢迎PHPers到北京发展 ￥