Nginx深度開發與客制化

——來自阿里巴巴的經驗

阿里巴巴集團-核心系統部
朱照遠（叔度）

2013-08-04

提綱

- 背景介紹
- 應用案例
- 開發與客制化
- 進行中的工作
- 開源總結

1. 背景介紹

關於阿里巴巴集團


Alibaba.com
Global trade starts here.™


阿里巴巴
1688.com


AliExpress


淘宝网
Taobao.com


天猫 TMALL.COM


一淘
淘宝旗下网站


聚划算
juhuasuan.com
品质团购每一天


阿里云计算
Alibaba Cloud Computing


支付宝
Alipay.com

面對的技術挑戰

- 2012年淘寶、天貓的交易額為11600億元人民幣
  - 超過Amazon與eBay之和
- 四家網站的網路流量在全球排名前100（Alexa統計）
  - taobao.com(**#10**) tmall.com(**#42**)
  - alibaba.com(**#72**) alipay.com(**#77**)
- 2012年雙11大促活動的一些數據
  - 雙11購物當天總銷售額191億人民幣
  - 第一分鐘超過1000萬人湧入，1分鐘成交19.2萬筆交易
  - 全天有2.13億獨立訪客，佔中國大陸網民總數40%
  - CDN最高峰值流量到達2100Gbps

使用Nginx的過程

- 2009年開始使用和探索
- 2010年開始開發大量模組以滿足業務
- 2011年開始修改Nginx的核心
- 2011年12月將修改過的Nginx項目開源
  - 項目名稱為Tengine

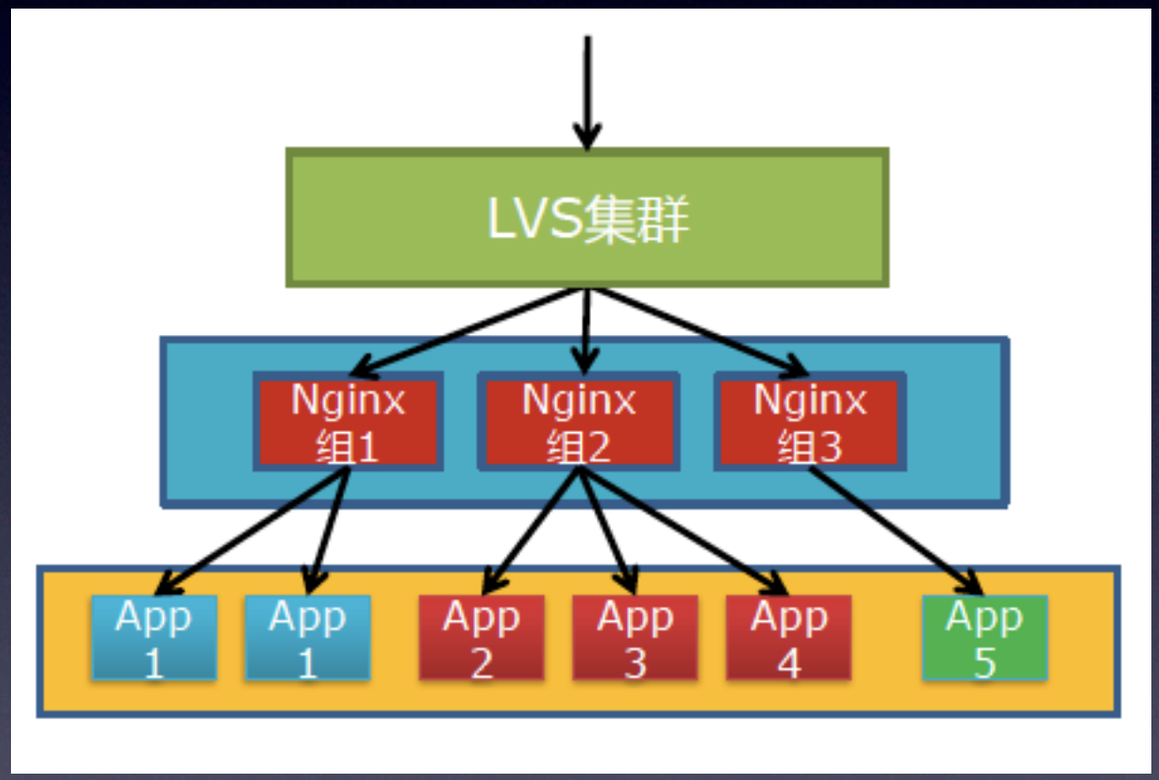
產生的收益

- Nginx使得業務更加穩定
  - 對大連接數目支援非常好
  - 佔用記憶體非常節省，更不會用swap
- Nginx使得應用的性能更高
  - QPS比Apache高
  - 節省伺服器數目
  - 基於Nginx的模組性能往往是之前業務的數倍

2. 應用案例

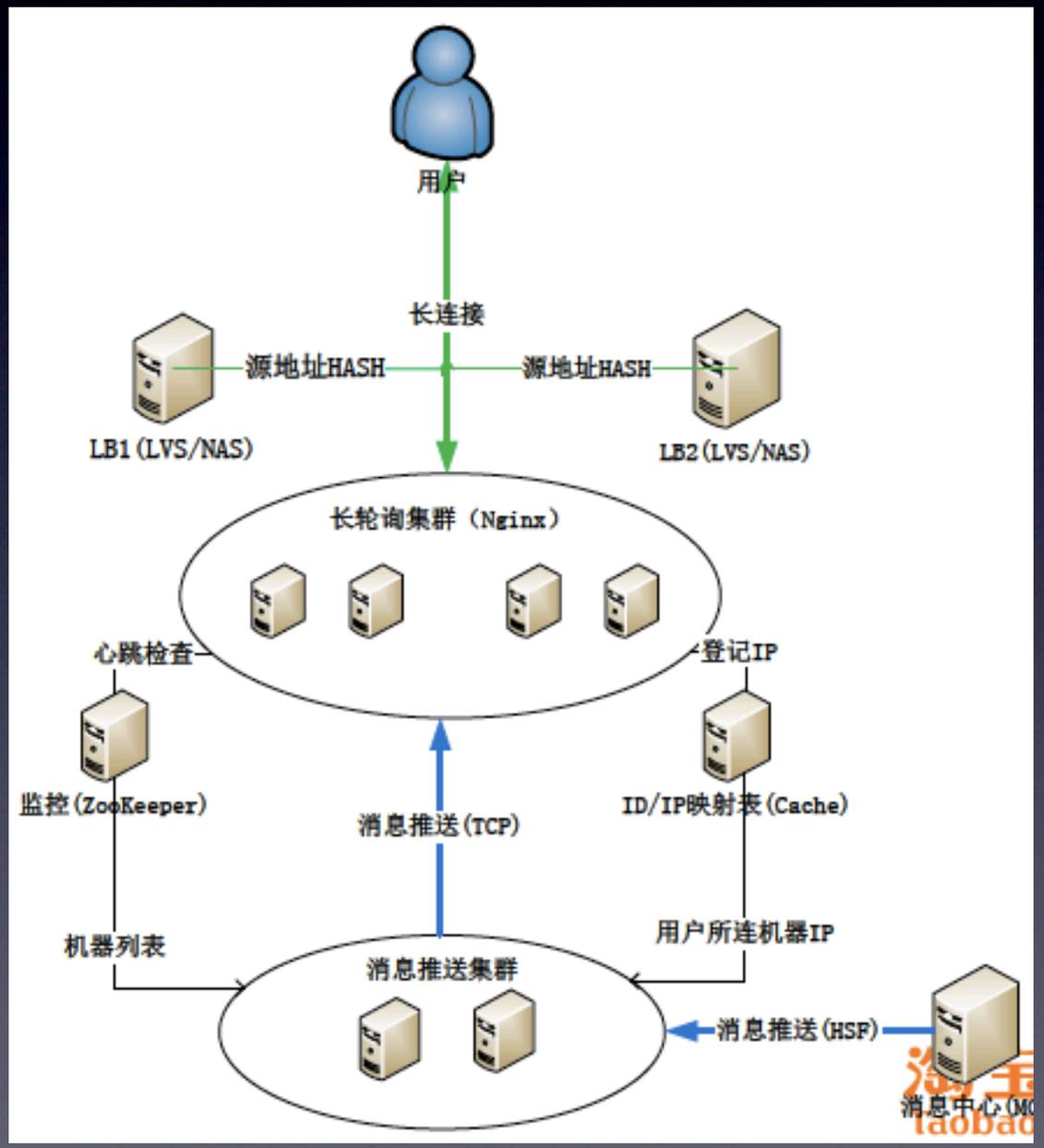
Web接入層

- 負載均衡
- SSL卸載
- 管理界面
- 安全防禦
- 灰度發佈
- 靜態化與cache



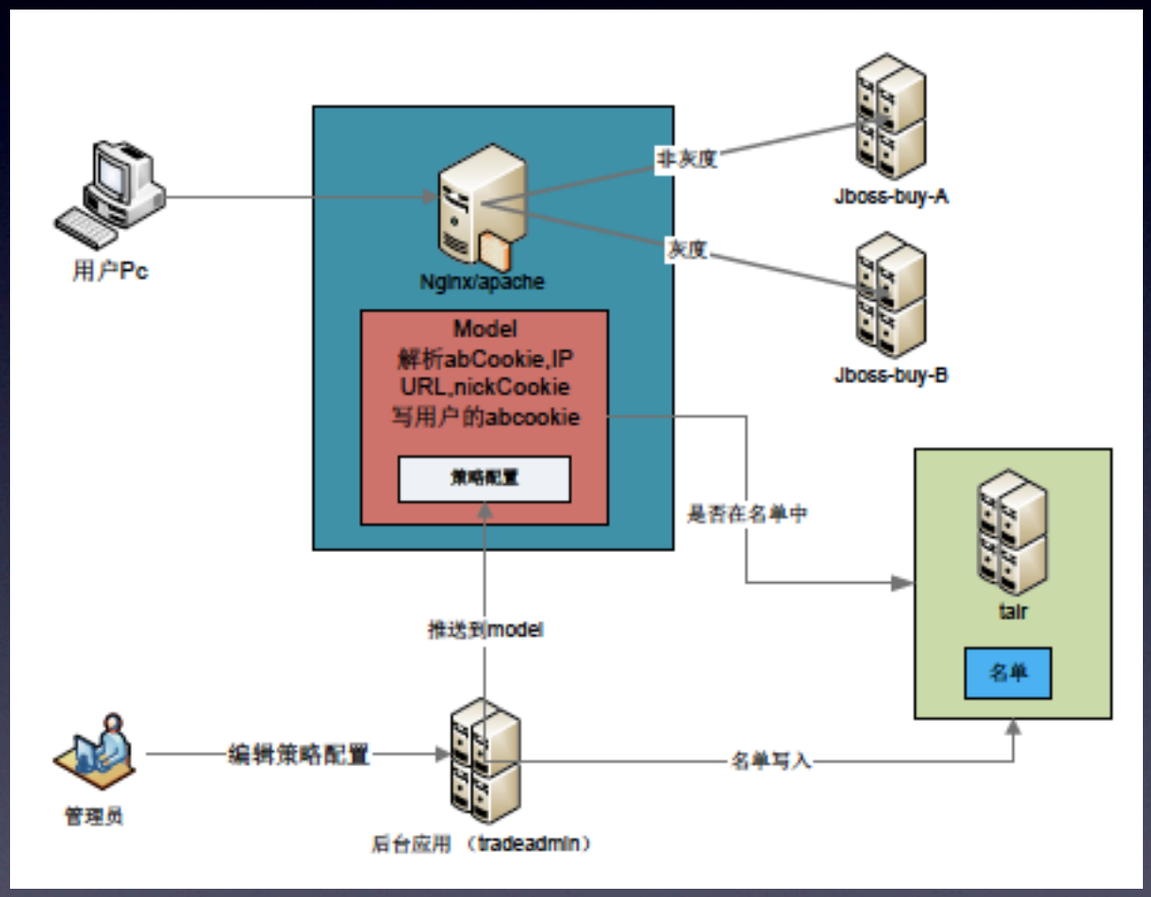
用戶群消息服務

- 提供comet服務
- 60萬連接每臺伺服器



灰度發佈與A/B Testing

- 灰度發佈
  - 逐漸放量測試
  - 方便的管理界面
- 規則
  - IP地址
  - cookie值
  - K/V存儲



日誌收集與統計系統

- 阿里的類似於Google Analytics系統
  - JavaScript植入
  - 收集日誌
  - 分析統計資訊
- 內部實做
  - Nginx模組
  - 分佈式傳輸系統
  - 在Hadoop上運行MapReduce統計
- 性能
  - 幾十臺伺服器每天處理數百億query
  - 單機處理能力4萬QPS（短連接）

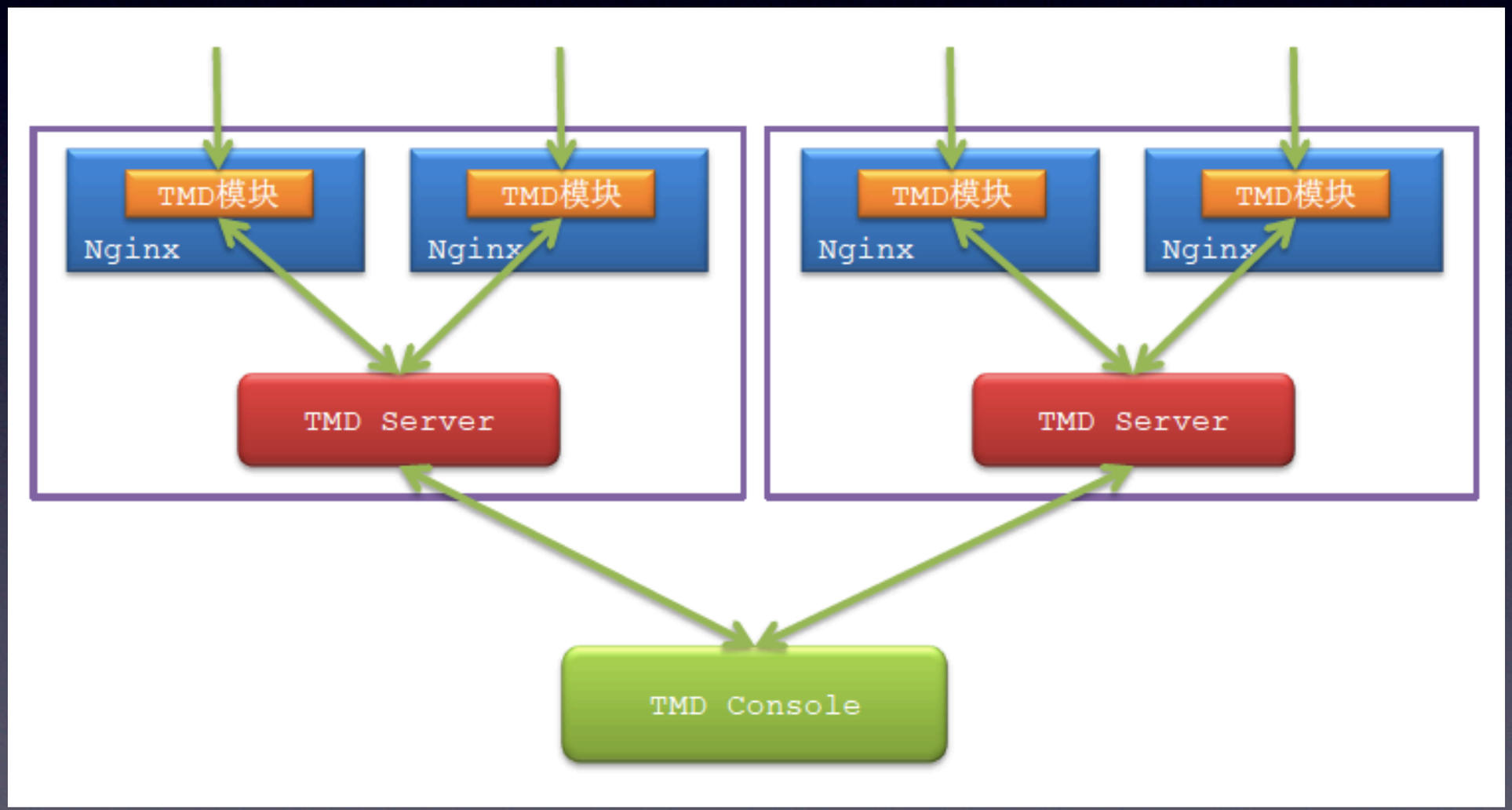
RESTful界面層

- RESTful界面支援
  - TFS（阿里分佈式檔案系統，類似GFS）
  - Tair（可以看作一個分佈式的memcached + Redis）
- 簡化應用開發
  - 可返回JSON格式直接讓瀏覽器處理
    - 從而不必在伺服器做組裝

分佈式防攻擊系統

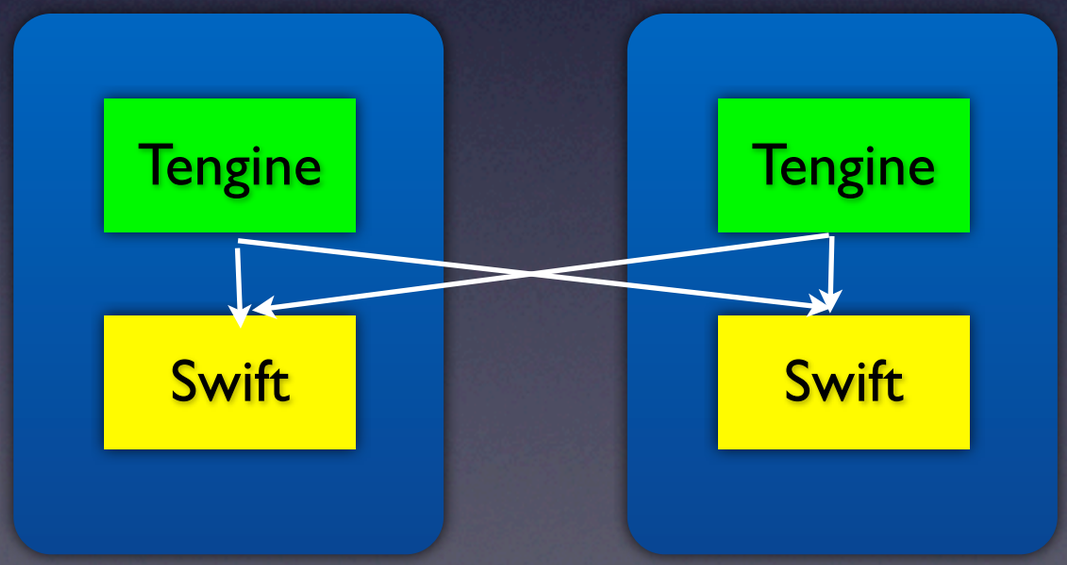
- 應對的問題
  - HTTP層面的DDoS攻擊
  - 惡意的爬蟲
  - 解決單機防護不知道全局資訊的問題
- TMD（Taobao Missle Defense）系統
  - Nginx作為防攻擊系統的客戶端
  - TMD Server做策略分析
  - TMD Console執行彙總和控制檯

TMD系統架構圖


TMD模块
TMD模块
TMD模块
TMD模块
Nginx
Nginx
Nginx
Nginx
TMD Server
TMD Server
TMD Console

CDN系統

- consistent_hash模組
  - 同一對象到同一台cache服務器
  - 不會cache抖動



3. 開發與客制化

動態模組加載

- DSO (Dynamic Shared Object) 支援
- 避免每次新加模組都要重新編譯binary

  ./configure --with-http_sub_module=shared

  ./dso_tool --add-module=/home/dso/lua-nginx-module
- 配置例子

```
dso {
    load ngx_http_lua_module.so;
    load ngx_http_memcached_module.so;
}
```

流式上傳

- Nginx的問題
  - 寫硬碟與記憶體佔用之間的權衡，避免IO
  - client_body_buffer_size的限制
- Tengine的實做
  - proxy_request_buffering
  - client_body_buffers
  - client_body_postpone_size

組合JavaScript和CSS

- 根據Yahoo!前端優化第一條原則
  - Minimize HTTP Requests
    - 減少三方握手和HTTP請求的發送次數
- 阿里CDN combo
  - concat模組，組合JavaScript和CSS

CDN combo的使用

- 以兩個問號（??）激活combo特性
- 多個檔案之間用逗號（,）分開
- 用一個?來表示timestamp
  - 用來突破瀏覽器cache
- 例子
  - http://example.com??1.js,2.js,3.js?t=20130805

為頁面瘦身

- trim模組
  - 刪除HTML和內嵌JavaScript、CSS的註釋和空白符號

```
location / {
    trim on;
    trim_jscss on;
}
```

系統過載保護

- 監控系統條目
  - 伺服器的load
  - 記憶體的使用（如swap的比例）
- sysguard模組

```
sysguard on;
sysguard_load load=4 action=/high_load.html;
sysguard_mem swapratio=10% action=/mem_high.html
```

- 可客制化保護返回的頁面

多種日誌支援的方式

- 本地和遠程syslog支援

  access_log syslog:user:info:127.0.0.1:514 combined;

- 管道支援

  access_log pipe:/path/to/cronolog combined;

- 抽樣支援（減少日誌的條目數目）

  access_log /path/to/file combined ratio=0.01;

伺服器資訊調試

- footer模組

  footer $host_comment;

- 例子（在頁面最後添加）

  <!-- joshua Fri, 03 Aug 2013 08:24:43 GMT -->

對於行程設置的簡化

- 可以通過auto選項來自動設置行程數目和CPU親和性

```
worker_processes 8;
worker_cpu_affinity 00000001 00000010 00000100
00001000 00010000 00100000 01000000 10000000;
```

```
worker_processes auto;
worker_cpu_affinity auto;
```

user_agent模組

- 將瀏覽器、爬蟲輸出成變量
- 具體實做
  - 使用trie樹，O(n)的複雜度
  - 對比Nginx的browser模組是O(n^3)

增強了limit_req模組

- 多變量支援
- 白名單支援
- 指定跳轉頁面支援
- 同location下多limit_req支援
- 支援off選項

```
location / {
    limit_req zone=one burst=5;
    limit_req zone=two forbid_action=@test1;
    limit_req zone=three burst=3 forbid_action=@test2;
}
```

主動健康檢查

- 發現後端伺服器失效的響應快
- L7的檢查使上線下線很方便
- 後端伺服器的狀態監控頁面
- 可以檢查多種後端伺服器
  - TCP/HTTP/HTTPS/AJP/MySQL/FastCGI

輸入體過濾器

- 流式地做安全過濾
- 標準Nginx的問題
  - POST體與記憶體之間的關係
  - 性能與硬碟
- 已應用場景
  - 防hashdos/SQL注入/XSS

動態腳本支援

- Lua語言的支援
  - 支持各種phase
  - 支持header、body filter
- 性能和彈性的完美結合
  - LuaJIT，和Java、C一個量級
  - 方便修改，不需編譯
- 不必開發C的模組

Lua支援的例子

```
location /http_client {
    proxy_pass $arg_url;
}
location /web_iconv {
    content_by_lua '
        local from, to, url = ngx.var.arg_f, ngx.var.arg_t, ngx.var.arg_u
        local iconv = require "iconv"
        local cd = iconv.new(to or "utf8", from or "gbk")
        local res = ngx.location.capture("/http_client?url=" .. url)
        if res.status == 200 then
            local ostr, err = cd:iconv(res.body)
            ngx.print(ostr)
        else
            ngx.say("error occured: rc=" .. res.status)
        end
    ';
}
```

會話保持

- session_sticky模組
  - 通過cookie實現客戶端與伺服器做負載均衡，後續訪問同一台伺服器

```
upstream foo {
    server 192.168.0.1;
    server 192.168.0.2;
    session_sticky;
}
server {
    location / {
        proxy_pass http://foo;
    }
}
```

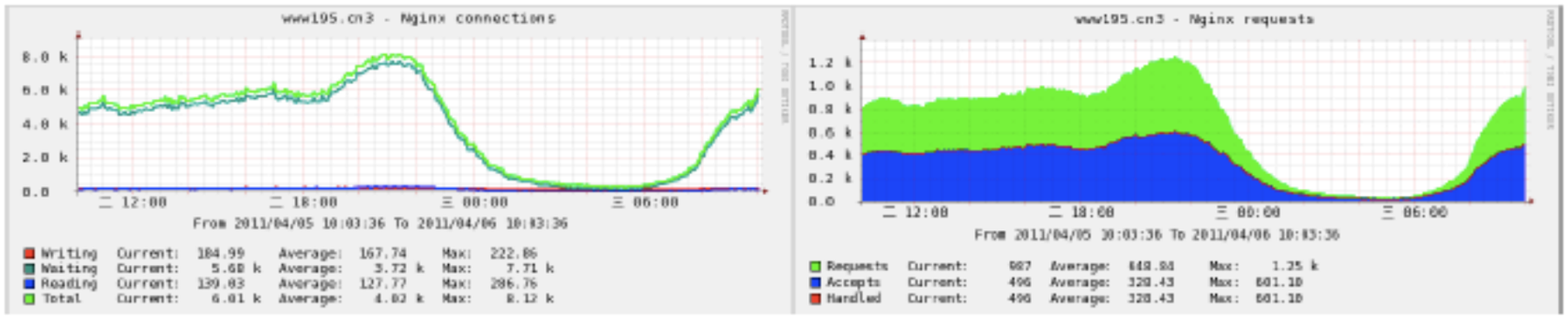
一致性hash

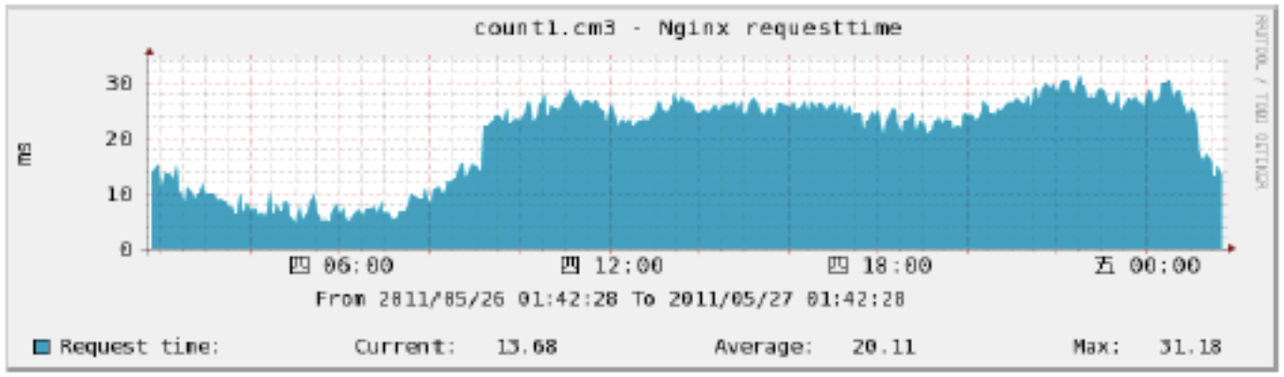
- consistent hash模組
- 防止抖動
- 可以根據變量來hash

命令行參數的完善

- 列出已經編譯的模組
  - -m選項
- 列出已支援的指令
  - -l選項
- 列出配置檔案的全部內容
  - -d選項

監控增強


www195.cn3 - Nginx connections
8.0 k
6.0 k
4.0 k
2.0 k
0.0
二 12:00
二 18:00
三 00:00
三 06:00
From 2011/04/05 10:03:36 To 2011/04/06 10:03:36
Writing Current: 184.99 Average: 167.74 Max: 222.86
Waiting Current: 5.60 k Average: 5.72 k Max: 7.71 k
Reading Current: 139.83 Average: 127.77 Max: 286.76
Total Current: 6.01 k Average: 4.02 k Max: 8.12 k
www195.cn3 - Nginx requests
1.2 k
1.0 k
0.8 k
0.6 k
0.4 k
0.2 k
0.0
二 12:00
二 18:00
三 00:00
三 06:00
From 2011/04/05 10:03:36 To 2011/04/06 10:03:36
Requests Current: 987 Average: 648.84 Max: 1.25 k
Accepts Current: 496 Average: 320.43 Max: 601.10
Handled Current: 496 Average: 320.43 Max: 601.10


count1.cm3 - Nginx requesttime
ms
30
20
10
0
四 06:00
四 12:00
四 18:00
五 00:00
From 2011/05/26 01:42:28 To 2011/05/27 01:42:28
Request time: Current: 13.68 Average: 20.11 Max: 31.18

4. 進行中的工作

計時器的優化

- 數據結構對比
  - rbtree
  - minheap
  - timer wheel

SO_REUSEPORT支援

- epoll的驚群問題（Thundering Hurd）
- Google的patch
- 性能提升

後端keep_alive優化

- 超時時間限制
- 針對後端伺服器的pool

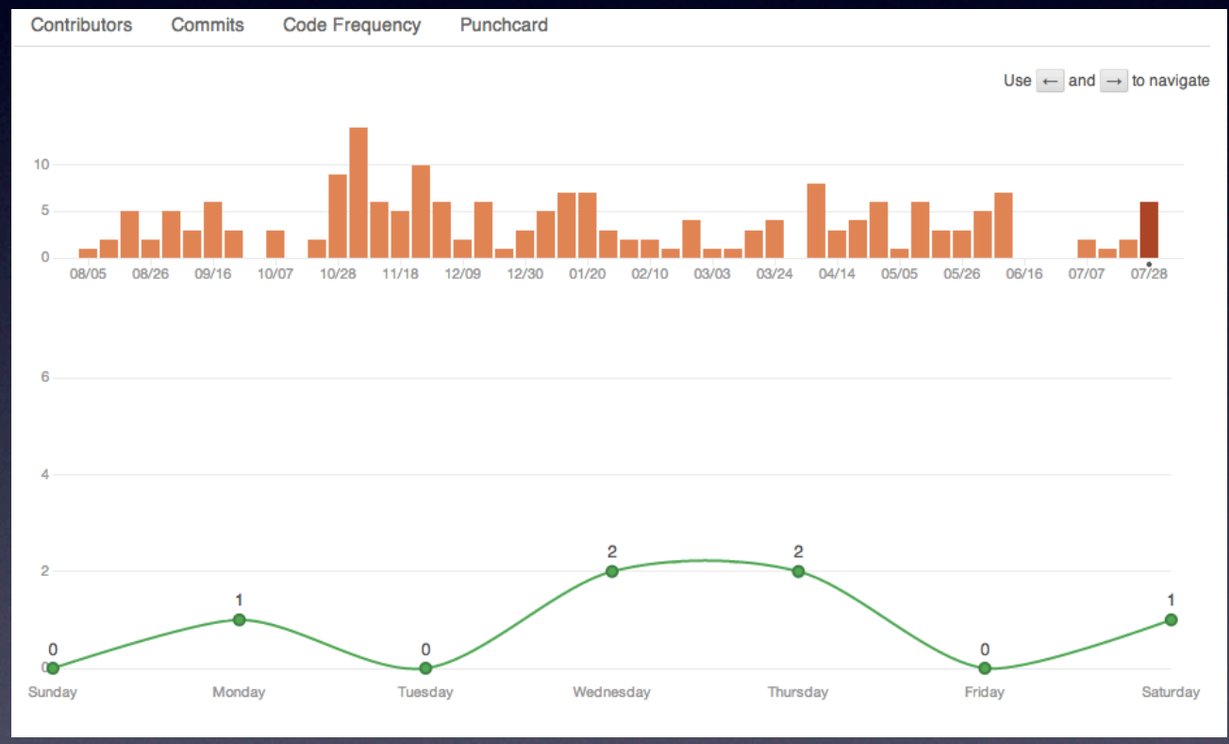
對10GE網路的良好支援

- 10GE網卡跑滿做Load Balancer
- CPU要求還有足夠的空閒

5. 開源總結

開源社區

- 發佈15個版本
- 核心開發者7人
- 貢獻者25人



非阿里的外部用戶

- Internet Archive
- 騰訊
- 土豆
- 京東
- 酷六
- PPLive
- 鳳凰網
- 開源中國
- ...

參考資源

- 前面講過的內容絕大部分已經開源!
- Tengine的主頁
  - http://tengine.taobao.org
- Tengine的GitHub
  - https://github.com/alibaba/tengine
- 歡迎參與Tengine開源項目!

阿里開源情況

- 中國大陸開源力度投入最大的企業!
- https://github.com/alibaba
- 60個開源項目左右
  - Linux kernel
  - Hadoop
  - LVS
  - TFS/Tair
  - ...

Linux內核全球patch貢獻排名

No.102	LG Electonics	242(0.07%)
No.104	Tao Bao	235(0.07%
No.106	P.A. Semi	231(0.07%)
No.107	OpenedHand	226(0.07%)
No.108	rPath	220(0.07%)
No.108	Embedded Alley Solutions	220(0.07%)
No.110		9(0.06%)
No.111	Bull SAS	215(0.06%)
No.112	Myricom	214(0.06%)
No.113	LWN	208(0.06%)
No.114	Hansen Partnership	198(0.06%)
No.114	STRATO	198(0.06%)
No.116	M&N Solutions	192(0.06%)
No.116	Avionic Design Development GmbH	192(0.06%)
No.116	igalia	192(0.06%)
No.119	Hauppauge	185(0.05%)
No.120	EXAR	184(0.05%)
No.121	Voltaire	180(0.05%)
No.122	CSR	179(0.05%)
No.123	SANPeople	178(0.05%)
No.123	Collabora Multimedia	178(0.05%)
No.125	Toshiba	174(0.05%)
No.126	Tuxera	173(0.05%)
No.127	Philosys Software	172(0.05%)
No.128	Bitmer	170(0.05%)
No.129	tieto	169(0.05%)
No.130	HuaWei	164(0.05%)
No.131	MathEmbedded Consulting	163(0.05%)

Q&A

- Thank you!