

]

A Firefox cluster driven by JavaScript, Perl, and PL/PgSQL

A Firefox **cluster** driven by *JavaScript, Perl, & PL/PgSQL*

☺ *agentzh@yahoo.cn* ☺

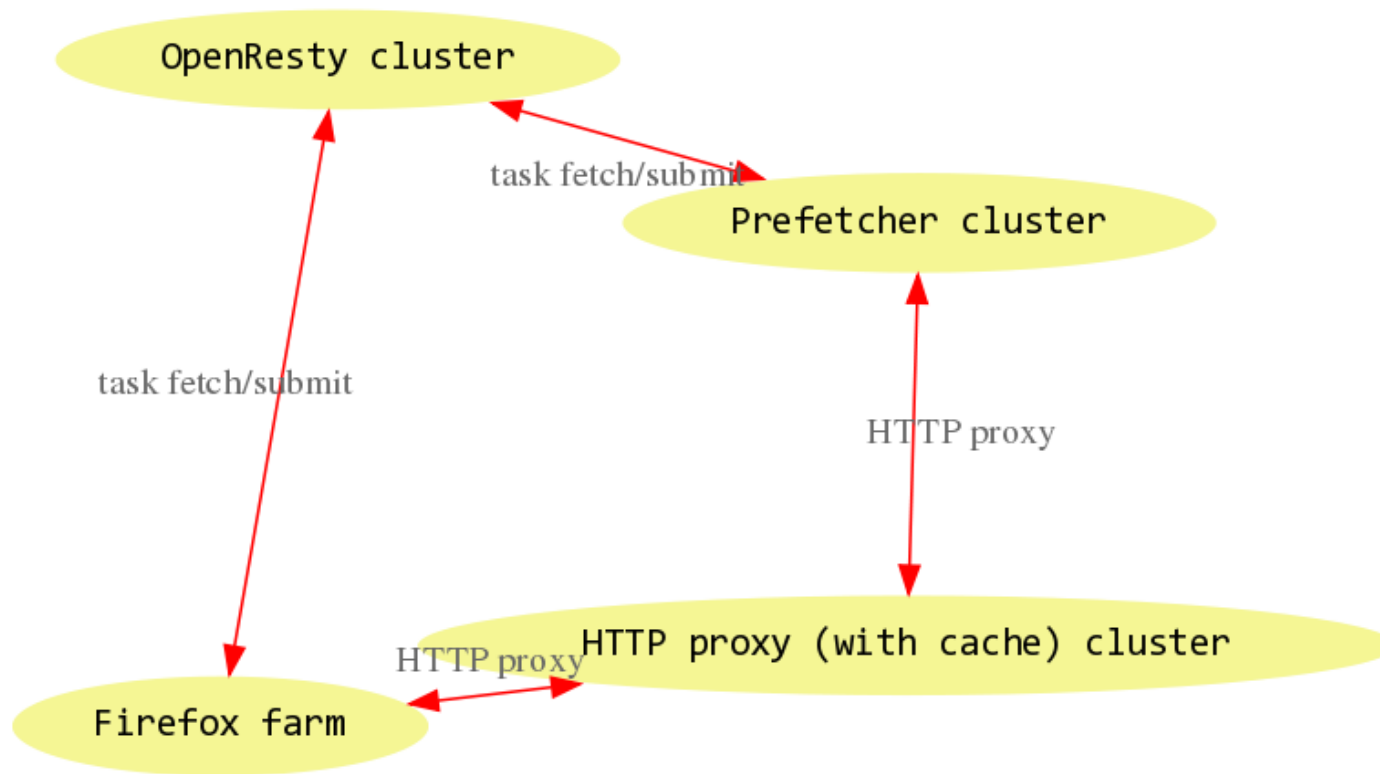
章亦春 (*agentzh*)

2009.2

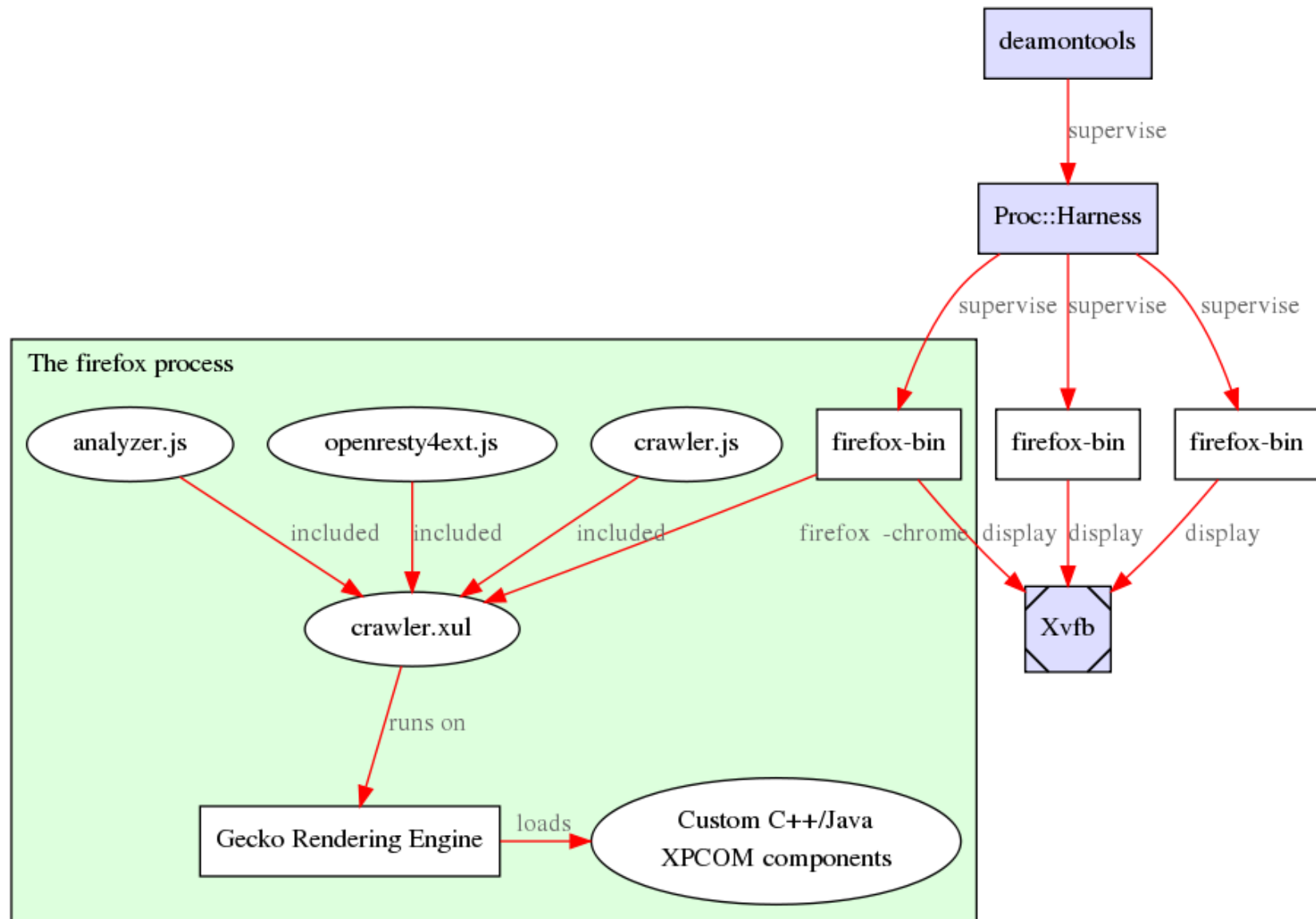
"How about using **Firefox** in a crawler *cluster*?"

"Man, you're **crazy**!"

- ✓ We're running **24** headless firefox processes on **8** production machines (Linux) and their load is around **3.0**.
- ✓ We get **100,000** web pages crawled and analyzed by my our Firefox cluster **every hour**.



Architecture Diagram for My Firefox Cluster



Firefox Farm Internals

☆ We use Firefox **extensions** to *control* Firefox's Gecko *from inside* rather than talk to it from outside.

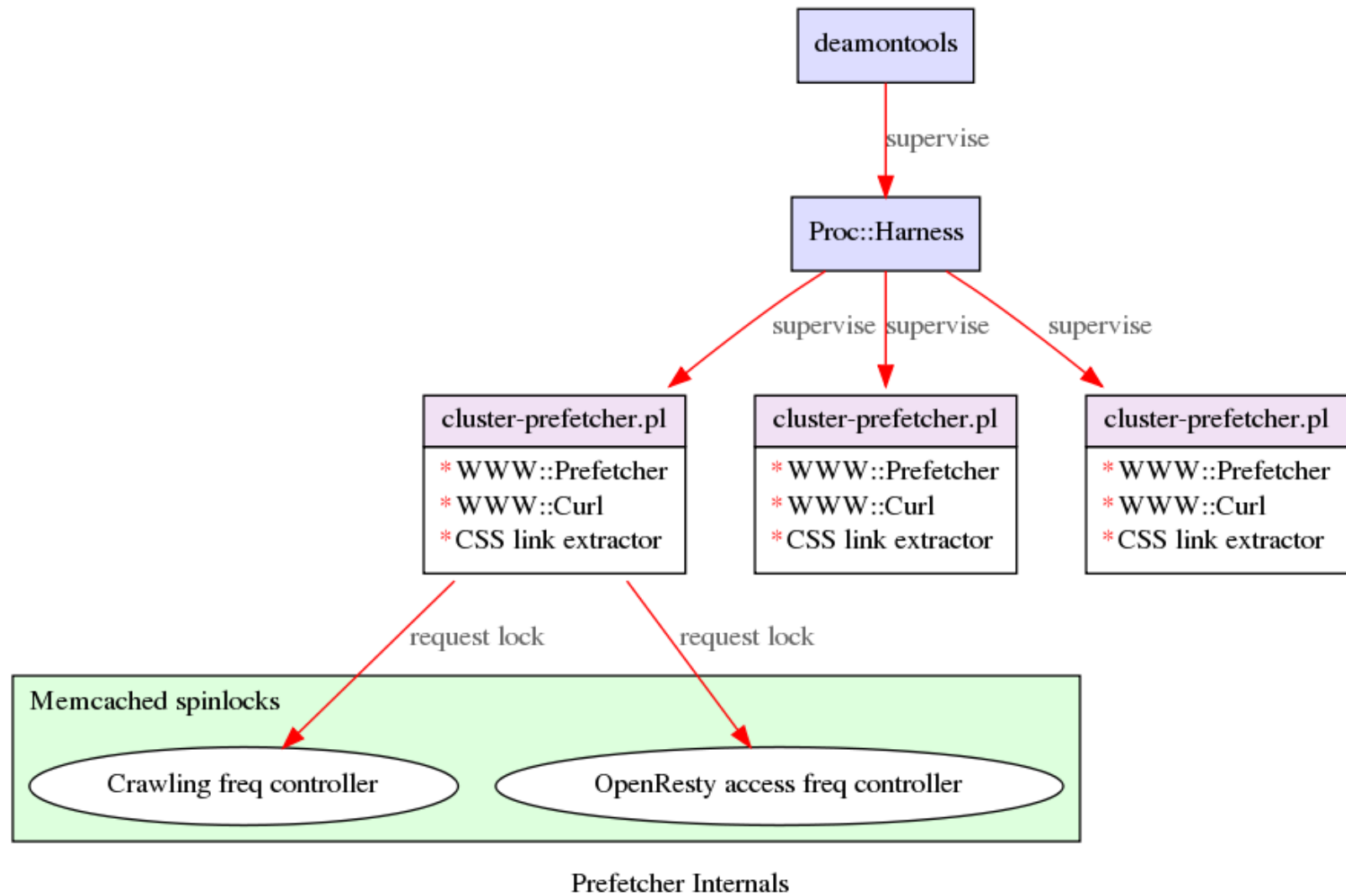
```
/* crawler.js */  
var browser = document.getElementById('my-browser');  
var browserListener = new BrowserListener(browser);  
browserListener.register();  
var openresty = new OpenResty.Client(  
    {  
        server: 'http://api.openresty.org',  
        user: 'listhunter.Firefox'  
    }  
);  
openresty.callback = doTasks;  
openresty.get('/=/view/FirefoxGetTasks/count/200');
```

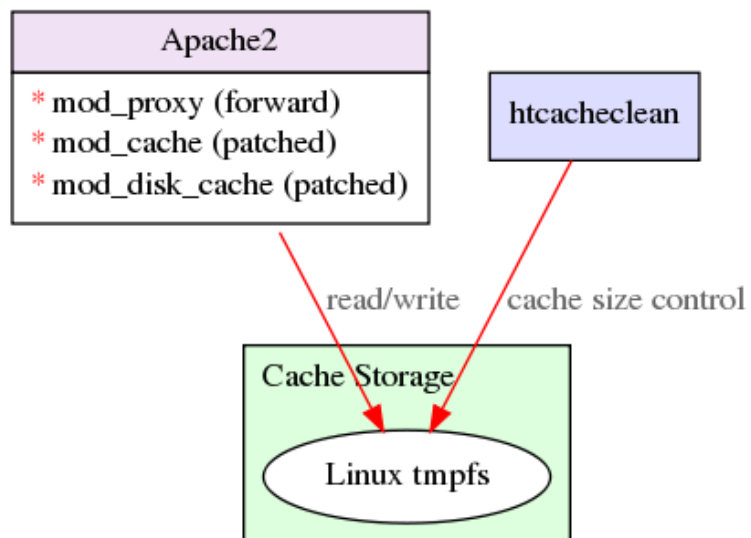
```
function doTasks(tasks, ind) {  
    if (ind == null) ind = 0;  
    var task = tasks[ind];  
    if (task == null) return;  
    browserListener.loadPage(  
        function (url, done) {  
            if (done) {  
                analyze(browser.contentDocument);  
            }  
            doTasks(tasks, ind + 1);  
        },  
        3 /* timeout in sec */  
    );  
}
```

😊 We did **NOT** patch Firefox
with only two small **exceptions**:

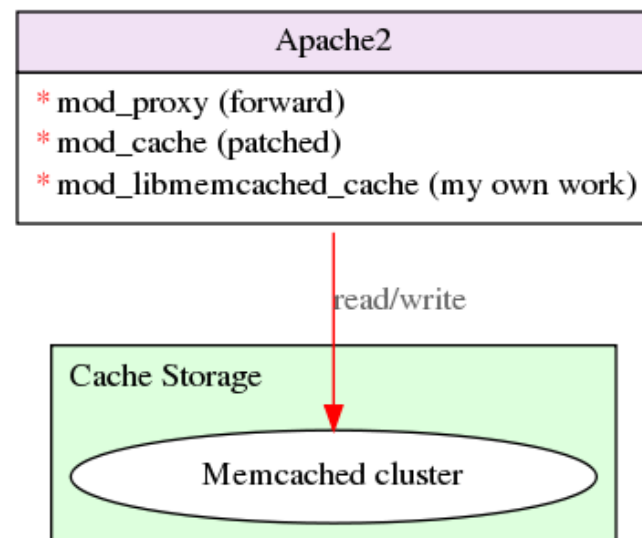
- ➔ Redirect **Error Console** outputs to *stderr*
- ➔ Ignore **CSS MIME** type *mismatch*

☆ The **prefetchers** *prefetch* the web page content via the **HTTP proxy** with cache so that **Firefox** can load stuffs from the *cache* directly.



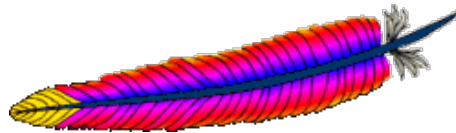


HTTP Proxy (with cache) Internals (Version 1)

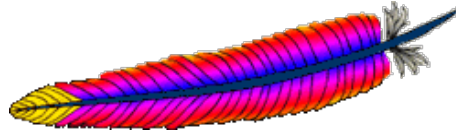


HTTP Proxy (with cache) Internals (Version 2)

😊 I added an *OverrideExpire* config directive to `mod_cache` so that it *forgets* everything about RFC.

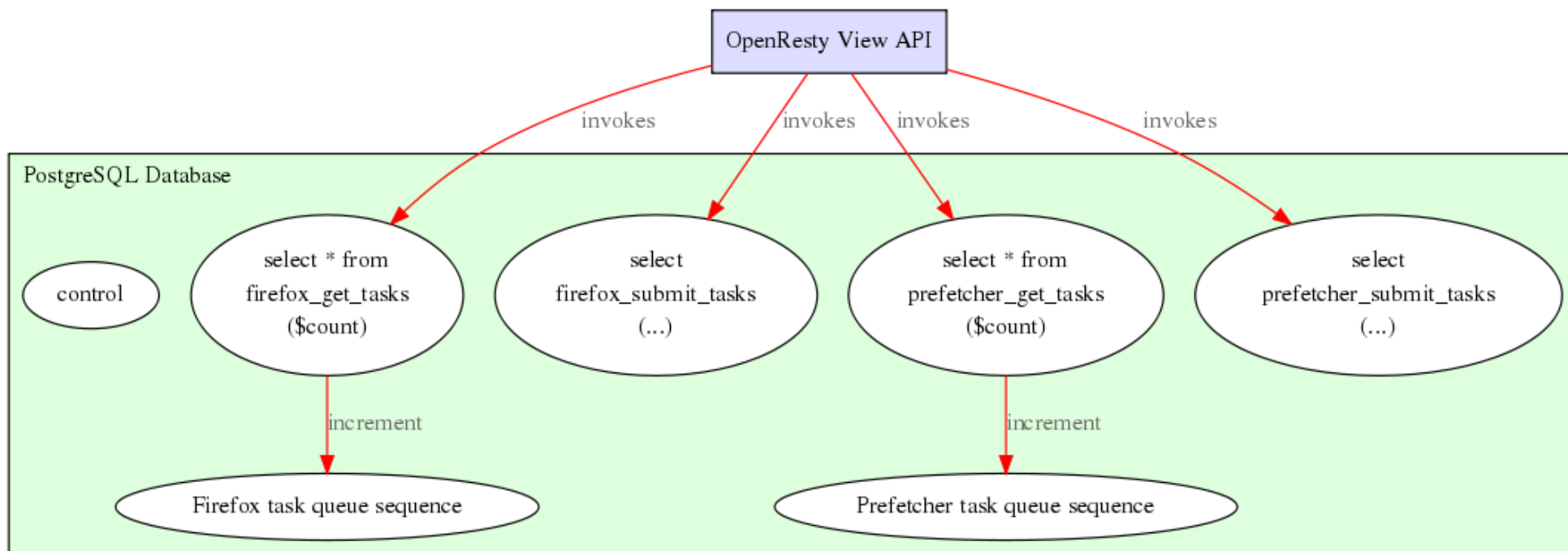


😊 I implemented a `mod_libmemcached_cache` module
so that we can have *distributive* cache storage for `mod_cache`



Sample benchmark with 59 URLs, 200 currency

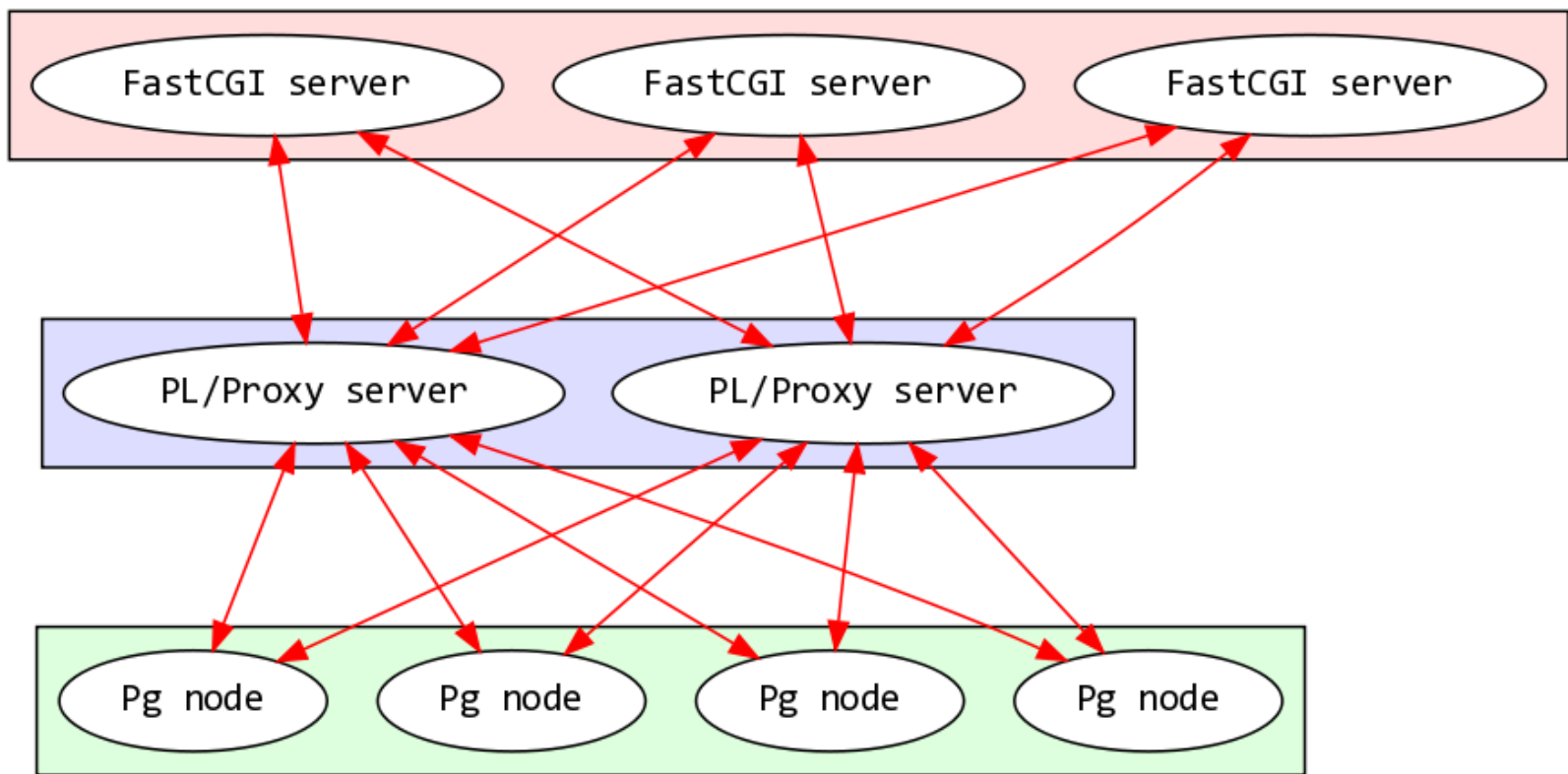
mod_disk_cache + SATA disk	200 ~ 300 QPS
mod_disk_cache + tmpfs	400 ~ 500 QPS
mod_libmemcached_cache	2200+ QPS



OpenResty (for Firefox Cluster) Internals

😊 OpenResty is a **REST** wrapper for PostgreSQL.
It is trivial to *expose* PL/PgSQL functions/stored procedures
to the outside world via *web services* without loosing security.





OpenResty Cluster Architecture

List Hunter

- Is the web page a **list page** or a **content page**?
- Extract links in the "*main list*" in list pages.

List Hunter

Address

Browser

Test Suite

Preferences

折扣信息 | 门票预定

资讯

折扣信息 | 广州折扣 | 北京折扣 | 上海折扣 | 深圳折扣 | 免费试用 | 品牌展示

您的位置：折扣信息网|我爱打折|折扣信息|折扣优惠|折扣信息|优惠券 >>> 深圳折扣

全部

 搜

资讯列表

SPORT100深圳万象城店单柜消费满520元加29.9得影券一张(2月13日-2月15日) (02-12 08:02)

SPORT100深圳南山东京基店情人节促销海报(2月13日-2月15日) (02-12 08:02)

"2009深圳欢乐谷情人节元宵节优惠门票-浪漫情人结系列活动" (02-06 09:02)

岁宝百货店庆优惠促销信息(1.17-22) (01-21 10:01)

岁宝百货景田店7周年庆折扣促销信息2009-01-17至2009-01-22 (01-14 10:01)

崇尚百货四周年店庆折扣促销信息 (01-10 07:01)

天虹商场 流行百货7折起，折后再买100送50消费卡(1.9-15) (01-10 06:01)

龙岗新城天虹：元旦·喜年 买100送100折扣券(12.31-1.3) (12-31 11:12)

天虹百货09元旦折扣促销信息：流行百货买100送100元折扣券(12.31-1.4) (12-31 10:12)

深圳好友多2009元旦折扣促销海报（附新春抵用券） (12-28 08:12)

东门天虹圣诞节折扣促销信息 买200送230(12.25止) (12-23 08:12)

常平天虹:200换400限时消费卡(12.19-21) (12-19 10:12)

REEL时尚生活百货圣诞促销信息 秋冬新品低至5折 (12-19 10:12)

岁宝百货三周年店庆折扣促销信息(12.15-28) (12-16 09:12)

深圳折扣

深圳折扣促销信息

月度关注热点

岁宝百货景田店7周年庆折扣促销信息2009-01-17至2009-01-22

2009深圳欢乐谷情人节元宵节优惠门票-浪漫情人结系列活动

岁宝百货店庆优惠促销信息(1.17-22)

SPORT100深圳南山东京基店情人节促销海报(2月13日-2月15日)

SPORT100深圳万象城店单柜消费满520元加29.9得影券一张(2月13日-2月15日)

Go!

Lists

Text

Debugger

body>div#wrap>div#head...
body>div#wrap>div.cont...
body>div#wrap>div.cont...

Settings

☒ Show recognized lists

☒ Show central box

Is this a list page?

YES

Lists in Central Area ratio

0.809

Texts in Central Area ratio

0

Total List Area ratio

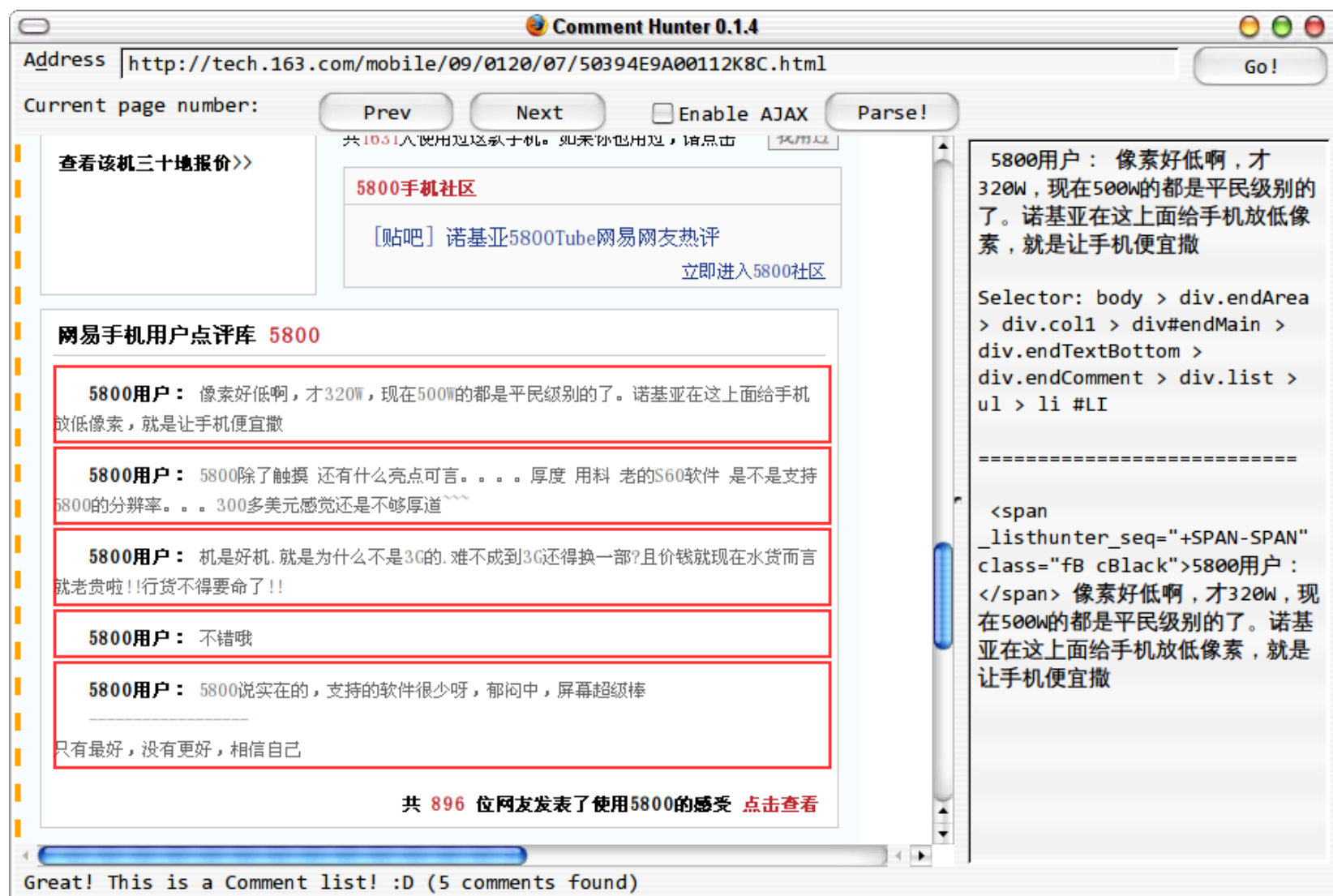
0.602

Total Text Area ratio

0

Comment Hunter

- ➔ Extract user **comments** from **arbitrary** web pages



Test results from our surfer girls
(with 100 random Chinese commercial sites):

Test results from our surfer girls
(with 100 random Chinese commercial sites):

Precision ratio: 97.6%

Test results from our surfer girls
(with 100 random Chinese commercial sites):

Precision ratio: 97.6%

Recall ratio: 91.2.%

😊 **Vision**-based filters to rule out
non-comment lists



element.offsetWidth * **element.offsetHeight** // node area

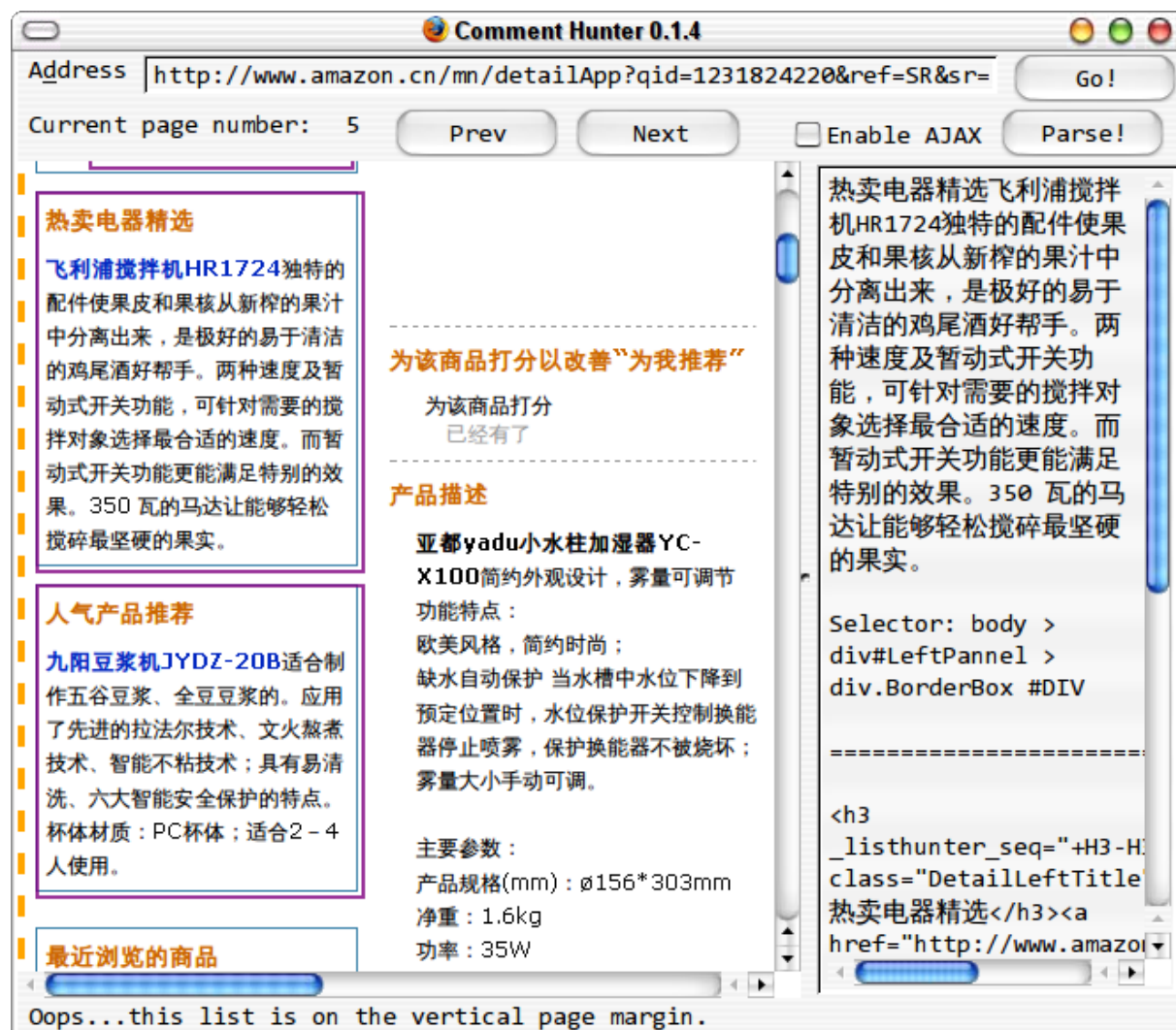
element.offsetWidth / **element.offsetHeight** // node shape

// x coordinate of element's left-upper corner

element.offsetLeft + absolute x coordiate of **element.offsetParent**

// y coordinate of element's left-upper corner

element.offsetTop + absolute y coordiate of **element.offsetParent**



Comment Hunter 0.1.4

Address Go!

Current page number: 5

Prev Next ☐ Enable AJAX Parse!

热卖电器精选

飞利浦搅拌机HR1724独特的配件使果皮和果核从新榨的果汁中分离出来，是极好的易于清洁的鸡尾酒好帮手。两种速度及暂动式开关功能，可针对需要的搅拌对象选择最合适的速度。而暂动式开关功能更能满足特别的效果。350 瓦的马达让能够轻松搅碎最坚硬的果实。

人气产品推荐

九阳豆浆机JYDZ-20B适合制作五谷豆浆、全豆豆浆的。应用了先进的拉法尔技术、文火熬煮技术、智能不粘技术；具有易清洗、六大智能安全保护的特点。杯体材质：PC杯体；适合2-4人使用。

最近浏览的商品

为该商品打分以改善“为我推荐”

为该商品打分
已经有了

产品描述

亚都yadu小水柱加湿器YC-X100简约外观设计，雾量可调节

功能特点：
欧美风格，简约时尚；
缺水自动保护 当水槽中水位下降到预定位置时，水位保护开关控制换能器停止喷雾，保护换能器不被烧坏；
雾量大小手动可调。

主要参数：
产品规格(mm)：ø156*303mm
净重：1.6kg
功率：35W

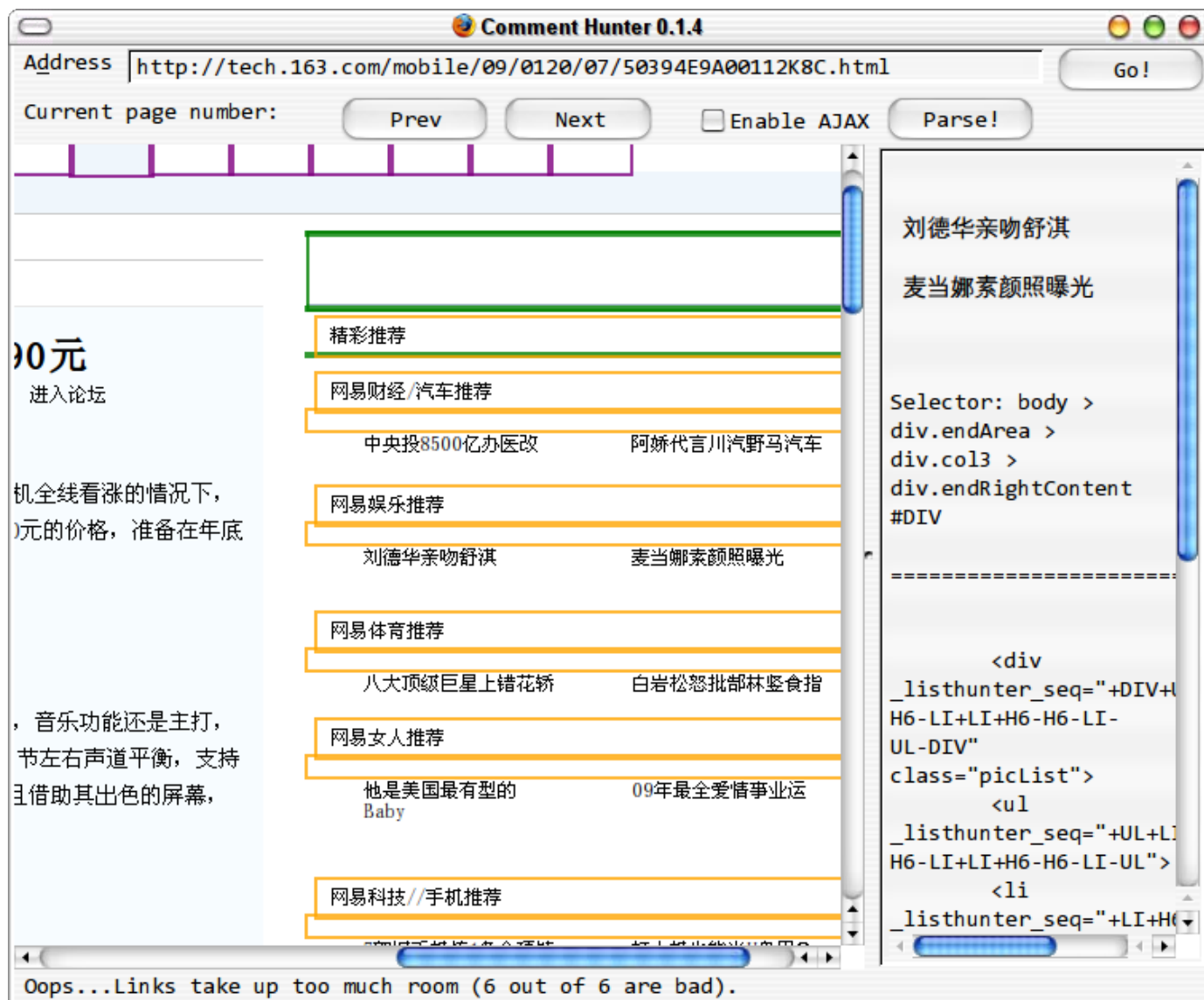
热卖电器精选飞利浦搅拌机HR1724独特的配件使果皮和果核从新榨的果汁中分离出来，是极好的易于清洁的鸡尾酒好帮手。两种速度及暂动式开关功能，可针对需要的搅拌对象选择最合适的速度。而暂动式开关功能更能满足特别的效果。350 瓦的马达让能够轻松搅碎最坚硬的果实。

Selector: body >
div#LeftPannel >
div.BorderBox #DIV

<h3
_listhunter_seq="+H3-H
class="DetailLeftTitle
热卖电器精选</h3><a
href="http://www.amazon

Oops...this list is on the vertical page margin.





☺ *Ranking testing* is *expensive*
but *necessary* for the last filter

Comment Hunter 0.1.4

Address Go!

Current page number: Prev Next ☐ Enable AJAX Parse!

成，随机附赠600万像素摄像头。网络方面，支持高速的3.1Gbps HSPA网络。诺基亚5800XpressMusic配备了320万像素的卡尔蔡司认证的AF镜头，最大可以拍摄2048×1536的照片和VGA分辨率的视频，支持两颗LED闪光灯。

规格参数	
型号：	5800XpressMusic
网络制式：	WCDMA(3G), GSM 850/900/1800/1900
体积：	111×51.7×15.5mm
重量：	104克
主屏参数：	640×360像素16:9长阔比例触摸屏，触控技术
系统：	Symbian 9.4, Series 60第5版
摄像头：	内置卡尔蔡司认证镜头
传感器类型：	CMOS传感器
像素：	
铃声：	可选MP3、WAV、AAC或和弦Midi铃声等格式
蓝牙：	支持，蓝牙v2.0, A2DP蓝牙立体声
WLAN(WiFi)：	

网络制式：
WCDMA(3G), GSM 850/900
/1800/1900

Selector: body >
div.endArea > div.col1
> div#endMain >
div#endText >
article_content > table
> tbody > tr #TR

=====

<td
_listhunter_seq="+TD-TD"
bgcolor="#f2f2f2"
height="30" width="93">
网络制式：</td>
<td
_listhunter_seq="+TD-TD"
bgcolor="#f2f2f2"
height="30"
width="354">WCDMA(3G), GSM
850/900/1800/1900</td>

Oops...failed text keywords ranking tests: 0 < 1.5

♡ Perl's Test::Simple *love* for
extension JavaScript



List Hunter

Address

http://www.17u.com/price_report/price_report_290138.html

Browser

Test Suite

Preferences

Run Regression

Run List Pages

Run Text Pages

Stop

Run custom URLs

Lists

Text

Debu

body>form#aspr

body>form#aspr

body>form#aspr

Settings

☒ Show recogn:

☒ Show centra:

Is this a list ;

YES

Lists in Centra:

0.749

Texts in Centra:

0

Total List Area

```

1..80
# Regression testing started with Firefox 3.0.6.
ok 1 - Result count ok - 5 - http://www.zhekou4u.com/?action-category-catid-93
ok 2 - Result content ok - 5 - http://www.zhekou4u.com/?action-category-catid-93
ok 3 - List area ratio ok - http://www.zhekou4u.com/?action-category-catid-93
ok 4 - Text area ratio ok - http://www.zhekou4u.com/?action-category-catid-93
ok 5 - Result count ok - 5 - http://www.zhekou4u.com/?action-category-catid-92
ok 6 - Result content ok - 5 - http://www.zhekou4u.com/?action-category-catid-92
ok 7 - List area ratio ok - http://www.zhekou4u.com/?action-category-catid-92
ok 8 - Text area ratio ok - http://www.zhekou4u.com/?action-category-catid-92
ok 9 - Result count ok - 5 - http://www.baidu.com/s?wd=perl
ok 10 - Result content ok - 5 - http://www.baidu.com/s?wd=perl
ok 11 - List area ratio ok - http://www.baidu.com/s?wd=perl
ok 12 - Text area ratio ok - http://www.baidu.com/s?wd=perl
ok 13 - Result count ok - 5 - http://www.google.com/search?hl=en&q=Perl&btnG=Google+Sea
ok 14 - Result content ok - 5 - http://www.google.com/search?hl=en&q=Perl&btnG=Google+S
ok 15 - List area ratio ok - http://www.google.com/search?hl=en&q=Perl&btnG=Google+Sear
ok 16 - Text area ratio ok - http://www.google.com/search?hl=en&q=Perl&btnG=Google+Sear
ok 17 - Result count ok - 5 - http://www.zhekou4u.com/?action-viewnews-itemid-3658
ok 18 - Result content ok - 5 - http://www.zhekou4u.com/?action-viewnews-itemid-3658
ok 19 - List area ratio ok - http://www.zhekou4u.com/?action-viewnews-itemid-3658
ok 20 - Text area ratio ok - http://www.zhekou4u.com/?action-viewnews-itemid-3658
ok 21 - Result count ok - 5 - http://www.5loff.com/inc_html/inc10363.html
ok 22 - Result content ok - 5 - http://www.5loff.com/inc_html/inc10363.html
ok 23 - List area ratio ok - http://www.5loff.com/inc_html/inc10363.html
ok 24 - Text area ratio ok - http://www.5loff.com/inc_html/inc10363.html
ok 25 - Result count ok - 3 - http://www.5loff.com/inc_html/inc10363.html
ok 26 - Result content ok - 3 - http://www.5loff.com/inc_html/inc10363.html
ok 27 - List area ratio ok - http://www.5loff.com/inc_html/inc10363.html
ok 28 - Text area ratio ok - http://www.5loff.com/inc_html/inc10363.html
# timed out...
# Completely failed to send requests. Maybe proxy broken?
ok 29 - Result count ok - 3 - http://www.perl.org/
ok 30 - Result content ok - 3 - http://www.perl.org/
ok 31 - List area ratio ok - http://www.perl.org/

```

List Hunter

Addresshttp://www.17u.com/price_report/price_report_290138.html

Browser

Test Suite

Preferences

Lists

Text

Debug

Run Regression

Run List Pages

Run Text Pages

Stop

Run custom URLs

```
#      '0.15'
ok 61 - Result count ok - 3 - http://www.xiaoxishu.com/5159/140201-409-3-1.html
ok 62 - Result content ok - 3 - http://www.xiaoxishu.com/5159/140201-409-3-1.html
not ok 63 - List area ratio ok - http://www.xiaoxishu.com/5159/140201-409-3-1.html
#      Failed test
#      '0.66'
#      ne
#      '0.77'
ok 64 - Text area ratio ok - http://www.xiaoxishu.com/5159/140201-409-3-1.html
ok 65 - Result count ok - 3 - http://zz.5ljz.cn/yellowpage_dangan.asp?vipname=duzheweng
ok 66 - Result content ok - 3 - http://zz.5ljz.cn/yellowpage_dangan.asp?vipname=duzheweng
not ok 67 - List area ratio ok - http://zz.5ljz.cn/yellowpage_dangan.asp?vipname=duzheweng
#      Failed test
#      '0.15'
#      ne
#      '0.04'
not ok 68 - Text area ratio ok - http://zz.5ljz.cn/yellowpage_dangan.asp?vipname=duzheweng
#      Failed test
#      '0.74'
#      ne
#      '0.87'
ok 69 - Result count ok - 3 - http://www.5lyala.com/list/list_678_1.Html
ok 70 - Result content ok - 3 - http://www.5lyala.com/list/list_678_1.Html
not ok 71 - List area ratio ok - http://www.5lyala.com/list/list_678_1.Html
#      Failed test
#      '0.03'
#      ne
#      '0'
not ok 72 - Text area ratio ok - http://www.5lyala.com/list/list_678_1.Html
#      Failed test
#      '0.32'
#      ne
#      '0.51'
ok 73 - Result count ok - 3 - http://www.17u.com/price_report/price_report_290138.html
ok 74 - Result content ok - 3 - http://www.17u.com/price_report/price_report_290138.html
not ok 75 - List area ratio ok - http://www.17u.com/price_report/price_report_290138.html
#      Failed test
```

body>form#aspne

body>form#aspne

body>form#aspne

Settings

☒ Show recogni:

☒ Show central

Is this a list p:

YES

Lists in Central

0.749

Texts in Central

0

Total List Area

0.305

```
agentz@agentz-office: ~/git/xhunter/ListHunter
File Edit View Terminal Tabs Help

MOZ_NO_REMOTE=1 firefox -jsconsole -P dev23 -chrome chrome://listhunter/content/test.xul
[2009-1-12 12:51:10] Using proxy agentzh.org with port 1080
[2009-1-12 12:51:10] [info] List Hunter Test 0.3.9 started.
1..80
# Regression testing started with Firefox 3.1b3pre.
# timed out...
not ok 1 - Result count ok - 5 - http://www.zhekou4u.com/?action-category-catid-93
#   Failed test
#     '4'
#       ne
#     '3'
not ok 2 - Result content ok - 5 - http://www.zhekou4u.com/?action-category-catid-93
#   Failed test
#     'body>div#wrap>div#header>div#menu>ul
body>div#wrap>div.contentR>div.sideR>div.block>d1
body>div#wrap>div.contentR>div.mainarea>div.block>ul.newslist
body>div#wrap>div#header>div
'
#       ne
#     'body>div#wrap>div#header>div#menu>ul
body>div#wrap>div.contentR>div.sideR>div.block>d1
body>div#wrap>div.contentR>div.mainarea>div.block>ul.newslist
'
not ok 3 - List area ratio ok - http://www.zhekou4u.com/?action-category-catid-93
#   Failed test
#     '0.47'
#       ne
#     '0.96'
ok 4 - Text area ratio ok - http://www.zhekou4u.com/?action-category-catid-93
ok 5 - Result count ok - 5 - http://www.zhekou4u.com/?action-category-catid-92
ok 6 - Result content ok - 5 - http://www.zhekou4u.com/?action-category-catid-92
ok 7 - List area ratio ok - http://www.zhekou4u.com/?action-category-catid-92
```

```
Test.GuiMode = false;  
Test.plan(2 * list.length);  
for (var i = 0; i < list.length; i++) {  
    Test.ok(i >= 0, 'i is always non-negative');  
    Test.is(i * 2, i + i, 'i x 2 = i + i');  
}  
Test.summary();
```

Comment Hunter: JavaScript & Perl code *only*



```
$ find js -name '*.js' | xargs wc -l
```

```
27 js/cli-prefs.js
```

```
332 js/main.js
```

```
3 js/test-data.js
```

```
374 js/haiway-miner.js
```

```
26 js/box.js
```

```
32 js/util.js
```

```
7 js/env.js
```

```
62 js/benchmark-timer.js
```

```
18 js/samples.js
```

```
160 js/test.js
```

```
329 js/filters.js
```

```
151 js/browser-listener.js
```

```
137 js/test-more.js
```

```
1658 total
```

```
$ find lib -name '*.pm' | xargs wc -l
```

```
39 lib/CommentHunter/View/Test.pm
```

```
106 lib/CommentHunter/View/Main.pm
```

```
34 lib/CommentHunter/View/Overlay.pm
```

```
52 lib/CommentHunter/App.pm
```

```
231 total
```

Powered by my XUL::App framework



A *Hello World* extension in XUL::App



```
# File lib/HelloWorld/App.pm
package HelloWorld::App;
our $$VERSION; BEGIN { $$VERSION = '0.01' }
use XUL::App::Schema;
use XUL::App schema {
    xulfile 'hellowin.xul' =>
        generated from 'HelloWorld::View::HelloWin',
        includes qw( jquery.js hellowin.js );

    xpifile 'helloworld.xpi' =>
        name is 'HelloWorld',
        id is 'helloworld@agentz.agentz-office', # FIXME
        version is $$VERSION,
        targets {
            Firefox => ['2.0' => '3.0a5'], # FIXME
        },
        creator is 'The HelloWorld development team',
        ...
    }
}
```

Ruby: "We have this *gorgeous* syntax!"

Perl: "Hey, we do **as well** ;)"

```
# File lib/HelloWorld/View/HelloWin.pm
package HelloWorld::View::HelloWin;
use base 'XUL::App::View::Base';
use Template::Declare::Tags 'XUL';
template main => sub {
    show 'header'; # from XUL::App::View::Base
    window {
        attr {
            id => "helloworld-hellowin",
            xmlns => $::XUL_NAME_SPACE,
            title => _('Hello World ') .
                $$HelloWorld::App::VERSION,
            ...
        }
        label { _("Hello, world!") }
    }
    ...
}
```

\$ *xulapp* bundle .

Writing file helloworld.xul

Writing bundle file ./helloworld.xpi

\$

Our **helloworld.xpi** bundle ➡

- ✓ contains **0** Perl
- ✓ has **0** dependencies
(except Firefox itself)
- ✓ runs happily *everywhere*
(Win32, Linux, Mac, and etc.)

The *future*

- ✓ *Opensource* everything we have :)
- ✓ More hunters, more fun: *Table Hunter*, *Title Hunter*,
Ranking Hunter, *Ads Hunter*, *Summary Hunter*, ...
- ✓ Automatic C/C++ **XPCOM** wrapper generator for XUL::App.
- ✓ Bring Firefox extension love to Apple's *WebKit*
(A WebKit crawler cluster?)

😊 *Any questions?* 😊

