Back-Office Web Traffic on the Internet

Enric Pujol  
TU-Berlin

Philipp Richter  
TU-Berlin

Balakrishnan Chandrasekaran  
Duke University

Georgios Smaragdakis  
MIT / TU-Berlin / Akamai

Anja Feldmann  
TU-Berlin

Bruce Maggs  
Duke University / Akamai

Keung-Chi Ng  
Akamai

IMC 2014  
Vancouver, BC, CANADA  
November 5-7, 2014
The Web for an end user

Front-office Web traffic:
Web traffic between end users and servers

End user
HTTP GET
Search engine
HTTP GET
CDN
HTTP GET
AdPublisher

The front-office
Behind the scenes...

Back-office Web traffic:
Machine-to-machine Web traffic

The front-office

The back-office
Search engines: **crawlers**

The diagram illustrates the process of how search engines interact with content. It shows the flow of data from the front-office to the back-office, with specific components such as search engines, crawlers, end users, CDN, and AdPublisher. The diagram also highlights the use of HTTP GET requests to fetch content.
Content delivery: proxies

The front-office

The back-office
AdExchanges: real-time bidding

The back-office

The front-office

6.11.2014
Internet Measurement Conference 2014
Agenda

1. Introduction
2. Methodology and datasets
3. Characteristics
   1. Traffic
   2. Patterns
   3. Inter-domain perspective
4. CDN back-office traffic
5. The end-user perspective
6. Summary and implications
## Vantage points (VP)

<table>
<thead>
<tr>
<th>Type</th>
<th>VP</th>
<th>Daily traffic</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>IXPs</td>
<td>L-IXP</td>
<td>11,900 TB</td>
<td>SFlow (1/16K)</td>
</tr>
<tr>
<td></td>
<td>M-IXP</td>
<td>1,580 TB</td>
<td></td>
</tr>
<tr>
<td>Transit</td>
<td>BBone-1</td>
<td>40 TB</td>
<td>Packet sampled (1/1K)</td>
</tr>
<tr>
<td></td>
<td>BBone-2</td>
<td>70 TB</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>CDN</td>
<td>350 TB</td>
<td>5 locations</td>
</tr>
<tr>
<td>Eyeballs</td>
<td>RBN</td>
<td>35 TB</td>
<td>Packet dumps</td>
</tr>
</tbody>
</table>

**Diverse vantage points:** multiple perspectives
Candidate IPs for the back-office

Dual role IPs are prime candidates
Candidate IPs for the back-office

Dual role IPs are prime candidates
Candidate IPs for the back-office

Many requests to many servers

Heavy hitter IPs are also prime candidates
Candidate IPs for the back-office

- Many requests to many servers

- Many requests to a few servers

Heavy hitter IPs are also prime candidates
Sources of back-office Web traffic
Sources of back-office Web traffic

- Heavy hitters
- Dual role IPs

L-IXP

- Crawling?
- Real-time bidding?
# Dual-role IPs: active measurements

<table>
<thead>
<tr>
<th>L-IXP</th>
<th>Client only (%)</th>
<th>Server only (%)</th>
<th>Dual-role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passive</td>
<td>96.90</td>
<td>2.74</td>
<td>0.36</td>
</tr>
<tr>
<td>Passive+Active</td>
<td>93.85</td>
<td>2.74</td>
<td>3.40</td>
</tr>
</tbody>
</table>

ZMap project: Internet-wide scan of Web Servers (scans.io)

**Observations:**

1. Most IPs have only client behavior
2. Many servers also show client behavior

Active measurements augment the number of servers
Candidates: **manual classification**

Crawlers:
- Reverse DNS + Origin AS

3.9K IPs, 74% in 2 orgs
Candidates: manual classification

Crawlers:
  • Reverse DNS + Origin AS

Auctioneers:
  • URL + Origin AS

3.9K IPs, 74% in 2 orgs

316 IPs, 4 orgs
Candidates: *manual classification*

Crawlers:
• Reverse DNS + Origin AS

Auctioneers:
• URL + Origin AS

Content Delivery Proxies:
• Origin AS + Reverse DNS (for caches)

- 3.9K IPs, 74% in 2 orgs
- 316 IPs, 4 orgs
- 36K IPs, 8 orgs
Candidates: manual classification

Crawlers:
- Reverse DNS + Origin AS

Auctioneers:
- URL + Origin AS

Content Delivery Proxies:
- Origin AS + Reverse DNS (for caches)

Other:
- Rest of dual-role IPs

- 3.9K IPs, 74% in 2 orgs
- 316 IPs, 4 orgs
- 36K IPs, 8 orgs
- 151K IPs, mostly in cloud prov.
Agenda

1. Introduction
2. Methodology and datasets
3. Characteristics
   1. Traffic
   2. Patterns
   3. Inter-domain perspective
4. CDN back-office traffic
5. The end-user perspective
6. Summary and implications
Traffic

At least 10% in our VPs

IXPs: from 10% to 20%

Transit links: different obs.
Traffic: Contribution per class

<table>
<thead>
<tr>
<th>L-IXP</th>
<th>CDPs</th>
<th>Auctioneers</th>
<th>Crawlers</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bytes</td>
<td>12.1 %</td>
<td>1.1 %</td>
<td>10.3 %</td>
<td>76.5 %</td>
</tr>
<tr>
<td>Requests</td>
<td>11.8 %</td>
<td>22.5 %</td>
<td>15.1 %</td>
<td>50.6 %</td>
</tr>
</tbody>
</table>

Observations:

1. CDPs          big players – significant share
2. Real-time bidding many but small transactions
3. Crawlers      a few orgs – significant share
4. Other         cloud service providers

All classes contribute. More to discover
Traffic patterns: bytes

% back-office Web traffic increases during off hours in IXPs
Traffic patterns: requests

Observations:

1. A multiplicative factor of human activity (e.g., RTB)
2. Non-human triggered activity (e.g., crawlers)
Inter-domain perspective

Back-office traffic appears in many peering links

Top 10 traffic carrying links:
- 4 x Cloud – Content
- 3 x Search – Hosters
- 2 x CDN – Content
- 1 x Content – Advertisement
Agenda

1. Introduction
2. Methodology and datasets
3. Characteristics
   1. Traffic
   2. Patterns
   3. Inter-domain perspective
4. CDN back-office traffic
5. The end-user perspective
6. Summary and implications
A CDN perspective

Three sub-classes of back-office traffic
A CDN perspective

Public: front-end back-end over the Internet
A CDN perspective

Private: within same cluster
A CDN perspective

Origin: inter-organization over the Internet
Back-office per location

CDNs heavily rely on back-office traffic
Agenda

1. Introduction
2. Methodology and datasets
3. Characteristics
   1. Traffic
   2. Patterns
   3. Inter-domain perspective
4. CDN back-office traffic
5. The end-user perspective
6. Summary and implications
The end-user perspective

Residential broadband network: backbone latency (no access)

A smaller front-office: but the back-office may be large
Summary

1. A back-office to support the Web

2. Significant traffic: bytes and requests

3. Different type of traffic patterns

4. Visible at multiple peering links

An important yet understudied class of traffic
Implications

**Feasibility to deploy new protocols:**

- It is easier to change the back office than the front office
Implications

Feasibility to deploy new protocols:
• It is easier to change the back office than the front office

Performance evaluation:
• Interactions with the back office
• More users than anticipated
Implications

Feasibility to deploy new protocols:
• It is easier to change the back office than the front office

Performance evaluation:
• Interactions with the back office
• More users than anticipated

Opportunities:
• ISPs: micro-data centers, virtualized services
• IXPs: co-location strategies
• NSPs: new services e.g., SLAs
Back-office traffic on the Internet

The back-office (some examples thereof)

The front-office

6.11.2014

Internet Measurement Conference 2014