Measuring Web Similarity from Dual-Stacked Hosts

Vaibhav Bajpai
Jacobs University, Bremen

RIPE 72, Copenhagen

Joint work with

Steffie Jacob Eravuchira
SamKnows Limited, London

Jürgen Schönwälder
Jacobs University, Bremen

Sam Crawford
SamKnows Limited, London

May 2016

Supported by:
Flamingo Project: flamingo-project.eu
Leone Project: leone-project.eu
Introduction | Research Questions

Recent work [1], [2], [3] has compared performance of dual-stacked websites over IPv4 and IPv6.

No study comparing web similarity over IPv4 / IPv6.

We want to know:

▶ How similar are webpages accessed over IPv6 to their IPv4 counterparts?
▶ What factors contribute to the dissimilarity over IPv4 and IPv6?
Introduction | Research Contributions

We measure against ALEXA top 100 dual-stacked websites.

1. **simweb**: A tool for measuring web similarity over IPv4 and IPv6.
2. Websites (27%) have some fraction of webpage elements failing over IPv6.
3. Failure rates over IPv6 are largely due to DNS resolution error on images, js and CSS.
4. Both same-origin and cross-origin sources contribute to the failure rates over IPv6.

---

*To the best of our knowledge, this is the first study to:*

- Measure webpage similarity over IPv4 and IPv6.
- Investigate IPv6 adoption that goes beyond the root page of a dual-stacked website.
Methodology
We use 2 well-known webpage complexity metrics from literature [4, 5]:

1. **Content Complexity**
   *The number & size of fetched webpage elements.*

2. **Service Complexity**
   *The number of same-origin & cross-origin sources.*
Methodology | Selection of Websites

We use the ALEXA top 100 dual-stacked websites as measurement targets [1].

1. www.google.com
2. www.facebook.com
3. www.youtube.com
4. www.yahoo.com
5. www.wikipedia.org
6. www.qq.com
7. www.blogspot.com
8. ...
The simweb test:

- runs twice (once for each AF).
- repeats every hour.
- uses user-agent string: Mozilla/4.0
Methodology | Measurement Trial

We measure from 80 dual-stacked SamKnows probes.
Data Analysis\footnote{Measurements conducted for 65 days between April 2015 and June 2015.}
Can we fetch all webpage elements over IPv6?

- 27% of websites show some rate of failure over IPv6.
- 9% exhibit more than 50% failures over IPv6.
- 6% show complete failure (0% success) over IPv6.
ALEXA top 100 dual-stacked websites:

- 6% show complete failure over IPv6.

<table>
<thead>
<tr>
<th>#</th>
<th>Webpage</th>
<th>Success Rate (%)</th>
<th>W6LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td><a href="http://www.bing.com">www.bing.com</a></td>
<td>0 100</td>
<td>✓</td>
</tr>
<tr>
<td>02</td>
<td><a href="http://www.detik.com">www.detik.com</a></td>
<td>0 100</td>
<td>✓</td>
</tr>
<tr>
<td>03</td>
<td><a href="http://www.engadget.com">www.engadget.com</a></td>
<td>0 100</td>
<td>✓</td>
</tr>
<tr>
<td>04</td>
<td><a href="http://www.nifty.com">www.nifty.com</a></td>
<td>0 100</td>
<td>✓</td>
</tr>
<tr>
<td>05</td>
<td><a href="http://www.qq.com">www.qq.com</a></td>
<td>0 100</td>
<td></td>
</tr>
<tr>
<td>06</td>
<td><a href="http://www.sakura.ne.jp">www.sakura.ne.jp</a></td>
<td>0 100</td>
<td></td>
</tr>
</tbody>
</table>

- Metrics that measure IPv6 adoption should account for changes in IPv6-readiness.
Results | Causality Analysis

*Where in the network does the failure occur?*

- CURLE_COULDNT_RESOLVE_HOST is the major contributor to failure rates.
- AAAA entries missing for these webpage elements in the DNS.
Results | Causality Analysis

Which type of objects fail more than others?

- image/*, */javascript, */json and */css content contribute to the majority of the failure over IPv6.
Where do the failing objects originate from?

- Both same and cross origin sources contribute to the failure of webpage elements over IPv6.
### Results | Causality Analysis

**What is failure contribution of same-origin sources?**

<table>
<thead>
<tr>
<th>#</th>
<th>Webpage</th>
<th>Same Origin (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td><a href="http://www.bing.com">www.bing.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>02</td>
<td><a href="http://www.detik.com">www.detik.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>03</td>
<td><a href="http://www.engadget.com">www.engadget.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>04</td>
<td><a href="http://www.nifty.com">www.nifty.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>05</td>
<td><a href="http://www.usps.com">www.usps.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>06</td>
<td><a href="http://www.qq.com">www.qq.com</a></td>
<td>100%</td>
</tr>
<tr>
<td>07</td>
<td><a href="http://www.sakura.ne.jp">www.sakura.ne.jp</a></td>
<td>100%</td>
</tr>
<tr>
<td>08</td>
<td><a href="http://www.comcast.net">www.comcast.net</a></td>
<td>85%</td>
</tr>
<tr>
<td>09</td>
<td><a href="http://www.yahoo.com">www.yahoo.com</a></td>
<td>83%</td>
</tr>
<tr>
<td>10</td>
<td><a href="http://www.terra.com.br">www.terra.com.br</a></td>
<td>74%</td>
</tr>
<tr>
<td>11</td>
<td><a href="http://www.marca.com">www.marca.com</a></td>
<td>70%</td>
</tr>
<tr>
<td>12</td>
<td><a href="http://www.wikimedia.org">www.wikimedia.org</a></td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.elmundo.es">www.elmundo.es</a></td>
<td>37%</td>
</tr>
<tr>
<td>14</td>
<td><a href="http://www.vk.com">www.vk.com</a></td>
<td>31%</td>
</tr>
<tr>
<td>15</td>
<td><a href="http://www.t-online.de">www.t-online.de</a></td>
<td>30%</td>
</tr>
<tr>
<td>16</td>
<td><a href="http://www.youm7.com">www.youm7.com</a></td>
<td>24%</td>
</tr>
<tr>
<td>17</td>
<td><a href="http://www.wiktionary.org">www.wiktionary.org</a></td>
<td>22%</td>
</tr>
<tr>
<td>18</td>
<td><a href="http://www.wikimedia.org">www.wikimedia.org</a></td>
<td>22%</td>
</tr>
<tr>
<td>19</td>
<td><a href="http://www.free.fr">www.free.fr</a></td>
<td>13%</td>
</tr>
<tr>
<td>20</td>
<td><a href="http://www.folha.uol.com.br">www.folha.uol.com.br</a></td>
<td>12%</td>
</tr>
<tr>
<td>21</td>
<td><a href="http://www.mozilla.org">www.mozilla.org</a></td>
<td>7%</td>
</tr>
<tr>
<td>22</td>
<td><a href="http://www.uol.com.br">www.uol.com.br</a></td>
<td>7%</td>
</tr>
<tr>
<td>23</td>
<td><a href="http://www.mobile.de">www.mobile.de</a></td>
<td>7%</td>
</tr>
<tr>
<td>24</td>
<td><a href="http://www.aol.com">www.aol.com</a></td>
<td>5%</td>
</tr>
<tr>
<td>25</td>
<td><a href="http://www.orange.fr">www.orange.fr</a></td>
<td>5%</td>
</tr>
<tr>
<td>26</td>
<td><a href="http://www.seznam.cz">www.seznam.cz</a></td>
<td>4%</td>
</tr>
<tr>
<td>27</td>
<td><a href="http://www.flipkart.com">www.flipkart.com</a></td>
<td>1%</td>
</tr>
</tbody>
</table>

- 12% of websites have more than 50% webpage elements that belong to the same origin source and fail over IPv6.
Results | Causality Analysis

What is failure contribution of cross-origin sources?

- Some of the cross-origin sources contribute to the failure of multiple websites.
Results | Causality Analysis

Which cross-origin sources span across multiple failing websites?

- **doubleclick.net** spans 5 websites with a 0.54% median contribution to failure rates.

- **creativecommons.org** has 76% median contribution to the failure rate of 3 websites.
Takeway

- Metrics that measure IPv6 adoption should account for changes in IPv6-readiness.
- Limiting to root webpage can lead to overestimation of IPv6 adoption numbers.
- Unclear whether websites with failure rates can be deemed IPv6-ready.
- Few cross-origin sources once IPv6 enabled will help large number of websites at once.

Graduating in 2016. Currently on the job market!

v.bajpai@jacobs-university.de | @bajpaivaibhav
Appendix
Introduction | Motivation

- 4/5 RIRs have exhausted available pool of IPv4 address space [6]

<table>
<thead>
<tr>
<th>RIR</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>APNIC</td>
<td>Apr’11</td>
</tr>
<tr>
<td>RIPE</td>
<td>Sep’12</td>
</tr>
<tr>
<td>LACNIC</td>
<td>Jun’14</td>
</tr>
<tr>
<td>ARIN</td>
<td>Sep’15</td>
</tr>
</tbody>
</table>

- Large IPv6 broadband rollouts\textsuperscript{2} since World IPv6 Launch Day in 2012 [7].

- Increased global adoption of IPv6 to 10.5% [8] (as seen by Google, March 2016).

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>40.49%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>27.38%</td>
</tr>
<tr>
<td>United States</td>
<td>23.62%</td>
</tr>
<tr>
<td>Germany</td>
<td>21.41%</td>
</tr>
</tbody>
</table>

\textsuperscript{2}Comcast, Deutsche Telekom AG, AT&T, Verizon Wireless, T-Mobile USA
Methodology | SamKnows wget

SamKnows [9] probes run wget\(^3\):

- DNS lookup time.
- Time to first byte.
- HTTP request time.
- Content size.
- Download speed

as a aggregated report for a website.

\(^{3}\)files.samknows.com/~gpl
We extend the SamKnows webget test to measure webpage similarity:

simweb in addition also reports:

- Content Type
- Content Size
- Resource URL
- IP endpoint
- CURL response code
- HTTP status code

for each webpage element of a website.

% SIMWEB_L=1 IPVERSION=6 webget 1 www.google.com
#: 1
version: SIMWEB.0
service: www.google.com
timestamp: 1427822156
af: 6
status: OK
curl_response_code: CURLE_OK
object_type: text/html;charset=ISO-8859-1
http_code: 200
resource_url: www.google.com
ip_endpoint: 2a00:1450:4008:801::1010;
size_bytes: 52674
#: 2
...
Results | Content Similarity

**Is there a difference in the number of fetched webpage elements?**

\[ \Delta n(u) = \frac{\hat{n}_4(u) - \hat{n}_6(u)}{\hat{n}_4(u)} \times 100\% \]

- 14% of websites exhibit dissimilarity in number.
- 6% showing more than 50% difference.

**Is there a difference in the object size of fetched webpage elements?**

\[ \Delta s(u) = \frac{\hat{s}_4(u) - \hat{s}_6(u)}{\hat{s}_4(u)} \times 100\% \]

- 94% of dual-stacked websites exhibit dissimilarity in size.
- 8% showing at least 50% difference.
Appendix | References I


