

Measuring Web Similarity from Dual-Stacked Hosts

Vaibhav Bajpai
Jacobs University, Bremen

CNSM 2016
Montréal, Canada

Joint work with

Steffie Jacob Eravuchira
SamKnows Limited, London

Jürgen Schönwälder
Jacobs University, Bremen

Sam Crawford
SamKnows Limited, London

Oct 2016

Supported by:
Flamingo Project: flamingo-project.eu
Leone Project: leone-project.eu

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

- ▶ 4/5 RIRs have *exhausted* available pool of IPv4 address space [1].

APNIC	Apr'11
RIPE	Sep'12
LACNIC	Jun'14
ARIN	Sep'15

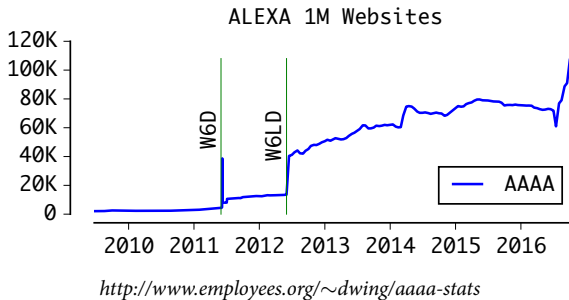
- ▶ Large IPv6 broadband rollouts¹ since World IPv6 Launch Day in 2012 [2].
- ▶ Increased global adoption of IPv6 to $\sim 14.9\%$ (native) [3] (Oct 2016).

Belgium	45.39%
United States	28.89%
Switzerland	26.73%
Germany	25.93%

¹Comcast, Deutsche Telekom AG, AT&T, Verizon Wireless, T-Mobile USA

Recent work [4], [5], [6] 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?

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.*

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

Methodology

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

SamKnows [7] probes run webget²:

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

as a aggregated report for a website.

```
% webget 1 www.google.com
version: WEBGETMT.2
endtime: 1427820219
status: OK
target: www.google.com
address: 2a00:1450:4008:801::1013
fetch_time: 145270
bytes_total: 194818
bytes_sec: 1848376
objects: 3
threads: 1
requests: 3
connections: 1
reused_connections: 2
lookups: 1
request_total_time: 128883
request_min_time: 12930
request_avg_time: 42961
request_max_time: 100458
...
```

²files.samknows.com/~gpl

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

- ▶ 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
...
```

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

We use 2 well-known webpage complexity metrics from literature [8, 9]:

1. **Content Complexity**

The number & size of fetched webpage elements.

2. **Service Complexity**

The number of same-origin & cross-origin sources.

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

- We use the ALEXA top 100 dual-stacked websites as measurement targets [4].

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. ...

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

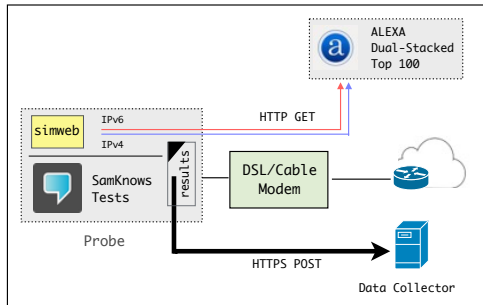
Success Rates

Causality Analysis

Takeway

The simweb test:

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



Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

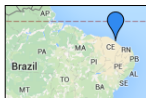
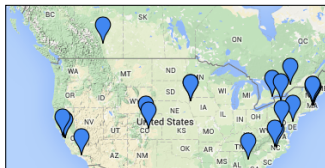
Results

Success Rates

Causality Analysis

Takeway

Methodology | Measurement Trial



NETWORK TYPE	#
RESIDENTIAL	55
NREN / RESEARCH	11
BUSINESS / DATACENTER	09
OPERATOR LAB	04
IXP	01

RIR	#
RIPE	42
ARIN	29
APNIC	07
AFRINIC	01
LACNIC	01

We measure from 80 dual-stacked SamKnows probes.

Measurement Trial

*Data Analysis*³

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

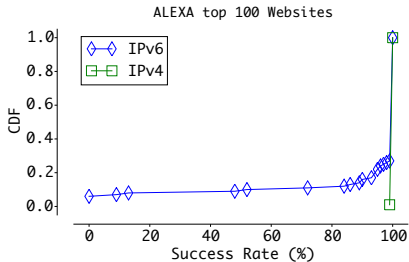
Takeway

³Measurements conducted for 65 days between April 2015 and June 2015.

Results | Success Rates

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.



#	Webpage	Success Rate (%)		W6LD
		IPv6(↓)	IPv4	
01	www.bing.com	0	100	✓
02	www.detik.com	0	100	✓
03	www.engadget.com	0	100	✓
04	www.nifty.com	0	100	
05	www.qq.com	0	100	
06	www.sakura.ne.jp	0	100	
07	www.flipkart.com	09	99	✓
08	www.folha.uol.com.br	13	100	
09	www.aol.com	48	100	✓
10	www.comcast.net	52	100	✓
11	www.yahoo.com	72	100	✓
12	www.mozilla.org	84	100	✓
13	www.orange.fr	86	100	✓
14	www.seznam.cz	89	100	✓
15	www.mobile.de	90	100	✓
16	www.wikimedia.org	90	100	
17	www.t-online.de	93	100	✓
18	www.free.fr	95	100	
19	www.usps.com	95	100	
20	www.vk.com	95	100	✓
21	www.wikipedia.org	95	100	✓
22	www.wiktionary.org	95	100	
23	www.elmundo.es	96	100	✓
24	www.uol.com.br	96	100	✓
25	www.marca.com	97	100	✓
26	www.terra.com.br	98	100	✓
27	www.youm7.com	99	100	

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

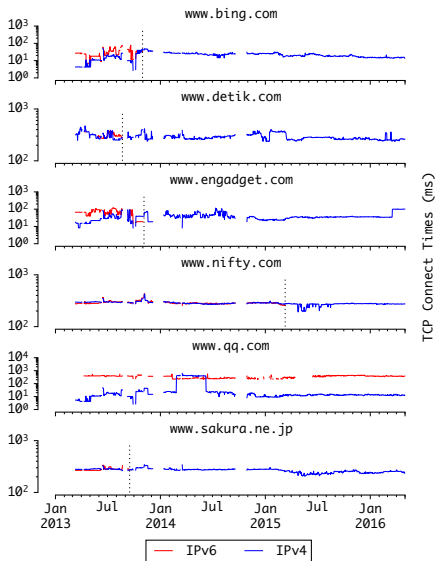
Results | Success Rates

ALEXA top 100 dual-stacked websites:

- 6% show complete failure over IPv6.

#	Webpage	Success Rate (%)		W6LD
		IPv6(↓)	IPv4	
01	www.bing.com	0	100	✓
02	www.detik.com	0	100	✓
03	www.engadget.com	0	100	✓
04	www.nifty.com	0	100	
05	www.qq.com	0	100	
06	www.sakura.ne.jp	0	100	

- Metrics that measure IPv6 adoption should account for *changes* in IPv6-readiness.



Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

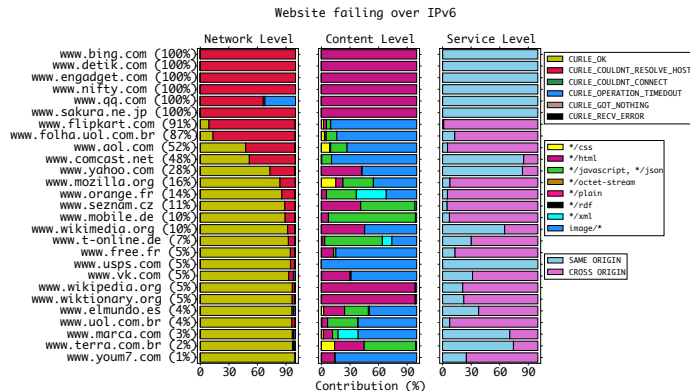
Success Rates

Causality Analysis

Takeway

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.

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

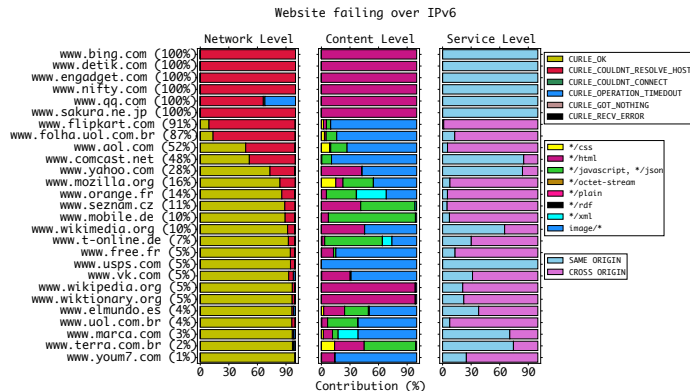
Success Rates

Causality Analysis

Takeway

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.

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

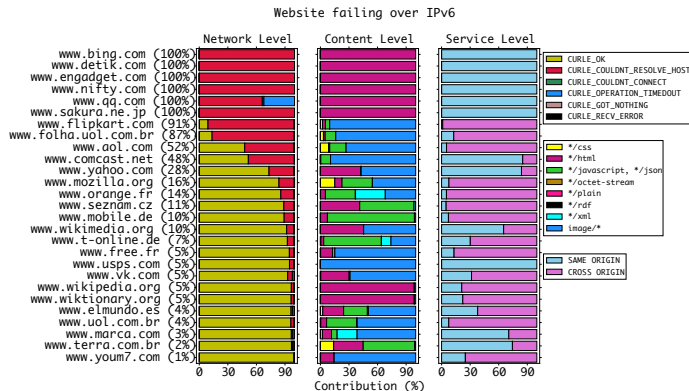
Success Rates

Causality Analysis

Takeway

Results | Causality Analysis

Where do the failing objects originate from?



- Both same and cross origin sources contribute to the failure of webpage elements over IPv6.

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

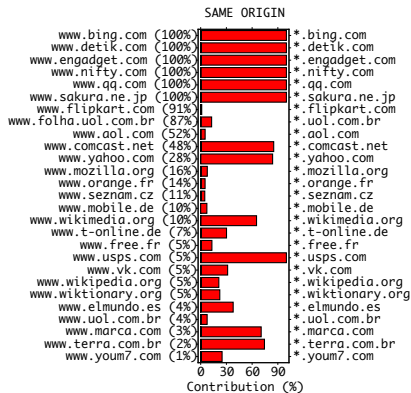
Results

Success Rates

Causality Analysis

Takeway

What is failure contribution of same-origin sources?



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

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

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

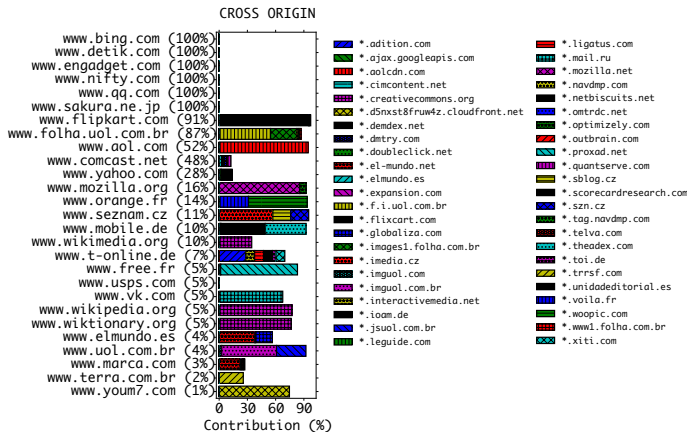
Results

Success Rates

Causality Analysis

Takeway

What is failure contribution of cross-origin sources?



- Some of the cross-origin sources contribute to the failure of multiple websites.

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

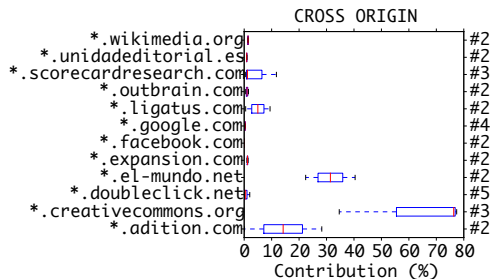
Results

Success Rates

Causality Analysis

Takeway

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.

CROSS ORIGIN	MEDIAN
*.creativecommons.org	76.33%
*.el-mundo.net	31.41%
*.adition.com	14.20%
*.ligatus.com	4.98%
*.wikimedia.org	1.40%
*.expansion.com	1.21%
*.scorecardresearch.com	1.19%
*.outbrain.com	1.06%
*.unidadeditorial.es	0.94%
*.doubleclick.net	0.54%
*.google.com	0.31%
*.facebook.com	0.06%

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

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.

www.vaibhavbajpai.com

v.bajpai@jacobs-university.de | @bajpaivaibhav

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

Appendix

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway

References

- [1] P. Richter, M. Allman, R. Bush, and V. Paxson, "A Primer on IPv4 Scarcity," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 2, pp. 21–31, Apr. 2015. [Online]. Available: <http://doi.acm.org/10.1145/2766330.2766335>
- [2] The Internet Society, "World IPv6 Launch," <http://www.worldipv6launch.org>, [Online; accessed 11-January-2016].
- [3] Google, "Google IPv6 Adoption Statistics," <http://www.google.com/intl/en/ipv6/statistics.html>, [Online; accessed 11-January-2016].
- [4] V. Bajpai and J. Schönwälder, "IPv4 versus IPv6 - who connects faster?" in *IFIP Networking Conference (IFIP Networking)*, 2015, May 2015, pp. 1–9. [Online]. Available: <http://dx.doi.org/10.1109/IFIPNetworking.2015.7145323>
- [5] M. Nikkhah, R. Guérin, Y. Lee, and R. Woundy, "Assessing IPv6 Through Web Access a Measurement Study and Its Findings," in *Proceedings of the Seventh Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '11. New York, NY, USA: ACM, 2011, pp. 26:1–26:12. [Online]. Available: <http://doi.acm.org/10.1145/2079296.2079322>
- [6] A. Dhamdhere, M. Luckie, B. Huffaker, k. claffy, A. Elmokashfi, and E. Aben, "Measuring the Deployment of IPv6: Topology, Routing and Performance," in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, ser. IMC '12. New York, NY, USA: ACM, 2012, pp. 537–550. [Online]. Available: <http://doi.acm.org/10.1145/2398776.2398832>
- [7] V. Bajpai and J. Schonwalder, "A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 3, pp. 1313–1341, thirdquarter 2015. [Online]. Available: <http://dx.doi.org/10.1109/COMST.2015.2418435>
- [8] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding Website Complexity: Measurements, Metrics, and Implications," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 313–328. [Online]. Available: <http://doi.acm.org/10.1145/2068816.2068846>
- [9] —, "Characterizing Web Page Complexity and Its Impact," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 943–956, Jun. 2014. [Online]. Available: <http://dx.doi.org/10.1109/TNET.2013.2269999>

Introduction

Motivation

Research Question

Research Contributions

Methodology

Metrics and Implementation

Selection of Websites

Measurement Setup

Measurement Trial

Results

Success Rates

Causality Analysis

Takeway