An Empirical View on Consolidation of the Web

Trinh Viet Doan   Technical University of Munich

Roland van Rijswijk-Deij   University of Twente, NLnet Labs

Oliver Hohlfeld   Brandenburg University of Technology

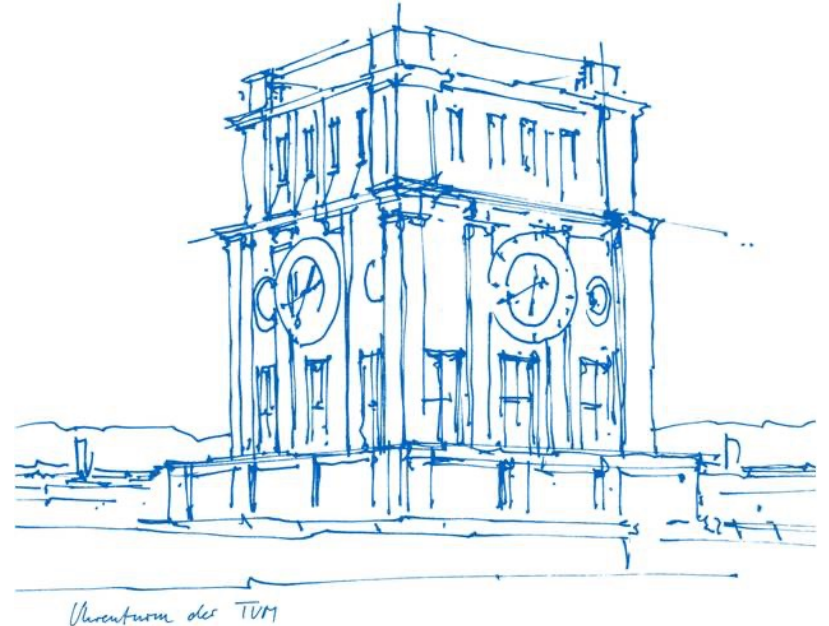Vaibhav Bajpai   CISPA, Technical University of Munich

IETF 113 Vienna

IAB Open Meeting

March 24, 2022

Uhrenturm der TUM

Motivation

Increasing demand and need for measurement studies on Internet centralization

Differing opinions about centralization
- Facilitates deployment of new protocols and extensions, streamlining services
- Risks for privacy, dependencies, concentration of control
- Difficult to define centralization and to assess extent and implications

Recent centralization studies on DNS, third-party dependencies, co-location, …

→ In this paper: Web consolidation around Content Delivery Infrastructures (CDIs, i.e., CDN + cloud hosts)

Longitudinal DNS Measurements

More than 160M landing Web pages of all .com/.net./.org domains
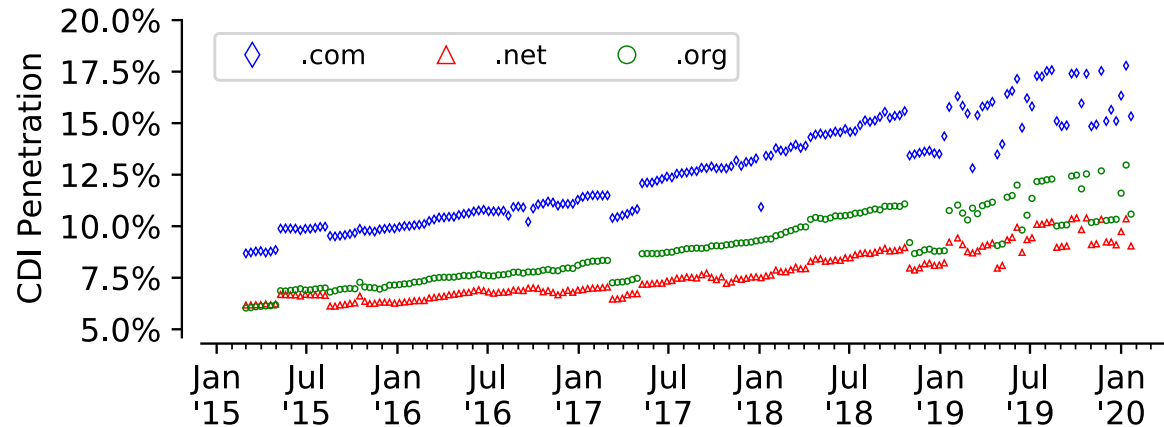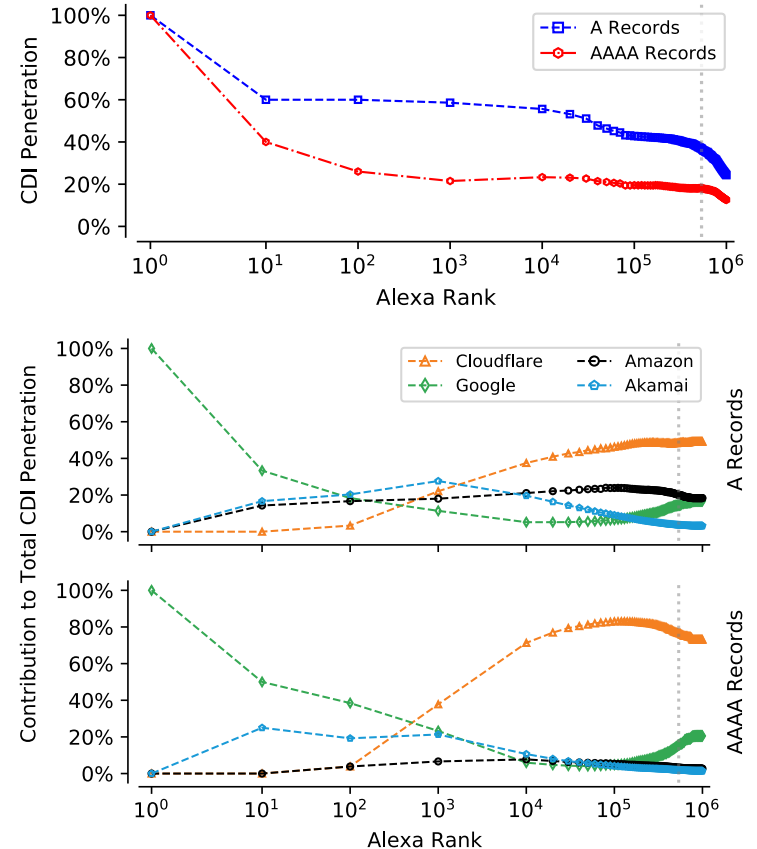

→ Increase of CDI penetration from 8% in 2015 to 15% in 2020

Popular Landing Pages

Based on DNS measurements for Alexa Top 1M

Overall penetration of roughly 24% (IPv4) and 12% (IPv6), higher among more popular domains

Handful of CDIs are major contributors to overall penetration, particularly over IPv6

Page Resources

Based on HTTP Archive data:
Page loads for 4.3M websites, incl. 392M resources

Overall CDI penetration
- Landing Pages: 32.1%
- Page Resources: 56.6%

Google + Amazon:
>50% of CDI-hosted resources

Especially high percentages for static resources

Case studies in paper:
High CDI penetration for ads/trackers, TLS 1.3, …

	Provider	# Resources (↓)	Share of CDI Resources	Share of All Resources
1)	Google	76.6M	34.5%	19.5%
2)	Amazon	38.9M	17.5%	9.9%
3)	Cloudflare	27.5M	12.4%	7.0%
4)	Facebook	17.7M	8.0%	4.5%
5)	Akamai	15.7M	7.1%	4.0%
6)	Fastly	10.8M	4.9%	2.7%
7)	WordPress	4.1M	1.9%	1.1%
8)	Twitter	4.0M	1.8%	1.0%
9)	Microsoft	3.8M	1.7%	1.0%
10)	NetDNA	3.6M	1.6%	0.9%

Resource Type	# CDI Resources	CDI Pen. of Type	# All Resources of Type	Share (All) (↓)
image	82,613,713	46.8%	176,660,130	45.0%
javascript	64,223,345	64.1%	100,195,949	25.5%
text	21,676,628	50.4%	43,017,071	11.0%
html	19,590,470	69.6%	28,148,091	7.2%
other	11,864,834	70.4%	16,847,204	4.3%
font	14,245,056	86.0%	16,569,827	4.2%
application	6,303,607	68.4%	9,220,762	2.4%
video	1,135,211	91.8%	1,236,756	0.3%
audio	265,302	62.2%	426,583	0.1%
Total	221,918,166	56.6%	392,322,373	100.0%

Conclusion

Increasing consolidation of Web pages around CDIs
→   CDI penetration roughly doubled since 2015
→   Most of the resources delivered from a few selected CDIs

Some CDIs significantly contribute to the deployment/support of protocols
→   IPv6
→   TLS 1.3

Open questions for discussion (see [arch-d] mailing list):
•   Motivation of users/customers?
•   Natural evolution of ecosystem?
•   Impact of/on IETF activities?