The Role of DNS in Residential Internet Use

Shuyue Yu, Thomas Koch, Gil Zussman, Ethan Katz-Bassett

Columbia University

Outline: DNS in the Context of Residential Traffic

Expectation

1. Services use DNS to direct users to desired servers
2. DNS uses caches to limit DNS overhead
3. DNS datasets are used to understand service popularity

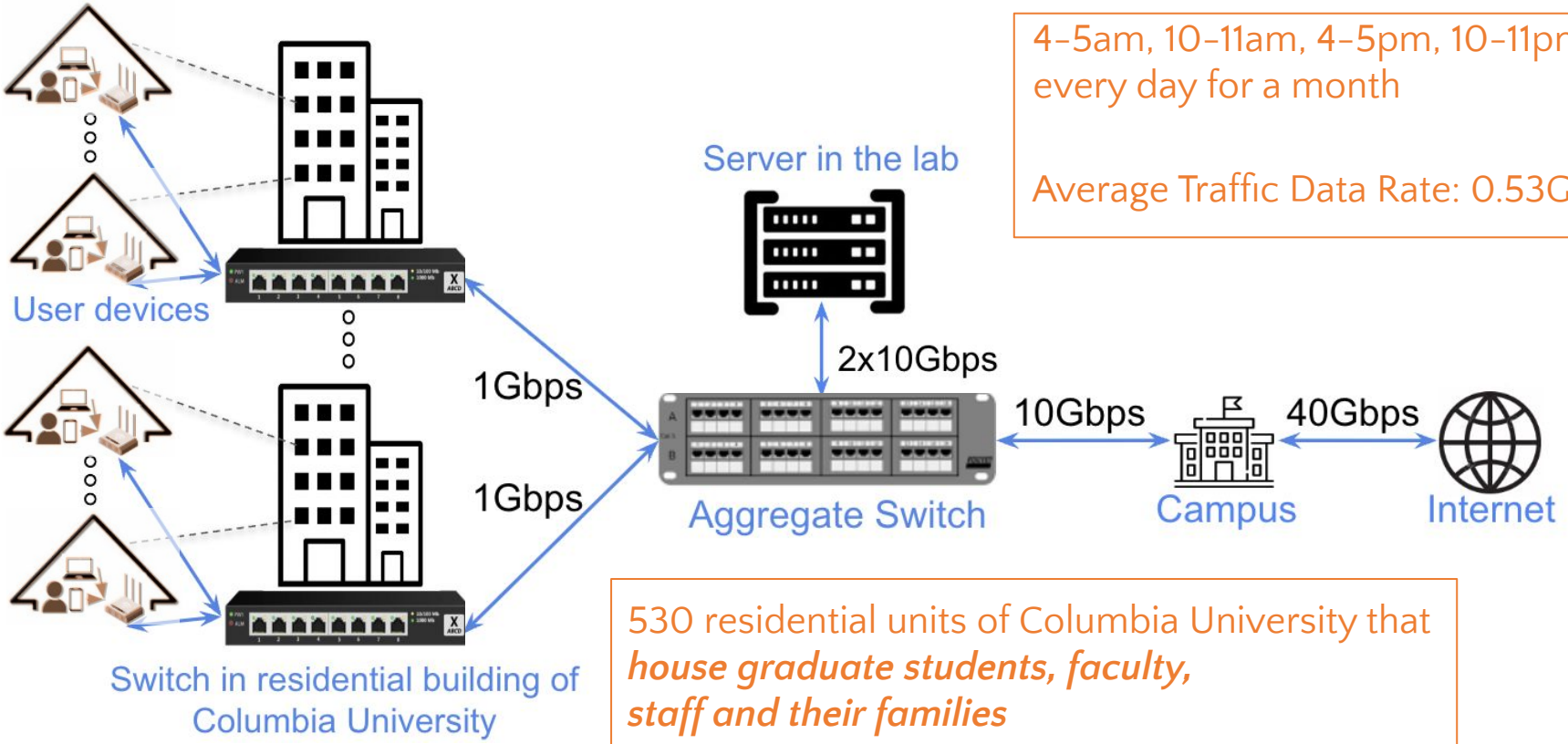
Reality

1. **Significant traffic volumes are not directed by DNS**
   12% of traffic cannot be paired with a DNS record
2. **The prevalent use of IP addresses from stale cached DNS records**
   13% of flows start after DNS records have expired
3. **Service popularity varies by metrics**
   Using traffic volume and the number of DNS queries lead to different service popularity estimations

Roadmap

1. Data Collection and Processing
2. The Prevalence of No-DNS Traffic
3. The Prevalent Use of Stale DNS Records
4. The Variability of Service Popularity

Data Collection

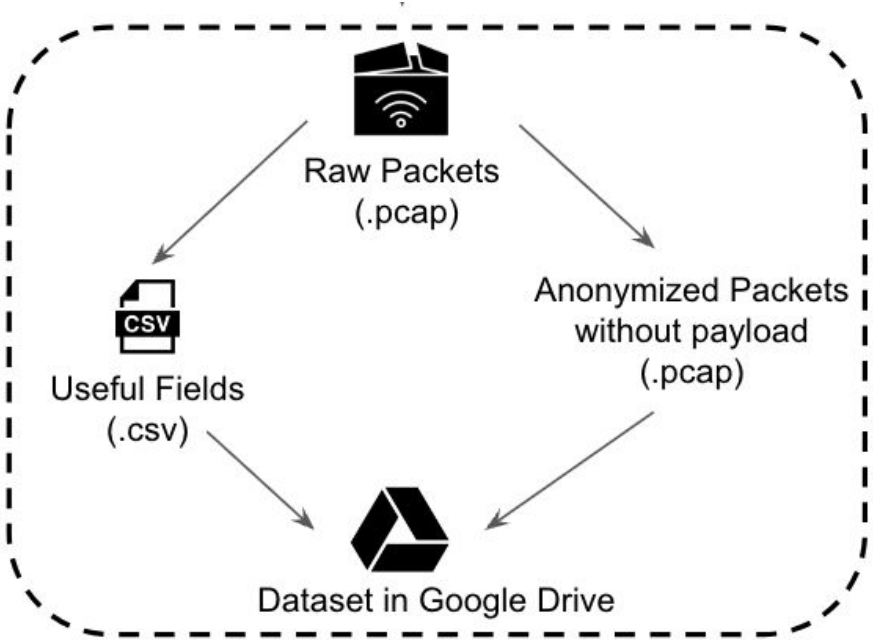


4-5am, 10-11am, 4-5pm, 10-11pm every day for a month

Average Traffic Data Rate: 0.53Gbps

530 residential units of Columbia University that ***house graduate students, faculty, staff and their families***

Data Collection



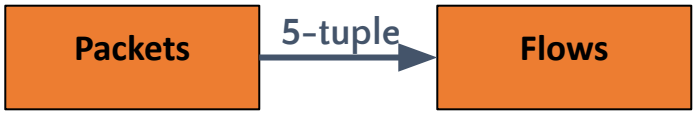
Data Collection Pipeline

Anonymization[1]: at collection time,

- save non-privacy sensitive fields (e.g., domains - last three suffixes)
- anonymize privacy sensitive fields
- discard any payload

[1] Non-human subjects project
Exempt from Institutional Review Board (IRB) review
Approved by the IT department of Columbia University

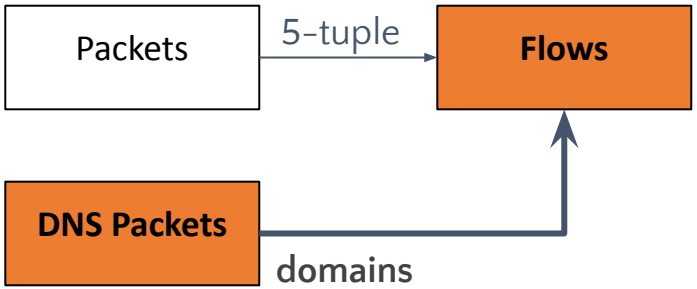
Data Processing

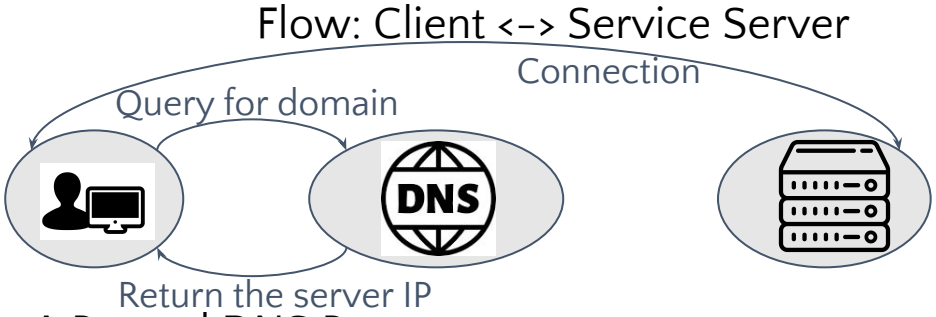


Aggregate packets into flows:

Flow – packets with the same **5-tuple in IP header** (transport protocol, source IP, destination IP, source port, destination port) within an hour

Data Processing





A Record DNS Response

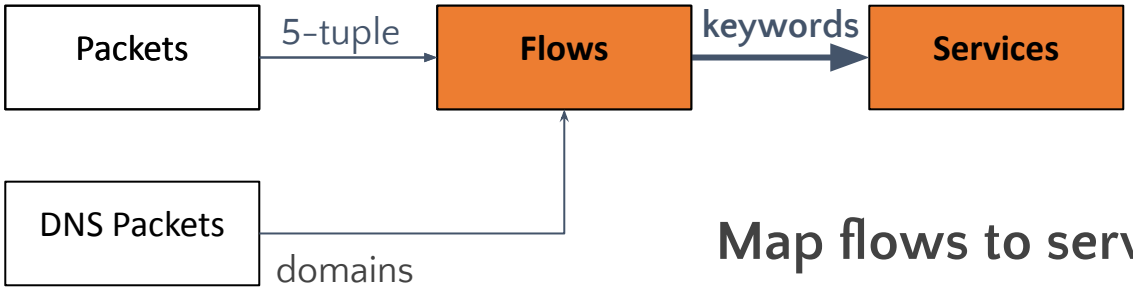
Pair flows with DNS domains:

Match the most recent A Record DNS query, up to 5 hours before the hour capture

- from the same apartment
- returned the result IP address that is the destination of the flow

Allman, Mark. "Putting DNS in context." *Proceedings of the ACM Internet Measurement Conference*. 2020.

Data Processing



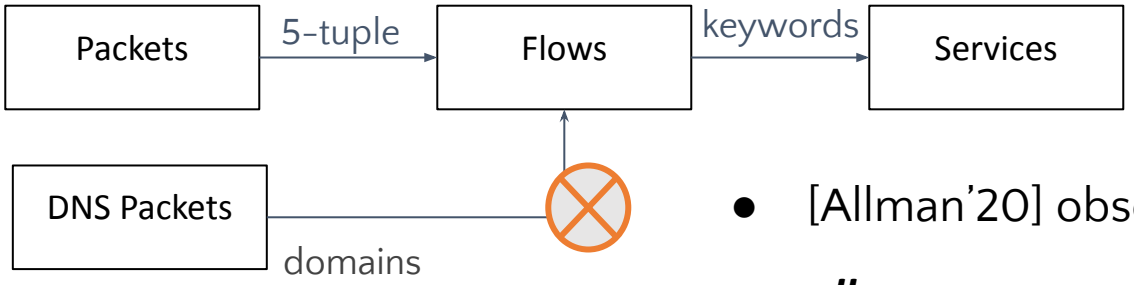
Map flows to services:

- 76% of traffic is mapped to 100 popular services, using ≈200 keywords from DNS domains
- 24% of traffic have the paired DNS domains as their own services

Roadmap

1. Data Collection and Processing
2. The Prevalence of No-DNS Traffic
3. The Prevalent Use of Stale DNS Records
4. The Variability of Service Popularity

No-DNS Traffic



No-DNS traffic: traffic not paired with a DNS domain

- [Allman'20] observed **7%** of *no-DNS connections*

  all were assumed as P2P

- Our measurements observed 12% of *no-DNS traffic*, **27%** of *no-DNS connections*

  three reasons

Allman, Mark. "Putting DNS in context." *Proceedings of the ACM Internet Measurement Conference*. 2020.

Reasons Behind No-DNS Traffic

1

2

3

Server Lookup without DNS Packets

- **Peer-to-peer (P2P)** traffic
- Variants of DNS (e.g., HTTPDNS by Tencent)
- Hardcode IP addresses

P2P Traffic

Inference technique based on three properties

1. Peers tend not to be accessed through DNS
2. Peers tend to be in ISPs that host human users
3. Peers tend to use non- privileged, high-numbered ports

6% of traffic, **17%** of connections as P2P

< 7% of connections as P2P in [Allman'20]

Rebirth of P2P

Allman, Mark. "Putting DNS in context." *Proceedings of the ACM Internet Measurement Conference*. 2020.

Reasons Behind No-DNS Traffic

1

Server Lookup without DNS Packets

- **Peer-to-peer (P2P)** traffic
- Variants of DNS (e.g., HTTPDNS by Tencent)
- Hardcode IP addresses

2

Encrypted DNS Protocols

- DNS-over-HTTPS (DoS)
- DNS-over-TLS (DoT)
- DNS-over-QUIC (DoQ)

3

DNS resolvers of Columbia University do not support encrypted DNS and there are few non-DNS packets to popular public resolvers.

Reasons Behind No-DNS Traffic

1

Server Lookup without DNS Packets

- **Peer-to-peer (P2P)** traffic
- Variants of DNS (e.g., HTTPDNS by Tencent)
- Hardcode IP addresses

2

Encrypted DNS Protocols

- DNS-over-HTTPS (DoS)
- DNS-over-TLS (DoT)
- DNS-over-QUIC (DoQ)

3

DNS Caching

Users cache DNS records that are inaccessible to us due to

- switch of networks
- beyond collection window

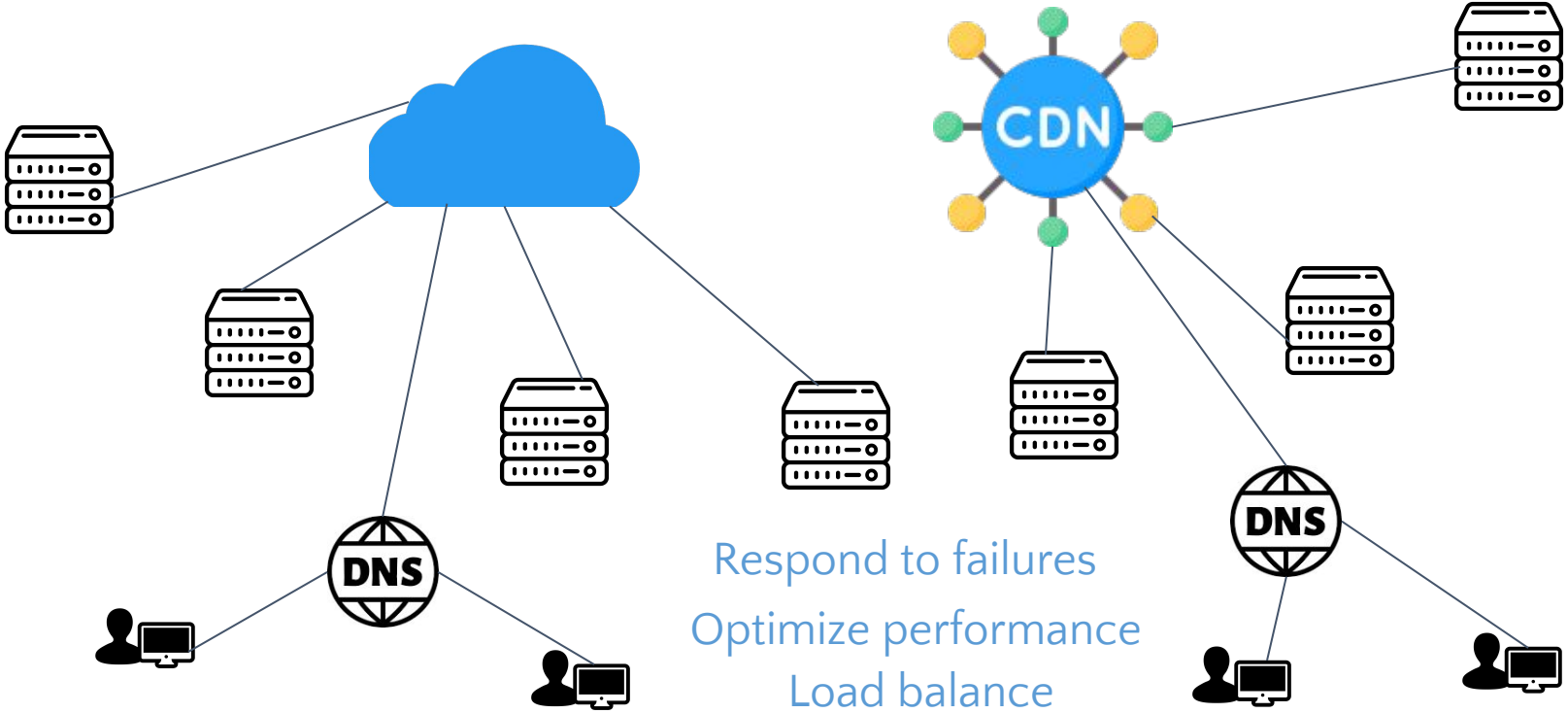
90% of the flows have DNS time-to-live (TTL) smaller than 300 seconds

However, TTL violations are prevalent

Roadmap

1. Data Collection and Processing
2. The Prevalence of No-DNS Traffic
3. The Prevalent Use of Stale DNS Records
4. The Variability of Service Popularity

CDNs and clouds use DNS to steer users



Respond to failures

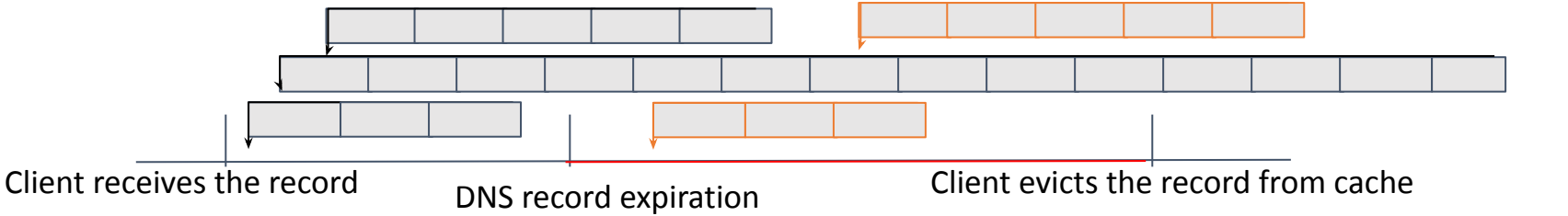
Optimize performance

Load balance

Limited Resilience to Failures

CDN cannot quickly update to a backup site due to TTL violations

significant traffic volumes



50% of TTL-violating flows start **more than 56 seconds** after the expiration

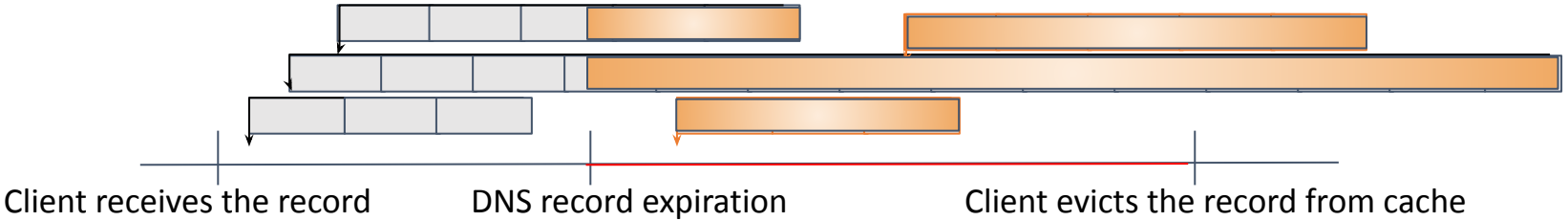
well after expiration

Limited Agility to Steer Traffic

Traffic may not immediately use uncongested paths updated by clouds

significant traffic volumes

50% of bytes originating from tne residential buildings occur after their DNS records have expired

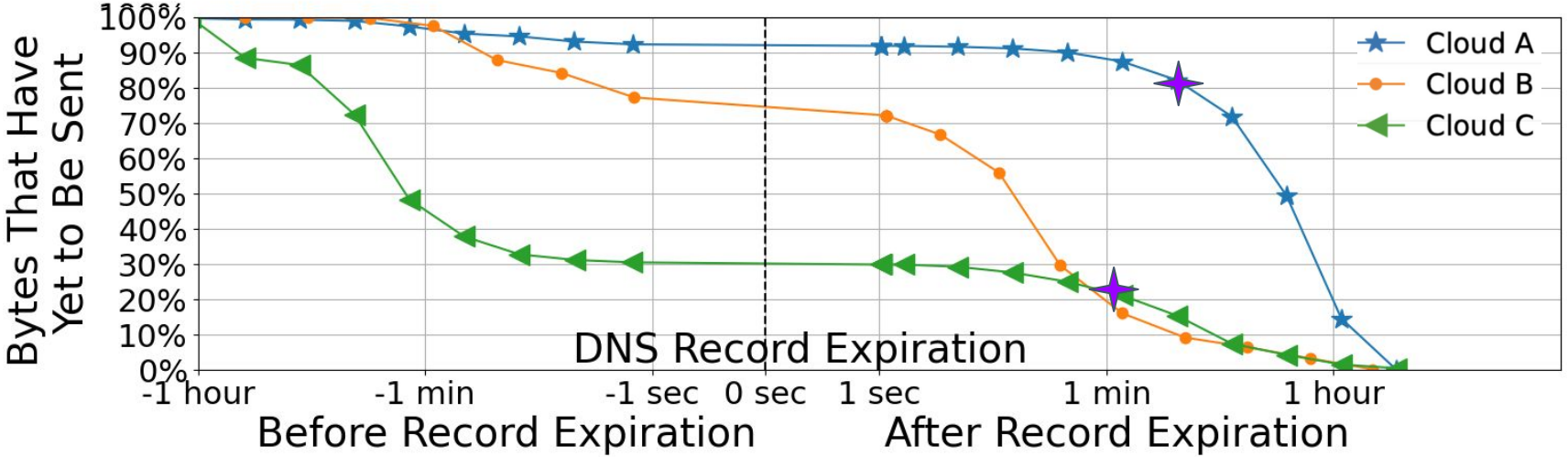


Limitation of TCP + DNS

Limited Agility to Steer Traffic

Bytes sent after the expiration

Limitation of TCP + DNS



80% of traffic to Cloud A – 5 minutes after the expiration

> 20% of traffic to Cloud B/C – a minute after the expiration

well after expiration

Service Redirection Techniques with Better Tradeoffs

Tom Koch, Shuyue Yu, Ethan Katz-Bassett, Sharad Agarwal, and Ryan Beckett. "PAINTER: Ingress Traffic Engineering and Routing for Enterprise Cloud Networks." In *ACM SIGCOMM*. 2023.

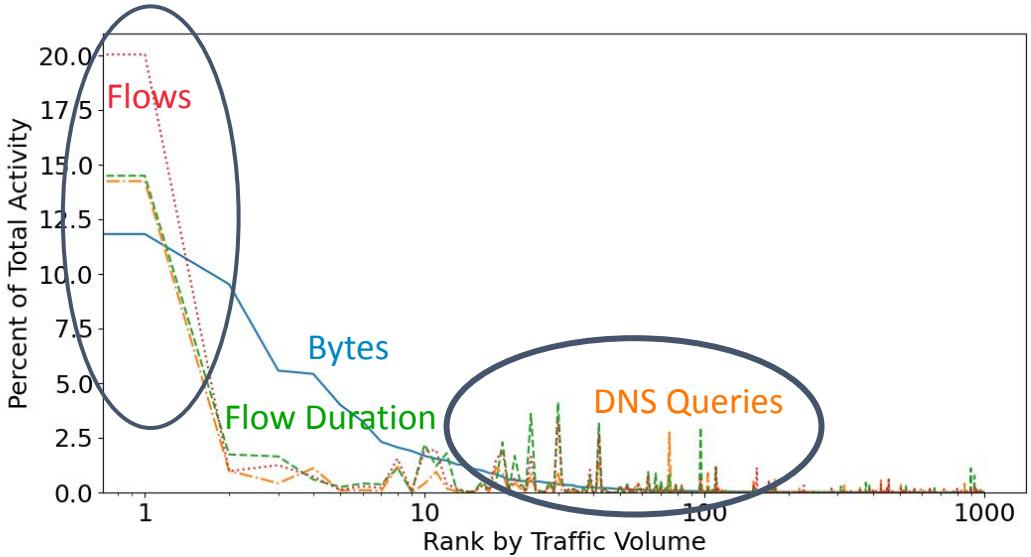
Jiangchen Zhu, Kevin Vermeulen, Italo Cunha, Ethan Katz-Bassett, and Matt Calder. "The best of both worlds: high availability CDN routing without compromising control." In *ACM Internet Measurement Conference*. 2022.

Roadmap

1. Data Collection and Processing
2. The Prevalence of No-DNS Traffic
3. The Prevalent Use of Stale DNS Records
4. The Variability of Service Popularity

Service Popularity Varies by Metric

- DNS Responses –
  Cisco Umbrella top list
  Secrank list
- #Flows – web page loads
- Flow Duration – time on the site
- #Bytes – traffic volume



All measures identify the top services (e.g., YouTube, Apple, Netflix)

They disagree on the importance of other services (e.g., Slack, Gmail)

Use DNS datasets carefully

Conclusions

1. The prevalence of no-DNS traffic and the rebirth of P2P
   New methods of mapping traffic to services
2. IP addresses from stale DNS records are heavily used
   Service redirection techniques with better tradeoffs
3. Service popularity varies by metrics
   Richer datasets to estimate website popularity

ACKNOWLEDGEMENT

The security, privacy and networking teams of Columbia University Information Technology
Columbia University

Arpit Gupta
University of California Santa Barbara

Sanjay Chandrasekaren
University of California Santa Barbara

Mahshid Gahsemi
Columbia University

wim.net
Wireless & Mobile Networking Lab

COLUMBIA UNIVERSITY

THANK YOU
for listening!

Shuyue Yu
Columbia University
Email: shuyue.yu@columbia.edu