Back-Office Web Traffic on the Internet

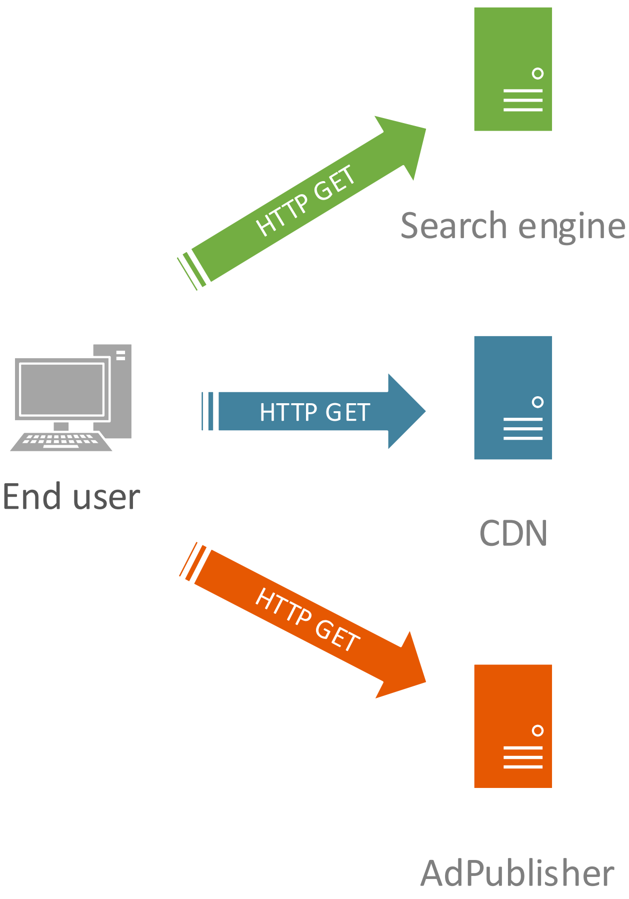
Enric Pujol	TU-Berlin
Philipp Richter	TU-Berlin
Balakrishnan Chandrasekaran	Duke University
Georgios Smaragdakis	MIT / TU-Berlin / Akamai
Anja Feldmann	TU-Berlin
Bruce Maggs	Duke University / Akamai
Keung-Chi Ng	Akamai

IMC 2014

Vancouver, BC, CANADA

November 5-7, 2014

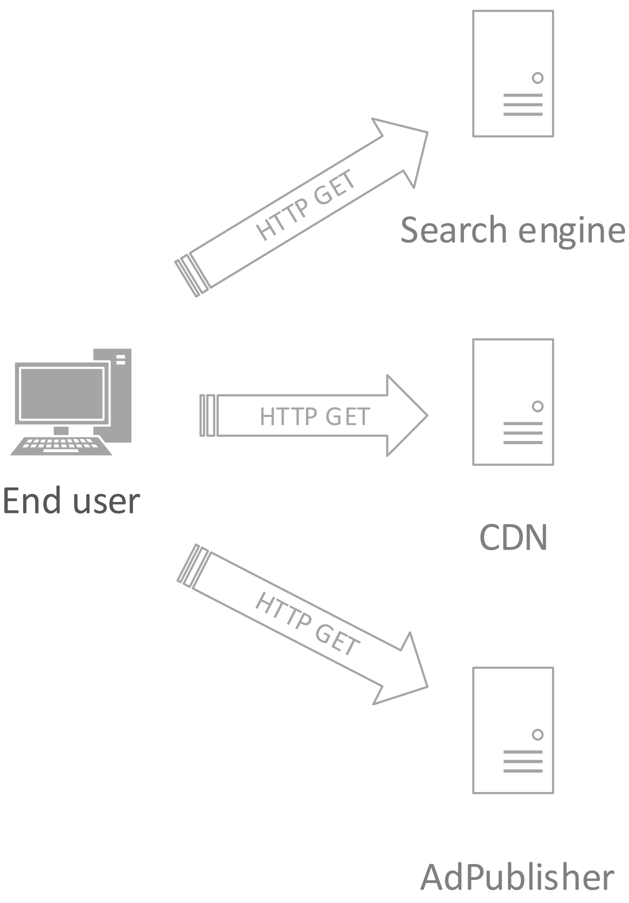
The Web for an end user



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

The front-office

Behind the scenes...



HTTP GET

Search engine

End user

HTTP GET

CDN

HTTP GET

AdPublisher

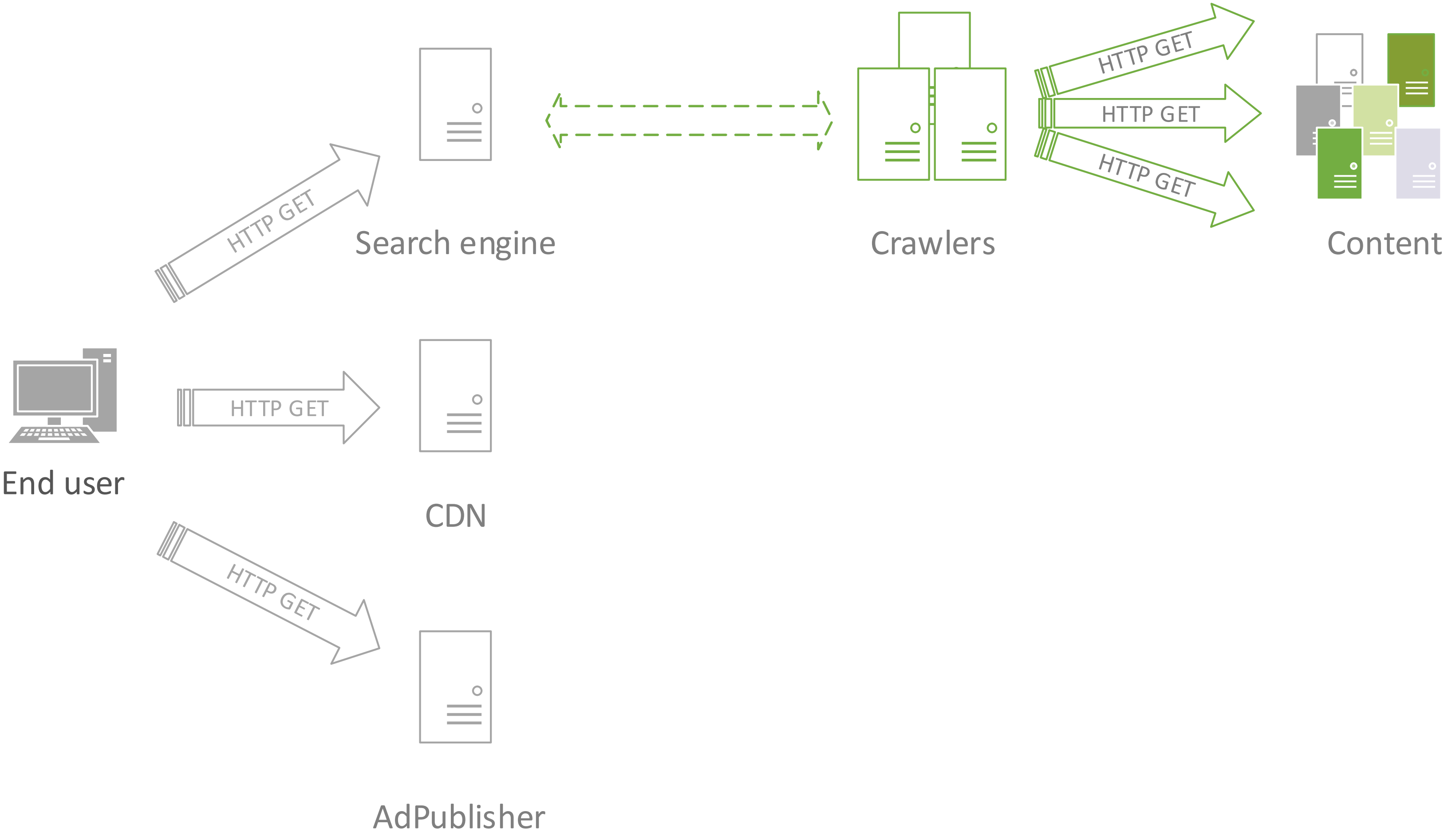
Back-office Web traffic:
Machine-to-machine Web traffic

?

The front-office

The back-office

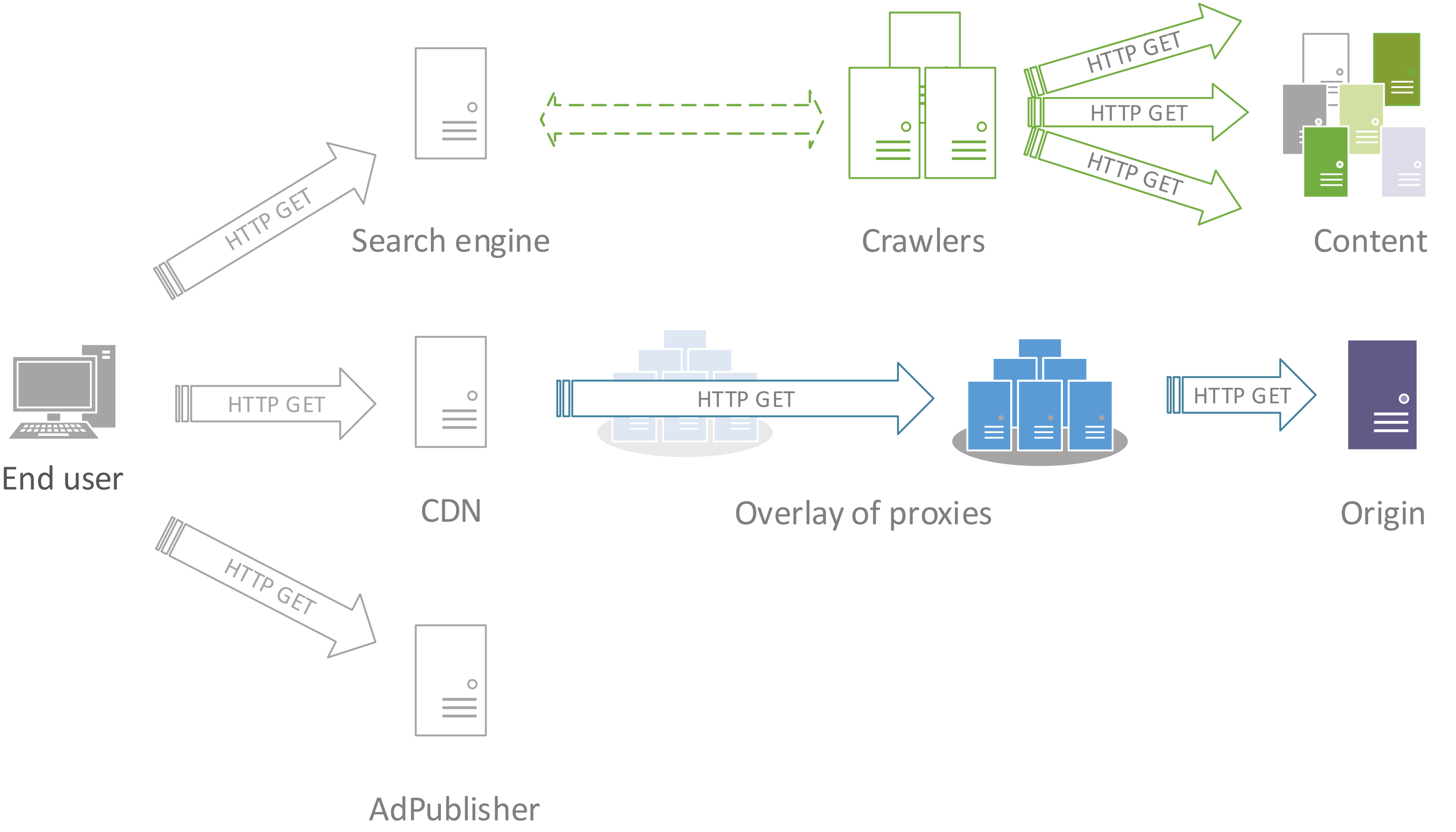
Search engines: crawlers



The front-office

The back-office

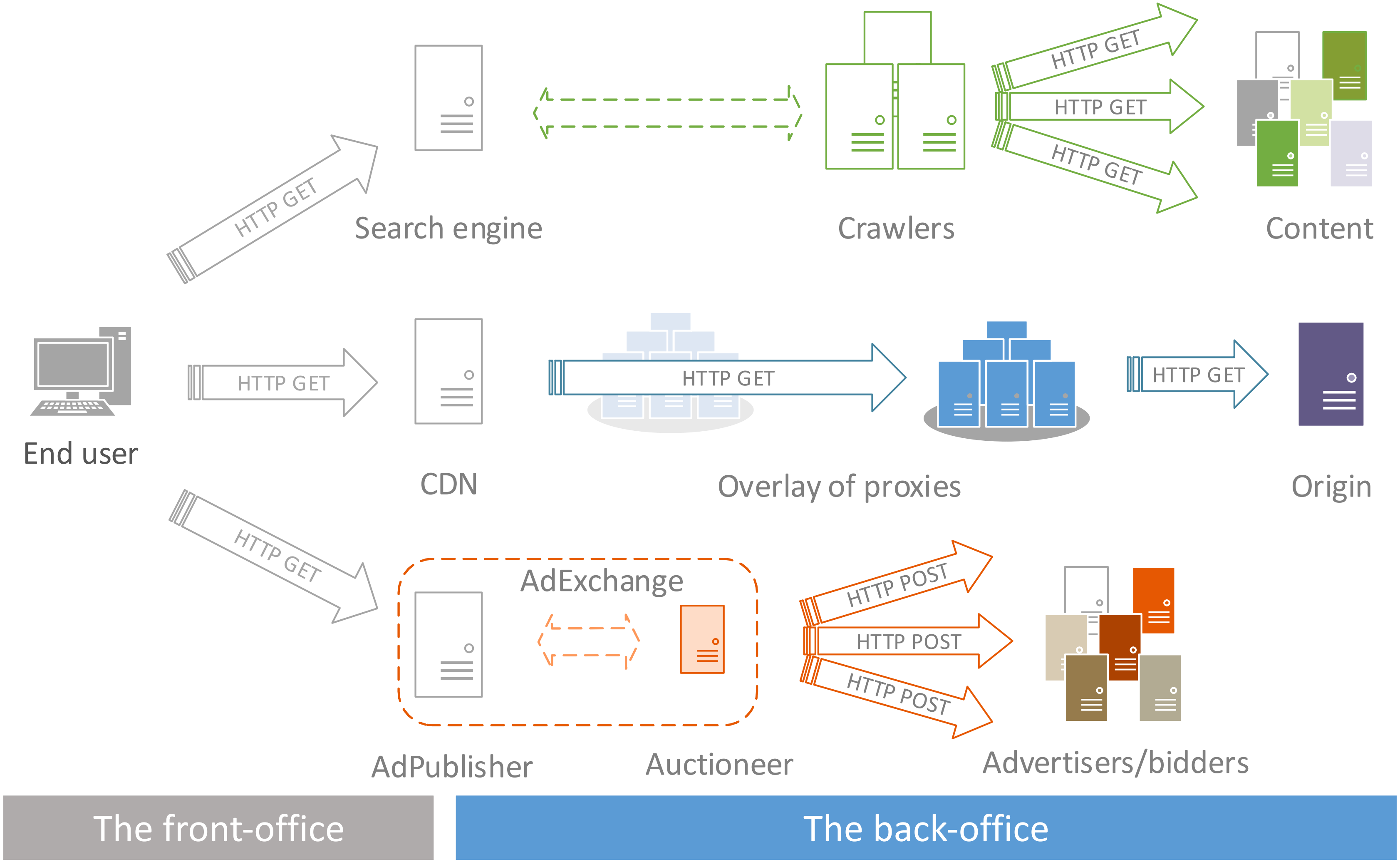
Content delivery: proxies



The front-office

The back-office

AdExchanges: real-time bidding



HTTP GET

Search engine

Crawlers

HTTP GET

HTTP GET

HTTP GET

Content

End user

HTTP GET

CDN

HTTP GET

Overlay of proxies

HTTP GET

Origin

HTTP GET

AdExchange

AdPublisher

Auctioneer

HTTP POST

HTTP POST

HTTP POST

Advertisers/bidders

The front-office

The back-office

Agenda

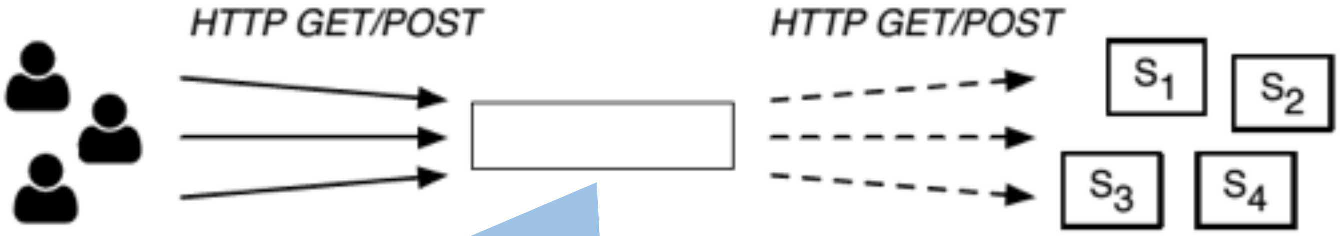
1. Introduction
2. Methodology and datasets
3. Characteristics
   1. Traffic
   2. Patterns
   3. Inter-domain perspective
4. CDN back-office traffic
5. The end-user perspective
6. Summary and implications

Vantage points (VP)

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

Diverse vantage points: multiple perspectives

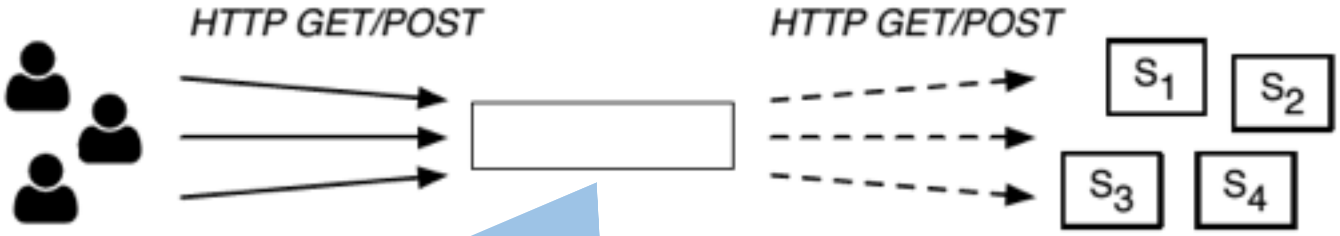
Candidate IPs for the back-office



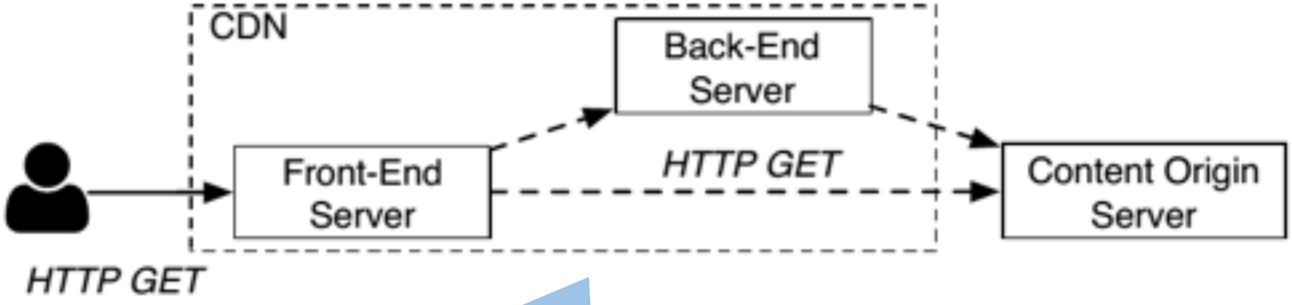
Send and receive requests

Dual role IPs are prime candidates

Candidate IPs for the back-office



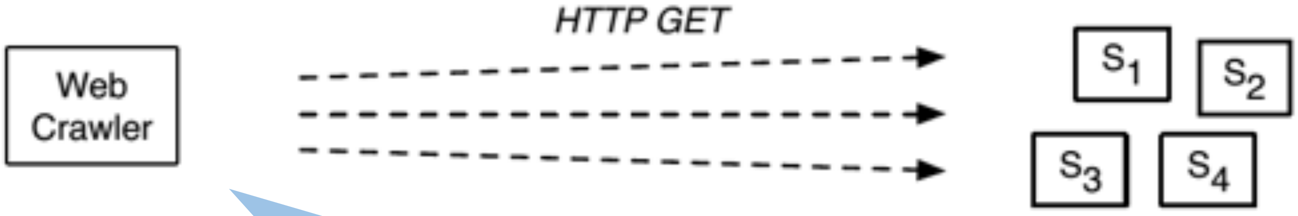
Send and receive requests



Send and receive requests

Dual role IPs are prime candidates

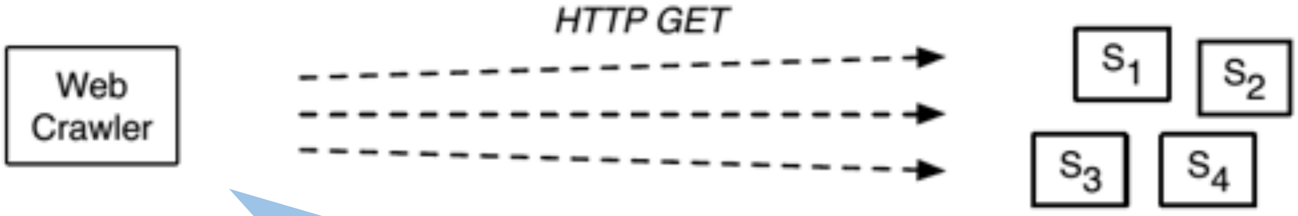
Candidate IPs for the back-office



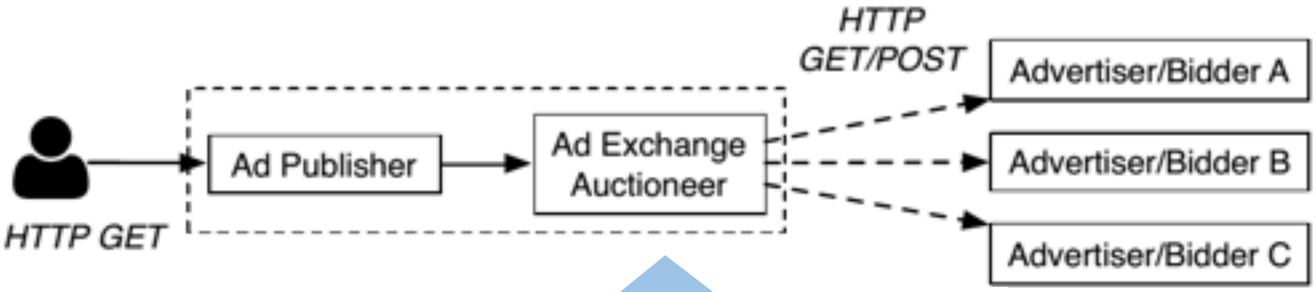
Many requests to many servers

Heavy hitter IPs are also prime candidates

Candidate IPs for the back-office



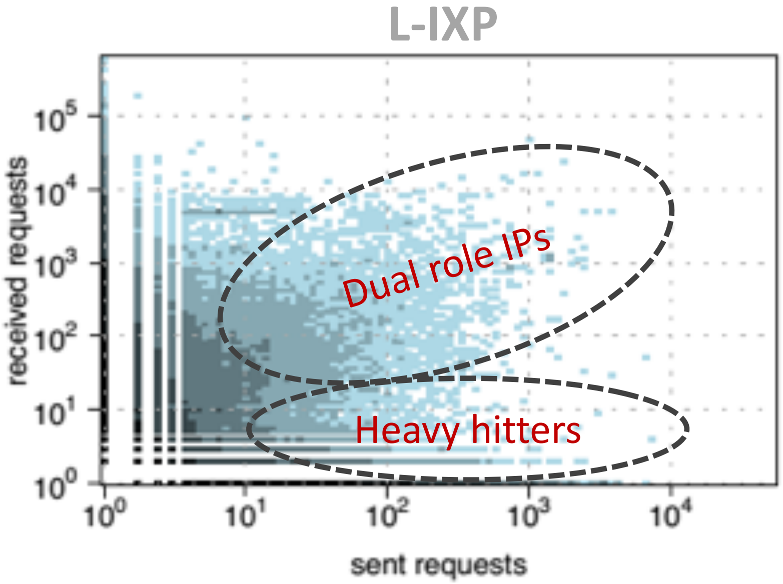
Many requests to many servers



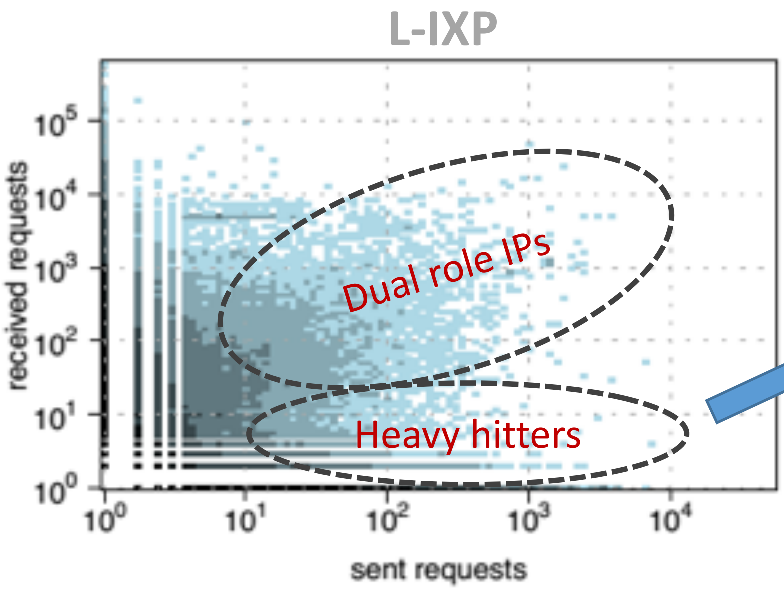
Many requests to a few servers

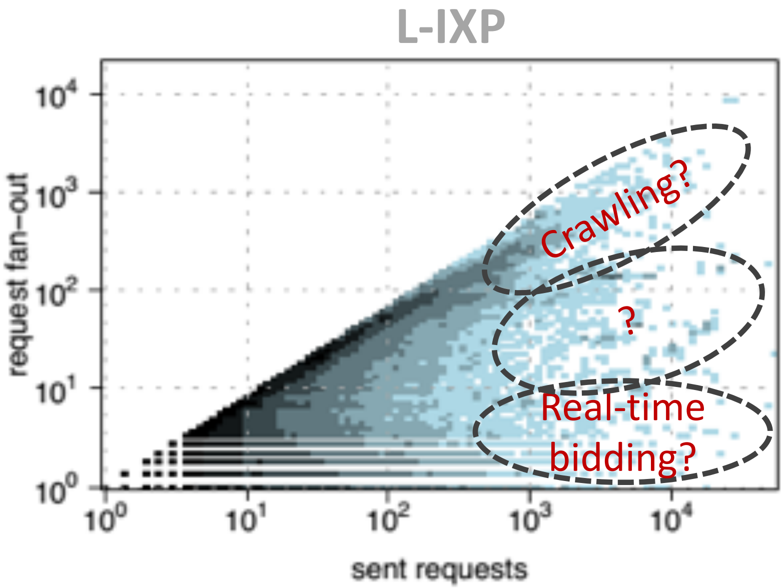
Heavy hitter IPs are also prime candidates

Sources of back-office Web traffic



Sources of back-office Web traffic


L-IXP


L-IXP

Dual-role IPs: active measurements

		Client only (%)	Server only (%)	Dual-role (%)
L-IXP	**Passive**	96.90	2.74	0.36
	Passive+Active	93.85	2.74	**3.40**

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

L-IXP

3.9K IPs, 74% in <u>2 orgs</u>

Candidates: manual classification

Crawlers:

- Reverse DNS + Origin AS

3.9K IPs, 74% in 2 orgs

L-IXP

Auctioneers:

- URL + Origin AS

316 IPs, 4 orgs

L-IXP

Candidates: manual classification

Crawlers:

- Reverse DNS + Origin AS

L-IXP

3.9K IPs, 74% in 2 orgs

Auctioneers:

- URL + Origin AS

L-IXP

316 IPs, 4 orgs

Content Delivery Proxies:

- Origin AS + Reverse DNS (for caches)

L-IXP

36K IPs, 8 orgs

Candidates: manual classification

Crawlers:

- Reverse DNS + Origin AS

L-IXP

3.9K IPs, 74% in 2 orgs

Auctioneers:

- URL + Origin AS

L-IXP

316 IPs, 4 orgs

Content Delivery Proxies:

- Origin AS + Reverse DNS (for caches)

L-IXP

36K IPs, 8 orgs

Other:

- Rest of dual-role IPs

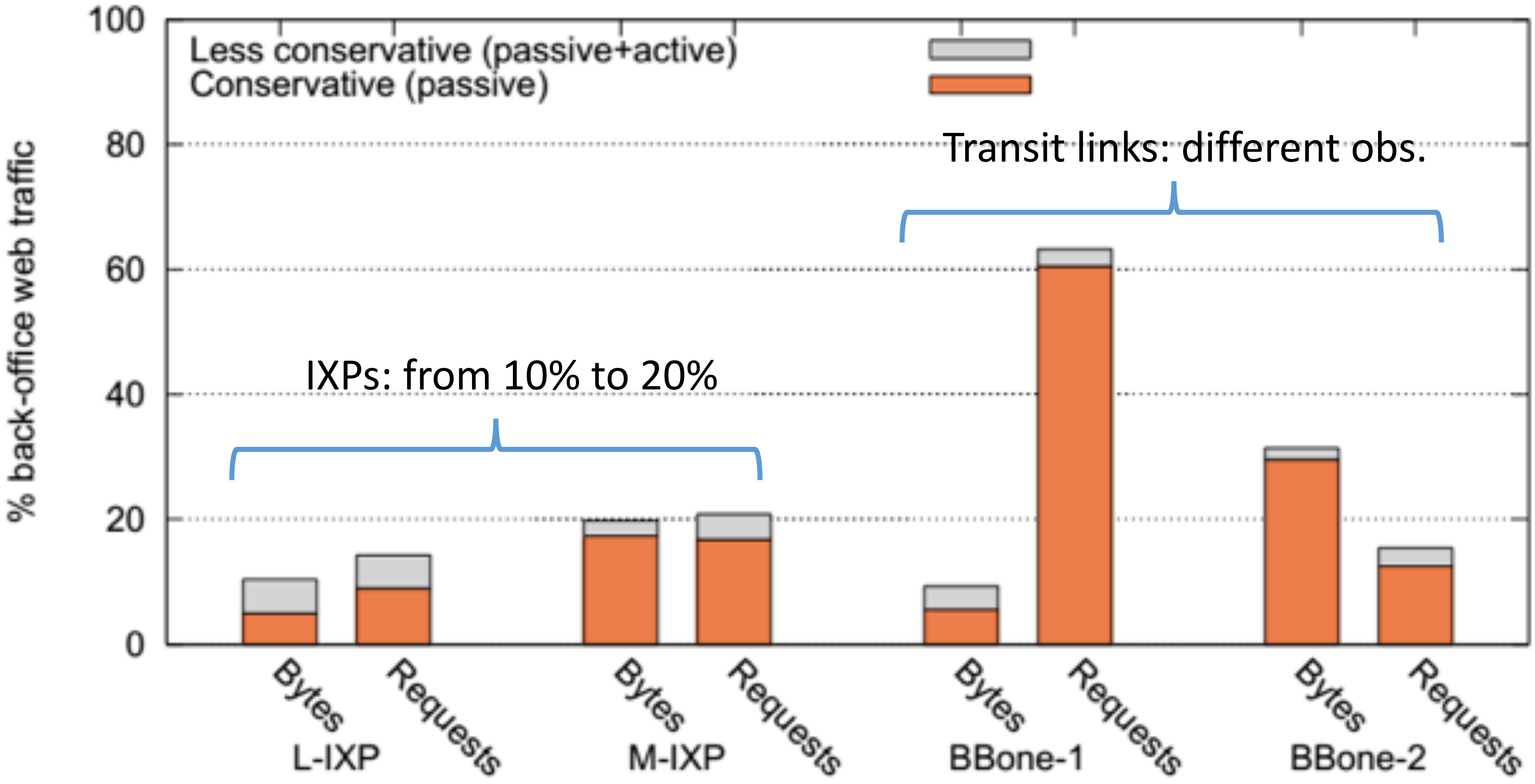
L-IXP

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



Transit links: different obs.

IXPs: from 10% to 20%

At least 10% in our VPs

Traffic: Contribution per class

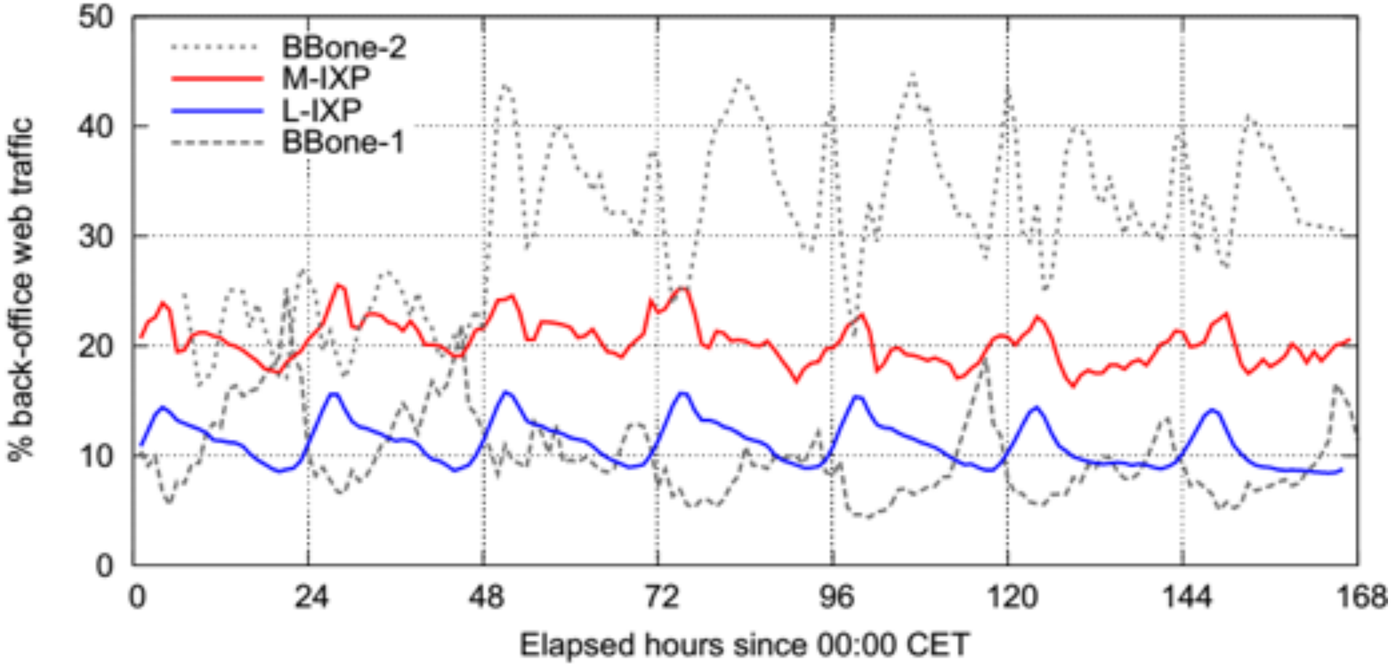
		CDPs	Auctioneers	Crawlers	Other
L-IXP	**Bytes**	12.1 %	1.1 %	10.3 %	76.5 %
	Requests	11.8 %	22.5 %	15.1 %	50.6 %

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

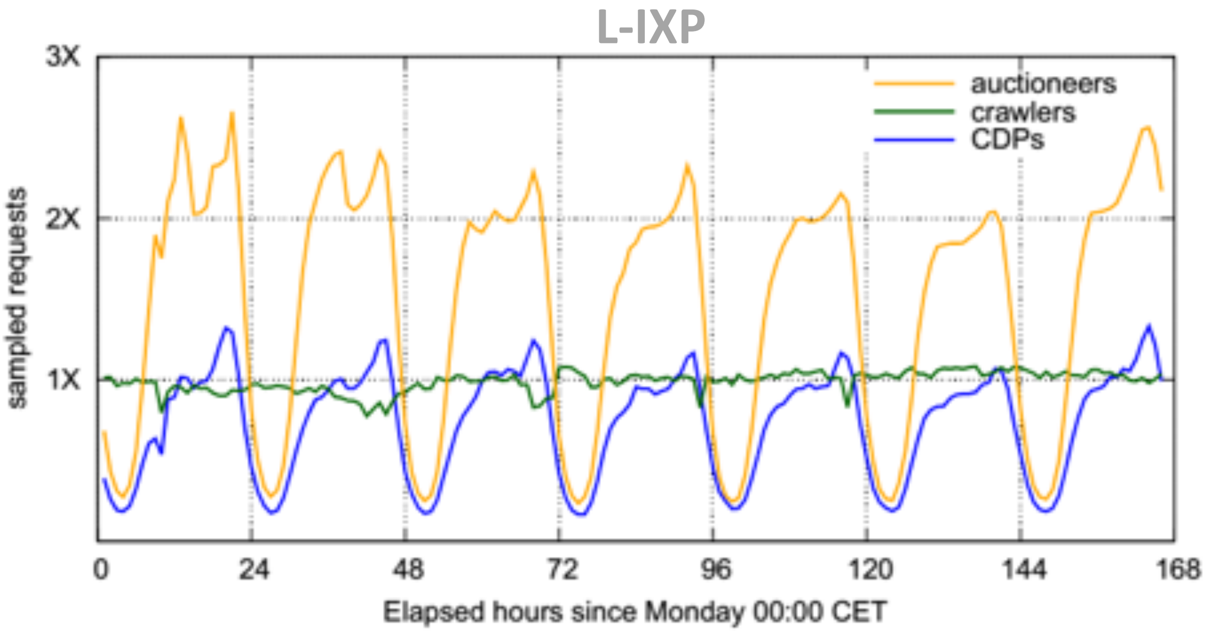
All classes contribute. More to discover

Traffic patterns: bytes



% back-office Web traffic increases during off hours in IXPs

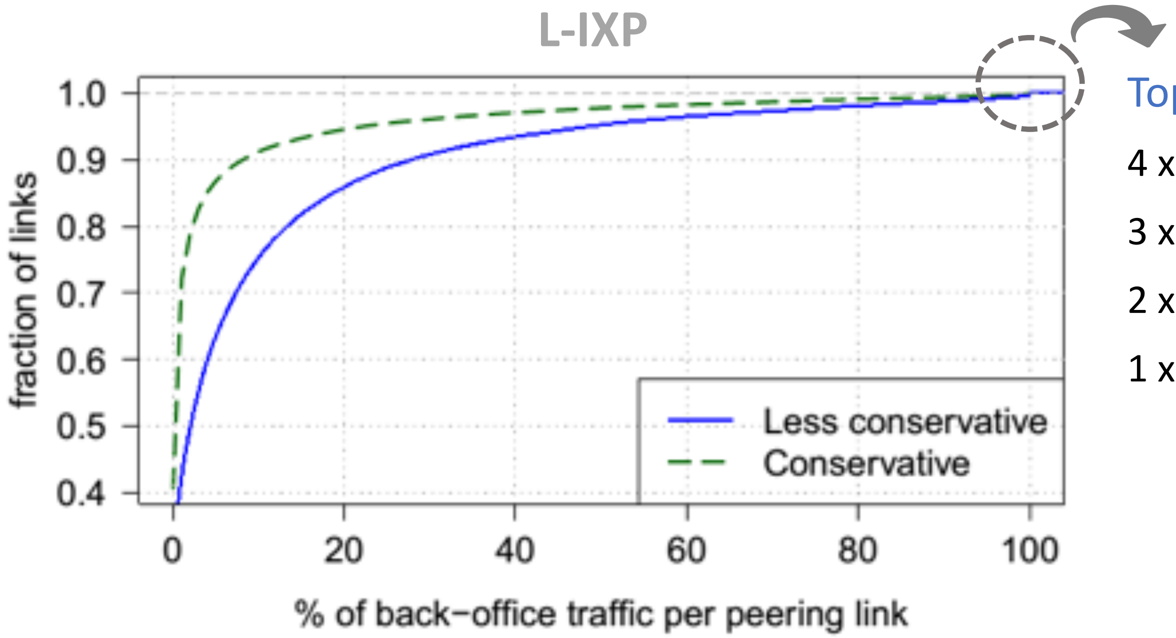
Traffic patterns: requests



Observations:

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

Inter-domain perspective



Top 10 traffic carrying links:

4 x Cloud – Content

3 x Search – Hosters

2 x CDN – Content

1 x Content – Advertisement

Back-office traffic appears in many peering links

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

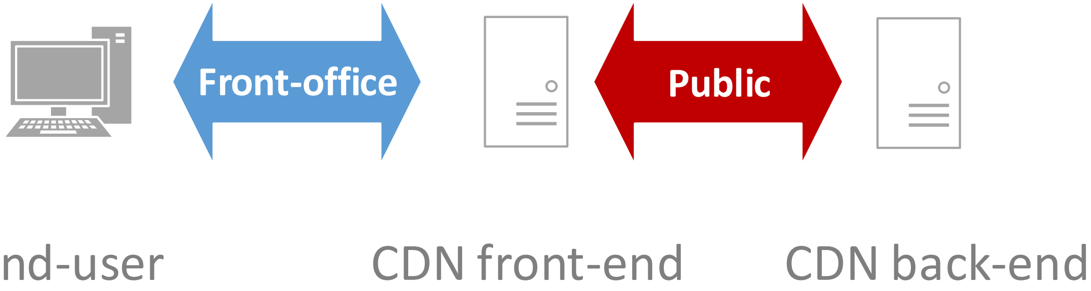
Front-office

End-user

CDN front-end

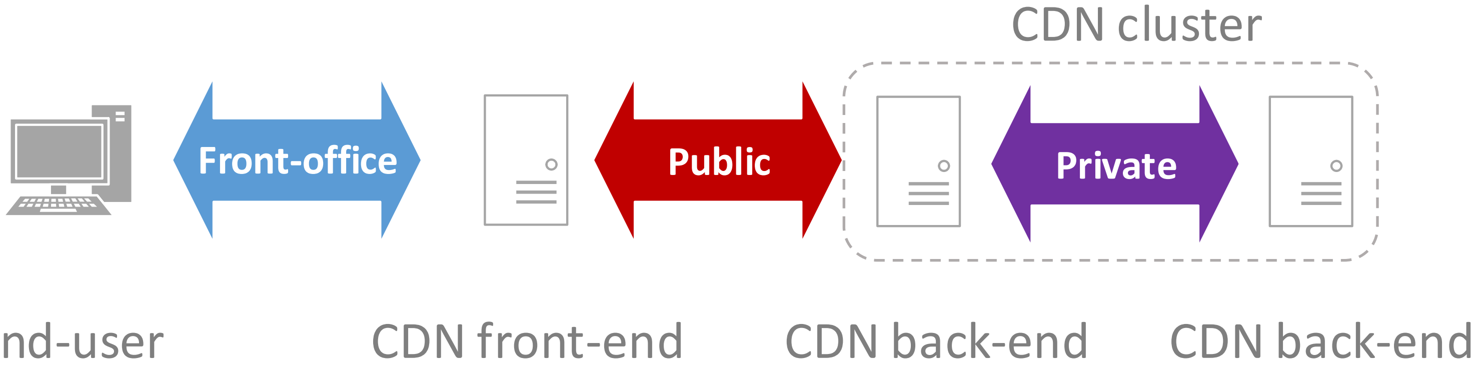
Three sub-classes of back-office traffic

A CDN perspective



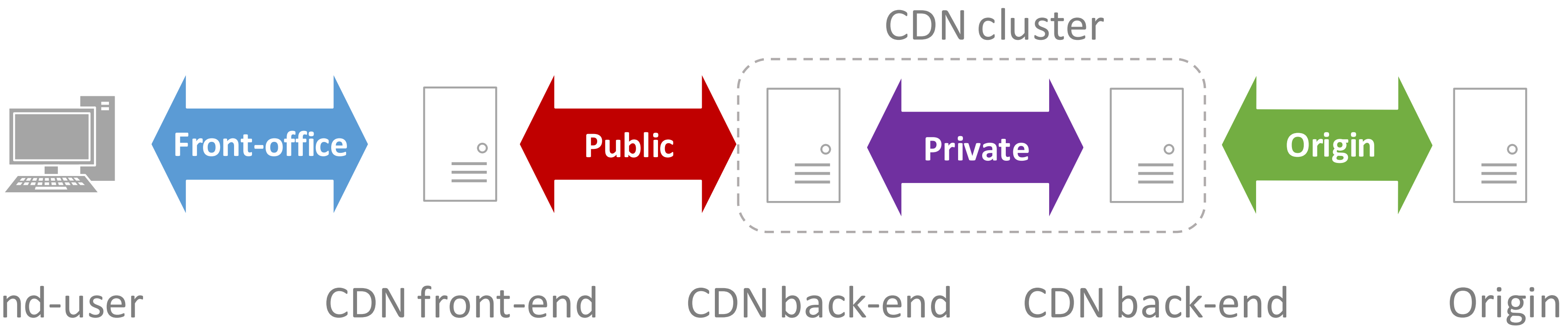
Public: front-end back-end over the Internet

A CDN perspective



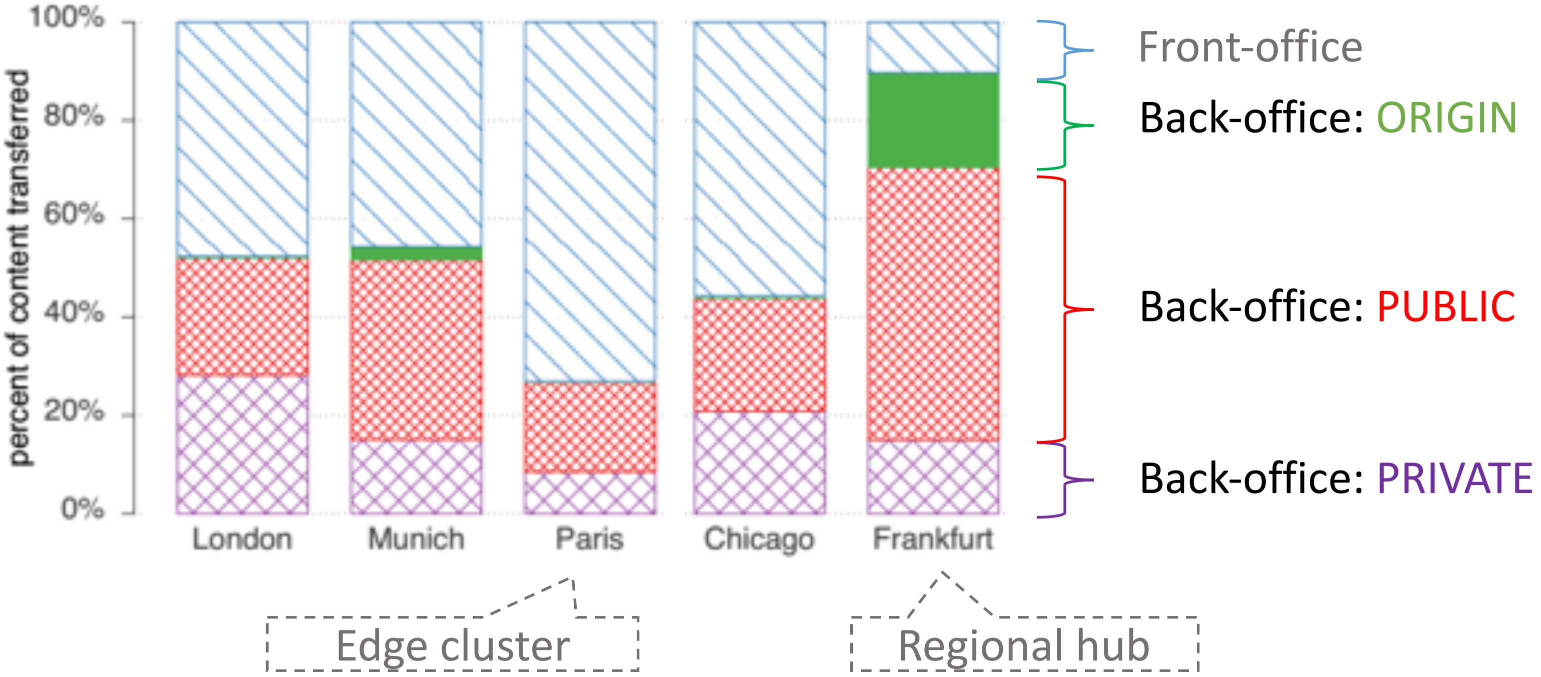
Private: within same cluster

A CDN perspective



Origin: inter-organization over the Internet

Back-office per location



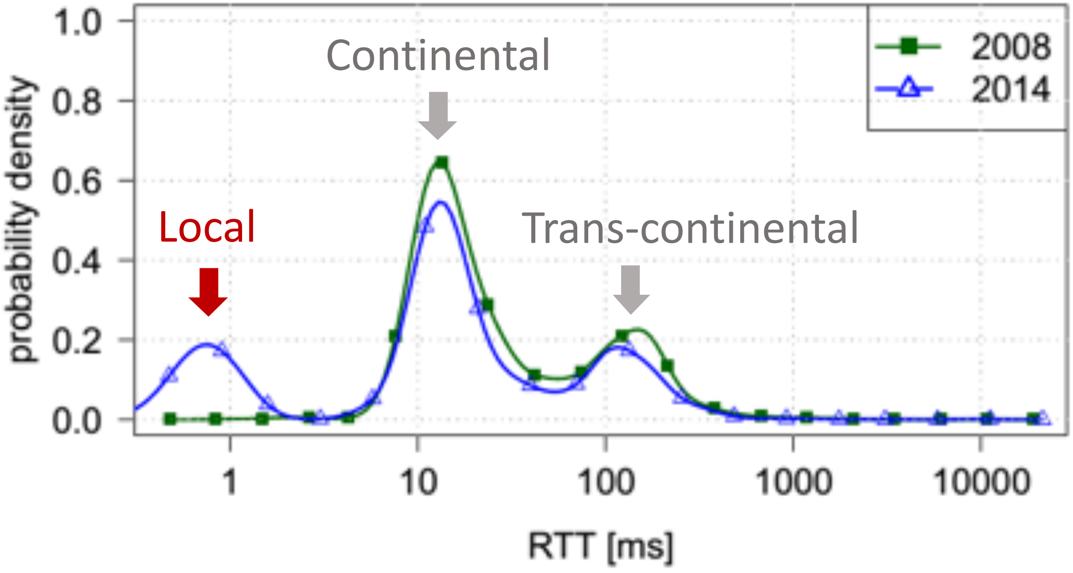
CDNs heavily rely on back-office traffic

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)



Continental

Local

Trans-continental

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

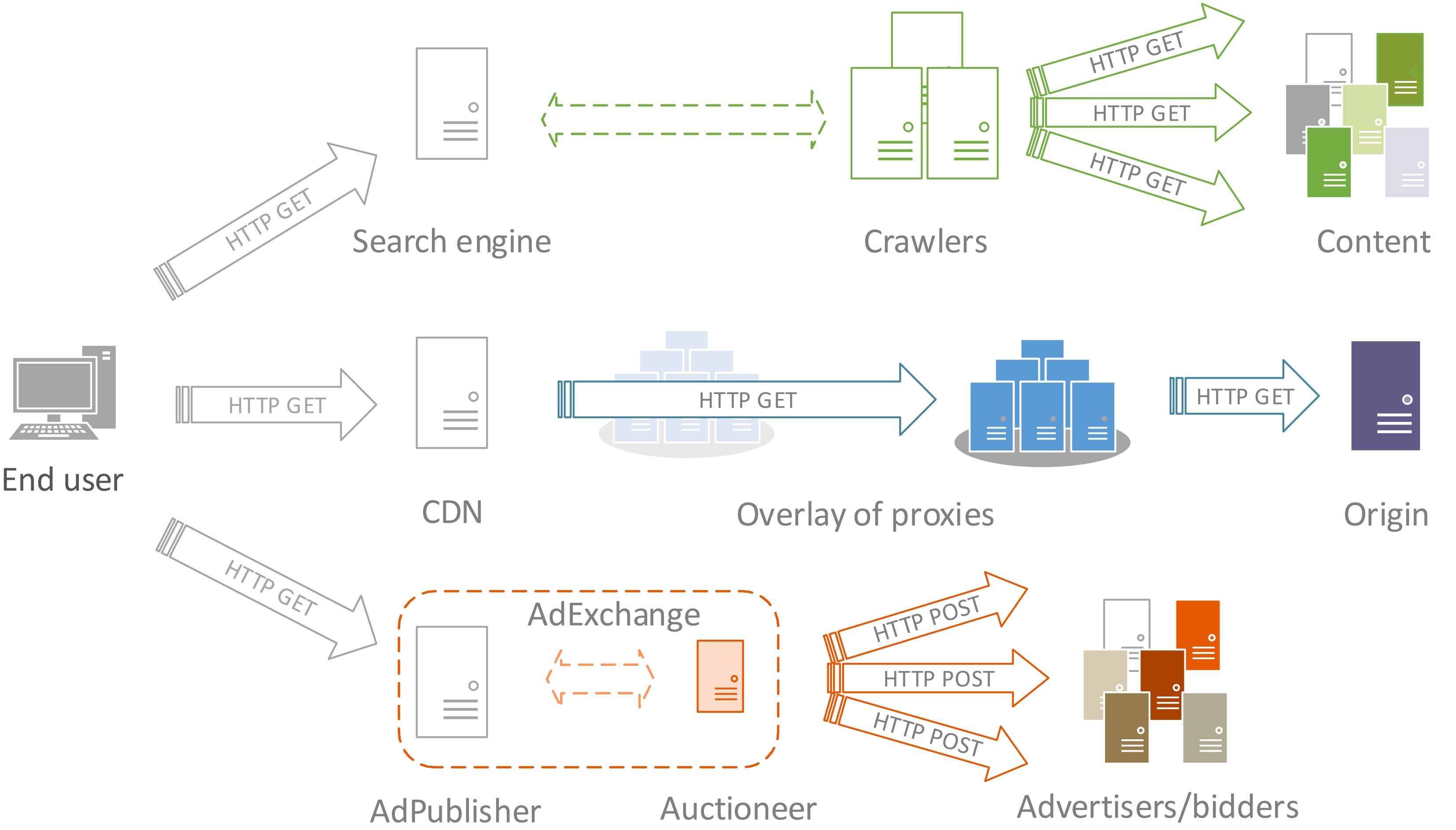
Performance evaluation:

- 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 front-office

The back-office (some examples thereof)