Back-Office Web Traffic on The Internet

Enric Pujol
TU Berlin
enric@inet.tu-berlin.de

Philipp Richter
TU Berlin
prichter@inet.tu-berlin.de

Balakrishnan Chandrasekaran
Duke University
balac@cs.duke.edu

Georgios Smaragdakis
MIT / TU Berlin / Akamai
gsmaragd@csail.mit.edu

Anja Feldmann
TU Berlin
anja@inet.tu-berlin.de

Bruce Maggs
Duke / Akamai
bmm@cs.duke.edu

Keung-Chi Ng
Akamai
kng@akamai.com

ABSTRACT

Although traffic between Web servers and Web browsers is readily apparent to many knowledgeable end users, fewer are aware of the extent of server-to-server Web traffic carried over the public Internet. We refer to the former class of traffic as *front-office Internet Web traffic* and the latter as *back-office Internet Web traffic* (or just front-office and back-office traffic, for short). Back-office traffic, which may or may not be triggered by end-user activity, is essential for today's Web as it supports a number of popular but complex Web services including large-scale content delivery, social networking, indexing, searching, advertising, and proxy services. This paper takes a first look at back-office traffic, measuring it from various vantage points, including from within ISPs, IXPs, and CDNs. We describe techniques for identifying back-office traffic based on the roles that this traffic plays in the Web ecosystem. Our measurements show that back-office traffic accounts for a significant fraction not only of core Internet traffic, but also of Web transactions in the terms of requests and responses. Finally, we discuss the implications and opportunities that the presence of back-office traffic presents for the evolution of the Internet ecosystem.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations.

Keywords

Network measurement; the Web; content delivery; online advertisements; real-time bidding; crawlers.

1. INTRODUCTION

The Web has not only revolutionized the way people publish, access, and search for content but, some would argue (e.g., [49]), has also evolved to become the new "narrow waist" of the Internet. Indeed, the HTTP protocol provides a common interface that many popular Internet applications rely on, including video, social networking, e-commerce, and software delivery. These applications are often supported by advertisements, which are also delivered via HTTP.

IMC '14, November 5–7, 2014, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2663716.2663756.

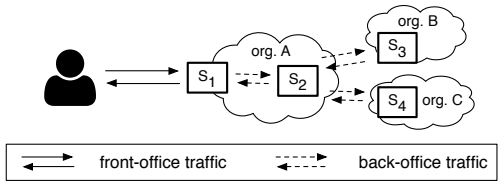


Figure 1: Front- vs. back-office Internet Web traffic.

Although an end user typically views a Web page as a unit, recent studies [17, 36] demonstrate that a single Web page often contains links to objects that are delivered by a large and diverse set of servers. For example, the creation of a single Web page may involve several Web companies as, e.g., parts of the Web page may be under the control of a content provider, a Web advertiser, a video streamer, a search engine, and/or a social network. Furthermore, even fetching an individual part of a Web page may involve many parties. For example, when an end user requests Web content, the delivery involves not only the servers that receive HTTP requests from the end user's browser, but also a whole service ecosystem consisting of proxies, content delivery networks (CDNs), ad-sellers, ad-bidders, back-end servers or databases, crawler bots, etc.

Thus, "there is more to content delivery than is visible to the eye," and, consequently, this paper explores the distinction between *front-office* and *back-office* Web traffic. The first refers to the traffic involving end users directly. The second refers to Web traffic exchanged between machines (e.g., the front-office servers and any other server which is part of the Web service ecosystem). Figure 1 depicts this distinction. Note that not all back-office Web traffic travels over the public Internet. Some is carried over private backbones or within data centers. In this paper, we focus on *back-office Web traffic on the public Internet*. For short, we henceforth use the term "front-office traffic" to refer to front-office Web traffic carried on the public Internet and similarly for "back-office traffic."

In contrast to back-office traffic, front-office traffic has long been studied, e.g., [8, 14, 17, 26, 27, 45, 47]. While there is some related work on Machine-to-Machine traffic in specific environments, e.g., in cellular networks [52] and within data centers [12, 13, 34], we are not aware of studies of back-office Web traffic on the public Internet. Liang et al. studied security-related aspects arising from CDN back-end communication [44] and for some specific other services, e.g., DNS, Gao et al. have characterized the corresponding Machine-to-Machine traffic [31].

The reason why previous work has focused mainly on front-office traffic is that end-user Quality of Experience (QoE) can be analyzed by observing front-office traffic, but back-office traffic is

often opaque. Today, more and more Web services also depend on some back-office communication over the public Internet e.g., to assemble Web pages, to perform search queries, to place advertisements, to conduct database transactions, to dynamically generate personalized content, Thus, service QoE now also depends critically on the architecture and performance of the back-office, which relies increasingly on the public Internet. This complexity makes measuring and characterizing back-office traffic challenging.

Among the difficulties faced in studying back-office traffic is that it is rarely present on the network links connecting end users to the Internet. Instead, back-office traffic can generally only be observed on backbone or inter-domain links. However, existing studies of inter-domain and/or backbone traffic [32, 42, 50] have not separated front-office and back-office Web traffic. Indeed, the back-office traffic component for any individual link depends highly on whether the link is on any of the routes between the involved servers. Thus, to observe this traffic requires a variety of vantage points. In this paper we analyze data collected from two IXPs, multiple links of a Tier-1 ISP, and a major CDN.

Web services that involve back-offices include content delivery, search, and advertisements. We focus on these because content delivery is responsible for a significant fraction of all Internet traffic, while advertisements (and in particular those in response to search) are responsible for a significant fraction of Internet revenues.

According to recent studies [32, 42, 48] CDN traffic accounts for more than 50% of Web traffic. This percentage is expected to further increase in part due to the increasing traffic volume attributed to video delivery [23]. Since CDNs operate sophisticated back-offices with distributed server infrastructures, some of this traffic is back-office traffic, which may or may not be routed via the public Internet depending on whether the CDN operates its own backbone [37] or not [54]. The rationale for this distributed infrastructure is the need to improve the end-user experience, react to flash crowds, mitigate attacks, and reduce the cost of content delivery using economies of scale [7, 43, 46]. CDNs are aware of the need to constantly improve communication between their front-end and back-end servers [28, 43, 54].

Search is one of the essential Internet Web services. Without search, the Internet would hardly be usable for most end users as their desired content would be difficult to locate. Search relies on a back-end database which is typically populated by crawling the Internet. For this purpose, search providers including Google and Microsoft operate distributed server infrastructures that crawl the Web. Web crawlers (also known as crawl bots), are orchestrated to partition the Web and index different parts of it to more efficiently cover it. Once the crawler bots have collected their data, they upload it to the back-end where it is processed in large data centers [11], where massively parallel indexing algorithms are used to enable fast search queries [19]. To deliver search results they rely on overlays and/or the above mentioned CDNs [22]. Thus, search engines contribute to back-office Web traffic.

To monetize their content, most Web sites rely on targeted online advertisements. In 2013, online advertising revenues in the United States were estimated to be $42.8 billion [2], an increase of 17% over the previous year. The increasing revenue stream of Web advertisement has given rise to another innovative part of the Web ecosystem: ad-sellers, ad-bidders, and ad-brokers—the ad-networks [10, 59]. These parties negotiate placement of advertisements in today's Web. In many instances, the selection of an advertisement does not take place until an end user visits a Web page where advertisement space is available. At this point, an auctioneer contacts the potential advertisers with information about the visitor's profile and a bidding process is initiated, but hidden from the user. Thus, a visit of an end user to a Web page may trigger a number of back-office connections. Advertisement content is often delivered via a CDN.

Thus, back-office Web Traffic is one of the principle but yet largely unexplored components of today's Web. The contributions of this paper are:

- We introduce the notion of back-office Web traffic and show that its contribution ranges on average from 10% to 30% per vantage point and can even exceed 40% for some time periods. The vantage points include two major IXPs, multiple backbone links from a major ISP, and a number of server clusters from a major CDN. We explore the reasons that different levels of contributions are seen at different vantage points.
- Our methodology allows us to identify and classify different types of back-office traffic including proxy services, crawling, and advertisement bidding.
- Our analysis demonstrates that back-office traffic characteristics differ from front-office characteristics. Moreover, they vary enough by service that individual services can be identified.
- We find, for example, that at one of the IXPs auctioneers have a 22% share of the back-office requests but only 1% of the bytes, while crawlers contribute respectively roughly 10% and 15% to both.
- Our analysis of data from a major CDN confirms what we observe in the wild: CDNs deploy sophisticated back-office infrastructures, and back-office Web traffic is significant.
- Given the volume of back-office traffic on the Internet and its importance for end-user QoE, we identify implications of our analysis on network protocols design and co-location strategies.

2. BACK-OFFICE COMMUNICATION

In this section we provide a brief overview of the typical (i.e., expected) communication patterns of Web services that create back-office traffic. Hereby, we distinguish four different cases: (a) proxies/intermediaries, (b) CDN services, (c) auctioneers, and (d) crawlers. Figure 2 provides an illustration of the expected exchange of HTTP messages. Note, however, that our analysis (Section 7) unveils richer and more complex communication patterns than those shown in the figure.

(a) Proxies/Intermediaries: An intermediary is a network entity that acts as both a client and a server. As shown in Figure 2(a), a Web proxy is an intermediary that acts as a client with the main purpose of forwarding HTTP(S) requests. Thus, Web proxies send and receive requests in a temporally correlated fashion. Forward and reverse Web proxies evaluate requests, check if they can be satisfied locally, and contact a remote server only if necessary. When intermediaries act as clients, they create back-office traffic, but when intermediaries act as servers, the traffic they create can be either front- or back-office traffic. We describe how to differentiate these two cases in Section 6.

(b) CDN Servers: CDNs typically operate front-end servers (i.e., reverse proxies) close to the end user as well as back-end servers. Back-end servers either host the content in data centers or are closer to the origin content server, depending on the CDN's deployment and operation strategy [46, 35, 57, 37, 18, 5, 3]. If the front-end does not have a requested object available locally, it fetches the object from another front-end, a back-end, or the origin server. Since the overall content delivery time has a direct impact on application performance, e-commerce revenue, and end user engagement [39, 23, 41], a number of optimizations for creating

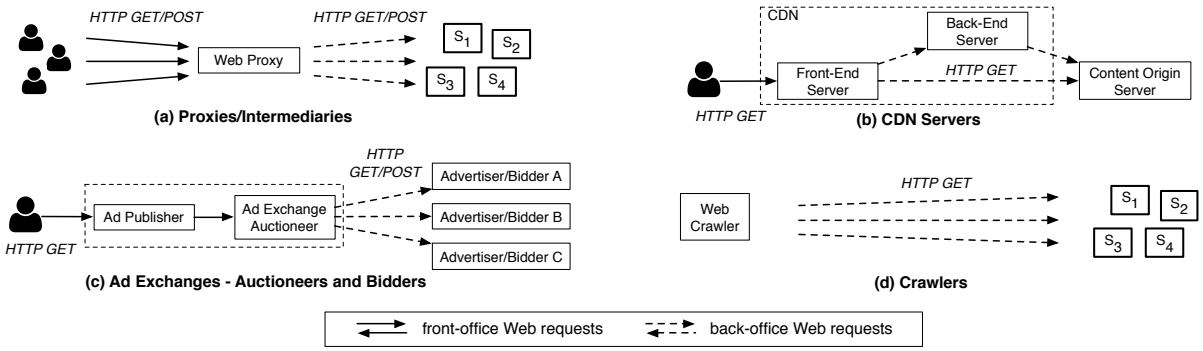


Figure 2: Back-office Web Traffic: typical HTTP requests made by Web proxies, CDNs, ad-exchanges, and crawlers.

overlays to improve end-to-end performance and for task sharing between front-end and back-end servers are deployed by today's CDNs [43, 54, 28, 40, 21].

(c) Ad Exchanges – Auctioneers and Bidders: As shown in Figure 2(c), advertisement exchanges consist of (i) publishers that sell advertisement space (ad space) on their Web pages, as well as (ii) advertisers that buy ad space on these Web pages. An exchange acts as a common platform to bring publishers and advertisers together. The matching between offered ad space on a Web site and interested advertisers is often performed using *real-time bidding* (RTB). Once an end user visits a Web page where ad space is available, the ad exchange auctioneer contacts the potential advertisers (i.e., the bidders), and provides information about the visitor to start a bidding process among the interested parties [10, 59, 55, 9].[1] A number of sophisticated tools together with visitor information suppliers optimize the bidding process for both the advertisers and the publishers. Hence, if RTB is used to place ads on a website, the visit of a Web page by an end user may trigger a large number of requests in the background. The final advertisement content is typically delivered via CDNs [9]. We note that today's Web advertisement ecosystem is complex and may involve many different types of back-office traffic, caused by a variety of different actors. In this paper, we solely focus on RTB-related activity, i.e., back-office traffic as a result of auctioneers interacting with bidders.

(d) Crawlers: Web crawlers continuously index the Web. To optimize crawling, each crawl bot is typically responsible for indexing a small part of the Web [11]. Indexing involves requesting the Web page as well as following embedded links [38, 16]. Web crawlers typically issue an order of magnitude more Web queries than regular end users. Best practices among the major search engines ensure that crawlers have appropriate reverse DNS entries along with well-specified user agents in order to avoid being blocked by Web sites.

Hereafter, we refer to back-office Web traffic as all Web traffic that is not exchanged between end users and servers. This includes traffic exchanged between intermediaries and Web servers (e.g., traffic between a CDN front-end server and a back-end server or between a Web proxy and a Web server), as well as traffic exchanged between automated hosts such as crawlers or auctioneers and any other Web server.

3. DATA SETS

In this work we rely on measurements collected at a diverse set of vantage points.

[1]The bidders may also contact other entities (i.e., *trackers*) to get information regarding the profile of the visitor [33, 59].

IXPs: Packet-sampled traces collected at two Internet eXchange Points (IXPs), which allow us to study back-office traffic in an inter-domain environment, as exchanged between hundreds of networks [6].

ISP: Anonymized packet-sampled traces collected from two transatlantic links from a Tier-1 ISP, providing a view of back-office traffic on long-distance links.

Access network: Anonymized packet dumps collected in a residential network of a Tier-1 ISP, revealing front-office traffic between end users and servers.

CDN: Web server logs from multiple servers in different locations within a large commercial CDN. These logs give us an inside view of back-office traffic created by a CDN.

Active measurements: Probes of IP addresses and DNS reverse lookups to identify Web servers.

This diverse set of traces allows us to study back-office traffic in a variety of locations, including between domains, on backbone links, and within a CDN. Table 1 summarizes the properties of our data sets.

The IXP traces are collected from the public switching infrastructure at two European IXPs. This includes a large IXP (L-IXP) with around 500 members and a medium-sized IXP (M-IXP) with around 100 members. Among the member ASes there are many CDNs, Web hosting services, cloud providers, and large commercial Web sites. We collect sFlow records [51] with a 1 out of 16K sampling rate. sFlow captures the first 128 bytes of each sampled Ethernet frame, providing us access to full network- and transport-layer headers and some initial bytes of the payload, allowing for deep packet inspection (DPI).

The ISP traces are collected from two transatlantic links on the backbone of a large European Tier-1 ISP. These links carry mainly transit traffic. We collect anonymized packet traces with a random packet sampling rate of 1 out of 1K. We also collect unsampled anonymized packet dumps in a residential network with about 20K end users of the same ISP.

The logs from the large commercial CDN encompass the activity of all servers at one hosting location in each of five large cities. Each log entry contains TCP summary statistics including endpoint IPs, number of bytes transferred, and initial TCP handshake round-trip latency. In addition, we received a complete list of all IP addresses used by the CDN infrastructure.

We also use active measurement data from the ZMap Project [25]. This data set contains a list of IPs, i.e., servers, that are responsive to `GET` requests on port 80 (HTTP) and SSL services on port 443 (HTTPS), spanning the time period from October 2013 to January 2014. In addition, we also make use of the data made public by the

Type	Name	Daily traffic rate	Collection	Period	% TCP	% Web of TCP
Exchanges	L-IXP	11,900 TB	sFlow (1/16K)	37th week 2013	84.40%	78.27%
	M-IXP	1,580 TB	sFlow (1/16K)	4th week 2014	97.22%	92.22%
Transit ISP	BBone-1	40 TB	Packet sampled (1/1K)	5th Feb. - 12th Feb. 2014	86.30%	64.58%
	BBone-2	70 TB	Packet sampled (1/1K)	4th week 2014	73.87%	78.74%
Eyeball ISP	RBN	35 TB	Packet dumps (unsampled)	9th. Jan. 2014	74.83%	79.45%
Server	CDN	350 TB	Server logs (unsampled)	24-25 Apr. 2014	95%	95%

Table 1: Summary of the vantage points and collected traces.

Name	#IPs	Method	C-O (%)	S-O (%)	Dual (%)
L-IXP	45.79M	DPI	96.90	2.74	0.36
		DPI+ZMap	93.85	2.74	3.40
M-IXP	1.9M	DPI	95.15	4.62	0.24
		DPI+ZMap	92.86	4.62	2.52
BBone-1	1.1M	DPI	92.26	7.56	0.18
		DPI+Zmap	86.62	7.56	5.82
BBone-2	4.5M	DPI	95.54	4.36	0.09
		DPI+ZMap	93.97	4.36	1.67

Table 2: Web activity of IPs: client-only (C-O), server-only (S-O), or both (dual) across vantage points.

authors of [18, 56] that disclose the set of IPs used by the Google infrastructure. When combining Google IPs with one of our packet traces, we use the snapshot of the Google IPs that corresponds to the last day of the trace.

4. IDENTIFYING BACK-OFFICE TRAFFIC

Given the above characteristics of back-office Web traffic, we next describe how we identify a significant fraction of it within our data sets. Our methodology involves three steps. First, we classify all IPs based on whether they are involved in any Web activity. Second, we classify their activities as either Web client, Web server, or both—client and server. Finally, we identify auctioneers and crawlers among the clients, bidders among the servers, and Web proxies among those that are acting as both clients and servers.

4.1 Web server endpoints

We focus only on those IPs for which we see Web activity in our traces, meaning either HTTP or HTTPS activity. For this we rely on well-known signatures to detect HTTP requests (`GET`, `POST`) and responses (`HTTP/1.{0,1}`) on any TCP port. For HTTPS we use signatures to match packets to/from port 443 that contain a SSL/TLS hand-shake, i.e., `Client Hello` and `Server Hello` messages [15]. We focus on IPv4, since IPv6 traffic accounts for less than 1% of the traffic across all data sets.

The result is a set of Web server endpoints, i.e., tuples that contain the IP address and the corresponding port number, as identified using the above-mentioned signatures. We then refer to all packets that are sent to/from one of the Web endpoints as Web traffic. Our methodology ensures that in the case of server IPs also hosting other applications on different ports (e.g., email), only their Web-related traffic is considered. Table 1 shows the percentages of Web traffic within our different data sets. As expected, this traffic constitutes a large fraction of the TCP traffic, ranging from 64% to 95% in each data set.

4.2 IP: Client/server/dual roles

Given that we have identified Web server endpoints, we next classify the Web activity of IP addresses to be client-only, server-only, or both (referred to as *dual* behavior in the following). We say that an IP address acts only as server if all of its traffic is related to its previously identified server-endpoint(s) (typically on port 80). If we see this IP address acting only as client, i.e., sending requests and receiving replies from other server-endpoints, it is classified as client only. If we see an IP address both acting as client and as server i.e., it runs a server on a specific port but also issues requests towards other servers, we classify its behavior as dual.[2]

Depending on the vantage point however, one may not see all Web activity a host is involved in. For example, a proxy server might exhibit only client-activity when monitored in the core of the Internet, and only server-activity when monitored in an access network. To tackle this limitation, we rely on a combination of passive and active measurements to uncover more IPs with dual-behavior as follows: we obtain a list of client IPs via DPI from our traces and then use the ZMap data set to check if these IPs respond to HTTP(S) queries. The ZMap data set provides lists of IPs that answer to a `GET` on port 80 or to an SSL handshake on port 443. Thus, if we see an IP address acting only as client, but we find it in the ZMap data set, we classify its behavior as dual.

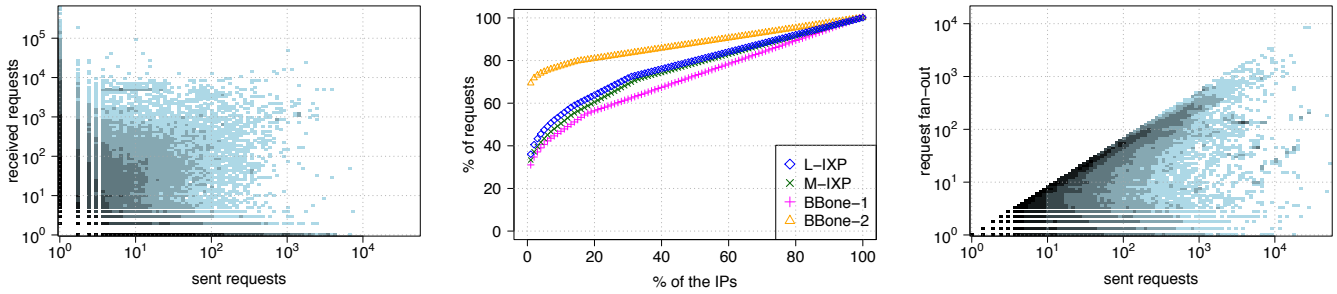
Table 2 shows the classification of IPs when only relying on DPI (first row for each vantage point), as well as after taking the ZMap data set into account (second row for each vantage point). We make three observations. First, with only DPI, roughly 90% of IPs are classified as client-only across all data sets. Second, a significant fraction of the server IPs also show client behavior e.g., with DPI we see in the L-IXP trace that 11% of the total number of servers also act as clients. Third, adding more information for identifying dual behavior helps e.g., with DPI+ZMap we see in the L-IXP trace that 55% of the servers behave also as clients. Indeed, the fraction of IPs acting both as clients and servers increases significantly across all vantage points when combining active and passive measurements.

There are two main caveats when using this classification approach, which likely result in some overcounting of dual hosts on the one hand, as well as some undercounting on the other hand.

One factor contributing to possibly overcounting dual hosts include the use of dynamically assigned IP addresses. If a dual host is assigned different IP addresses at different times, then each of those IP addresses may be classified as a dual host, even though at other times the same IP addresses act only as servers or, more commonly, only as clients. Dynamically assigned IP addresses are typically found in residential networks. Due to bandwidth limitations, these addresses do not serve a significant fraction of Web traffic or a significant fraction of Web requests. Nevertheless, to minimize the impact of dual hosts with dynamically assigned addresses on our statistics, we only count as servers IP addresses that appear in two consecutive snapshots of the ZMap data set, i.e., they replied to HTTP requests issued two weeks apart.

On the other hand, our methodology may undercount dual hosts because is not able to detect more complex cases of dual behav-

[2]Recall that requests are typically issued with *ephemeral* source port numbers.



(a) Scatterplot: # sent vs. received requests sampled per IP in the L-IXP trace.

(b) Cumulative sum: # of requests sampled per IP, ranked by activity.

(c) Scatterplot: # sampled sent requests vs. fan-out per IP in the L-IXP trace.

Figure 3: Web IP activity: request frequency and fan-out.

ior, e.g., if a server has multiple interfaces and thus sends/receives traffic from multiple IP addresses, each acting only as a client or server. Furthermore, while we can uncover the server activity of a host that acts only as a client in our traces using the ZMap data set, we lack the tools to uncover the opposite case i.e., to uncover the client activity of a host that acts only as a server in our traces.

Along with the methodology previously described, we also exploit a list of servers provided to us from a large CDN. Henceforth, we distinguish three different classifications of IP addresses, *IPs-CDN*, *IPs-DPI*, and *IPs-ZMap*, based on our confidence in their accuracy. We are most confident in the accuracy of the classifications of the CDN addresses. Based on the DPI performed on a trace, we can also be very sure that an IP at that point in time was used in the identified fashion. The ZMap data set comes with the largest degree of uncertainty, but covers a wide range of IPs. Note that the same IP address may appear in more than one of these sets of classifications.

4.3 IP activity

Figure 3(a) shows a scatter plot for all IPs seen in the L-IXP trace, where we plot each IP according to the number of sampled Web requests it sent vs. received on a log-log scale. This plot highlights the different classes of IPs. The server-only IPs only receive requests and are scattered along the y-axis. The client-only IPs are scattered along the x-axis. The dual-role IPs are scattered across both axes. While the IPs with significant activity in both roles are likely to be intermediaries/proxies, we first take a closer look at some of the other heavy-hitters, i.e., those that only act as clients.

Figure 3(b) shows the cumulative percentage of the number of sampled `GET` and `POST` requests issued per IP, sorted by the number of observed requests. The most striking observation from this plot is that less than 1% of all the IPs are responsible for more than 30% of all requests across the IXP and backbone vantage points. Indeed, we estimate from the sampled data that the IP that contributes most requests at the L-IXP contributes more than 5M requests per hour on average, while at BBone-2 this number is roughly 310K requests per hour. These are unlikely to be humans — rather this behavior points towards a crawler, a proxy, or an auctioneer.

Accordingly, Figure 3(c) shows for each IP the number of sampled requests it sent vs. the number of IPs to which these are sent, namely the *fan-out*, on a log-log scale. While some of these clients contact a huge number of servers, there are also clients that send an enormous number of requests to a rather small set of server IPs (bottom right). As shown in Table 3, when inspecting the types of the requests we find an unexpectedly large number of `POST` requests. These can be attributed to the intensive use of protocols for Web services (e.g., SOAP). Closer inspection shows that for clients with an extraordinary high number of requests the fraction of `POST` requests is larger when compared to clients with low numbers of requests. Based on these observations, we now differentiate between proxies, auctioneers, crawlers, and bidders.

Name	Total	GETs	POSTs	CHellos
L-IXP	76.36M	71.6%	11.5%	16.9%
M-IXP	2.65M	78.9%	3.7%	17.4%
BBone-1	1.81M	88.3%	5.2%	6.5%
BBone-2	2.92M	58.1%	8.2%	33.7%

Table 3: Web IP activity: sampled HTTP/HTTPS requests.

4.4 CDNs, proxies, and other intermediaries

Typical examples of dual-behavior hosts are proxies such as those deployed by some institutions and forward and reverse proxies such as those operated by CDNs and content providers (CPs). However, intermediaries can, in fact, serve many other purposes. For instance, there are many kinds of proxies in the wild that affect a significant number of connections [58] and which are not operated by CDNs or CPs. In addition, intermediaries at hosting or or cloud service provider networks may not necessarily operate for the single purpose of request forwarding. While keeping this in mind, we focus on identifying some of the intermediaries that are operated by CPs or CDNs, which we hereafter refer to as *Content Delivery Proxies* (CDPs).

These are the steps we follow to classify CDPs. Along with the IPs in the *IPs-CDN* set, we select as potential candidates those intermediaries for which we sampled more than 5 requests (heavy-hitters).[3] We then check the origin AS of the corresponding subnet, and manually inspect if the WHOIS information reveals that the address is registered to a known CP or CDN. Since this check is not sufficient to reveal cases in which front-end servers and caches are embedded in ISP networks and use the address space registered to those networks, e.g., Akamai and Google Global Cache, we also check for DNS host-names and use the techniques reported in [18, 56] to attribute IPs to content providers and CDNs.

Based on the previous manual identification, we are able to classify among the list of intermediaries some of the front-ends of 8 well-known organizations such as Google, Akamai, Limelight or EdgeCast. We find more than 36K (15K) IPs in the L-IXP (M-IXP)

[3]Note that 5 sampled requests correspond to an estimated number of roughly 80K requests per week for the IXP traces and to 5K requests per week in the BBone links, respectively.

traces. We also find CDPs that are active on the transatlantic links i.e., 9K and 19K for the BBone-1 and BBone-2 traces.

4.5 RTB: Auctioneers and bidders

The bidding process between auctioneers and bidders is generally done using Web services, such as Google AdExchange or Facebook Exchange. Bidders register with the auctioneer and provide a URI on which they accept offers from the auctioneer and reply with their corresponding bid after a maximum time threshold, often around 100ms [1].

Thus, one of the distinguishing characteristics of auctioneers is that they typically send their offers via `POST` to the potential bidders. The bidders, in turn, receive a large number of `POST` requests from a relatively small number of hosts (the auctioneers). This fits nicely with our earlier observation that there are many more `POST` requests in today's Internet than were observed in the past. In particular, an examination of traces from L-IXP over the past three years shows that the fraction of requests of type `POST` has increased by 80% over that time. Indeed, for each user request which involves an advertisement there may be multiple bidders contacted.

Given that the market for real-time bidding (RTB) is heavily concentrated among a relatively small number of players accounting for the major share of the RTB activity [59], prime candidates for auctioneers are IPs sending large numbers of requests to a comparably small set of IPs (bidders), which in turn receive a large number of requests from a relatively small number of IPs. As bidders can provide customized URIs to auctioneers, we cannot identify auctioneers and bidders in a straightforward manner using payload signatures. Instead, we identify auctioneers and bidders based on partially available URL strings as follows: we first obtain a list of partial URLs sent by the heavy-hitters and select those IPs whose HTTP requests contain in the absolute path and/or query parts of the URL strings such as `ad`, `bid`, or `rtb`. Then, for each of these IPs, we check if its corresponding subset of requests has a fixed structure, i.e., we observe many repetitions of the same partial URL (only the value fields in the query part change), and we mark these IPs as candidates for potential auctioneers. We then manually validate that these IPs are operated by organizations offering RTB services by relying on meta-information such as reverse DNS, WHOIS, and publicly available API documentation of some Ad exchanges. Having a list of auctioneers, we now inspect the corresponding destination IPs, further manually verify the corresponding URL strings to be bidding-related, and mark these IPs as bidders.

With this method we are able to manually identify 316 IPs used by auctioneers and 282 IPs used by bidders in the L-IXP trace. We were not able to identify bidding activity in the M-IXP trace, and also did not identify any ad exchanges co-located at M-IXP. Nor did we find any bidding activity in the backbone traces, perhaps because of the high delay in transatlantic links.

4.6 Web crawlers

One of the distinguishing characteristics of Web crawlers is that they issue many Web queries and then upload their results to the search infrastructure. The queries constitute a large number of `GET` requests sent to a diverse set of servers belonging to different organizations.

For each data set we use the heavy hitters, in terms of `GET` requests, as candidates. Specifically, we pre-select IPs for which we sample at least five queries and then try to verify that they are Web crawlers as follows. It is a best common practice to make crawlers clearly identifiable by setting up reverse DNS entries and including an appropriate user-agent with each request.[4] Thus, we search for host-names that include indicative strings such as *bot, crawl, spider* and select those that can be either automatically validated via the user-agent or validated with a manual inspection of the corresponding reverse DNS entry. scp

Name	CDPs	Bidders	Auctioneers	Crawlers	Other
L-IXP	36054	282	316	3920	151095
M-IXP	15116	0	0	541	4417
BBone-1	9330	0	0	81	1214
BBone-2	19890	0	0	894	2669

Table 4: Back-office: IP classification.

With this method, we identify 3920 and 541 crawlers in the L-IXP and M-IXP traces. Surprisingly, we also find crawlers in the backbone traces: 81 and 894 for the BBone-1 and –respectively– BBone-2 traces. We see activity from well-known search engines e.g., Google, Bing, Yandex, and Baidu, as well as from smaller search engines and research institutions. To put these numbers into perspective, we use the ZMap reverse DNS data set (i.e., all IPv4 PTR records) and search for host-names of crawlers belonging to three major search engines, which are well-defined in publicly available documents provided by these engines. The percentage of crawler IPs for which we see activity in the L-IXP trace is 7%, 23%, and 51% for three of the major search engines.

Summary

We find that most IPs that are part of the Web ecosystem are clients, but there are a substantial number of Web intermediaries across the vantage points e.g., just by inspecting the L-IXP trace with DPI, we find that 11% of the Web servers also act as Web clients. However, after combining our passive data with active measurements, we discover that many of the client IPs in the traces also act as servers, which is not visible when purely relying on passive data. As a consequence, the number of IPs with dual behavior increases e.g., for the L-IXP trace more than 50% of the server IPs exhibit also client behavior. As shown in Table 4, after we inspect the heavy-hitter IPs, we are able to find activity from content delivery proxies, ad auctioneers, ad bidders, and crawlers in the L-IXP trace, as well as crawling and intermediary activity as seen from the other vantage points.

5. WEB BACK OFFICE: A CLOSE LOOK

To better understand which players are involved in back-office services, we next take a closer look at its components in the L-IXP trace.

Auctioneers and bidders: We identify more than 300 IPs that are auctioneers. These IPs are operated by four different organizations that offer real-time bidding: Two search engines, an online social network, and a major Web portal. Each of these organizations operates at least one AS and the IPs are hosted in the corresponding AS. With regards to the number of IPs per AS we see a rather uneven distribution: The top one hosts 83% of the IPs. The same holds for the distribution of the number of requests: the top organization is involved in 55% of the bids, the others in 32%, 10%, and 3%.

These auctioneers communicate with a rather small set of bidders (282 IPs). The IXP data shows that many of the auctioneers

[4]See, for example Google `https://support.google.com/webmasters/answer/80553?hl=en`, and for Microsoft Bing `http://www.bing.com/webmaster/help/how-to-verify-bingbot-3905dc26`

are co-located with the bidders (both the AS of the auctioneer and the AS hosting the bidder are members of the IXP), with the bidders residing in 42 different ASes. This confirms that bidders are following the recommendations by the auctioneers to locate their servers close by in order to adhere to the strict deadlines of auction-based advertisements. Bidder IPs are typically contacted from all four identified auctioneering organizations. Thus, bidders do not use different IPs for different auctioneers and often cooperate with all of them. The likely motivation is that advertisers try to maximize their bidding opportunities (i.e., receiving offers from all organizations). Moreover, at first glance the number of bidders may appear small but this is mainly due to aggregation. Indeed, most IPs belong to advertisement aggregators.

With regards to the ASes that host the bidders, we find, surprisingly, that a very large hosting service provider dominates with a share of 34%. Indeed, even the remaining bidders are mainly located in the ASes of other Web hosting providers. This finding indicates that today's major players in the Web ecosystem often do not operate their own infrastructure either in terms of an AS or in terms of a data center. They rely instead on cloud services and Web hosting companies. The second AS with the most bidders belongs to, surprisingly, a company that operates a search engine. Indeed, this search engine is involved in all services: it is a search engine, an auctioneer, and even bids on other ad-network auctions. This finding illustrates the complexity of the advertising ecosystem, where different types of business relationships exist between organizations that offer multiple services to both advertisers and publishers, and who may also partner with some of them.[5]

Crawlers: We identify more than 3K crawler IPs from 120 different ASes. Among the ASes, there are two hosting more than 72% of the crawler IPs. These are related to two popular Web search engines. We also see crawlers of 3 other well-known search engines, each with roughly 100 crawlers. Then there is a gap with regards to the number of crawlers per AS as the remaining crawler IPs are hosted in many different ASes. Inspecting the user agent and the reverse DNS entries allows us to conclude that these are mainly associated with specialized search engines.

With regards to the number of requests, the four top contributors all belong to major search engines. The top three/four account for 94/96% of the requests. The fourth accounts for only 2% of the requests. Even though the crawling activity is directed towards more than 4.2K ASes, a single AS receives more than 43% of the requests — a Web hosting provider. The second most popular AS is another hosting provider and both are members of the IXP. Overall, the members account for more than 80% of the received crawling requests. In terms of request popularity, the first AS that is not a member of this IXP is yet another hosting provider and receives roughly 1% of the requests. Overall, the top crawling search engine AS and the top receiving hosting provider AS account for more than 20% of all crawling-related requests.

Content delivery proxies: We identify more than 30K intermediary IPs from 8 well-known CPs and CDNs, scattered across hundreds of different ASes, interacting with IPs from more than 1K different ASes. The CDPs are responsible for roughly 17% of the requests from heavy-hitter IPs. While one expects many of the front ends to be located in ISP networks, a close inspection of destination IPs reveals that some of the back-end servers are also located within the ASes of ISPs, and not in the AS of the CDN. In fact, we observe requests for the content of a major online social network (OSN) where both source and destination IPs are operated by a major CDN, but neither of the endpoints is located within the AS of the CDN or OSN. We also find other more typical scenarios, such as CDNs fetching content directly from OSNs and from large-scale hosting provider ASes.

Other intermediaries: The rest of IPs in the intermediary list (roughly 151K) are located in more than 7K ASes. They contact 399K servers in 10K different ASes. While we strongly suspect that most of these are indeed Web proxies, we cannot be certain. Indeed, on the one hand, one of the heavy-hitters IPs in this set — an oddball — is hosted in an unexpected AS. This oddball IP is serving both ad-related images to a CDN and acting as a Web auctioneer. On the other hand, we see several organizations that use resources from cloud services to set up their own virtual CDNs. A close analysis of which ASes are hosting the heavy hitters shows that most of these ASes are hosting and/or cloud service providers (8 out of 10). There is, however, more diversity in the destination ASes: we find hosting providers, CPs, OSNs and CDNs. We see that a single hosting/cloud service provider is responsible for 21% of the requests issued by IPs in this set. This observation highlights the importance of cloud service providers in the back office of the Web ecosystem once again.

6. BACK-OFFICE TRAFFIC: ESTIMATION

The raw numbers of IPs involved in back-office activity do not tell us much about how much back-office traffic we observe from our vantage points — estimating the volume of this traffic is the topic of this section.

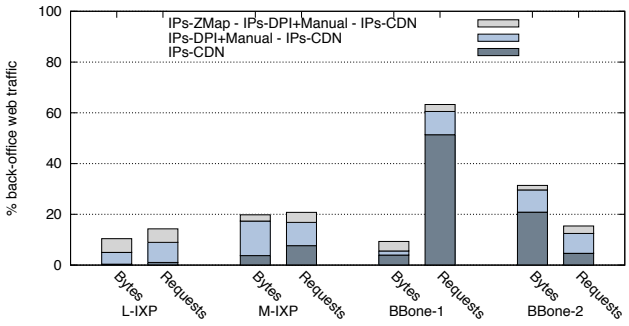
6.1 Across vantage points

The methodology presented in Section 4 allows us to classify IPs as Web crawlers, auctioneers, bidders, and intermediaries such as CDPs. Thus, we can now quantify the amount of back- and front-office Web traffic. For a packet to be classified as back-office traffic, we require that both the source as well as the destination were previously identified as part of the back-office ecosystem. More specifically, we require that the source IP address was previously identified as belonging to an intermediary, crawler, or auctioneer, and the destination IP:port pair matches one of our identified Web server endpoints. We then tag this packet as back-office traffic, issued by the source, namely a crawler, auctioneer, CDP, or other intermediary.

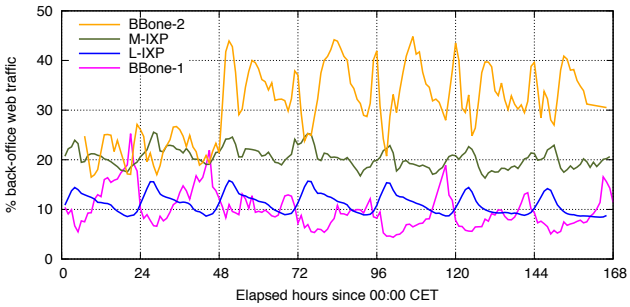
Recall from Section 4 that we rely on passive and active measurements to uncover intermediaries, as well as on manual identification of crawlers and auctioneers, and lastly on a list of CDN servers. To account for the varying degrees of certainty when using these data sets, we distinguish between three different classes when quantifying back-office traffic. In particular, we consider back-office traffic caused by servers in our CDN data set (*IPs-CDN*), caused by servers we identified using DPI and manual inspection (*IPs-DPI+Manual*) and the back-office traffic caused by servers identified using the ZMap data set (*IPs-ZMap*).

Figure 4(a) shows the percentage of back-office traffic of the total Web traffic for each vantage point as a stacked bar plot. Thus, we depict the volume of back-office traffic found with the different methods: (a) information from the CDN only (the bottom bar), (b) information from the *IPs-CDN* and *IPs-DPI+Manual* (the sum of the bottom and the middle bar), (c) all information including ZMap (the sum of all bars). Across all vantage points we see at least 5% back-office Web traffic using the *IPs-CDN* and *IPs-DPI+Manual* set of IPs, confirming that back-office Web traffic is a significant contributor to today's Internet Web traffic. Even when only using the *IPs-CDN* data set, we see at least 4% back-office traffic at all vantage points except for L-IXP. This does not mean that the

[5]For an example of such a complex relationship see `http://ir.yandex.com/releasedetail.cfm?releaseid=828337`



(a) % of Web traffic which is back-office across vantage points.



(b) % of Web traffic which is back-office over time (IPs-ZMap).

Figure 4: Fraction of back-/front-office Web traffic across vantage points.

CDN is not active here but most of its traffic is front-office traffic. In terms of requests, the fraction of requests associated with back-office traffic is even larger with a minimum of 9% when using the *IPs-CDN* and *IPs-DPI+Manual* sets. This points out that some components of back-office traffic are associated with smaller transactions. But asymmetric routing—meaning the forward and return path do not use the same link— are likely the explanation for the extreme difference at BBone-1, where we see a huge number of back-office requests but only a relatively small percentage of back-office bytes. When we include the ZMap server IPs, the percentages of back-office traffic increases to more than 10% across all vantage points.

We next dissect the back-office traffic by type of activity using the *IPs-DPI+Manual* and the *IPs-ZMap* information. We illustrate our findings in Table 5, where we attribute back-office traffic to the entity that acts as client. We find that CDPs contribute 11% and 12% to the back-office requests and bytes in the L-IXP trace. The crawlers contribute 15% and 10% to the back-office requests and bytes, respectively. Surprisingly, the auctioneers are the big contributors to the number of requests with a share of 22% but only 1% of the bytes. The rest of intermediaries contribute more than 76% and 50% of the back-office bytes and requests. The situation differs for the other vantage points, where CDPs clearly dominate the share of bytes and requests with at least 50% of the bytes and 65% of the requests.

Figure 4(b) shows how the percentages of back- and front-office bytes change over time using time bins of one hour. The percentages never decrease below 5% but can even exceed 40%. While some traffic is triggered by end-user action, activities such as crawling and cache synchronization are not. We see that, particularly for the two IXPs, the percentage of back-office traffic increases during the off-hours. The percentage of back-office traffic for BBone-2 increases on the third day of the trace by more than 10%. This increase may be due to (a) a routing change or (b) a change in the operation of the application infrastructure or (c) a change in the popularity of a Web application. In addition, we see more variability for the individual backbone links than for the IXPs. A likely explanation for this is that the IXPs aggregate the information from thousands of different peering links. Similar observations hold for the percentages of back-/front-office requests and responses (not shown).

Name	% of	CDPs	Auctioneers	Crawlers	Other
L-IXP	Bytes	12.1%	1.1%	10.3%	76.5%
	Requests	11.8%	22.5%	15.1%	50.6%
M-IXP	Bytes	73.3%	-	1.5%	25.2%
	Requests	65.7%	-	3.2%	31.1%
BBone-1	Bytes	50.7%	-	<0.1%	49.2%
	Requests	95.3%	-	<0.1%	4.6%
BBone-2	Bytes	93.6%	-	<0.1%	6.3%
	Requests	73.7%	-	4.3%	22%

Table 5: Classification of back-office Web traffic.

6.2 Across peering links

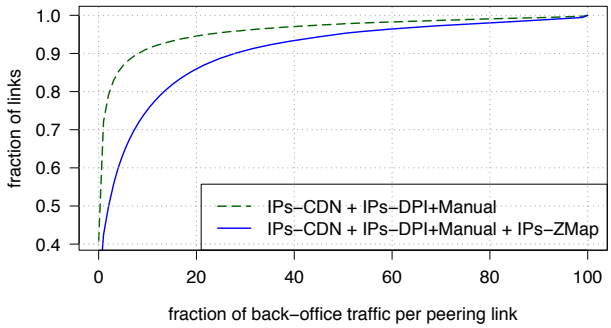
The two backbone traces illustrate that there can be notable differences in terms of percentages of back-office bytes and requests on different links, suggesting that links should be examined individually. Hence, we now take advantage of our ability to dissect the traffic seen on hundreds of individual AS-AS links at L-IXP.

Figure 5(a) shows the fractions of back-office traffic per AS-AS link (of the total traffic carried over it), where we sort them by the fraction of back-office Web traffic. We see that the fractions vary drastically from 100% to 0%. Indeed, 18.2% (10.9%) of the peering links carry more than 15% (7%) back-office bytes when relying on the *IPs-ZMap + IPs-DPI+Manual* (*IPs-DPI+Manual*) data set. On the other hand, 25.5% (40.8%) of the peering links carry no back-office traffic at all. In order to get a better understanding of the most important AS-AS links, we inspect more closely the top-10 traffic-carrying links that have a fraction of back-office traffic larger than 95%. We find four links between cloud providers and content providers, three links between search engines and hosting providers, two links between CDNs and content providers, and one link between a content provider and an online advertisement company. This analysis illustrates the diversity of the players contributing to the back-office Web traffic.

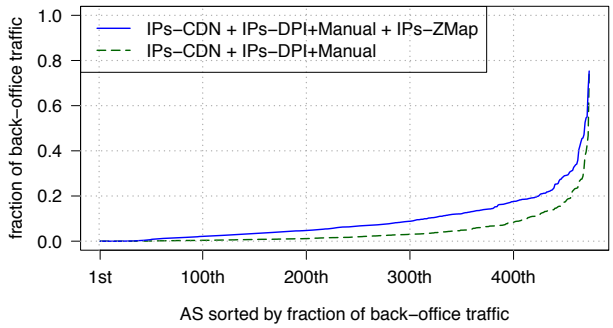
If we aggregate the information to the level of IXP member-ASes, the overall observation does change a bit, as shown in Figure 5(b). We do not see member ASes that exchange only back-office Web traffic. They all have at least 20% front-office Web traffic. Nevertheless, most have some fraction of back-office traffic. There are 19.2% (18.0%) of the members with more than 15% (7%) back-office bytes for the *IPs-ZMap + IPs-DPI+Manual* (*IPs-DPI+Manual*) data set. Among the networks with the highest share of back-office traffic are cloud providers, hosting providers, a search engine, and an advertisement company.

Summary

We find a significant percentage of back-office Web traffic in our traces, yet this percentage varies from vantage point to vantage point. Indeed, the back-office traffic carried over the backbone links is mostly dominated by CDPs. The picture differs when looking at IXPs, where we can monitor multiple links at once. While most of the back-office traffic there is also due to CDPs and other intermediaries, real-time-bidding and crawling also contribute a significant share of bytes and Web requests. Our analysis illus-



(a) ECDF: fraction of back-office traffic—per link.



(b) Fraction of back-office traffic—per member AS.

Figure 5: Back-office traffic across peering links—L-IXP.

trates that a significant part of the back-office traffic is not triggered by end users. We see that besides the expected players such as CDNs, search-engines and advertisement companies, cloud service providers and Web hosting providers are also responsible for a large fraction of back-office Web traffic. Indeed, they play an important role since they provide the resources necessary to build back-office infrastructure. Thus, back-office traffic is present in multiple links on the AS-level, revealing a complex ecosystem of players in support of the Web.

7. BACK-OFFICE: CHARACTERISTICS

Next, we examine some of the temporal and spatial characteristics of the different types of back-office Web traffic. In this section, we focus exclusively on back-office traffic caused by hosts that we manually identified to be CDPs, crawlers, auctioneers, or bidders.

7.1 Temporal behavior

To illustrate the temporal characteristics of some of the key players in the Web back-office, Figure 6 provides a time series plot of the number of requests seen at L-IXP and issued by content delivery proxies (CDPs), auctioneers, and crawlers, where we normalize the number of issued requests by the average number of crawler requests.

On the one hand, crawlers display rather constant activity throughout the week, which is the reason we use them for normalization. This constancy is to be expected because the request activity is not triggered by humans. The request patterns of the CDPs and auctioneers, on the other hand, display a diurnal pattern due to their connection to end-user activity. Interestingly, the rate of decrease between peak and off hours is larger for the auctioneers than for the CDPs. A possible explanation for the larger decrease is that the bidding process is a multiplicative factor of the end-user activity, i.e., one page visit triggers an auction involving multiple bidders. In terms of traffic volume (not shown), both CDPs and auctioneers exhibit typical diurnal variation while crawlers do not. While crawlers and CDPs dominate in terms of traffic contribution, auctioneers only contribute a tiny share of traffic. This is expected, as the bidding process involves numerous, but small, transactions.

7.2 Spatial behavior: Request forwarding

Noticing that many HTTP requests include the `Via` header, we next take a closer look at Web request forwarding. There are two HTTP header fields that are especially relevant for proxies: `Via` and `X-Forwarded-For`. The former indicates if a request has been forwarded through a proxy or a gateway; multiple `Via` field

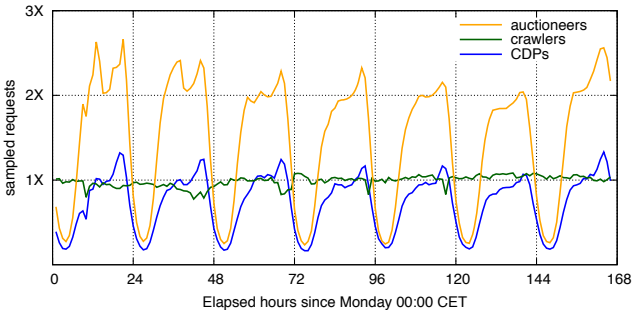


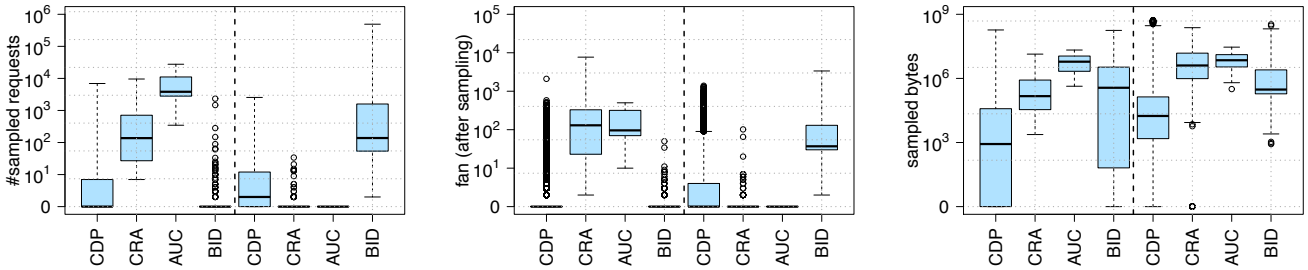
Figure 6: Time series: requests per hour by CDPs, auctioneers, crawlers (normalized by crawler requests).

values represent each host that forwarded the HTTP request. The `X-Forwarded-For` field contains the IP address of the original client, followed by the IP addresses of the proxies on the path to the server. Thus, if present and correctly set, this header includes the original client IP as well as all proxy IPs that forwarded the request. This allows us to elucidate the complexity of the back-office by showing how far requests are forwarded and via how many Web proxies.

Inspecting these headers requires the ability to inspect the complete payload of a packet. We have full payloads for the BBone-1 and BBone-2 traces, and we extract from them the requests according to the previous *IPs-CDN+IPs-DPI+Manual* classification. Recall that a significant fraction of the requests in these traces are issued by IPs in *IPs-CDN*. Thus, the following analysis may be biased by the behavior of this particular CDN.

The `Via` header field indicates that while 12% of the requests traversed one proxy, another 77% traversed two proxies. We even observed a few requests that traversed seven proxies. With the `X-Forwarded-For` header field we now reconstruct the paths of the requests, i.e., for each request we extract the list of proxy IPs and append the destination IP at the end. Perhaps surprisingly, we find many private IP addresses among the IPs in the `X-Forwarded-For` headers, suggesting that either (a) end users use proxies on their premises and/or (b) proxies are located within data-center networks, where private IP address space is used. We argue that the second case dominates as the first IP in the list is typically publicly routable, e.g., belonging to an end user.

Out of the 1M requests we consider, we find 766K different client IPs appearing in the first position of the reconstructed paths.



(a) Sampled requests sent/received (left/right). (b) Fan-out/in (left/right). (c) Sampled bytes sent/received (left/right).

Figure 7: IP characteristics: Content delivery proxies (CDPs), crawlers (CRA), auctioneers (AUC), and bidders (BID) — L-IXP.

These IPs map to 7.9K different ASes, and around 94% of these IPs appear only once. The last IP in a reconstructed path may be an origin server or another proxy. We observe 5.9K different IPs appearing in this position, and they map to 350 ASes. Note that an observed request may be further forwarded. Finally, for the subset of IPs that do not appear at the beginning or end of a path (i.e., forwarding proxies), we find 16.5K different IPs scattered across 885 ASes. Notably, around 2.7K of the IPs are not publicly routable, yet they sum up to 40% of the occurrences.

To conclude this section, we take a look at the geographical characteristics of request forwarding. For this exercise we focus on the subset of requests detected via the *IP-CDNs* set. We then use validated information about the geographical coordinates of these CDN servers and rely on this CDN's commercial geolocation product to geolocate end users. We observe the following typical scenario for CDN activity, as seen from these backbone links: an end user issues a request to a front-end server (at 10 to 1000 km distance), this front-end contacts a back-office server within a CDN cluster (0 km distance). This back-office server in turn forwards the request to another back-office server that is still on the same continent (10-1000 km). Then, this proxy forwards the request to an origin server or to another back-office proxy across the Atlantic.

7.3 Communication patterns

Next, we return to examine the activity of the IPs in the L-IXP trace. Figure 7(a) shows, for the crawlers, auctioneers, bidders, and CDPs, a box plot of the number of sampled back-office HTTP/HTTPS requests we observed. Note that we only analyze back-office traffic characteristics here, and do not, for example, consider any requests related to the front-office activity of CDPs. We separate sent and received requests on the left and right sides of the plot. Accordingly, Figure 7(b) shows the observed fan-out (i.e., to how many different hosts were requests sent) and fan-in (i.e., from how many hosts were requests received). Figure 7(c) shows the number of sampled bytes received/sent.

Auctioneers and bidders: Auctioneers are the most active in terms of number of requests sent. From our sampled data we estimate that the average number of bid requests/hour issued by these IPs is roughly 232 million. This estimate implies that an average auctioneer IP issues more than 700K bid requests/hour. Overall, auctioneers also contribute significant numbers of bytes in both directions. Indeed, as Figure 7(c) shows, the number of bytes sent and received are of the same order of magnitude. This balance is reasonable given the underlying bidding protocol (e.g., [1]). Note, that the auctioneers only contact a limited set of servers, as highlighted in Figure 7(b). Correspondingly, the bidders are also contacted only by a limited set of auctioneers. However, in terms of received requests, not all bidder IPs are equally active – some of them receive just a few bidding requests while others see more than 450K sampled requests. Indeed, many bidders receive requests from different organizations simultaneously. Given the sampling ratio of this vantage point, we estimate that the most active bidders receive more than 42 million requests for bids per hour!

Crawlers: Crawler IPs are the second most active group of IPs in terms of requests sent. We estimate that, at this vantage point, crawling accounts for roughly 155 million requests/hour and that the most active crawlers issue up to 910K requests/hour. Naturally, the number of bytes received is larger than the number sent. Overall, we estimate that all together crawlers fetch roughly 3.8 TB per hour. However, not all are equally active, and we even see some fetching content from only a single IP.

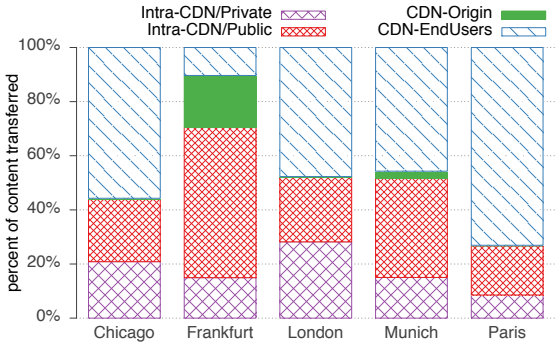
Content delivery proxies: On average, the proxies show the lowest activity per individual IP. This observation applies to both bytes and traffic. However, due to their large number, they contribute significantly to back-office traffic. This category of IPs exhibits the largest variation in behavior, and some of the heavy hitters in this category compete with those in the other categories.

Summary

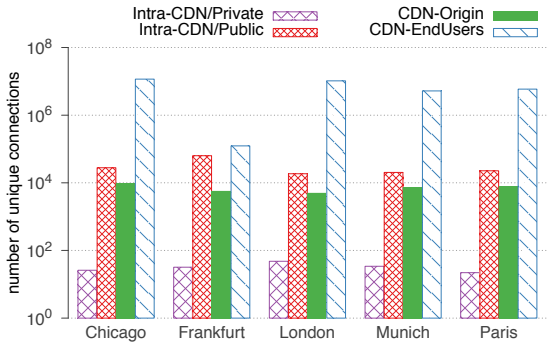
Real-time bidding is very prominent and relies on many small transactions involving a fairly small set of organizations and hosts. As each end-user request may trigger multiple bid requests, RTB significantly contributes to the number of back-office transactions. Crawling, on the other hand, happens on a coarser-grain time scale and is executed by a limited number of organizations that constantly fetch content from a diverse set of mainly Web hosting providers. While CDPs have a diverse profile, our analysis illustrates that a single end-user request to a CDN front-end server can involve a chain of proxies. These connections remain entirely hidden to the end users.

8. A CDN'S PERSPECTIVE

Until now, we have analyzed back-office Web traffic from our vantage points in ISPs and IXPs. In this section, we present a complementary perspective provided by vantage points inside a commercial CDN. A CDN can be viewed as a high-bandwidth low-latency conduit that facilitates data exchanges between end users and different points of origin. As seen in previous sections, they



(a) Traffic volumes.



(b) CDN-EndUsers Fan in, Others Fan Out.

Figure 8: Front-office and back-office traffic at a large CDN.

are one of the major contributors to back-office Web traffic. This section delves into the details of the data set provided by a large commercial CDN and highlights a few ways this data can used to characterize back-office Web traffic.

8.1 CDN Dataset

The analysis in this section is based on server logs from the CDN's *edge*, or front-end, servers. Each log line records the details of an exchange of data where the edge server is one of the endpoints. Thus, the logs capture the interactions between the edge server and the end users, i.e., front-office Web traffic, as well as the interactions with other CDN servers and origin servers, i.e., back-office Web traffic.

We obtained the server logs from all servers at one cluster in each of five different cities: Chicago, Frankfurt, London, Munich, and Paris.[6] Note that there may be multiple clusters at each city, and we selected only one of the larger clusters in each city. CDNs also deploy multiple servers at each cluster, e.g., for load-balancing, and servers at each cluster offer a diverse set of services ranging from Web-site delivery to e-commerce to video streaming. We selected clusters of servers configured to handle Web traffic, and our logs measure Web traffic of more than 350TB in volume.

8.2 Front-office vs. back-office CDN traffic

The primary focus of a CDN is to serve content to the user as efficiently as possible. Therefore, one should expect CDN front-office traffic to dominate CDN back-office traffic in volume. As not all content is cacheable [8], up to date, or popular, some content has to be fetched from other servers. Moreover, many CDNs, e.g., Akamai [54], create and maintain sophisticated overlays to interconnect their edge servers and origin servers to improve end-to-end performance, to by-pass network bottlenecks, and to increase tolerance to network or path failures. Hence, a CDN edge server may contact, besides origin servers, other CDN servers located either in the same cluster, with back-office Web traffic routed over a private network, or in a different cluster at the same or different location, with the back-office Web traffic routed over a private or public network.

With the knowledge of the IP addresses used by the CDN's infrastructure, we can differentiate the intra-CDN Web traffic from the traffic between the CDN servers and end users (CDN-EndUsers), and CDN servers and origin servers (CDN-Origin). Furthermore, within the class of intra-CDN Web traffic, we can differentiate the traffic between servers in the same cluster from that between servers in different clusters; traffic between servers in the same cluster uses high-capacity low-latency links and is routed over a private network (Intra-CDN/Private). We note that this traffic does not qualify as back-office Web traffic routed over the public Internet, which is the main focus of this paper. But in order to properly account for the publicly-routed back-office traffic that we are interested in, we must be able to separate out the Intra-CDN/Private traffic. Note also that since this category of back-office Web traffic is not routed via the public Internet it does not accrue any peering cost or hosting cost. Our classification scheme partitions the Web traffic identified via the logs into four categories: (1) CDN-EndUsers, (2) Intra-CDN/Public, (3) Intra-CDN/Private, and (4) CDN-Origin.

[6]A small fraction of servers at each location did not respond to our requests to retrieve the logs, but this should not affect the analysis.

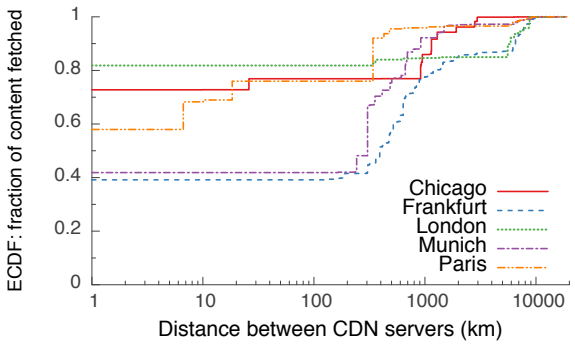


Figure 9: Content volume by distance (Intra-CDN) at a large CDN.

Figure 8(a) shows the proportion of traffic observed in each of the above four categories at the five different locations (or clusters). Not surprisingly, we see that most traffic is, as expected, CDN-EndUsers traffic. We still observe at least 25% back-office traffic at each location. Of the five clusters, Paris is the most efficient from the perspective of the content provider, with more than 70% of the traffic in the CDN-EndUsers category, and CDN-Origin traffic very low (around 1%).

Frankfurt is an oddball. At Frankfurt, the end-user traffic accounts for less than 12%. After discussions with the CDN operator, we learned that servers in the Frankfurt cluster cache content from origin servers for other edge servers in nearby clusters. The high-volume of Intra-CDN/Public traffic (about 55%) is indicative of this role for the servers in the Frankfurt cluster. Besides reducing the latency involved in fetching the content from the origin servers, this practice limits the number of servers that have to fetch content from the origin servers. The traffic at other locations show significant volumes in both the Intra-CDN/Public and Intra-CDN/Private categories. These statistics are indicative of the reliance on cooperative caching with the CDN.

Recall from Section 7.2 that there is a wide range of diversity in the number of hops over which an HTTP request is forwarded, as well as the distances to the final server. Using the actual locations of each CDN server as ground truth, we computed the distances for all Intra-CDN data exchanges. Figure 9 plots the resulting ECDF of the distances for the Intra-CDN/Public traffic weighted by the content size. The cluster in Frankfurt, in addition to serving end-user traffic, acts as a caching hub, as explained previously. Figure 9 provides further evidence of Frankfurt's role as a caching hub. About 20% of the traffic to the cluster in Frankfurt is being transferred over trans-continent links.[7] Contrast this with the cluster in Munich which receives around 2% of it's intra-CDN traffic via trans-continent links; discussion with the CDN operator confirmed that Munich does not serve as a caching hub. Figure 9 also reveals that a substantial fraction of the traffic travels only a short distance. This is expected, since in large metropolitan areas, like those selected for our study, edge servers are located at multiple data centers in the same city.

8.3 CDN back-office: Characteristics

Previously, we observed that the hosts' fan out, i.e., the number of hosts contacted by a host, can vary significantly. Accordingly, we may ask if fan out varies among the different classes of back-office CDN traffic. Not surprisingly, it turns outs that the number of unique end-user addresses to which the edge servers deliver content, i.e., the fan in, is larger than the combined number of CDN and origin servers from which they fetch content, i.e., the fan out.

Figure 8(b) shows the number of unique connections observed in the different traffic categories at each location. From the figure, we see that the number of unique connections in the back-office traffic categories (Intra-CDN/Private, Intra-CDN/Public, and CDN-origin) is two orders of magnitude less than that in the CDN-EndUsers category; note that the y-axis is plotted using a log scale. Moreover, the Intra-CDN/Private category mainly captures intra-cluster data exchanges and thus the fan out is even smaller. Finally, although the number of unique connections in the CDN-Origin category is smaller, it is equivalent in order of magnitude to the connection count in the Intra-CDN/Public category.

Aggregating the traffic volume by server addresses in both the CDN-Origin as well as the Intra-CDN/Public category reveals that the traffic is not uniformly distributed across all servers; rather there are heavy hitters. 20% of the origin servers contribute to more than 96% of the content delivered to the CDN's edge servers. A similar trend manifests in the Intra-CDN/Public category; 20% of the CDN's servers account for over 94% of the traffic volume moved from different servers in the CDN's infrastructure to the front-end, or edge, servers. These figures hint at the impact of the varying popularity and cacheability of content on the traffic patterns within the CDN infrastructure.

9. AN END-USER'S PERSPECTIVE

Improving the end-user experience can lead to a significant increase in revenues [39] and drive up user engagement [23, 41]. These benefits have catalyzed the competition among service companies to offer faster access to content in order to improve the end-user experience. Two "straightforward" ways of improving the end-user experience that can be implemented by ISPs are to (a) upgrade access networks, and to (b) improve the Internet's middle mile (backbones, peering points, and/or transit points); both approaches, however, are expensive.

[7]We assume that distances of 6000 km or more indicate trans-continent links.

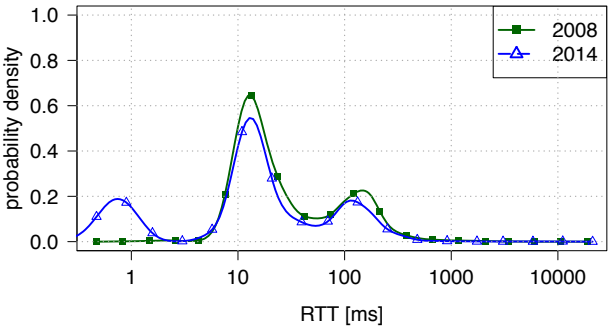


Figure 10: Delay between access aggregation point and front-end servers in 2008 [45] vs. 2014.

Content and application providers, on the other hand, can provide better service by installing servers deep inside ISPs and/or through increased peering at IXPs or colocation facilities. These approaches have been reported in [28, 46, 20] and have led to an increase in back-office Web traffic. Next, we analyze our data sets to see if these approaches have indeed decreased the latencies between end users and front-end servers. The analysis is based on anonymized packet-level traces captured in a residential network within a large European ISP in 2008 and in January 2014. For more details on data capturing, processing and anonymization, we refer to [45].

Figure 10 shows the backbone delay between the aggregation point of the ISP and the front-end servers that end users of the *same* residential network experienced in two different years — 2008, and 2014.[8] While the access technology, and the corresponding delay between end users and the access aggregation point, have not changed over the last six years (not shown), the delay between the aggregation point and servers has seen a significant shift towards shorter delays. A significant fraction of the requests are now terminated by servers within the same city (see the new bell curve around 1 ms delay which was not present in 2008); pushing the content closer to the end users has been successful.

10. IMPLICATIONS

In sections 2 through 4, we defined back-office Web traffic, discussed some methods to identify this traffic from different data sets, and presented some key insights from our analysis. We showed that back-office traffic is responsible for a significant fraction of today's Web traffic, and offered two key reasons to explain this phenomenon: (1) sustained deployment of front-end servers close to end users, and (2) substantial data exchanges conducted by service providers as part of their operations viz., coordination and synchronization of components of their distributed infrastructure. In this section, we discuss a few implications that affect researchers and operators.

This point of view allows us to differentiate the deployed servers into two broad categories, namely front-office and back-office servers that support many of the popular Web applications which run on top of the Internet. The front office is devoted to serving end users, and its performance is mostly influenced by the access technology and, to a different extent, by back-office performance. The back office, in turn, operates on the faster and wider pipes of the Internet backbone, involves multiple organizations and remains invisible to the end-users, yet virtually connected to them via the front-office. However, there is more in the back office than content delivery,

[8]We note that not necessarily all traffic is exchanged with front-end servers e.g., some traffic is due to peer-to-peer applications.

which includes, but is not limited to, real-time bidding and crawling.

With the increasing trend to offload "computations" to the cloud to support enterprise applications [53] or desktop application, e.g., the new release of Microsoft Office, Apple iCloud, and DropBox [24], it is clear that the volume of back-office Web traffic will continue to increase. First, application servers at different data centers will have to exchange data with each other for data synchronization and replication. Second, as edge servers attempt to offer more services [43, 28], viz., personalized Web experiences for end users, they will need to communicate with other edge servers and back-office servers, at different locations, to retrieve relevant content for the end users.

The Web ecosystem as presented in this work, manifesting a rich interplay between front-office and back-office servers, has many *practical* implications that affect not only end users, but also researchers and operators.

For researchers, the implications are two-fold. First, it is of utmost importance to differentiate between back-office and front-office Web traffic as they exhibit different characteristics and operations. Follow-up works may study the correlation of traffic dynamics, topology aspects of the Internet, and performance evaluation with each of the two classes of Web traffic. Second, researchers should be aware that the infrastructure deployed for the back-office presents a unique opportunity to develop and deploy new protocols that are better suited to the characteristics of the front- or the back-office.

For operators, the implications are multi-fold. First, operators should be aware of the importance of links that carry back-office traffic because they may (1) help in monetization of services, or (2) affect the front-office operations, viz., impairing the QoE of more end-users than anticipated. Second, operators should be aware that back-office traffic may have different requirements than front-office traffic. This, in turn, creates new opportunities for operators to provide customized services. In addition, operators may also want to offer different Service Level Agreements (SLAs) for different traffic classes.

10.1 Deployment of new protocols

Deploying new protocols or protocol variants may be very slow or even infeasible if it requires changes in the core infrastructure of the Internet or all end Internet end systems, for instance, consider the adoption of a new TCP variant or IPv6. However, restricting the change to one organization makes this otherwise infeasible effort much easier. Indeed, most Web service companies have the ability to roll out updates to their infrastructure and upgrade it on a regular basis. Thus, if end users connect only to the front-end servers and the front-end and back-end servers are managed by a single entity, it is feasible to rapidly deploy changes and optimize back-office Web communications; these changes are restricted to the servers involved in carrying the back-office traffic. Examples include but are not limited to use of persistent TCP connections between servers, IPv6, multi-path TCP [29], on-the-fly Web page assembly [43], object pre-fetching, compression and delta-encoding, and offloading of computations. Indeed, even the TCP version and parameters can be chosen according to the tasks [28], e.g., the initial window size can be increased. Moreover, it is possible to adopt new networking paradigms, such as Software Defined Networking [37], much more quickly to handle the back-office traffic. As most end-user communication involves some back-office communication, these improvements can directly affect the end-user experience. Lastly, more efficient use of the networking resources can also reduce the cost of content delivery.

10.2 New service opportunities

Network operators i.e., ISPs, can supply CDNs or content providers with the infrastructure that is specifically tailored to handle back-office traffic. For ISPs, this opens up additional opportunities for monetization of services and also provide a better experience to their end users. But, this comes at a cost: ISPs must invest in and enable micro data centers or virtualized services [30, 4] in their networks to harness the opportunity.

On the other hand, ISPs can also customize their traffic engineering policies to better accommodate the two different classes of traffic. They can also offer a different set of SLAs targeted at organizations operating a back office. Along the same line, IXPs can provide value-added services for time-critical applications e.g., bidding and financial applications. IXPs can also incentivize co-location at their peering locations using arguments such as the proximity of third-party servers, the pricing model for exchanging traffic with servers in other networks, and the ability to access the resources that cloud providers offer.

11. CONCLUSION

The Web and its ecosystem are constantly evolving. This paper takes a first step towards uncovering and understanding one component of this ecosystem that is increasing in complexity, but remains understudied: the back-office. The back-office includes the infrastructure necessary to support Web search, advertisements, and content delivery. By using a diverse set of vantage points, we've shown that back-office traffic is responsible for a significant fraction not only of today's Internet traffic but also today's Internet transactions.

Improvement in the back-office has been a major contributor to reducing the delays experienced by end users. Yet the back-office architecture still exposes many opportunities for deploying new protocols or versions of protocols for specialized tasks. The complexity of the back-office, however, also poses new questions. In future work we plan to improve our current methodology and extend it to identify other use-cases for back-office Web traffic, further dissect the interactions of the different services and better understand the performance implications thereof. Finally, we also plan to study the non-Web back-office of the Internet.

Acknowledgments

We would like to thank Oliver Spatscheck (our shepherd), the anonymous reviewers, and Paul Barford for their constructive feedback. This work was supported in part by the EU project BigFoot (FP7-ICT-317858). Georgios Smaragdakis was supported by the EU Marie Curie International Outgoing Fellowship "CDN-H" (PEOPLE-628441).

12. REFERENCES

[1] Google AdExchange. `http://developers.google.com/ad-exchange/rtb/getting_started`.

[2] Internet Advertising Bureau (IAB). 2013 Internet Advertising Revenue Report. `http://www.iab.net/AdRevenueReport`.

[3] Netflix Open Connect. `https://signup.netflix.com/openconnect`.

[4] Network Functions Virtualisation. SDN and OpenFlow World Congress, 2012.

[5] V. K. Adhikari, S. Jain, Y. Chen, and Z. L. Zhang. Vivisecting YouTube: An Active Measurement Study. In *IEEE INFOCOM*, 2012.

[6] B. Ager, N. Chatzis, A. Feldmann, N. Sarrar, S. Uhlig, and W. Willinger. Anatomy of a Large European IXP. In *ACM SIGCOMM*, 2012.

[7] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig. Web Content Cartography. In *ACM IMC*, 2011.

[8] B. Ager, F. Schneider, J. Kim, and A. Feldmann. Revisiting Cacheability in Times of User Generated Content. In *IEEE GI*, 2010.

[9] S. Angel and M. Walfish. Verifiable auctions for online ad exchanges. In *ACM SIGCOMM*, 2013.

[10] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan. Adscape: Harvesting and Analyzing Online Display Ads. In *WWW*, 2014.

[11] L. A. Barroso, J. Dean, and U. Holzle. Web Search for a Planet: The Google Clustering Architecture. *IEEE Micro*, 23, 2003.

[12] T. Benson, A. Akella, and D. A. Maltz. Network traffic characteristics of data centers in the wild. In *ACM IMC*, 2010.

[13] T. Benson, A. Anand, A. Akella, and M. Zhang. MicroTE: Fine Grained Traffic Engineering for Data Centers. In *CoNEXT*, 2011.

[14] I. Bermudez, M. Mellia, M. Munafà, R. Keralapura, and A. Nucci. DNS to the Rescue: Discerning Content and Services in a Tangled Web. In *ACM IMC*, 2012.

[15] L. Bernaille and R. Teixeira. Early recognition of encrypted applications. In *PAM*, 2007.

[16] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *WWW*, 1998.

[17] M. Butkiewicz, H. V. Madhyastha, and V. Sekar. Characterizing Web Page Complexity and Its Impact. *IEEE/ACM Trans. Networking*, 22(3), 2014.

[18] M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan. Mapping the Expansion of Google's Serving Infrastructure. In *ACM IMC*, 2013.

[19] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. 2006.

[20] N. Chatzis, G. Smaragdakis, A. Feldmann, and W. Willinger. There is More to IXPs than Meets the Eye. *ACM CCR*, 43(5), 2013.

[21] Y. Chen, S. Jain, V. K. Adhikari, and Z. L. Zhang. Characterizing Roles of Front-End Servers in End-to-End Performance of Dynamic Content Distribution. In *ACM IMC*, 2011.

[22] Y. Chen, R. Mahajan, B. Sridharan, and Z. L. Zhang. A Provider-side View of Web Search Response Time. In *ACM SIGCOMM*, 2013.

[23] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stoica, and H. Zhang. Understanding the Impact of Video Quality on User Engagement. In *ACM SIGCOMM*, 2011.

[24] I Drago, M. Mellia an M. Munafo, A. Sperotto, R. Sadre, and A. Pras. Inside Dropbox: Understanding Personal Cloud Storage Services. In *ACM IMC*, 2012.

[25] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-Wide Scanning and its Security Applications. In *USENIX Security Symposium*, 2013.

[26] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware Forward Caching. In *WWW*, 2009.

[27] A. Feldmann, N. Kammenhuber, O. Maennel, B. Maggs, R. De Prisco, and R. Sundaram. A methodology for estimating interdomain web traffic demand. In *ACM IMC*, 2004.

[28] T. Flach, N. Dukkipati, A. Terzis, B. Raghavan, N. Cardwell, Y. Cheng, A. Jain, S. Hao, E. Katz-Bassett, and R. Govindan. Reducing Web Latency: the Virtue of Gentle Aggression. In *ACM SIGCOMM*, 2013.

[29] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar. Architectural guidelines for multipath TCP development. Internet Draft, rfc-6182.

[30] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. Maggs, J. Rake, S. Uhlig, and R. Weber. Pushing CDN-ISP Collaboration to the Limit. *ACM CCR*, 43(3), 2013.

[31] H. Gao, V. Yegneswaran, Y. Chen, P. Porras, S. Ghosh, J. Jiang, and H. Duan. An Empirical Reexamination of Global DNS Behavior. In *ACM SIGCOMM*, 2013.

[32] A. Gerber and R. Doverspike. Traffic Types and Growth in Backbone Networks. In *OFC/NFOEC*, 2011.

[33] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the Money: Understanding Economics of Online Aggregation and Advertising. In *ACM IMC*, 2013.

[34] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: A Scalable and Flexible Data Center Network. In *ACM SIGCOMM*, 2009.

[35] C. Huang, A. Wang, J. Li, and K. Ross. Measuring and Evaluating Large-Scale CDNs. In *ACM IMC*, 2008.

[36] S. Ihm and V. S. Pai. Towards Understanding Modern Web Traffic. In *ACM IMC*, 2011.

[37] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Holzle, S. Stuart, and A. Vahdat. B4: Experience with a Globally-Deployed Software Defined WAN. In *ACM SIGCOMM*, 2013.

[38] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *ACM/SIAM SODA*, 1998.

[39] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. In *ACM KDD*, 2007.

[40] R. Krishnan, H. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao. Moving Beyond End-to-end Path Information to Optimize CDN Performance. In *ACM IMC*, 2009.

[41] S. S. Krishnan and R. K. Sitaraman. Video Stream Quality Impacts Viewer Behavior: Inferring Causality using Quasi-Experimental Designs. In *ACM IMC*, 2012.

[42] C. Labovitz, S. Lekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In *ACM SIGCOMM*, 2010.

[43] T. Leighton. Improving Performance on the Internet. *Communications of the ACM*, 52(2):44–51, 2009.

[44] J. Liang, J. Jiang, H. Duan, K. Li, T. Wan, and J. Wu. When HTTPS Meets CDN: A Case of Authentication in Delegated Service. In *IEEE Symp. on Security and Privacy*, 2014.

[45] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *ACM IMC*, 2009.

[46] E. Nygren, R. K. Sitaraman, and J. Sun. The Akamai Network: A Platform for High-performance Internet Applications. *SIGOPS Oper. Syst. Rev.*, 2010.

[47] I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving Content Delivery using Provider-aided Distance Information. In *ACM IMC*, 2010.

[48] I. Poese, B. Frank, G. Smaragdakis, S. Uhlig, A. Feldmann, and B. Maggs. Enabling Content-aware Traffic Engineering. *ACM CCR*, 42(5), 2012.

[49] L. Popa, A. Ghodsi, and I. Stoica. HTTP as the Narrow Waist of the Future Internet. In *SIGCOMM HotNets*, 2010.

[50] F. Qian, A. Gerber, Z. M. Mao, S. Sen, O. Spatscheck, and W. Willinger. TCP Revisited: A Fresh Look at TCP in the Wild. In *ACM IMC*, 2009.

[51] InMon – sFlow. `http://sflow.org/`.

[52] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. A First Look at Cellular Machine-to-Machine Traffic – Large Scale Measurement and Characterization. In *ACM SIGMETRICS*, 2012.

[53] J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratsanamy, and V. Sekar. Making Middleboxes Someone Else's Problem: Network Processing as a Cloud Service. In *SIGCOMM*, 2012.

[54] R. K. Sitaraman, M. Kasbekar, W. Lichtenstein, and M. Jain. *Overlay Networks: An Akamai Perspective*. John Wiley & Sons, 2014.

[55] K. Springborn and P. Barford. Impression Fraud in Online Advertising via Pay-Per-View Networks. In *USENIX Security Symposium*, 2013.

[56] F. Streibelt, J. Boettger, N. Chatzis, G. Smaragdakis, and A. Feldmann. Exploring EDNS-Client-Subnet Adopters in your Free Time. In *ACM IMC*, 2013.

[57] S. Triukose, Z. Wen, and M. Rabinovich. Measuring a Commercial Content Delivery Network. In *WWW*, 2011.

[58] N. Weaver, C. Kreibich, M. Dam, and V. Paxson. Here Be Web Proxies. In *PAM*, 2014.

[59] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: measurement and analysis. In *ADKDD*, 2013.