Uncovering the Big Players of the Web

Vinicius Gehlen[1], Alessandro Finamore[2], Marco Mellia[2], and Maurizio M. Munafò[2]

[1] CNIT - UoR c/o Politecnico di Torino
[2] Dept. of Electronics and Telecommunications - Politecnico di Torino
lastname@tlc.polito.it

Abstract. In this paper we aim at observing how today the Internet large organizations deliver web content to end users. Using one-week long data sets collected at three vantage points aggregating more than 30,000 Internet customers, we characterize the offered services precisely quantifying and comparing the performance of different players. Results show that today 65% of the web traffic is handled by the top 10 organizations. We observe that, while all of them serve the same type of content, different server architectures have been adopted considering load balancing schemes, servers number and location: some organizations handle thousands of servers with the closest being few milliseconds far away from the end user, while others manage few data centers. Despite this, the performance of bulk transfer rate offered to end users are typically good, but impairment can arise when content is not readily available at the server and has to be retrieved from the CDN back-end.

1 Introduction

Since the early days the Internet has continuously evolved. With the aggregation of different technologies, the convergence of telecommunication services and the birth of innovative services, nowadays understanding the Internet is very complicated. Measurements are thus becoming more and more a key tool to study and monitor what the users do on the Internet and how can the Internet meet users demand.

The last years witnessed the convergence of services over the web thanks to the increasing popularity of Social Networks, File Hosting, Video Streaming services. All of them involve a large amount of content that has to be delivered to a humongous number of users. Delivering such impressive volume of content in an efficient and scalable way has become a key point for the Internet service and content providers. Content Delivery Network (CDN) and data centers have become the vital elements to support such growth. It is indeed not surprising that 40% of the inter-domain traffic is generated by few *organizations*[1], with Google leading the group, and well-known CDN players closely following [1]. Companies

[1] In this paper we use the generic term "organization" to state commercial companies that manages web servers, caches, CDNs, and offering services directly to end users or from third parties.

A. Pescapè, L. Salgarelli, and X. Dimitropoulos (Eds.): TMA 2012, LNCS 7189, pp. 15–28, 2012.

like Akamai, Limelight and Level3, which have been deployed in the last decade, are now crucial for the majority of the web-based services. Several studies in the literature have focused on CDN architectures, trying to reverse engineer the server placement and load balancing algorithms [2–5]. These works usually focus on big players as Akamai [3, 4] or, more recently, YouTube [5]. However, besides CDN companies there are many others smaller organizations which provide web services as well [6], and overall their role is not marginal. Another common treat among previous works is the extensive usage of active experiments, i.e., abusing of synthetic traffic which does not reflect the real usage and users' habits, and thus does not allow to gauge the actual performance offered to the end users.

Our work goes in a different direction. We consider actual traffic originated from the activity of more than 30,000 residential customers monitored from three different vantage points of an Italian ISP. Analysing the traffic they generated for one week, we aim at identifying i) which services they access, and ii) how organizations handle it. We neither focus on single organization, nor service. We instead consistently compare and quantify the traffic users exchange with server managed by top players handling web traffic by looking for both what kind of service and performance they offer, and how dissimilar their architecture looks like.

It is important to highlight that we are not aiming at a complete characterization of all the possible services the big organizations provide. We instead would like to start understanding what type of traffic they have to handle, how do they handle it and what kind of performance they offer. Moreover, given we are considering only one ISP, we take a picture from the same point of view of the ISP customers. Clearly, we expect differences when observing from other ISPs.

In the following we report some of the key findings:

– the top 10 organizations handle 65% of the total web traffic. This is invariant respect to the location of the users, the time of the day, and the day of the week;
– nearly all of the top 10 organizations provide the same type of content, ranging from video services, to software download, to hosting generic content;
– only Google, Facebook and Akamai handle service over HTTP/SSL. Encryption is rapidly increasing its share due to users enabling security features in social network systems;
– the organizations adopt different architectures, with more than 10,000 IP addresses managed by Akamai but only 338 by Facebook according to our data sets;
– some organizations serve traffic within few milliseconds far away from ISP customers, while other organizations manage few data centers that are placed in other countries or even continent;
– traffic balanced among servers is the same over time for all organizations except for Google which instead routes traffic to different data centers at different time of the day;
– bulk download rate offered by different organizations is good, with Limelight taking the lead. Only Akamai shows some poor performance when the requested content is not readily available at the cache.

Table 1. Data sets used in this paper

Name	Volume [GB]	Flow [M]	# Servers	# Clients
VP1	1745 (35%)	16 (63%)	77,000 (0.14%)	1534 (99%)
VP2	10802 (44%)	84 (53%)	171,000 (0.6%)	11742 (97%)
VP3	13761 (35%)	125 (52%)	215,000 (0.5%)	17168 (98%)

In summary, our work aims at unveiling some aspects of modern web browsing. We believe this information is useful for both the service providers and the research community to shed some light on the role big organizations have on the current Internet.

2 Methodology and Data Sets

The data sets has been collected using Tstat [7], the Open Source packet sniffer developed in the last 10 years by the Telecommunication Network Group (TNG) at the Politecnico di Torino. Tstat rebuilds TCP connections by monitoring traffic sent and received by hosts. The connections are then monitored as to provide several types of Layer-4 statistics. Using Deep Packet Inspection (DPI) and statistical techniques [8, 9], each connection is further classified based on which application has generated it.

We collected data sets from three Point of Presence (PoP) aggregating thousands of customers of an Italian ISP. At each vantage point (VP), we installed a probe consisting of an high-end PC running Tstat and monitoring all the traffic sent and received by hosts inside the PoP. This paper focuses on one-week long data sets collected simultaneously at the three vantage points starting from 12:00 AM of June 20th, 2011. Each data set is composed of TCP flow-level logs. Each line logs the statistics of a specific TCP flow while the columns report specific features[2].

In this paper we analyze organizations handling content related to web-based services. Thus, we focus only on HTTP (and HTTPS) traffic since the remaining part of the traffic is based on applications and protocols which usually are not handled by these organization. The commercial MaxMind [10] "Organization" database has been used to associate each server IP address with the organization which owns it.

Tab. 1 summarizes the data sets reporting the name used in the remaining of the paper, the volume due to HTTP traffic in terms of bytes and flows, the number of distinct HTTP servers contacted by clients and the number of distinct internal hosts that generated HTTP traffic during the whole week. In brackets we report the corresponding fraction with respect to the total traffic. For example, consider VP1 data set. It contains 1.7 TB of HTTP traffic carried in 16 millions of TCP flows. It corresponds to 35% of the total volume exchanged

[2] A description of all statistics is available from `http://tstat.tlc.polito.it/measure.shtml`

Table 2. Characterization of the Top 10 organizations

	Volumes			Most known services		
Organization	%B	%F	%Clients	Video Content	SW Update	Adv. & Others
Google	22.7	12.7	97.1	YouTube	-	Google services
Akamai	12.3	16.7	97.2	Vimeo	Microsoft, Apple	Facebook static content, eBay
Leaseweb	6.3	1.1	64.3	Megavideo	Mozilla	publicbt.com
Megaupload	5.5	0.2	15.6	Megavideo	-	File hosting
Level3	4.7	1.9	79.7	YouPorn	-	quantserve, tinypic, photobucket
Limelight	3.9	1.6	72.5	Pornhub, Veoh	Avast	betclick, wdig, trafficjunky
PSINet	3.2	0.2	44.6	Megavideo	Kaspersky	Imageshack
Webzilla	2.9	0.3	13.2	Adult Video	-	filesonic, depositfiles
Choopa	1.5	0.01	5.7	-	-	zShare
OVH	1.0	0.7	63.1	Auditude	-	Telaxo, m2cai
Facebook	0.9	4.2	90.6	Facebook	-	Facebook dynamic content
total	64.9	39.6	-			

by hosts in that vantage point, i.e., 63% of TCP flows. The remaining part of the traffic is due to other applications like email, chat, and most of all, peer-to-peer (P2P) applications. 77000 distinct HTTP servers have been contacted by 1534 distinct customers hosts. HTTP servers are a small fraction of contacted hosts (the majority of them are due to P2P applications), but almost any PoP internal customer generated HTTP traffic. VP2 and VP3 data sets aggregate a larger number of local clients which is reflected in terms of volumes and number of contacted servers.

3 Traffic Characterization by Organization

We now group HTTP traffic based on the organization that owns the server IP address. We start by answering simple questions like which organizations are responsible for the largest majority of traffic, which kind of content they handle, and how the traffic changes when considering different vantage points or time of the day.

3.1 Share and Popularity

Tab. 2 shows the top 10 organizations ranked with respect to the volume of HTTP traffic they account for. On the left, after each organization name, we report the fraction of bytes and flows the organization handled, and the fraction of customers that generated more than 100 flows with servers of that organization during the week. Results reported refer to VP2 data set and are similar on other vantage points. The rightmost columns report the kind of services offered by the organizations, identifying three coarse categories, namely Video content, Software update and Other services as advertisement or File Hosting.

Several considerations hold. Beside the expected presence of big players as Google or Akamai, there are less known organizations as Choopa or PSINet responsible for a non negligible fraction of traffic. Even if this result is specific for the operator we are considering, knowing which are the top organizations the customers exchange traffic with is a key information. Interestingly, the top ten organizations account for 65% of the HTTP traffic volume. Overall, we have found more than 10,000 organizations; yet, 90% of the volume is managed by 70 companies only. This testifies that in the nowadays Internet there are plenty of companies providing an high number of services but the majority of the traffic is generated by very few of them, as already noted in [1].

Considering the popularity of the organizations among ISP customers, we notice that some are handling niche services, like the one supported by Choopa or Webzilla, while others are practically used by most users. It is not surprising that 97% of users have contacted Google, Level3 or Akamai administered servers. Instead it is more surprising that 91% of users contacted one Facebook server or 63.1% used some service handled by OVH. Considering Facebook, most of its static content like pictures and videos, is provided by the Akamai CDN. Indeed, we have found that more than 73% of the bytewise volume related to Facebook is outsourced to Akamai and overall Facebook content corresponds to 15% of the total volume handled by Akamai. The Facebook dynamic content, including all embedded objects found in other web pages as the "Like" button, is directly handled by the Facebook servers. The pervasiveness of these objects on the web causes a high probability of connecting to a Facebook server, even without accessing to the Social Network site. For OVH we believe that the presence of several advertisement services that it hosts causes most of customers to contact the OVH servers.

For the remaining part of this work, we will focus only on the most important organization listen in Tab. 2, namely Google, Akamai, Leaseweb, Level3, Megaupload, Limelight and Facebook.

3.2 Spatial and Temporal Difference

We now investigate how the traffic changes considering i) different vantage points, ii) different time of day and iii) different days. Are those results invariant to the spatial and temporal choice of observation? Fig. 1 reports the daily evolution of the HTTP bitrate breakdown among the considered organizations for the VP2 data set. Each bar refers to a 2 hour time window. Fractions are derived by averaging the traffic in all days of the data set. As expected, traffic volume follows the typical day/night pattern. However, in each period of the day, the fraction of traffic handled by the different organization is very stable, so that practically in all period they handle 60% of the web traffic. Focusing on peak hour time from 22:00 to 00:00, Fig. 2 (left) compares the traffic share considering different days of the week. Also in this case, marginal changes are

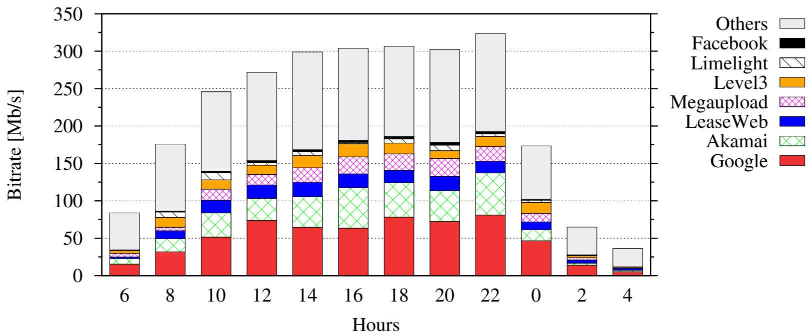

Fig. 1. HTTP volume breakdown for VP2 with respect to the organizations

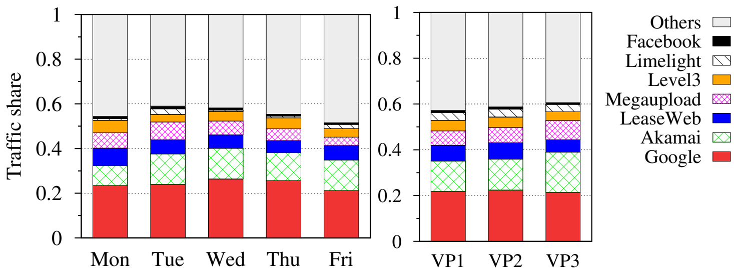

Fig. 2. Comparison of the HTTP traffic share during peak hours considering different days (left) and data sets (right)

observed. This is also true when considering different vantage points as reported in Fig. 2 (right).

Overall, we therefore notice that there is a limited impact on this kind of analysis when changing time, day, or vantage point. This has interesting implications on the possibility of monitoring the traffic from a few vantage points instead of deploying a capillary monitoring system. However, we expect the picture to be very different when changing ISP or nation, given we expect users' habits and cultural interest play an important role.

3.3 Content Characterization

To investigate and compare the characteristics of the content served by each organization we rely on the DPI capabilities of Tstat. In particular, we define five coarse classes of content: HTTPS/SSL, Facebook, File Hosting, Video Content and Generic web content.

Fig. 3 reports the breakdown of the volume of each organization with respect to the five classes, sorting organization in increasing fraction of Generic web. Most of the organizations serve Video Content. More precisely, Akamai hosts Vimeo, Megaupload and Leaseweb serve Megavideo, while Level3 and Limelight

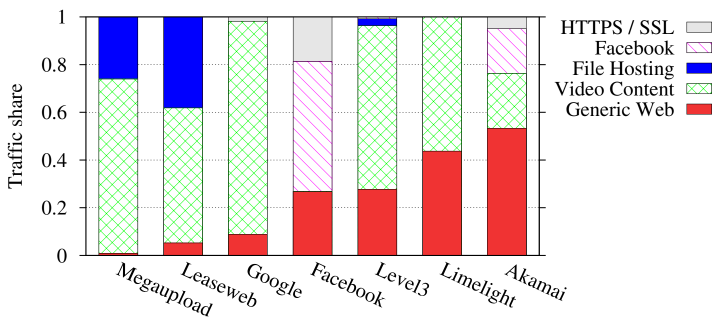

Fig. 3. Breakdown of the HTTP volume downloaded from each organization with respect to the type of content. Results refer to VP2 data set.

Table 3. Fraction of HTTPS/SSL traffic for each organization

	%Bytes		%Flows	
Organization	Jun	Oct	Jun	Oct
Google	1.75	2.67	11.1	17.34
Akamai	3.46	2.95	10.55	17.6
Leaseweb	0.46	0.26	0.55	1.85
Level3	1.22	0.34	1.49	2.48
Limelight	0.05	0.78	1.82	2.16
Megaupload	-	-	-	-
Facebook	13.3	26.8	15.35	24.3

host several adult video sites, and Google hosts all YouTube videos. Considering File Hosting services, Level3 serves RapidShare while Megaupload and Leaseweb serve the files for the Megaupload site.

Interestingly, a small fraction of traffic is encrypted. To give more insight, Tab. 3 reports the fraction of volume and flows related to HTTPS/SSL traffic for each company. The results refer to VP2, for which a more recent data set has been collected during October 2011. Results show that encrypted traffic is significant only for Google, Facebook and Akamai. Both Google and Facebook are known to provide services for which the end user is allowed to opt for HTTPS as the default application protocol. In case a user has opted to encrypt its communications with Facebook, then all Facebook content will be provided over HTTPS, including content coming from the Akamai servers. This explains the relative large fraction of HTTPS/SSL flows served by Akamai.

Comparing the June and October data sets, we notice that encrypted traffic is increasing its share. For Facebook it doubled, with 26% of bytes and 24% of flows encrypted in October. This suggests that Facebook users are increasingly opting to enable the privacy/security functionality offered by the Social Network. The server cost to support encrypted traffic is thus increasing as well.

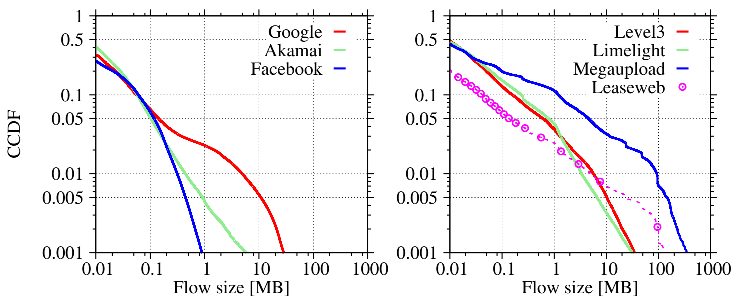

Fig. 4. CCDF of the server-to-client flow size for VP2

Finally, we focus our attention on the length of content that each organization has to handle, which reflects also the storage capacities they support. Fig. 4 reports the Complementary Cumulative Distribution Function (CCDF) of the amount of bytes the server sent to the client for the VP2 data set. As usual, results are identical for the other data sets too. As you can see, most of the flows are mice, i.e., they carry less than 10 kB. The fraction of mice flows ranges from 80% (Leaseweb) to 50% (Level3). Facebook, which hosts only small dynamic objects, has a negligible fraction of elephants, i.e., flows that carry more than 1 MB. For other organizations, more than 1% of the flows carry more than 1 MB. For Google, the larger number of elephants is due to the YouTube video flows. Megaupload is the organization that carries the largest number of elephants, since most of the content downloaded from its server is either due to complete movies or large archives. Surprisingly, 50% of the content served by Megaupload is less than 10 kB long, much smaller than average chunk size used by download managers. Overall, the flow length follows an heavy tailed distribution (note the log-log scale).

Considering the flow duration (not reported here due to the lack of space), for all the organizations but Megaupload and Facebook, the flows are shorter than 5 minutes, with more than 83% of the flows shorter than 1 minute, while Facebook and Megaupload present an heavy tail with more than 2% of the flows longer than 30 minutes. For Megaupload this reflects the length of the content downloaded from the server, while for Facebook this is due to an intentional use of persistent HTTP connections that are used to carry some services like chat or Ajax applications that update the web page dynamically.

4 How Organizations Handle Traffic

We now investigate on how large is the number of servers that have been contacted in our data set for each organization, and how traffic is balanced among servers. To augment the visibility on each organization servers, in this section all three data sets are aggregated and analyzed.

Table 4. Number of IPs and /24 subnets for each organization. Results refer to the aggregation of all data sets.

	IP addresses		/24 Subnets	
Organization	No.	(%)	No.	Top5 %Bytes
Google	3678	(0.76)	135	94.1
Akamai	10445	(2.16)	1255	86.1
Leaseweb	3833	(0.79)	546	80.0
Level3	1868	(0.39)	572	65.8
Limelight	1179	(0.24)	115	97.2
Megaupload	808	(0.17)	15	64.0
Facebook	338	(0.07)	27	74.5
Total	33798	(7.00)	4596	-

Tab. 4 reports the number of unique server IP addresses belonging to each organization that have been contacted by any customer. For each organization, the second column shows the total number of server IP addresses (referred as servers in the following for simplicity), the third column details their fraction over the total number of HTTP servers, and the forth column reports the number of /24 subnets the IP addresses belong to[3]. Considering each organization individually, we have also computed the traffic volume generated by each subnet. Ranking the subnets according to this measure, the last column shows the percentage of volume delivered by the top 5 subnets of each organization.

First of all, the total number of servers belonging to the selected organizations accounts to 7% of all HTTP servers. Yet, those 7% of servers are responsible for more than 60% of the HTTP traffic volume (see Tab. 2). Considering both the number of IP addresses and /24 subnets and how they are used, we can observe that the organizations pursue different approaches. Akamai is the organization with the highest number of servers and subnets, respectively 10445 and 1255. Google, which serves a similar amount of traffic as Akamai, uses one third of the servers, which are aggregated into one tenth of /24 subnets. This reflects the different architecture Akamai and Google have: small data centers spread over many ISPs [4] for Akamai, few large data centers for Google [5]. Limelight, another large CDN organization, follows a more concentrated deployment as well.

Facebook results the organization with the smallest number of IPs. Being most of Facebook static content actually served by Akamai, Facebook can handle the dynamic content of his social network by managing 338 servers only. Megaupload, another very specialized organization targeting file hosting services, shows a small number of servers and subnets as well.

Considering the last column of Tab. 4, we see that most of the traffic is actually served using only few subnets, with the five most used subnets accounting for more than 64% of the volume of each organization. Interestingly, the top 5

[3] The aggregation of servers in /24 subnets may not coincide with actual subnets. We consider it as an instrumental aggregation.

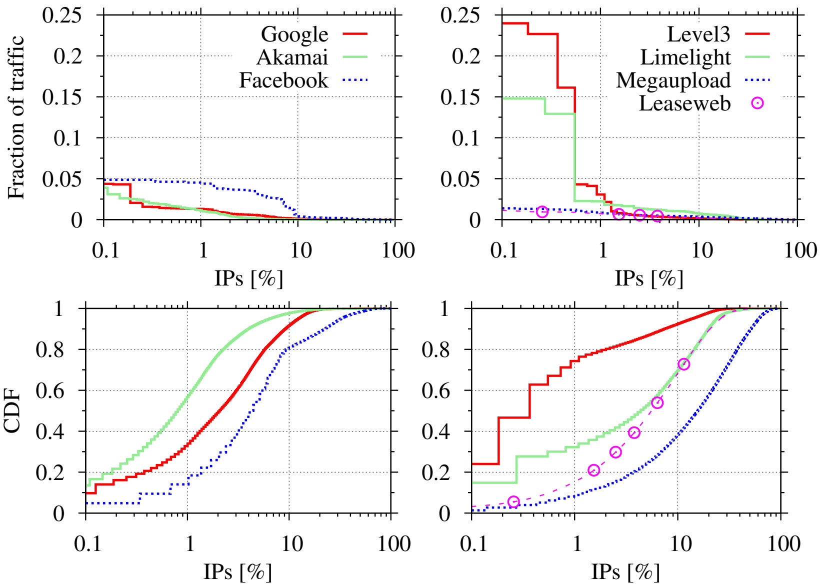

Fig. 5. Fraction of volume served by each IP (top) and the associated CDF (bottom) for each organization in VP2 data set

subnets of Limelight serves 97% despite we have found 115 subnets hosting its servers and being contacted by the customers.

To give more insight about how the traffic is concentrated, Fig. 5 shows the fraction of traffic generated by each single server. Servers have been ordered in decreasing fraction of traffic, and their rank has been normalized to the total number of servers in each organization. Top plots report the fraction of volume served for each IP, while bottom plots report the CDF. Several differences arise. Akamai tends to concentrate traffic on the same subset of server more than Google. Consider for example the fraction of servers that account for 80% of traffic. Akamai handle this with 2% of its servers. Google uses 5% of servers, while Facebook uses 10% of servers. In all cases, the top server handles about 5% of total volume. Considering other organizations, Level3 and Limelight follow a even more biased distribution of traffic, with the top three servers accounting for 63% and 31% of total traffic volume respectively. Finally, Megaupload and Leaseweb spread the traffic more uniformly among their servers.

The different concentration of traffic among the servers can be related to different causes: i) we observed that more than 3% of the users use third-party DNS resolvers as OpenDNS (1%) or Google's DNS (0.5%) which do not manage to redirect the users to the preferred data center, contrary to the ISP DNS servers [11]; ii) load balancing schemes are used to cope with day/night traffic shift redirecting the users to different servers according to the time of the day (more on this later on); iii) the content requested might be not available at the server contacted leading to application-layer redirections (e.g., YouTube [5]).

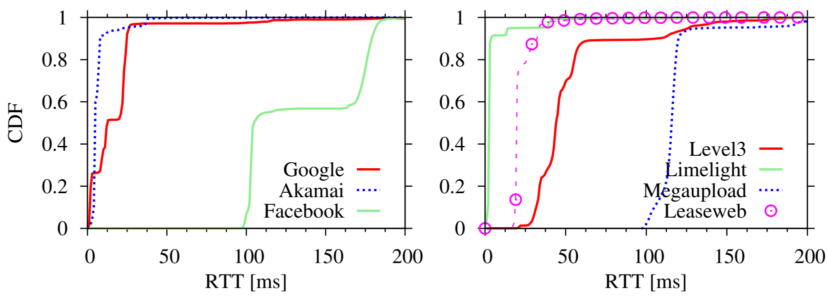

Fig. 6. Distribution of the minimum external RTT in VP2

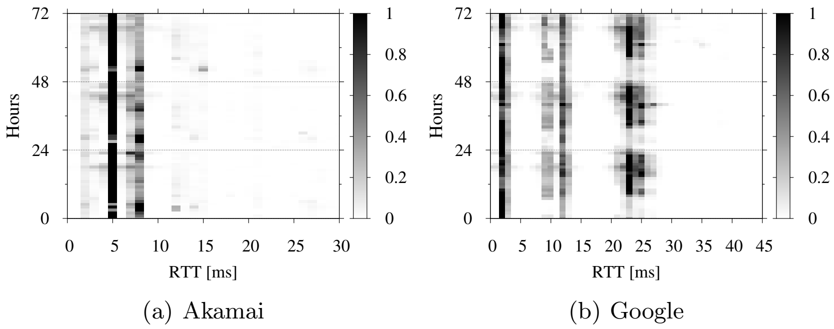

Fig. 7. Variation of the minimum external RTT over three days for VP2 data set

5 Performance Analysis

In this section we investigate the performance offered by the considered organizations. We start analyzing the Round Trip Time (RTT) as a measure of the server distance from the users hosts. We then focus on the bulk download rate for elephants.

5.1 Round Trip Time

The RTT is a typical measure that impacts both the TCP performance and the network cost. The smaller the RTT, the faster the TCP three-way handshake, the larger the download rate, and the less the network must handle the packet. Following the same methodology as in [5], we consider the *minimum external RTT* observed for each TCP flow over all valid RTT samples during the flow lifetime[4]. Only flows with at least 5 valid RTT samples are considered.

[4] The external RTT is measured considering the time elapsed between the data packet sent by a host inside the vantage point and the corresponding acknowledgement sent by the server.

Fig. 6 reports the CDF of the minimum external RTT observed from VP2 considering the whole week long data set. As expected the distributions present sharp knees corresponding to different servers positioned at different data centers. It follows immediately that different organizations are located at different distances from the vantage point. Flows going to Akamai, Limelight and about one third of Google servers are within few milliseconds far from the vantage point, i.e., very close to the ISP peering points; Level3 and the majority of Leaseweb flows are handled by servers in [20,50] ms range, i.e., possibly in some European country; at last, Megaupload and Facebook servers are found above 100ms, i.e., outside Europe. The long tail reflects the chance to contact one server which is not among the preferred location (recall Tab. 4).

More interestingly, some organizations leverage more than one server set (or data center) to handle the traffic. This is clearly the case for Facebook, for which two data centers are highlighted, the first at about 100ms, the second at about 160ms. Manually investigating their position, we uncover that the first data center is located in the New York area, while the second is located in the Bay Area. Interestingly, each of them serves about 50% of the traffic. Limelight and Level3 also show more than one (preferred) location. While Level3 balance the traffic among the different data locations, Limelight shows a higher preference to serve the requests from the closest data center. Finally, most of Akamai traffic is served by two data centers very close each other and to the ISP customers; recall indeed that 2% of Akamai servers manages 80% of traffic, see Fig. 5.

To give more insights about the traffic fluctuations during the day, we consider each hour long data set, computing the histogram of minimum external RTT samples in bins of 1 ms. Each histogram is then normalized with respect to its largest value. The heat maps reported in Fig. 7 shows the density of each bin for each histogram considering three days. The darker the color, the higher the fraction of flows that falls in that RTT bin. Consider first the left plot which refers to Akamai. We notice little variations over time, with most of the flows experiencing either 5 or 7 ms RTT, i.e., corresponding to the two data centers. Other organizations but Google show similar results with no dependency over time of the RTT.

Consider instead the right plot which refers to Google. It presents three possible values centered at about 3, 12, 23 ms, each corresponding to a different data center. Those correspond to the knees in Fig. 6. Interestingly, we notice that at different hours a different fraction of traffic is sent to different data center. This unveils the presence of routing policies that direct the traffic based not only on the distance but also on the time of day, i.e., the current system load. In fact, during the night when the number of requests to be served is the lowest, the "preferred" servers are the closest ones. This is not true during the day. A similar effect has been found in [5] when dealing with YouTube traffic. According to our findings, this dynamic routing affects all Google CDNs and not only YouTube traffic.

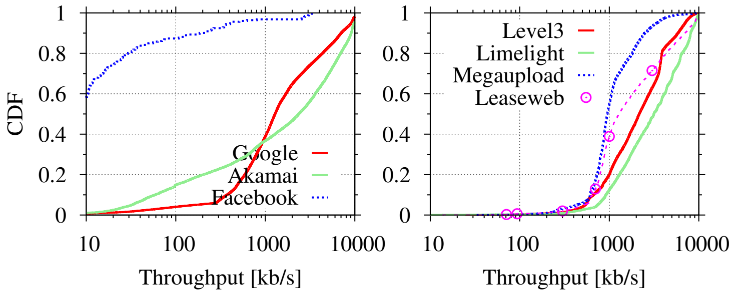

Fig. 8. Download rate of the flows with more than 1 MB for VP2

5.2 Download Throughput

We now focus on the bulk download rate defined as the total amount of bytes received from the server over the time elapsed between the first and last data segment. Only flows that carry more than 1 MB are considered. Fig. 8 shows the CDF aggregating the whole week in VP2 data set.

All the organization, except for Facebook and Akamai, present a similar distribution with a knee around 300-500 kb/s and a heavy tail. Considering the 60th percentile, Limelight offers the best performance, followed by Level3, Google and Leaseweb with 3 Mb/s, 2 Mb/s and 1 Mb/s respectively. Megaupload presents a shorter tail with a higher fraction of flows between 700 kb/s and 1 Mb/s. This is due to the throttling imposed by Megaupload on download rate to push customers to buy a premium access [6]. Finally, Facebook flows are much slower than others, with only 5% of the flows having more than 1 Mb/s. This is because Facebook is serving small objects using persistent HTTP connections for which the bulk download rate is a meaningless index. Most of Facebook flows are indeed short lived.

Interestingly, Akamai throughput follows a different distribution with 40% of flows that cannot exceed 1 Mb/s, and 40% exceeding 4 Mb/s instead. Investigating further, we have seen that this might be due to the probability of finding the content on the contacted server. If the server has already the content available, then the download is very fast. Instead, if the server has to retrieve the content, the download rate is much more limited. This suggests some form of congestion on the cache back-end rather than on the path from the server to the end customer. We have investigated if there is any time correlation, e.g., if the throughput increases during off-peak time, but we did not observe any correlation.

In particular, for organizations specialised in video distribution, i.e., Google, Akamai, Level3, the values of throughput can be biased by videos delivered with progressive download techniques that results in a bandwidth throttling by the video server.

6 Conclusions

In this paper we presented a first set of measurements that focus on the observation of HTTP traffic generated by the top player organization in the current Internet. We consider a one-week long data set collected from three different vantage points inside the same operator, analyzing the traffic from the customers' point of view.

After identifying the top players, we dug into the characteristics of each organization. We have unveiled how large is the number of servers each organization manages, how traffic is routed to which server or data center, how far those servers are from the customers, and which performance they offered. Results we collected are an interesting starting point to understand how today Internet traffic is delivered to customers.

References

1. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F.: Internet Inter-domain Traffic. In: Proceedings of the ACM SIGCOMM 2010 Conference, pp. 75–86. ACM, New York (2010)
2. Su, A.-J., Choffnes, D.R., Kuzmanovic, A., Bustamante, F.E.: Drafting Behind Akamai (Travelocity-based Detouring). SIGCOMM Comput. Commun. Rev. 36, 435–446 (2006)
3. Triukose, S., Wen, Z., Rabinovich, M.: Measuring a Commercial Content Delivery Network. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 467–476. ACM, New York (2011)
4. Nygren, E., Sitaraman, R.K., Sun, J.: The Akamai Network: a Platform for High-performance Internet Applications. SIGOPS Oper. Syst. Rev. 44, 2–19 (2010)
5. Torres, R., Finamore, A., Kim, J., Mellia, M., Munafò, M., Rao, S.: Dissecting Video Server Selection Strategies in the YouTube CDN. In: IEEE ICDCS (2011)
6. Mahanti, A., Williamson, C., Carlsson, N., Arlitt, M., Mahanti, A.: Characterizing the File Hosting Ecosystem: A View from the Edge. Perform. Eval. 68, 1085–1102 (2011)
7. Finamore, A., Mellia, M., Meo, M., Munafò, M.M., Rossi, D.: Experiences of Internet Traffic Monitoring with Tstat. IEEE Network 25(3), 8–14 (2011)
8. Bonfiglio, D., Mellia, M., Meo, M., Rossi, D., Tofanelli, P.: Revealing Skype Traffic: When Randomness Plays with You. SIGCOMM Comput. Commun. Rev. 37, 37–48 (2007)
9. Finamore, A., Mellia, M., Meo, M., Rossi, D.: Kiss: Stochastic Packet Inspection Classifier for UDP Traffic. IEEE/ACM Transactions on Networking 18(5), 1505–1515 (2010)
10. Maxmind Geoip Tool, `http://www.maxmind.com/app/organization`
11. Ager, B., Mühlbauer, W., Smaragdakis, G., Uhlig, S.: Comparing DNS Resolvers in the Wild. In: ACM IMC (2010)