
Measurement Based Analysis of One-Click File Hosting Services

Josep Sanjuàs-Cuxart · Pere Barlet-Ros ·
Josep Solé-Pareta

Abstract It is commonly believed that file sharing traffic on the Internet is mostly generated by peer-to-peer applications. However, we show that HTTP based file sharing services are also extremely popular. We analyzed the traffic of a large research and education network for three months, and observed that a large fraction of the inbound HTTP traffic corresponds to file download services, which indicates that an important portion of file sharing traffic is in the form of HTTP data. In particular, we found that two popular one-click file hosting services are among the top Internet domains in terms of served traffic volume. In this paper, we present an exhaustive study of the traffic generated by such services, the behavior of their users, the downloaded content, and their server infrastructure.

Keywords traffic analysis · HTTP · file sharing · web applications · RapidShare · Megaupload · peer-to-peer

1 Introduction

The Hypertext Transfer Protocol (HTTP) [1] is the most popular and well-known application-level protocol in the Internet, since it is the basis of the World Wide Web. The HTTP protocol follows the classic client-server architecture, and it was originally designed to transfer content, typically in Hyper-

Josep Sanjuàs-Cuxart · Pere Barlet-Ros · Josep Solé-Pareta
Universitat Politècnica de Catalunya, Departament d'Arquitectura de Computadors.
Campus Nord UPC – D6 Building. Jordi Girona 1-3, 08034 Barcelona, Spain
E-mail: jsanjuas@ac.upc.edu

Pere Barlet-Ros
E-mail: pbarlet@ac.upc.edu

Josep Solé-Pareta
E-mail: pareta@ac.upc.edu

text Markup Language (HTML) format, from a web server to a user running a web browser.

However, many alternative applications and services based on the HTTP protocol have recently emerged, such as video streaming or social networking, and have rapidly gained extreme popularity among users, mostly thanks to the explosion of the Web 2.0 [2] and the development of technologies such as Asynchronous JavaScript and XML (AJAX) [3] and Flash [4]. Given this increasing popularity, recent research works have been entirely devoted to the study of the particular characteristics of the traffic generated by these novel web-based sources [5–7].

Traditionally, file sharing has never stood out among the multiple and diverse usages of the HTTP protocol given that, from a technological point of view, peer-to-peer (P2P) protocols are superior. However, while P2P accounts for a large percentage of the total Internet traffic, HTTP-based file sharing has recently gained popularity and is already responsible for a significant traffic volume, even comparable to that of P2P applications. Several popular *one-click file hosting* services, also known as *direct download* (DD) services, such as RapidShare [8] and Megaupload [9], are mostly contributing to this phenomenon.

Unlike P2P file sharing, DD services are based on the traditional HTTP client-server model. These sites offer a simple web interface open for anyone to upload any file to their servers. The sites then host the content and provide the uploader with a Uniform Resource Locator (URL), which can be freely shared without restriction (e.g., on public forums) thus enabling massive file sharing. These links are often called *direct download links* (DDL), to emphasize the difference with the links to P2P content, where downloads are usually queued and, therefore, do not start immediately.

Even though they rely on a simpler technological model, DD services offer a competitive advantage over P2P in that they are over-provisioned in terms of bandwidth whereas, in P2P applications, speeds are constrained by the availability of the content and the upstream bandwidth of the peering nodes. In order to generate revenue, one-click file hosting services rely on advertisements to users and, more interestingly, offer *premium accounts* that can be purchased to increase download speeds and bypass many of the limitations imposed by the sites to non-premium users.

In contrast, revenue in the P2P file sharing communities is generated by web sites where popular content is announced. For example, BitTorrent search engines and trackers rely on mechanisms including advertisements, users' donations and even spreading malware to generate revenue [10]. However, under this paradigm, individuals have no direct monetary incentive to host content nor, more importantly, the possibility of increasing their download speeds via payment.

Our study reveals that DD services are far more popular than expected. In the analyzed network, DD traffic constitutes more than 22% of HTTP traffic, and around 25% of all file sharing traffic. We also observed that the two most popular DD services, Megaupload and RapidShare, are in the Top-3 list (Top-1

and 3 respectively) in terms of served bytes. Perhaps surprisingly, we also found that a significant amount of users paid for premium membership accounts for unlimited, faster downloads, which is a market that the P2P community fails to exploit.

The popularity of one-click file hosting services was not apparent until very recently and, despite the large associated transfer volumes, little is known about them. The main objective of this work is to gain an understanding of the main characteristics of DD traffic, and the usage patterns of such services. Such an understanding can aid network managers in deciding whether to limit the access to such sites, understanding if traffic shaping policies would be effective or technologically easy to implement, and its potential impact on nowadays' networks.

To the best of our knowledge, so far, only one work exists [11] that specifically targets one-click file hosting traffic. The study compares DD traffic to P2P, and analyzes the infrastructure of one of the major DD providers (RapidShare). The detection of premium accounts is heuristic, based on the observation that premium users obtain more bandwidth. Our study extends the existing to another network environment, and complements its results by including another major provider (Megaupload), featuring an accurate premium account detection, and providing an analysis of the download speeds obtained by users and the kind of contents available on DD sites. We also discuss differences between HTTP and P2P based file sharing traffic that are of relevance to network managers. We find that DD tends to use more transit link bandwidth, since it does not exploit locality of user interest on files. Another important finding is that DD traffic often abuses Transmission Control Protocol (TCP) congestion control mechanisms to gain an unfairly large share of network resources under congestion.

In this work, we analyzed the HTTP traffic of a large research and education network during three months without interruption (from March to June 2009). Our measurement-based study mainly focuses on DD services and spans four major dimensions: traffic properties (Section 4), user behavior (Section 5), downloaded content (Section 6), and server infrastructure and geographical distribution of the traffic (Section 7).

2 Related Work

The study of the Internet traffic is an important research topic. While several works have been devoted to the analysis of traffic workloads of the most popular Internet applications during the last years (e.g., [3, 5, 6, 12–20]), comparatively few have studied the long term evolution of the Internet traffic. One of the most recent ones analyzed the traffic of a trans-oceanic link from 2001 to 2008 [21]. The study confirms that HTTP and P2P are currently the most dominant applications on the Internet. They have progressively replaced other protocols and applications, such as File Transfer Protocol (FTP) and

e-mail, that carried a large part of the Internet traffic at the beginning of the nineties [22].

Ipoque has released three of the most extensive Internet traffic studies existing so far, which cover years 2006 [23], 2007 [24] and 2008-2009 [25]. While the 2006 and 2007 editions of the study highlighted P2P as the most dominant application on the Internet in terms of traffic volume, the 2008-2009 study reported a clear growth of HTTP traffic compared to P2P, which was mostly attributed to file sharing over HTTP.

Many applications run nowadays over HTTP, such as web-based e-mail clients, instant messengers, video streaming or social networks. Therefore, they have also been the main subject of study of several research works. For example, a measurement-based analysis of YouTube, which is currently the most popular video streaming service, was presented in [6], while [5] examined the particular traffic characteristics of several popular AJAX [3] based applications. Other previous works were devoted to characterize the behavior of web users (e.g., [12–14]), which is crucial for the proper design and provisioning of web-based services or to enable network optimizations.

A preliminary analysis of the application breakdown that runs over HTTP was presented in [7], which used two short traces of HTTP traffic collected in 2003 and 2006 at a research institution network. The study shows an increasing trend in the use of HTTP for activities other than traditional web browsing.

P2P protocols are widely used for file sharing, and are considered one of the main sources of traffic on the Internet. Therefore, they have been extensively studied in the literature. In [17], an analysis of the P2P traffic from a large Internet Service Provider (ISP) was presented. Traffic features similar to those studied in this work were analyzed for P2P applications, such as the number of hosts, traffic volumes, duration of connections and average bandwidth usage. Several measurement studies have concentrated in particular P2P networks. Ref. [16] focuses on Gnutella and Napster traffic. While both are P2P file sharing protocols, Gnutella is purely decentralized, and Napster relies on a centralized file location service. A significant fraction of peers are found not to share, and a high degree of peer heterogeneity is reported. An analysis of Kazaa traffic [26] attempts to model its workload, which is found to be very different to that of the Web. Currently popular P2P protocols have also been studied, including BitTorrent [19] and eDonkey [18]. Finally, Skype [20] is another widespread P2P application that provides Voice over Internet Protocol (VoIP) services. Ref. [27] shows that P2P file sharing can significantly save bandwidth in transit links when locality is exploited, while a recent study investigates the incentives that drive content publication specifically in BitTorrent trackers [10].

While research works have already studied in depth the traffic of several HTTP and P2P applications, so far, only one work [11] specifically targets DD traffic and, in particular, one of the major one-click file sharing services (RapidShare) through a combination of passive and active measurements. Ref. [11] also features an interesting comparison with BitTorrent download rates, an

Table 1 Principal features of the analyzed traffic

feature	value
Duration of study	3 months (March 20 to June 20 2009)
Traffic	148 TB in, 294 TB out
Avg. bw. usage	156 Mbps in, 310 Mbps out
Estimated p2p traffic	50.32 TB in, 235 TB out
HTTP traffic	78.85 TB in, 30.30 TB out
(DD only)	17.71 TB in, 0.89 TB out
HTTP queries	307.40 M in, 1321.56 M out
(DD only)	0 in, 7.59 M out
# Internet domains accessed	2.7 M

analysis of RapidShare’s load balancing policies, and an examination of the direct download links available in popular content indexing sites.

Our study complements the results of [11] using passive measurements on the traffic of a large academic ISP, with thousands of users of such services. Compared to [11], our study features a larger network scenario, accurate detection of premium accounts (instead of heuristic), and includes the top two DD services: Megaupload and RapidShare. Additionally, our study provides a deeper analysis of the data rates attained by users that reveals that premium users obtain higher download speeds abusing TCP congestion control mechanisms, and analyzes a large sample of the contents that were downloaded from DD sites.

3 Scenario and Methodology

3.1 Network scenario

The data presented in this work have been collected at the access link of a large research and education network with around 50000 users.¹ Our measurement point consists of a full-duplex Gigabit Ethernet link, with an average rate of 442 Mbps including both incoming and outgoing traffic.

Table 1 summarizes the most relevant features of the analyzed traffic. We have continuously ran our analysis between March 20 and June 20 2009. During these 3 months, our traffic analysis software has tracked well over 1.6 billion HTTP requests and responses. Around 81% of these requests were made by clients inside the network under study.

In order to reduce the impact of local particularities of our measurement scenario, we restrict our study to outgoing HTTP queries and incoming responses. We thus choose to ignore HTTP traffic served by the monitored network, as well as the corresponding requests, in the rest of this study.

¹ We avoid disclosing the name of the organization for privacy concerns.

Our data will still be biased by the geographical location and the profile of the users of this network. Removing such bias would require an open worldwide measurement infrastructure, which is currently not available to conduct this kind of research. Our measurements are thus necessarily constrained to the scenario where the data has been collected.

It is important to note that this study cannot be done with publicly available traces for two primary reasons. First, public traces are usually anonymized and, thus, accessed web servers cannot be recovered. Second, such traces do not contain the HTTP headers, which are needed in this study.

3.2 Measurement methodology

In this work, we have employed a passive traffic analysis approach. While the measurement methodology in itself is not a novelty of this work, we include several details that help understand how the measurement data was gathered. We were provided with access to a copy of the traffic that traverses the link between the described network and the Internet. In order to set up a monitoring node and quickly develop software tools for traffic analysis, we have used the CoMo [28] general-purpose passive network monitoring system. In particular, we have developed a set of CoMo modules that anonymously track and analyze all HTTP traffic.

In order to identify direct download (DD) traffic², we have compiled a large list of one-click file hosting domains that contains 104 entries, gathered by analyzing the source code of JDownloader [29] (a popular download manager that has support for several DD services) and browsing Internet forums where direct download links are published. We omit the full list of domains in the interest of space.

As will be shown, two download services stand out in popularity: Megaupload and RapidShare. For both these services, we have instrumented our measurement code to detect whether downloads correspond to paid visits by analyzing HTTP and HTML headers. However, we do not capture information that links such traffic to individuals.

In order to gain perspective on the proportion of file downloads, we also track the rest of the traffic. However, we do not collect detailed statistics. We limit data collection to the aggregate number of bytes per unique combination of IP protocol, source port and destination port in intervals of 10 minutes. This data is used to calculate the proportion of HTTP traffic compared to the total, as well as to compare the data volumes of HTTP traffic to that of P2P activity.

Traffic classification is technologically complex and requires access to packet payloads to obtain accurate results, while we only collect HTTP headers. This is an active topic of research (e.g., see [30]), and the application of sophisticated methods for accurate, on-line P2P identification was outside the scope of

² Throughout this work, we use the terms direct download and one-click file hosting to refer to HTTP-based file sharing services interchangeably.

this work. Our analysis is centered on measuring file hosting services, and we only require an estimation of P2P traffic volumes for comparative purposes. Therefore, we elect to identify P2P with a deliberately simple well-known ports approach. We use the Internet Assigned Numbers Authority (IANA) list of ports to clearly identify non P2P traffic, and assume the rest corresponds to P2P.

Our approximate approach has two primary sources of inaccuracy. First, P2P traffic can mask as regular traffic running on well-known ports and, in particular, on TCP port 80. Second, not all traffic running on unknown ports is necessarily P2P. To verify the significance of these sources of inaccuracy in our network, we obtained access to 13 full-payload, bidirectional traces from the same network under study, which we fed to Ipoque’s state-of-the-art Protocol and Application Classification Engine (PACE) traffic classifier [31]. According to PACE, only 5.7% of the identified P2P traffic used well-known ports, and only around 0.6% specifically TCP port 80. Conversely, 91.61% of the classified traffic not involving well-known ports was P2P. Thus, it can be concluded that our approach provides reasonable traffic volume estimates in our scenario.

3.3 Privacy concerns

Although the HTTP headers of each request and response are analyzed in order to extract relevant information such as the domain name being accessed, the URL being queried or the file sizes, we preserve the privacy of the individual users by implementing the following policies. First, all collected traffic is processed and discarded on-the-fly and only highly aggregated statistics are stored to disk. Second, the IP addresses of the hosts in the network under study are always anonymized prior to the analysis. We do not keep information that permits the recovery of original IP addresses or that can link individual users or hosts to their network activities.

3.4 Representativeness of the Data Set

Our dataset is inevitably, as almost any traffic analysis work, biased by the scenario where the data has been collected. A large majority of the traffic in our setting is generated by students. This might slightly overestimate the popularity of DD. However, in terms of traffic volume, we believe that our findings are still representative to a high degree, since independent traffic analysis highlight the growth of HTTP traffic, both in absolute terms and, in particular, relative to P2P. In particular, ref. [25] cites direct download as one of the main causes of this growth, together with the rise in popularity of video streaming. In this network, no traffic filtering policies are applied to P2P or HTTP traffic that can bias the results. In the cases where we observe a regional bias in the data set, we indicate it explicitly in the text. Our findings are consistent with prior work and, in particular, with the findings of [11].

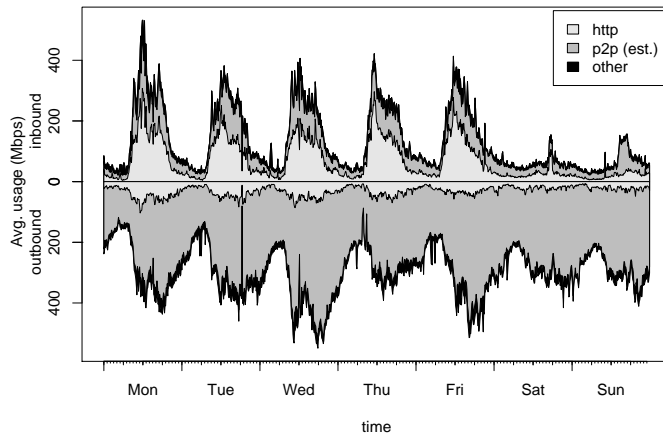


Fig. 1 Traffic profile for one week

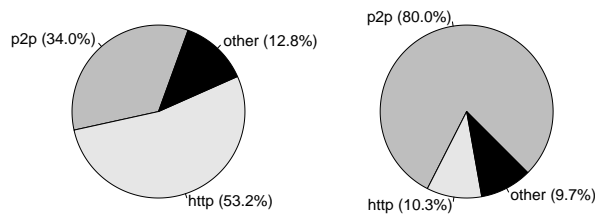


Fig. 2 Application breakdown for the inbound (left) and outbound (right) traffic

4 Traffic Analysis

4.1 Traffic profile

Figure 1 presents a profile corresponding to one week of the overall traffic we have observed. As expected, the traffic profile is extremely time dependent. The traffic spikes during work hours, but it is significantly lower at nights and during weekends. A residential network would most likely show a slightly different profile, spiking outside work hours, but still low during nights.

The majority of inbound traffic, especially during working hours, corresponds to HTTP (53.2%). Note however the different profile of outgoing traffic, where peer-to-peer (P2P) is clearly dominant, taking 80.0% of the bytes. Our intuitive explanation for this fact is that, since this network has a huge up-link capacity, P2P nodes can serve large volumes of traffic, but cannot achieve comparable downstream speeds, given that downloads are constrained by the capacities of remote nodes (e.g., ADSL lines). In the downlink, P2P represents a comparatively smaller 34.0% of the incoming data. Figure 2 presents the protocol breakdown of the traffic.

Downloads from one-click file hosting sites are taking 22.46% of all incoming HTTP traffic. Therefore, the common hypothesis that all file sharing is in

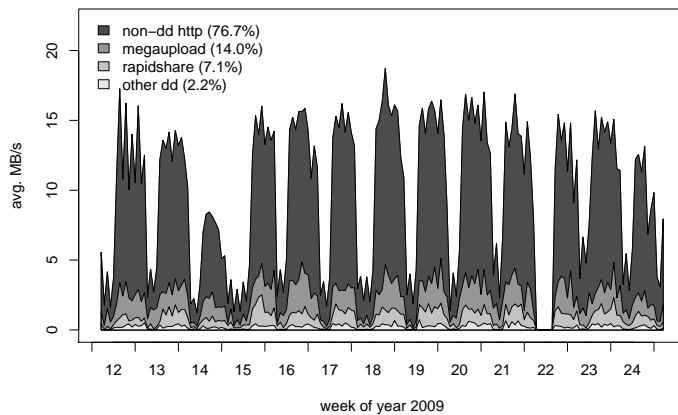


Fig. 3 Aggregate HTTP traffic volume (12-hour averages)

the form of P2P traffic leads to a severe underestimation of the incoming data volumes that correspond to file sharing. Under the assumption that all P2P traffic corresponds to file sharing, 24.91% of all incoming file sharing traffic is in the form of HTTP responses. Note that this is an estimate, given that not all P2P traffic necessarily corresponds to file sharing, and that we perform heuristic P2P detection, as explained in Section 3.2.

The observed data volumes are not particular to our scenario, and have been also recently observed in commercial networks [25], where an important increase of HTTP, together with a decrease of the relative volumes of P2P traffic, is reported.

Figure 3 shows a time series of the volume corresponding to the HTTP traffic we have analyzed, including both requests and replies, averaged across intervals of 12 hours. As expected, traffic volume significantly decreases during weekends, where most users of the network are away from the campuses. Week 14 also shows a significant drop in the traffic volumes, since it corresponds to a vacation period in the country where the network under study is located. During week 22, our network monitor suffered a service cut during approximately three days, and we therefore lack data for the corresponding time period. Finally, the downward trend at the end of the measurement period is a consequence of the diminishing academic activities associated to the end of semester. The figure also shows that RapidShare and Megaupload account for most HTTP file sharing traffic which, in turn, constitutes a large portion of all HTTP.

4.2 Popular Web Domains

As discussed in Section 3, we focus our analysis of HTTP traffic to only outgoing requests and incoming responses. By ignoring HTTP traffic served by

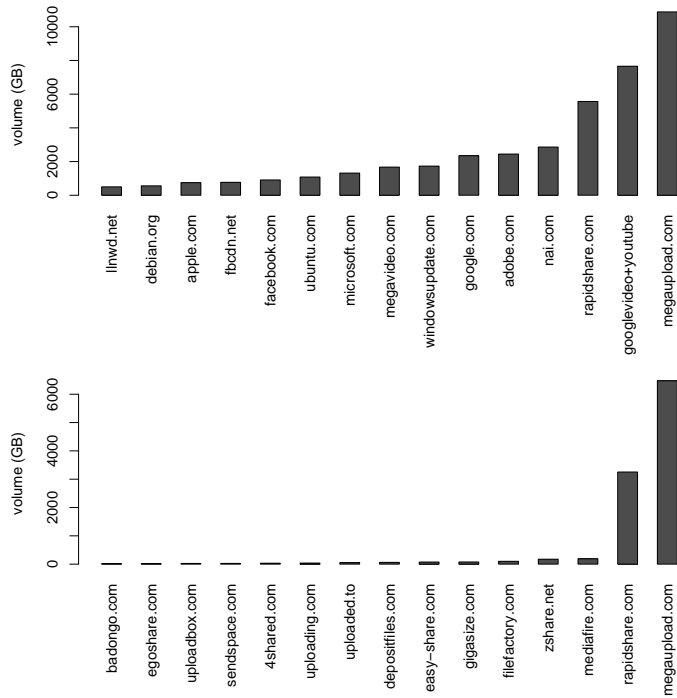


Fig. 4 Top 15 sites by served volume (top) and by served volume including only file hosting domains (bottom)

the monitored network, we avoid obvious biases in our results (e.g., the list of the most popular domains would be dominated by those served locally).

Table 1 showed that our traffic analysis software tracked approximately 1321 million outgoing HTTP requests. Figure 4 (top) shows information about the top Internet domains in terms of served volume of data over HTTP. The key observation that motivated this study was that the first and third top domains correspond to one-click file hosting services (Megaupload and Rapid-Share), even above domains as popular as search engines, social networks or video streaming services.

Even though one-click file hosting sites generate huge volumes of data, they cannot compete in popularity with the mentioned domains, as can be observed in Figure 5. In particular, Figure 5 (top) shows the top 15 domains ranked by number of hits. This rank is populated by domains that tend to attract a larger number of clients, but also favors sites that serve the content over multiple connections (e.g., using AJAX to incrementally update page contents). Figure 5 (bottom) contains the top 15 domains in number of unique IP client addresses. This list introduces sites that serve content to automatic software update agents as well as domains that are linked to by a large number of websites, such as statistics collection or advertising services.

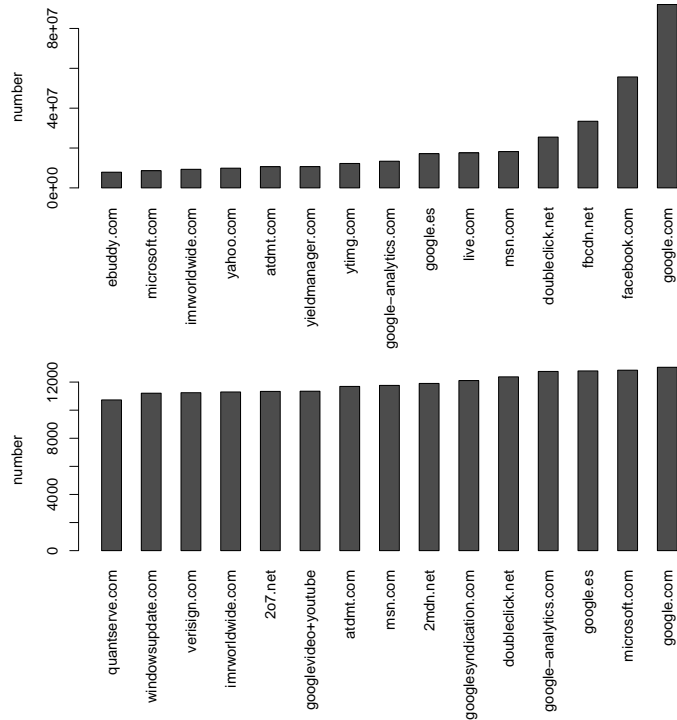


Fig. 5 Top 15 sites by number of hits (top) and number of clients (bottom)

Table 2 shows the ranks of the domains found in Figure 4 (top) according to several criteria. We have also included, as a reference, the Alexa “traffic ranks” [32] of each of the domains for the country where the network is located. Alexa is a company that collects information about the popularity of web sites. They offer a toolbar that users can install as an add-on to their web browser, which collects information about their web browsing activities. This service constitutes the best available source of information on website popularities to the best of our knowledge.

Figure 4 (bottom) focuses only on one-click file hosting services. While the amount of companies that offer file hosting services is rather large, the market appears to be dominated by the already mentioned Megaupload and RapidShare by an overwhelming margin. Together, they account for more than 90% of the one-click file hosting traffic volume and over 20% of all incoming HTTP traffic.

5 User Behavior

So far, HTTP data have been presented aggregated or on a per-domain basis. In this section, we focus only on direct download (DD) traffic and, in particular,

Table 2 Ranks of the top content serving domains

site	DD	volume	Ranks		Alexa
			#hits	#clients	
megaupload.com	✓	1	31	262	30
googlevideo+youtube	-	2	21	10	6
rapidshare.com	✓	3	49	304	25
nai.com	-	4	146	50	48104
adobe.com	-	5	83	17	89
google.com	-	6	1	1	4
windowsupdate.com	-	7	100	14	28
megavideo.com	-	8	239	467	15
microsoft.com	-	9	17	2	28
ubuntu.com	-	10	278	290	1897
facebook.com	-	11	2	31	5
fbcdn.net	-	12	3	33	3620
apple.com	-	13	44	148	108
debian.org	-	14	472	754	4328
llnwd.net	-	15	118	47	7108

on the behavior of the users of these services. Although a large number of sites exist that offer DD services, we restrict our analysis to Megaupload and RapidShare, since as shown in Section 4, they account for more than 90% of all DD traffic in our network.

We extended our traffic analysis software to include HTTP session tracking capabilities. In particular, we instrumented our code to detect logins to Megaupload and RapidShare, and to detect whether the users have signed up for premium membership accounts. In order to protect the privacy of the users, while still enabling us to track user behavior, we anonymize the user names prior to the analysis, and we avoid storing any information linking connections to individuals. Ref. [11] instead performs heuristic account detection based on the download speeds attained by clients; we therefore expect our measurements to be more accurate.

We did not enable our login tracking feature until May 28. Since then, we observed around 0.88 million accesses to Megaupload, which resulted in 2.79 TB of downstream traffic and 0.30 TB of upstream traffic, and 0.74 million accesses to RapidShare that served 1.50 TB and received 0.12 TB.

5.1 Paid membership

File hosting services generally offer downloads for free. However, they artificially limit the performance of their service and, in order to generate revenue, offer paid *premium* accounts that bypass such restrictions. These limitations include bandwidth caps, restrictions on the data volumes that can be downloaded per day, as well as delays in the order of 30 seconds before users can start the downloads. While this limited version of the service can still appeal

to a casual user, heavy users are compelled to obtain a premium account for several reasons:

1. they can download an unlimited or very large number of files per day (e.g., up to 5GB per day in the case of RapidShare);
2. they are able to simultaneously download several files;
3. they are allowed to pause and resume downloads using HTTP file range requests;
4. downloads can be accelerated by fetching a single file using multiple locations, using software download managers;
5. they can upload larger files (e.g., the limit in RapidShare increases from 200MB to 2GB and, in the case of Megaupload, from 500MB to an unlimited file size);
6. they can automate downloads, e.g, users are not required to solve CAPTCHAs [33] prior to each download;
7. premium accounts bypass download delays.

RapidShare premium accounts can be easily detected by checking for the presence of a cookie in the HTTP request.³ It is slightly more complex in the case of Megaupload, since they offer two types of accounts: “premium” and “member” accounts (that have some over advantages over unauthenticated access). In their case, our analyzer tries to match two patterns characteristic to each of the account types to the first bytes of each HTTP response from Megaupload servers.

According to our data, 470 (around 10%) users of DD services in our network have paid for premium accounts, an observation that is consistent with [11]. Although premium users are a minority, they contribute to a large portion of the download traffic on both Megaupload and RapidShare. In our data, premium users account for 29.6% of the total download traffic.

5.2 Bandwidth

The primary reason why DD links are popular is that downloads start immediately upon request (especially for premium users) and at a sustained high speed, compared to peer-to-peer systems, where download speeds are highly influenced by the number of peers that have pieces of the content and their available uplink bandwidth. Figure 6 plots the Cumulative Distribution Function (CDF) of the download speeds that unregistered users reach when downloading from Megaupload (top) and RapidShare (bottom). It is apparent that RapidShare throttles connections. Two bandwidth limitations can be clearly appreciated in the figure: one around 125KB/s and another around 225KB/s, due to RapidShare changing their policies on the constraints applied to non-premium users during this study [34]. Ref. [11] does not identify the lower bandwidth cap, since it was introduced during May 2009, after their study

³ According to our data, RapidShare enhanced their login procedure on June 8th. As of this writing, sessions cannot be tracked with this simple approach.

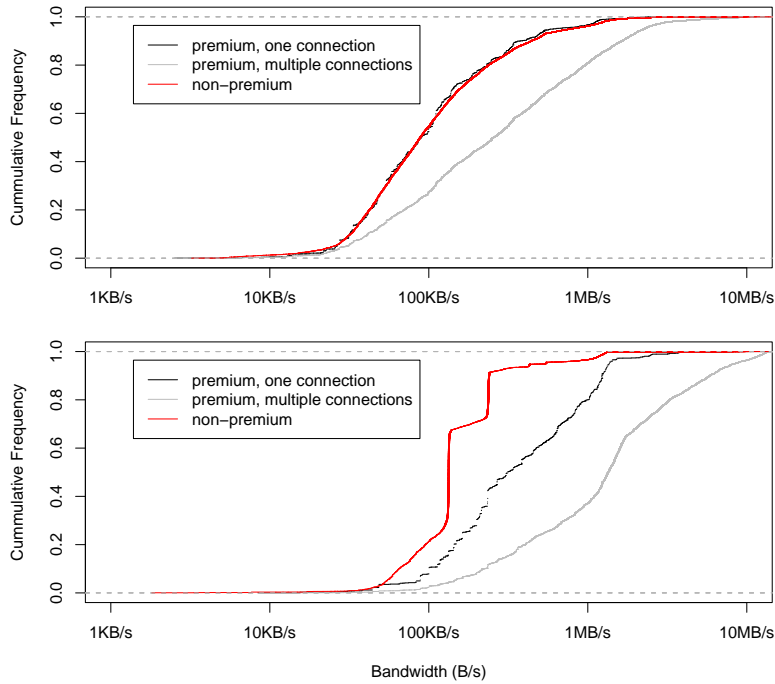


Fig. 6 Bandwidth obtained by Megaupload users (top) and RapidShare users (bottom)

had finished. No such throttling appears to be done by Megaupload, where download speeds average around 200KB/s. Usually, peer-to-peer networks can only achieve such download speeds on extremely popular content. However, the popularity of a file does not directly affect the rate at which it can be downloaded from DD sites. An interesting comparison of the data rates obtained using RapidShare versus BitTorrent can be found in [11].

As expected, premium users obtain considerably higher download rates, as can be seen in the figures. Premium downloads average around 600KB/s from Megaupload and around 2200KB/s from RapidShare. These higher data rates are achieved for two main reasons. First, in the case of RapidShare, premium downloads are not artificially throttled and are served at the highest possible rate. Second, in both services, premium users are allowed to open multiple connections to the servers, thus increasing their chances to get a larger share of network resources.

The gain obtained by opening multiple, simultaneous connections is caused by TCP congestion control mechanisms [35], which are designed to allocate an equal share of network resources to each connection. In other words, TCP is fair to connections, but not directly to users. As a result, users can unfairly increase their overall bandwidth share by parallelizing downloads, at the expense of users who do not resort to this practice. This is a known problem to network

managers [36] and a widespread practice among DD users, as will be seen in section 5.3.

Figure 6 also shows the CDF of the data rates achieved by premium users when opening a single connection. This way, we can isolate the throttling mechanisms used by each site from the effects of TCP congestion control algorithms. The figure confirms that Megaupload does not give special treatment to connections from premium users, since, when opening a single connection, their data rates are equal to non-premium users (who are restricted to one connection). One can therefore conclude that, at Megaupload, the performance gains that premium users obtain over non-premium ones are exclusively a consequence of their ability to parallelize downloads. On the contrary, RapidShare treats connections differently, serving premium connections more than twice as fast. However, RapidShare premium users can still benefit from better data rates by opening several concurrent connections, as can be observed in the figure. Thus, in DD services, users have a clear incentive to open multiple connections, often with the aid of software download managers, which we discuss in the next section.

5.3 Use of downloaders

Mass file downloading from DD services is time-consuming and ineffective. Unlike P2P downloads, where the software usually handles download queues, in the case of DD the client is a regular web browser, which does not have such capabilities by default. This inconvenience is aggravated because of the fact that, very often, large files are broken into several pieces by the uploader before reaching the DD servers, in an attempt to overcome the upload file size limitations. Such files must be independently downloaded by the user. Thus, even if a user is interested in one particular content, she is often forced to download several files.

For this reason, software download managers exist that are specialized on DD services. For example, RapidShare and Megaupload not only do not discourage their use but, instead, they offer their own download managers to their user base for free. Third party download managers that are compatible with several DD sites are also available (e.g., JDownloader [29]). The principal reason to use them is that they automate download queues. However, another important reason to use a download manager is they often provide download acceleration by simultaneously retrieving several files, or even opening several connections in parallel to download a single file, achieving noticeably higher download speeds by abusing TCP congestion control mechanisms to gain an unfair share of network resources, as discussed in section 5.2.

Figure 7 presents the distribution of the number of simultaneous connections for premium users. While in more than 50% of the cases only one download was active per user, there is a non negligible amount of parallel downloads.

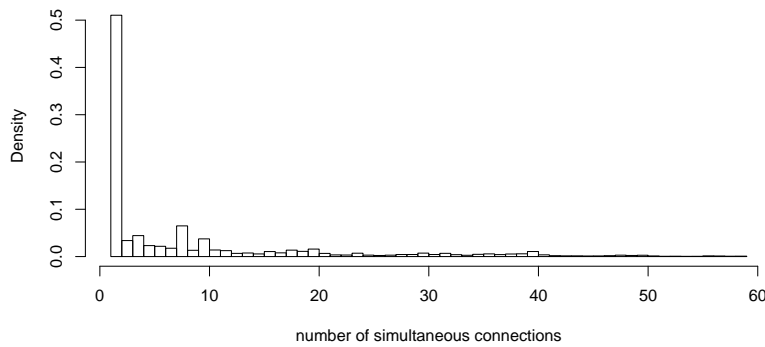


Fig. 7 Number of simultaneous connections for premium users

6 Downloaded Content

This section is devoted to the analysis of the contents that are downloaded from direct download (DD) services. In order to protect the privacy of the users, while we analyze file names, their extension, and size, we discard any information on which user has downloaded each file.

6.1 File types

We obtained a list of file names being downloaded, which is trivially found in each HTTP request. We also recorded their associated download volumes and number of requests. In total, our traffic analysis module has captured 181249 unique file names, all of them hosted at Megaupload or RapidShare.

Table 3 presents a list of the most prevalent file extensions. Notably, it can be observed that RAR archives are overwhelmingly the most prevalent (almost 75% of files), followed by the video extension AVI (around 9%).

We classified files into the following categories: *video*, *music*, *software*, *document*, and *image*. An additional *unknown* category was introduced for the cases where we were unable to classify a file.

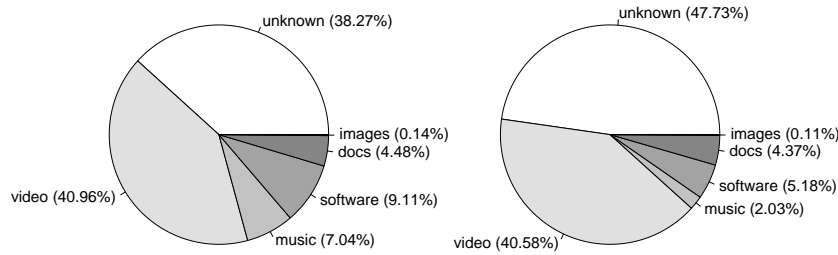
In an effort to identify the kind of contents that are being distributed within RAR archives, we stripped the ‘partX’ sub-string of each RAR file name (if present) and removed duplicates. Then, we randomly drew 3772 samples (around 3% of all RAR archives) out of the resulting list and manually classified their content type. Still, we were unable to classify around 43% of the RAR archives of our sample, due to file names being too cryptic or generic to determine the category of the content.

For the rest of the files (i.e., all but RAR archives), we classified their content type according to their extensions. In this case, we easily identified the content type of 80% of the files, since well known file extensions are prevalent.

Figure 8 (left) shows the content breakdown in terms of number of files, assuming that the sample of RAR files we classified is representative. We

Table 3 Most prevalent file extensions

extension	#files	%	extension	#files	%
pps	66	0.04%	rev	358	0.20%
iso	74	0.04%	rmvb	534	0.29%
divx	79	0.04%	exe	540	0.30%
djvu	112	0.06%	mp3	667	0.37%
doc	121	0.07%	wmv	835	0.46%
cab	125	0.07%	cbr	1474	0.81%
mkv	128	0.07%	mp4	2201	1.21%
dlc	131	0.07%	pdf	2220	1.22%
par2	135	0.07%	r(number)	3664	2.02%
flv	148	0.08%	zip	4237	2.34%
7z	149	0.08%	(number)	11025	6.08%
jdu	154	0.08%	avi	16596	9.16%
srt	222	0.12%	rar	133263	73.52%
mpg	324	0.18%	(other exts.)	1667	<1%

**Fig. 8** Content by number of files (left) and by number of hits (right)

classified around 40% of the files as video, 9% as software, 7% as music and 4.5% as documents. Figure 8 (right) presents the percentage of hits per file type, which provides a measure of the popularity of each category.

Our results differ from those of [11] since, in our case, we collect file names from the traffic generated by the users of the network and, therefore, our results reflect the popularity of the users of this network. Instead, ref. [11] gathers a list of files from content indexing sites, which is less prone to geographical bias, but omits part of the traffic of files interchanged directly by users (hence the large percentage of unknown files in our study). We find a greater amount of video and, most notably, a much smaller proportion of software compared to [11].

6.2 Fragmented content

It is a very common practice for users to manually split a large file in several parts prior to sharing them. This is a practice that is not found in P2P file sharing, since it does not bring any benefits. However, in this environment it does provide several advantages.

First, it circumvents upload and download file size restrictions. Second, it allows unpaid users, who are usually not allowed to pause and resume downloads, to download the file in smaller pieces. Third, the content provider gathers more *reward points*. In order to gain market share, DD sites usually reward the uploaders of popular content; for example, as of this writing, one of the major sites rewards any user with \$10,000 every 5 million downloads of the content he has uploaded.

This practice explains the popularity of RAR archives, which can contain any kind of content, conveniently partitioned in smaller fragments. This is an extremely common practice, to the point that 85.8% of the RAR files we have observed (and therefore almost two thirds of *all* the files) were of the form *filename.partX.rar*, where X denotes a number. For example, WinRAR [37] uses this exact file naming scheme when fragmenting an archive.

6.3 File sizes

We obtained the file sizes of the downloaded files by extending our measurement software to parse additional fields of the HTTP request. We enabled the analysis of the downloaded files in our measurement software on May 28, so the results presented in this section correspond to a period of three weeks. Overall, our traffic analysis module has collected 66,330 unique file names and sizes, all of them hosted at Megaupload or RapidShare. Note that in [11], file sizes are not directly analyzed. Instead, their analysis is centered around flow sizes, which tend to be smaller.

Figure 9 (top) plots the CDF of the observed file sizes. The most notable observation from this graph is that, according to our sample, around 60% of the files have a size of around 100MB. Figures 9 (middle and bottom) show the same CDF only for Megaupload and RapidShare files. Interestingly, while 60% of the files hosted at RapidShare are also around 100MB, this service hosts a smaller number of larger files compared to Megaupload. This difference is a result of the different upload file size restriction policies of each site.

In the case of Megaupload, the limit is 500MB, while at RapidShare, the limit is currently set at 200MB. This raises the question of why the 100MB file size is the most prevalent. The explanation we have found for this fact is two-fold. First, until July 2008, the limit for RapidShare was 100MB. Often, content uploaders publish the content in more than one direct download service for unpaid users to be able to parallelize downloads, and in order to provide a backup plan in the case a file is removed (e.g., in the event of a copyright claim over the content). Therefore, files uploaded to Megaupload were also of the same size. Second, as explained, WinRAR [37] appears to be the software being used to partition files in multiple fragments. WinRAR permits partitioning in any file size, but offers several presets that match popular storage media: floppy disks, Zip100 disks, CD-ROM and DVD. In particular, the Zip100 preset is exactly 98,078KB, and that is the most prevalent file size across RAR files

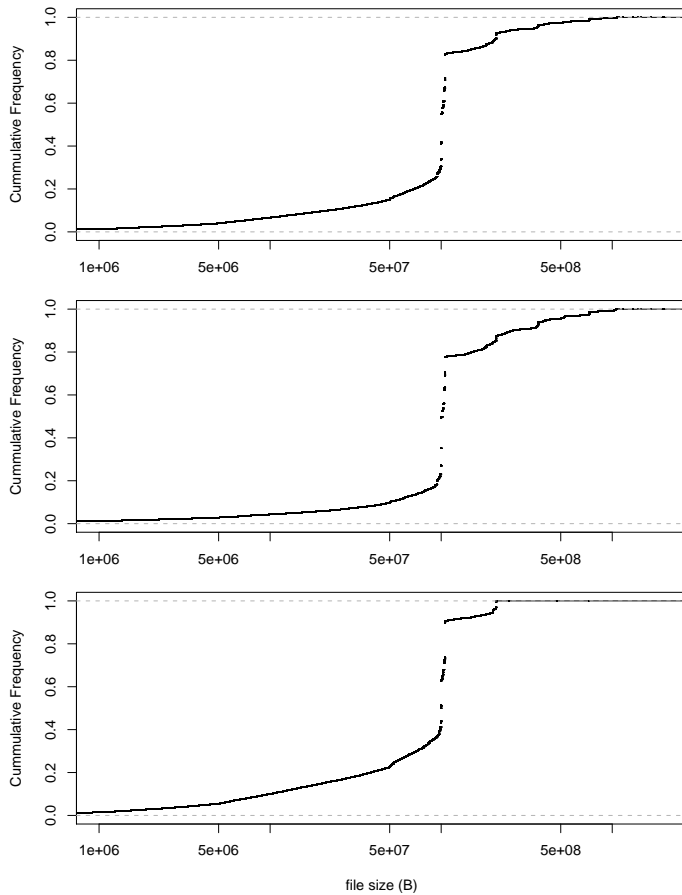


Fig. 9 Observed file sizes: overall (top), Megaupload (middle), RapidShare (bottom)

(12%), closely followed by exactly 100MB (around 9%) and 100 million bytes (around 6%).

It is apparent that, while the 100MB file limit was initially enforced by the sites, it has become the *de facto* standard file fragment size.

7 Server Infrastructure

As shown in previous sections, one-click file hosting domains serve extremely large data volumes compared to other web-based services. Therefore, it is interesting to analyze the server infrastructure that supports such services and compare it to that of other popular domains.

We collected a list of web servers that served HTTP traffic for each domain. For each HTTP connection, we also analyzed the Host header field of the requests [1]. We obtained a lower bound on the number of mirrors of each

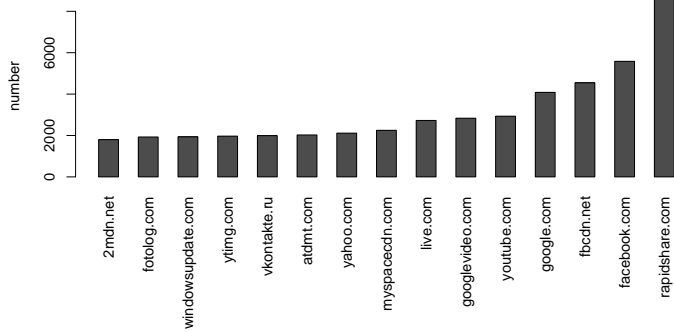


Fig. 10 Internet domains with the highest observed number of servers

domain by calculating the number of unique IP addresses that served content for each particular domain. The distance of this lower bound to the actual number of mirrors that serve a particular domain will vary depending on several factors, such as the amount of client traffic (the more popular, the larger the part of its servers that will surface) or whether sites dedicate part of the mirrors to serve requests from particular geographical areas.

The results we have obtained are summarized in Figure 10. Surprisingly, rapidshare.com is the Internet domain for which the largest number of mirrors has been observed (almost 9000), even above domains as popular as facebook.com (around 5500 servers) and google.com (around 4000). Next direct download services in number of observed mirrors are Megaupload and Mediafire, with around 800 each.

Alexa.com traffic ranks [32] indicate that, worldwide, RapidShare is the most popular file hosting service. This explains why their server infrastructure is larger compared to Megaupload's. However, according to Alexa, Megaupload and RapidShare rank in very close positions in the country where the analyzed network is located.

In Figure 11 we present the cumulative number of servers as a time series (top) and as a function of the number of accesses (bottom). During weeks 13 and week 23, we observe the addition to Megaupload of new IP address blocks belonging to four different /24 subnets. In the case of RapidShare, we did not observe new IP address blocks during the analysis, but their infrastructure appears to be considerably larger. We find a larger number of IP addresses compared to [11], which suggests that RapidShare have kept upgrading their infrastructure.

Table 4 exposes the DNS naming scheme of both RapidShare and Megaupload's server infrastructure. We used the IP address to Autonomous System (AS) mappings provided by Team Cymru [38] and the MaxMind GeoLite Country IP geolocation database [39] to approximate the geographical location of the mirrors. Both services have server infrastructure hosted in several ASs. While [39] reports IP addresses to belong to several different countries, ref. [11] features a more accurate geolocation study based on packet round-trip

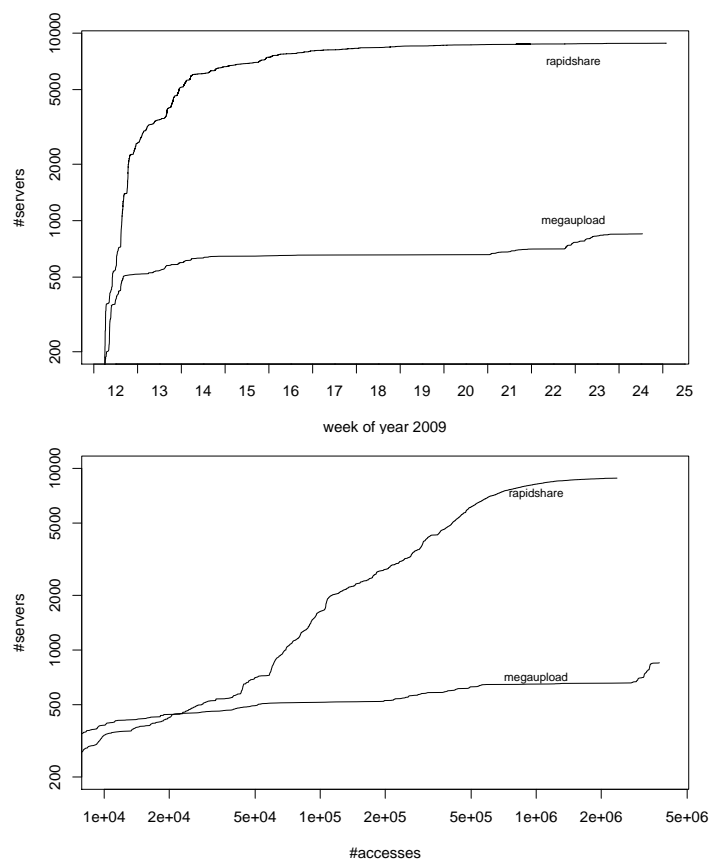


Fig. 11 Number of servers and accesses observed for Megaupload and RapidShare

times from a number of Planetlab nodes that pinpoints RapidShare servers to a single location in Germany.

In both the cases of Megaupload and RapidShare, as can be observed in Table 4, the content servers are easy to identify from their DNS name. This result suggests that, even though new servers are being deployed by both, traffic shaping of these services (to either reduce their bandwidth consumption or even block them) is relatively easy. In the case of P2P, traffic identification (and hence, policing) is more complex given its de-centralized architecture and the use of obfuscated protocols.

DD server infrastructures appear to run in a few specific data centers. In contrast, P2P architectures are distributed around the globe. As a consequence, P2P tends to use fewer backbone Internet link bandwidth compared to DD, primarily as an effect of the locality of user interest on files (additionally, [27] finds that P2P could exploit locality to achieve greater bandwidth savings in transit links). Thus, if HTTP-based file sharing were to continue

Table 4 DNS naming patterns, associated Autonomous Systems, and approximate geographical location of Megaupload and RapidShare mirrors

pattern	AS	AS name	#IPs	location (#IPs)
wwwqX.megaupload.com	29748	CARPATHIA	33	US (33)
wwwX.megaupload.com	38930	FIBERRING	39	NL (39)
	46742	CARPATHIA-LAX	133	US (133)
	35974	CARPATHIA-YYZ	135	US (135)
	16265	LEASEWEB	259	UNK (198) NL (61)
	29748	CARPATHIA	265	US (265)
rsXcg2.rapidshare.com	174	COGENT	182	DE (182)
rsXtl4.rapidshare.com	1299	TELIANET	193	EU (193)
rsXtl3.rapidshare.com	1299	TELIANET	383	DE (191) EU (192)
rsXl34.rapidshare.com	3356	LEVEL3	559	DE (182) GB (377)
rsXcg.rapidshare.com	174	COGENT	557	DE (557)
rsXgc2.rapidshare.com	3549	GBLX	557	US (557)
rsXtg2.rapidshare.com	6453	GLOBEINET	567	DE (191) UNK (184) EU (192)
rsXdt.rapidshare.com	3320	DTAG	618	DE (618)
rsXgc.rapidshare.com	3549	GBLX	748	US (748)
rsX.rapidshare.com	3356	LEVEL3	749	DE (557) GB (192)
rsXl3.rapidshare.com	3356	LEVEL3	749	DE (557) GB (192)
rsXl32.rapidshare.com	3356	LEVEL3	749	DE (373) GB (376)
rsXtg.rapidshare.com	6453	GLOBEINET	749	DE (375) GB (182) EU (192)
rsXl33.rapidshare.com	3356	LEVEL3	750	DE (374) GB (376)
rsXtl.rapidshare.com	1299	TELIANET	750	DE (558) EU (192)
rsXtl2.rapidshare.com	1299	TELIANET	752	DE (559) EU (193)

Table 5 Exchanged P2P traffic per country

country code	traffic	%	country code	traffic	%
ES	38.15 TB	46.02%	PL	778.28 GB	0.94%
(unknown)	7.20 TB	8.69%	NO	702.11 GB	0.85%
US	5.29 TB	6.38%	CN	679.21 GB	0.82%
FR	4.70 TB	5.67%	PT	638.57 GB	0.77%
IT	2.93 TB	3.54%	MX	552.43 GB	0.67%
GB	2.56 TB	3.10%	GR	534.08 GB	0.64%
SE	2.31 TB	2.79%	HU	520.60 GB	0.63%
DE	1.67 TB	2.02%	AR	513.09 GB	0.62%
NL	1.49 TB	1.80%	RO	497.28 GB	0.60%
CA	1.36 TB	1.65%	RU	445.78 GB	0.54%
JP	900.72 GB	1.09%	CL	426.71 GB	0.51%
BR	853.57 GB	1.03%	CH	399.04 GB	0.48%
AU	850.68 GB	1.03%	(others)	5.91 TB	7.12%

growing, traffic load in transit links should be expected to rise noticeably in the future.

To confirm this, we have also geolocalized the P2P traffic we have observed during the month of September 2009. The results of this experiment are summarized in Table 5. Most notably, we find that around 46% of the P2P traffic

was exchanged within the country of the network under study (ES), and more than 72% within the same continent. In contrast, all DD traffic comes from abroad.

8 Conclusions

In this paper, we analyzed the HTTP traffic of a large research and education network during three months (from March to June 2009). Our measurement-based study, which includes data from over 1.6 billion HTTP connections, reveals that the increase of HTTP traffic on the Internet reported in recent studies can be mostly attributed to two single Internet domains: RapidShare and Megaupload. The popularity of these sites is surprising considering that, for file sharing purposes, P2P based architectures are considered technologically superior. We performed an exhaustive packet-level analysis of the traffic of these two popular one-click file hosting services (a.k.a. direct download (DD) services) at four different levels: traffic properties, user behavior, content distribution and server infrastructure.

We can summarize the main results of this study in the following findings: (1) DD services generate a large portion of HTTP traffic, (2) DD services are among the Internet domains that generate the largest amount of HTTP traffic, (3) a significant fraction of file sharing is in the form of HTTP traffic, (4) DD services rely on a huge server infrastructure even larger than other extremely popular Internet domains, (5) a non-negligible percentage of DD users paid for premium accounts, (6) premium users proportionally generate a larger amount of DD traffic than unregistered users, (7) restriction policies applied to non-premium users differ significantly between different DD sites, (8) DD users can highly benefit from TCP congestion control mechanisms by parallelizing their downloads using specialized software download managers, (9) the prototypical download from a DD site is a 100MB RAR archive fragment.

The aim of this study was to investigate the characteristics of a service that is responsible for a large percentage of the Internet traffic volume. We also discussed practical implications for network management, and found that this kind of traffic is easy to identify and hence police, especially relative to P2P. Additionally, we found that direct download services utilize more transit link bandwidth compared to P2P, which is able to leverage locality.

Our plans for future work include the study of uploads from users, and the accurate detection of software download managers to obtain an estimate of their popularity.

Acknowledgements The authors thank the anonymous research institution for allowing the collection and analysis of their traffic for research purposes. We are also grateful to Ismael Castell-Uroz for his assistance and feedback. This work includes GeoLite data created by MaxMind, available from <http://www.maxmind.com/>. We acknowledge Ipoque for kindly providing access to their PACE [31] traffic classification engine for this research work. This research has been partially funded by the *Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya* (ref. 2009SGR-1140).

References

1. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext Transfer Protocol – HTTP/1.1. RFC 2616 (Draft Standard) (1999)
2. Anderson, P.: What is Web 2.0?: ideas, technologies and implications for education. JISC Technology and Standards Watch pp. 2–64 (2007)
3. Garrett, J.: Ajax: A new approach to web applications. Adaptive path (2005). <http://www.adaptivepath.com/ideas/e000385>
4. Adobe Flash: <http://www.adobe.com/>
5. Schneider, F., Agarwal, S., Alpcan, T., Feldmann, A.: The new web: Characterizing ajax traffic. In: Proceedings of the 9th International Conference on Passive and Active Network Measurement (2008)
6. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (2007)
7. Li, W., Moore, A., Canini, M.: Classifying HTTP traffic in the new age. In: ACM SIGCOMM, Poster session (2008)
8. RapidShare AG: <http://www.rapidshare.com/>
9. Megaupload: <http://www.megaupload.com/>
10. Cuevas, R., Kryczka, M., Cuevas, A., Kaune, S., Guerrero, C., Rejaie, R.: Is Content Publishing in BitTorrent Altruistic or Profit-Driven? In: Proceedings of ACM CoNext (2010)
11. Antoniadis, D., Markatos, E.P., Dovrolis, C.: One-click hosting services: a file-sharing hideout. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement (2009)
12. Feldmann, A., Rexford, J., Caceres, R.: Efficient policies for carrying Web traffic over flow-switched networks. *IEEE/ACM transactions on Networking* **6**(6), 673–685 (1998)
13. Catledge, L., Pitkow, J.: Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN systems* **27**(6), 1065–1073 (1995)
14. Barford, P., Bestavros, A., Bradley, A., Crovella, M.: Changes in web client access patterns: Characteristics and caching implications. *World Wide Web* **2**(1), 15–28 (1999)
15. Gummadi, P., Saroiu, S., Gribble, S.: A measurement study of Napster and Gnutella as examples of peer-to-peer file sharing systems. *ACM SIGCOMM Computer Communication Review* **32**(1), 82–82 (2002)
16. Saroiu, S., Gummadi, P., Gribble, S., et al.: A measurement study of peer-to-peer file sharing systems. In: Proceedings of Multimedia Computing and Networking (2002)
17. Sen, S., Wang, J.: Analyzing peer-to-peer traffic across large networks. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement (2002)
18. Tutschku, K.: A measurement-based traffic profile of the eDonkey filesharing service. *Lecture notes in computer science* pp. 12–21 (2004)
19. Pouwelse, J., Garbacki, P., Epema, D., Sips, H.: The bittorrent p2p file-sharing system: Measurements and analysis. *Lecture Notes in Computer Science* **3640**, 205 (2005)
20. Guha, S., Daswani, N., Jain, R.: An experimental study of the skype peer-to-peer voip system. In: Proceedings of IPTPS (2006)
21. Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., Cho, K.: Seven Years and One Day: Sketching the Evolution of Internet Traffic. In: Proceedings of INFOCOM (2009)
22. Claffy, K., Braun, H., Polyzos, G.: Tracking long-term growth of the NSFNET. *Communications of the ACM* **37**(8), 34–45 (1994)
23. Schulze, H., Mochalski, K.: P2P Survey 2006. <http://www.ipoque.com/resources>
24. Schulze, H., Mochalski, K.: Internet study 2007. <http://www.ipoque.com/resources>
25. Schulze, H., Mochalski, K.: Internet study 2008-2009. <http://www.ipoque.com/resources>
26. Gummadi, K.P., Dunn, R.J., Saroiu, S., Gribble, S.D., Levy, H.M., Zahorjan, J.: Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: Proceedings of ACM SOSP (2003)
27. Karagiannis, T., Rodriguez, P., Papagiannaki, K.: Should internet service providers fear peer-assisted content distribution? In: Proceedings of the 5th ACM SIGCOMM conference on Internet measurement (2005)

28. Barlet-Ros, P., Iannaccone, G., Sanjuà-Cuxart, J., Amores-López, D., Solé-Pareta, J.: Load shedding in network monitoring applications. In: Proceedings of USENIX Annual Technical Conference, pp. 59–72. Usenix Association (2007)
29. JDownloader: <http://www.jdownloader.org>
30. Nguyen, T., Armitage, G.: A Survey of Techniques for Internet Traffic Classification using Machine Learning. *IEEE Communications Surveys and Tutorials* **10**(4) (2008)
31. ipoque Protocol and Application Classification Engine: <http://www.ipoque.com/products/pace-application-classification>
32. Alexa: <http://www.alexa.com>
33. Von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: using hard AI problems for security. In: Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques (2003)
34. RapidShare news: <http://www.rapidshare.com/news.html>
35. Allman, M., Paxson, V., Blanton, E.: TCP Congestion Control. RFC 5681 (Draft Standard) (2009)
36. Briscoe, B.: Flow rate fairness: Dismantling a religion. *ACM SIGCOMM Computer Communication Review* **37**(2), 63–74 (2007)
37. WinRAR archiver: <http://www.rarlab.com/>
38. Team Cymru: <http://www.team-cymru.org>
39. MaxMind GeoLite Country: http://www.maxmind.com/app/geoip_country

Author Biographies

Josep Sanjuà-Cuxart is a Ph.D. student at the Computer Architecture Department of the Universitat Politècnica de Catalunya (UPC), where he received a M.Sc. degree in Computer Science in 2006. He was a visiting student at Intel Research Cambridge (2006) and Berkeley (2009). His research interests are centered around traffic analysis algorithms and the design of network monitoring systems.

Pere Barlet-Ros received his M.Sc. and Ph.D. degrees in Computer Science from Universitat Politècnica de Catalunya (UPC) in 2003 and 2008 respectively. He is currently an assistant professor and researcher at the Computer Architecture Department of the UPC. He was also a visiting researcher at NLANR/Endace (2004), Intel Research Cambridge (2004) and Berkeley (2007). His research interests are in the fields of network monitoring, traffic classification and anomaly detection.

Josep Solé-Pareta obtained his M.Sc. degree in Telecom Engineering in 1984, and his Ph.D. in Computer Science in 1991, both from the Universitat Politècnica de Catalunya (UPC). Currently he is Full Professor with the Computer Architecture department of UPC. He did Postdoc stages at Georgia Tech (1993, 1994), and later co-founded UPC-CCABA. He has published book chapters and more than 100 papers in research journals. His current research interests are in nanonetworking, traffic monitoring and analysis and high speed and optical networking.