Characterizing the File Hosting Service Ecosystem

Aniket Mahanti[⊠], Niklas Carlsson[⊛], and Carey Williamson[⊠]
[⊠] University of Calgary, Canada
[⊛] Linköping University, Sweden

ABSTRACT

File Hosting Services (FHS) such as Rapidshare and Megaupload have recently become popular. The decline of P2P file sharing has prompted various services including FHS to replace it. We propose a comprehensive multi-level characterization of the FHS ecosystem. We devise a measurement framework to collect datasets from multiple vantage points. To the best of our knowledge, this work is the first to characterize the FHS ecosystem. The work will highlight the content, usage, performance, infrastructure, and quality of service characteristics of FHS. FHS can have significant implications on Internet traffic, if these services were to supplant P2P as the dominant content sharing technology.

1. INTRODUCTION

File Hosting Services (FHS) were originally designed for file backup purposes and for uploading files that were too big to be sent as email attachments. FHS allow their users to upload a file to their servers in easy to follow steps. Once the file has been successfully uploaded, the site generates a unique URL that can be used for downloading the file. The user may then publish the link online for sharing content with other users.

FHS have recently received attention from networking researchers. Maier *et al.* [3] found that FHS account for 16% of all HTTP traffic in a large residential network. Labovitz *et al.* [2] report a decline in P2P traffic, but growth in traffic for FHS. The apparent decline of P2P file sharing points to a paradigm shift in how users share content. Despite the wide adoption of FHS, not much is known about their infrastructure, content characteristics, and user-perceived performance. One characterization study on FHS was done by Antoniades *et al.* [1] who studied traffic, usage, and performance characteristics of a single FHS, namely Rapidshare.

We propose a comprehensive characterization study of FHS workloads. We study four popular FHS: Rapidshare, Megaupload, Hotfile, and Mediafire. Using a year-long dataset of HTTP transaction summaries collected from a large university edge network, we characterize usage behaviour, content properties, service infrastructure, and performance of these services. To get a global picture, we use a large crawl dataset and compare and contrast the content properties of the services with locally observed characteristics.

A distinguishing feature of our work is the use of detailed Web transactions that allowed us to distinguish free and premium services based on user clickstreams. We present a case study comparing FHS with P2P, and show preliminary results highlighting content properties and performance of FHS.

2. FHS VERSUS P2P

FHS are different from traditional P2P file sharing and other content sharing services. FHS offer differentiated forms of service. Free users have to wait for a set amount of time before their download can begin, can only download a limited number of files within a given time window, and do not receive the best throughput rates for their downloads. These limitations are removed for (paying) premium users.

FHS offer several advantages over P2P technologies such as higher availability of active files, improved privacy for users, hosting diverse content, and economic incentive mechanisms for frequent uploaders [1]. We next present a case study comparing the dissemination of popular content via FHS and P2P. Specifically, for both services we analyze how quickly content is made available once the content has been broadcast, and how many copies of the content are available on the Web.

Publishing and Replication Case Study: We tracked the postings of a popular crime television drama on FHS and P2P for one season in 2009-10. Figure 1(a) shows that a small fraction (20%) of FHS content is available for download within 100 minutes of the pro-

ACM CoNEXT Student Workshop, November 30, Philadelphia, USA.

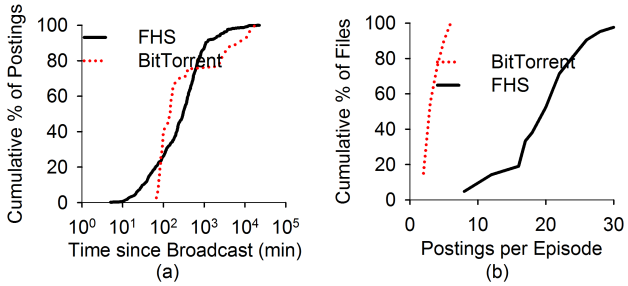

Figure 1: Content Publishing and Replication

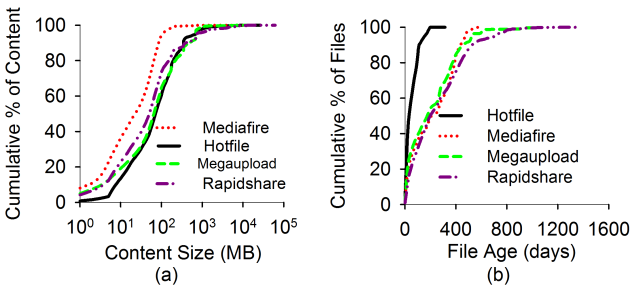

Figure 2: Size and Age of FHS Content

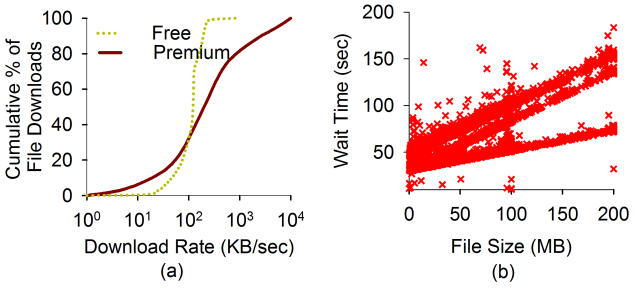

Figure 3: Rapidshare Performance

gram broadcast. Users with high-speed connections can promptly upload content to the FHS and distribute the URLs on the Web. Thus, the FHS content is ready for consumption sooner when compared to BitTorrent (BT). We found that additional reposting of episodes on FHS caused the median to be higher than that of BT as exemplified by the high number of FHS postings in Figure 1(b). Note that BT files may be available even before the content is fully seeded. These results show that *FHS are an easy media for users to make content available quickly and there are many content replicas on FHS, when compared to BitTorrent.*

3. CHARACTERIZATION RESULTS

Content Properties: We study what type of content is being hosted on FHS. After analyzing over one million files by crawling an FHS indexing search engine, we found that majority of them were archive files. This is due to the file upload size limitations imposed by FHS. Users split large content into smaller parts that comply with the FHS rules and then upload them part-by-part. Figure 2 highlights the differences among the FHS based on content size (after all parts have been assembled) and file age. Except for Mediafire, which primarily hosts MP3 files, the other three FHS host mostly very large content (over 100 MB on average). As we observe in Figure 2(b), the two older FHS, namely, Megaupload and Rapidshare, have the oldest files. We also observe files hosted on Rapidshare that dated back to its inception. Hotfile, a newer service, had files that were less than a year old. These results show that *FHS are generally being used to host very large content and the active files are being hosted for a long period of time.*

Performance: We study the download rates for free and premium users at our campus network. We restrict our attention to Rapidshare. From Figure 3(a) we observe that premium users receive an order of magnitude higher download rates than free users. We found that Rapidshare used a fixed throughput throttling rate for free users. Megaupload free users had significantly higher download rates since it did not use fixed throttling. Figure 3(b) shows the relationship between the file size and the wait time imposed by Rapidshare before free users can start their download. We observe three distinct regions where the wait times increase linearly with file size. Rapidshare is the only FHS that imposes variable wait times. These results show that *FHS offer an order of magnitude faster downloads for premium users, although there are some FHS that provide very fast downloads even for free users.*

4. CONCLUSIONS

We presented highlights from a comprehensive characterization study of FHS workloads. We showed that users can disseminate popular content more rapidly using FHS than P2P. We analyzed content properties and performance of FHS. We highlighted differences between individual FHS and showed that FHS were generally used to host very large content. Using HTTP user clickstreams, we distinguished free and premium FHS download instances, and showed that premium downloads received higher throughput than free downloads. Our results will aid in understanding the evolution of the new Web, provisioning future ISP networks, and designing better content distribution systems.

5. REFERENCES

[1] D. Antoniades, E. Markatos, and C. Dovrolis. One-click Hosting Services: A File-sharing Hideout. *IMC*, 2009.
[2] C. Labovitz, S. Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-domain Traff. *SIGCOMM*, 2010.
[3] G. Maier, A. Feldmann, V. Paxson, and M. Allman. Charac. of Resid. Broadband Internet Traff. *IMC*, 2009.