Classifying HTTP Traffic in the New Age

Wei Li, Andrew W. Moore
Computer Laboratory
University of Cambridge
{first.last}@cl.cam.ac.uk

Marco Canini
Department of Communication, Computer and System Sciences, University of Genoa
marco.canini@unige.it

ABSTRACT
HTTP has been a great success, used by many applications and provided a useful request / response paradigm. We set out to answer two questions: What kinds of purposes is HTTP used for? What is actually transmitted over HTTP?

Using full-payload data we are able to give answers to the two questions and a historical context conducting analysis over a multi-year period. We show that huge increase in http for non-browsing – notably web applications, news feeds and IM have occurred and give a quantitative analysis.

Categories and Subject Descriptors
C.2.3 **[Computer Communications-Networks]**: Network Operations, Network Monitoring

General Terms
Measurement

Keywords
HTTP, web applications

1. INTRODUCTION
The Hyper-Text Transfer Protocol (HTTP) is perhaps the most significant protocol used on the Internet today. Traditionally, the HTTP has played a fundamental part in the web browsing activities (e.g. publishing, searching, advertising), however the growth of network computing, web applications and services have expanded its role beyond user-driven web browsers, while increasing the number of applications that use HTTP for a variety of functions.

This is largely thanks to the request / response paradigm between a client and a server for which HTTP has defined. Such a paradigm is robust and flexible enough to allow most kinds of data transmission tasks besides the fundamental web browsing activities. Also, a HTTP client is handy to implement and can be easily embedded in any software to display advertisements or to send or retrieve data in a light-weighted way. Such dual property of HTTP has led to a rise of several phenomena, such as:

Non-Web Activities over HTTP. HTTP has accommodated remarkably more different kinds of activities than web-browsing, for example sending and receiving email, file downloading and sharing, instant messengers and multimedia streaming, as long as they can follow the request / response paradigm. Some of them were using a web interface, while many others blend the traffic into HTTP traffic in order to gain market advantages, data transmission priority or at least to be allowed by firewall rules.

Embedded HTTP Clients. Applications not related to web-based activities would also use HTTP for a secondary function, for example: advertiser sponsored software retrieving advertisements or an operating system automatically checking for updates.

Unlike works that characterise an individual new component in the HTTP traffic such as [1], we are interested to explore the increasing variety of application activities over HTTP, such as advertising, crawlers, file downloads, web-based applications and news feeds, and analyse the evolution of the usage of HTTP. From these analyses, we aim to (1) understand the growth of the HTTP traffic, (2) illustrate in depth how HTTP is being used on the current Internet and (3) describe the use of payload indicators that can reveal the actual purpose of an HTTP connection.

We consider this work as a contribution to the understanding of the Internet evolution through a measurement study. A conclusion in common with [1] is to do this work on a continual basis. A clear opportunity exists to spot the future trend in the traffic models, social network models and economic models of the next generation Internet.

SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.
ACM 978-1-60558-175-0/08/08.

2. METHODOLOGY

2.1 Data
To analyse the application activities within the HTTP "traffic-mix" as a whole, we collected a half-hour trace at around 11AM of a working day in late 2006 (referred to as Day3) from the edge of a research-institute network and compare it with a trace collected at a similar time from the same network in 2003 (called Day1). The research facility has about 1,000 users and is connected to the Internet via a full-duplex Gigabit Ethernet link. We monitored the bi-directional traffic on its link to the Internet.

We study the connections over HTTP, excluding connections which use an extended HTTP protocol specific for some P2P file sharing applications. The HTTP traffic is identified based on packet content and host knowledge in a similar way to [2], regardless of port numbers. The traces are listed in Table 1.

	Total web servers	# Types of client app(1)	HTTP connections	Total bytes in HTTP
Day3	2509	232	105,813	2,045,241,469
Day1	2137	113	50,002	728,782,686

Table 1 Number of connections and bytes in the traces. (1)- This is the number of different types of applications, estimated from the user-agent field and excluding different versions of the same application.

2.2 Classification
Within the traces, we discovered many different applications and purposes over HTTP. We try to categorise the traffic into a number of classes based on its purpose.

There are several key fields in HTTP headers that can reveal the client and the activity information of a connection, including the request method, host name, URL of the target object (which often includes the file name of the object), content type, and the user-agent [3]. An individual field may not be sufficient: for example, each different type of user-agents tells us of a different application, but there are many applications which tend to use default browser settings. However, through manual payload inspection, we are able to derive a number of heuristics to classify the majority of the traffic using a combination of the signatures in these fields. These signatures are manually derived from trace payload and online resources such as [6] and will have minimumfalse-positives but may contain some false-negatives for other categories than web-browsing. Table 2 describes the activity classes as well as the signatures we use to identify them.

Class	Activities	Signatures from
Web browsing	visiting web pages using a web browser	Method + Url + User Agent + Content Type
Web app	applications via web interface: e.g. java applets, web gadgets, and a variety of software	Url + User Agent
Crawler	bots crawling web pages	User Agent
File download	file downloading over HTTP	Url + Content Type + User Agent
Webmail	web-based e-mail service	Hostname
Advertising	advertisements on a webpage or embedded in software	Hostname + Content Type
Multimedia	streaming media or viewing media files on web pages	Url + User Agent
SW update	software update over HTTP (both automated and manual)	Content Type + User Agent
News feeds	RSS feeds	Url + Content Type
Link validator	automated link validators	User Agent
Calendar	calendar application based on web, e.g. ical, gcal	User Agent, Hostname
Attack	malicious traffic over HTTP	User Agent + Url
IM	MSN messenger	User Agent
Monitoring	network monitoring	User Agent

Table 2 Heuristics to identify different HTTP traffic classes.

3. RESULTS

We compared the total bytes in each HTTP traffic classes in the two traces, as shown in Table 3. While the whole HTTP traffic increased by 180%, Web browsing and Crawler both increased by about 108%. However, several classes have seen a sharp rise: Web apps, File download, Advertising, Web mail, Multimedia, News feeds and IM. Further, the Day3 HTTP traffic breakdown is shown in Figure 1, in which the Web browsing component only constitutes of no more than 53% of the HTTP traffic.

4. CONCLUSIONS AND FUTURE WORK

Based on analysis of the structural change of HTTP traffic from a research institute, we highlight a few characteristics of how the HTTP traffic has evolved. The study is unique: based on payload inspection, we analysed and classified the usage of HTTP in a much greater detail. In comparison, current standard traffic filters and classification approaches such as L7-filter [4] are limited to identifying HTTP, irrespective of anything based on port numbers.

Class	Day1 KB (%)	Day3 KB (%)	Day1:Day3
Web browsing	506,109 (71.11)	1,052,879(52.71)	1:2.08
Web app	22,063 (3.10)	317,204 (15.88)	1:14.38
Crawler	143,440 (20.15)	296,048 (14.82)	1:2.06
File download	24,817 (3.49)	232,878 (11.66)	1:9.38
Webmail	5,418 (0.76)	43,319 (2.17)	1:8.00
Advertising	3,804 (0.53)	28,962 (1.45)	1:7.61
Multimedia	1,949 (0.27)	10,627 (0.53)	1:5.45
SW update	3,583 (0.50)	8,140 (0.41)	1:2.27
News feeds	336 (0.05)	6,108 (0.31)	1:18.18
Link validator	157 (0.04)	439 (0.02)	1:2.79
Calendar	0 (0)	324 (0.02)	1:∞
Attack	0 (0)	232 (0.01)	1:∞
IM	1 (0.0001)	211 (0.01)	1:222.82
Monitoring	24 (0.003)	26 (0.001)	1:0.92

Table 3 Traffic size comparison between Day1 and Day3. The value in brackets is the proportion of bytes in that trace.

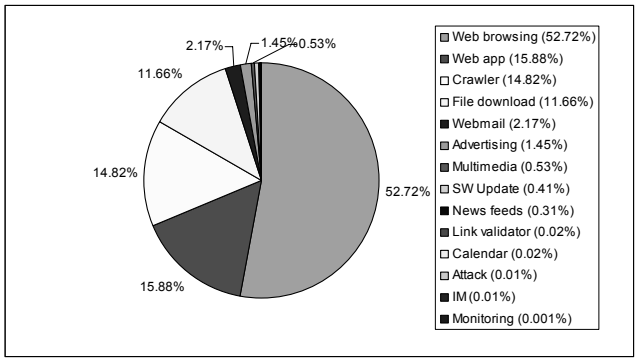

Figure 1 Day3 HTTP Traffic Breakdown

Although intrusion detection systems such as Bro [5] are extensible and have such capability, we did not find any documented work with the similar purpose.

Finally, our work is ongoing and we are working with more traces from different sites and keep on improving our classification database.

5. REFERENCES

[1] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann. The New Web: Characterizing AJAX Traffic. In *Proceedings of Passive and Active Measurement Conference 2008 (PAM 2008)*, April 2008, Cleveland, OH

[2] A. Moore, K. Papagiannaki. Toward the Accurate Identification of Network Applications. In *Proceedings of Sixth Passive and Active Measurement Workshop (PAM 2005)*, March/April 2005, Boston, MA

[3] RFC 2616 - Hypertext Transfer Protocol -- HTTP/1.1

[4] L7-filter. http://l7.sourceforge.net

[5] V. Paxson: Bro intrusion detection system (2007). http://www.bro-ids.org.

[6] http://www.user-agents.org/