Obtaining Internet Flow Statistics by Volunteer-Based System

Jens Myrup Pedersen[1] and Tomasz Bujlow[2]

[1] Department of Electronic Systems, Aalborg University, Denmark
jens@es.aau.dk
[2] Department of Electronic Systems, Aalborg University, Denmark
tbu@es.aau.dk

Summary. In this paper, we demonstrate how the Volunteer Based System for Research on the Internet, developed at Aalborg University, can be used for creating statistics of Internet usage. Since the data are collected on individual machines, the statistics can be made on the basis of both individual users and groups of users, and as such be useful also for segmentation of the users into groups. We present results with data collected from real users over several months; in particular we demonstrate how the system can be used for studying flow characteristics - the number of TCP and UDP flows, average flow lengths, and average flow durations. The paper is concluded with a discussion on what further statistics can be made, and the further development of the system.

1 Introduction

Understanding the behavior of Internet traffic is crucial in order to model traffic correctly, and to create realistic scenarios of future Internet usage. In particular, understanding the behavior of different kinds of traffic makes it possible to create scenarios of increasing/decreasing particular amounts of traffic. The models can then be used for the analysis and/or simulations of distribution and backbone networks under different scenarios. The application of different provisioning and traffic engineering techniques can be tested as well.

Traffic statistics are today often made by Internet Service Providers (ISPs), who monitor the activity in their networks. However, often ISPs consider these data to be private and are not keen on sharing with researchers. Some traces are publicly available, such as the Caida data sets[1]. Even with access to traces from ISPs or other, traffic monitored in the network cores suffers from missing important statistics that can only be known accurately at the sources - such as inter-arrival times between packets and flow durations. It should be noted that the literature covers a number of interesting studies, where researchers have gained access to real data [2, 3]. In the latter, the data are

collected at the broadband access router, which is quite close to the generating source.

There are also a large number of commercial tools for monitoring traffic in Local Area Networks (e.g. on Internet gateways), such as Cisco Netflow [4]. These can provide useful insights to traffic, but without collecting traffic from many different networks it does not give a good overview of how the Internet traffic looks like.

The open-source Volunteer Based System (VBS) for Research on the Internet, developed at Aalborg University, seeks to avoid these problems by collecting the traffic from a large number of volunteers, which are agreeing to have their Internet usage monitored and statistics collected for research purposes. This provides statistics from the point where the traffic is generated, meaning that quite precise statistical data can be obtained. Moreover, it is also possible to monitor which applications are opening the sockets, and thus get the precise picture of the behavior of different applications. The general idea was first described in [5] and a preliminary limited prototype was implemented in [6]. The current system design was announced in [7], while more technical details on later refinements can be found in [8]. Other papers ([9, 10, 11]) demonstrate various applications of our system.

In this paper, we show how the system can be used for generating statistics at the flow level. The paper is organized as follows: First, in Section 2, we describe how the data collection is made and how the statistics are extracted. In Section 3, we present the results, and in Section 4, we conclude the paper and discuss the further work.

The authors would like to stress that the system is based on open source software, published on SourceForge [12]. We would like to take the opportunity to encourage other researchers to use the system for collecting Internet traffic information, and to the widest possible extend share the data traces with the research community.

2 Data Collection and Extracting Statistics

In this section, we briefly describe the fundamentals of VBS, with a particular focus on the parts that influence the monitoring, data collection, and extraction of statistics. For more details, please refer to our previous paper [8].

For each volunteer, the system monitors both ingoing and outgoing traffic on his/her computer. Storing all these data, and transferring them back to the server, would be a huge task, if not impossible given the limited upload capacity available on many standard Internet connections. Therefore, the data are saved as follows, and transmitted to our central server as:

- For each flow, information is stored about e.g. source and destination IP addresses and port numbers, flow start time, flow end time, number of

packets, protocol (TCP or UDP) as well as the flow ID. Moreover, the information about the process name, which has opened the socket is collected. This feature provides valuable information allowing us to characterize the traffic created by different applications.
- For each packet, the main information is the packet size and the relative time stamp. Moreover flags from the header are stored, as well as the information about the packet direction and flow ID.

In this way, all relevant information is stored, while the requirements in the terms of memory and network usage are kept at a minimum. Also, no payload is stored at any time, which is an advantage with respect to privacy and security. One privacy concern has been the transfer of source and destination IP addresses. In the current implementation, the IP addresses are hashed before being transferred to the server. However, since the hash function is known (open source), and since the number of IP addresses in IPv4 is limited, it is not difficult to determine the original IP address.

The purpose of this particular study was to demonstrate the usage, so focus was on obtaining and presenting data from a limited number of users prior to run more large-scale experiments. The statistics were obtained from 4 users during the period from January to May 2012. One of the four users (User 4) did not join the system until late April, and thus only participated for the last 2-3 weeks of the study. Due to being a heavy user, the amount of data collected from this machine is higher than from any of the other participants, despite the shorter participation. During the time of study, all the traffic from the users were collected by the system as described above, and the data stored into our central database.

The four users can be described as follows:

- User 1 - Private user in Denmark
- User 2 - Private user in Poland
- User 3 - Private user in Poland
- User 4 - Private user in Denmark

With the data collected, a wide variety of studies can be conducted. For this paper, we chose to analyze only the flow data (not the packet data), since the amount of data makes it more manageable. As the main purpose is to demonstrate the usefulness of the system, we chose to derive the following statistics:

- Amount of TCP and UDP flows
- Average flow lengths for TCP and UDP flows
- Average flow durations for TCP and UDP flows
- Top 5 applications (measured on the number of flows)

The statistics are done for the individual users as well as for the users altogether.

3 Results

3.1 TCP and UDP Flows

The distributions of TCP and UDP flows are shown in Table 1. It can be seen that both the number of flows and the distribution between TCP and UDP vary quite a bit between the different users.

Table 1. The numbers of TCP and UDP flows for all users as well as for the individual users. The number in parenthesis shows the distribution.

User	#UDP flows	#TCP flows
All	4770315 (55%)	386530 (45%)
1	446692 (35%)	820927 (65%)
2	3142581 (60%)	2084590 (40%)
3	693389 (52%)	642740 (48%)
4	487653 (60%)	315273 (40%)

3.2 Flow Lengths and Durations

The distributions of flow lengths (the number of packets per flow) for TCP and UDP for the different users are shown in Table 2. It is quite interesting to observe that the flow lengths for both TCP and UDP are so different between the different users, indicating a different Internet usage. It should be noted that with the data in the system, it is possible to make a more detailed analysis of the distribution of flow lengths, not only for the different users but also for each application used by each user.

Table 2. Average flow lengths for TCP and UDP flows.

User	Avg. UDP length	Avg. TCP length
All	72	81
1	5	110
2	90	80
3	34	29
4	72	114

The distribution of flow durations (in seconds) is shown in Table 3. It seems that for users 1-3 the users who have longer average flows also have longer average flow durations. However, user 4 seems to have quite short flow durations even though the flows are quite long. Even though a more thorough

analysis is required to explain this in detail, we assume it is due to the user being on a fast Internet connection. However, the type of traffic with generally longer flows probably also plays a role.

Table 3. Average flow durations for TCP and UDP flows.

User	Avg. UDP duration	Avg. TCP duration
All	33	26
1	1	32
2	41	25
3	28	7
4	19	55

3.3 Top 5 Applications

Analyzing the applications is more challenging than deriving the other parameters. First, we did not manage to collect the socket names for a substantial number of the flows. This is mainly concerning very short flows, where the opening time of the socket is so short that it is not captured by the socket monitor. Secondly, what we obtain in order to identify an application is really the process name. For this study, 240 different process names were identified. Further work is needed in order to group these into applications, and for this study, we just list the top 5 process names. It should be noted that it is not a trivial task to determine how e.g. browser plugins should be grouped and categorized. The top application names for the different users are shown in Tables 4–8.

Based on the information obtained by the system, it is possible to make additional statistics, taking e.g. the flow lengths of different applications into account. Also packet statistics (e.g. packet lengths) can be taken into account, providing a quite precise picture of what applications are taking up most bandwidth for the different users.

Table 4. Top 5 applications for all users.

Application name	Number of flows	% of all flows
uTorrent	6399336	74.12
Unknown	948497	10.99
Chrome	441953	5.12
Firefox	361213	4.18
Svchost	103757	1.2

Table 5. Top 5 applications for user 1.

Application name	Number of flows	% of all flows
Unknown	729868	57.56
Firefox	330498	26.07
Chrome	138105	10.89
Amule	18863	1.49
Ntpd	17929	1.41

Table 6. Top 5 applications for user 2.

Application name	Number of flows	% of all flows
uTorrent	4674545	89.43
Chrome	227616	4.35
Unknown	136387	2.61
Svchost	101633	1.94
Java	38704	0.74

Table 7. Top 5 applications for user 3.

Application name	Number of flows	% of all flows
uTorrent	1220728	91.36
Chrome	76062	5.69
Unknown	21764	1.63
SoftonicDownloader	15035	1.13
Java	2169	0.16

Table 8. Top 5 applications for user 4.

Application name	Number of flows	% of all flows
uTorrent	504063	62.78
Moc	90358	11.25
Iexplore	64407	8.02
Unknown	60478	7.53
Firefox	26896	3.35

Without going into a more detailed data analysis, we did an observation regarding the unknown flows, which is worth highlighting. The *unknown* flows account for a large number of the total flows. However, the flows have the average length of 2 seconds and the average of 11 packets, indicating that it is not such a large share of the total traffic. These *unknown* flows are almost equally shared between TCP(53%) and UDP(47%).

3.4 Cumulative Number of Flows

The distribution of the cumulative number of flows for the 4 users during the time of our experiment is shown in Figure 1.

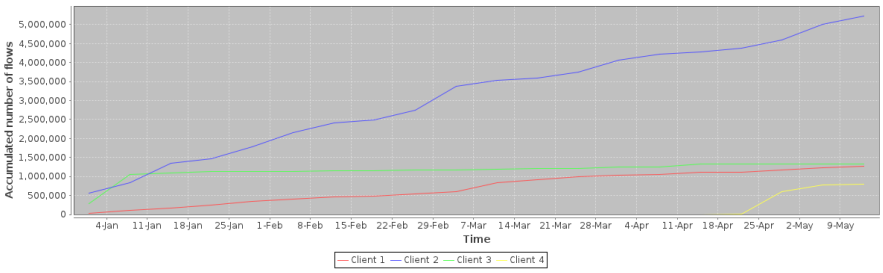

Fig. 1. Cumulative number of flows for all users.

4 Conclusion and Discussion

In this paper, we have demonstrated how the Volunteer Based System for Research on the Internet developed at Aalborg University can be used for creating statistics of Internet traffic, specifically within the studies of flows and their properties.

Future research will focus on developing efficient methods for extracting relevant information from the packet statistics. This can provide even more valuable information about the flows, for example, on average packet sizes of different flows (and the distribution of packet sizes), inter-arrival times between packets, and the number of successful vs. unsuccessful connections for different kinds of traffic. Moreover, particularly interesting statistics can be derived from the combined flow and packet statistics, such as the average size of flows of different kinds of traffic, and eventually how much traffic is created by different applications for individual users. The challenge is that it is large amounts of data, so efficient ways of handling these has to be developed.

Another important part is the recruitment of more volunteers, in order to collect larger amounts of data. Also, having appropriate background information about the users could be useful. This includes both the data about the users themselves, such as age, occupation, if the computer is shared etc., but also information about the connection, e.g. speeds and technologies.

In order to obtain more data, other researchers are invited to join the project and use it for the collection of data for scientific purposes. The code is available as open-source, and can be found together with a comprehensive documentation on our project homepage [12] located on SourceForge.

References

1. CAIDA Internet Data – Passive Data Sources, 2012. [Online]. Available: `http://www.caida.org/data/passive/`.
2. Chuck Fraleigh, Sue Moon, Bryan Lyles, Chase Cotton, Mujahid Khan, Deb Moll, Rob Rockell, Ted Seely, and S. Christophe Diot. Packet-Level Traffic Measurements from the Sprint IP Backbone. *IEEE Network*, 17(6):6–16, November 2003.
3. Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 90–102. ACM New York, Barcelona, Spain, August 2009.
4. Cisco IOS NetFlow, 2012. [Online]. Available: `http://www.cisco.com/go/netflow`.
5. Kartheepan Balachandran, Jacob Honoré Broberg, Kasper Revsbech, and Jens Myrup Pedersen. Volunteer-Based Distributed Traffic Data Collection System. In *Proceedings of the 12th International Conference on Advanced Communication Technology (ICACT 2010)*, volume 2, pages 1147–1152. IEEE, Phoenix Park, PyeongChang, Korea, February 2010.
6. Kartheepan Balachandran and Jacob Honoré Broberg. Volunteer-Based Distributed Traffic Data Collection System. Master's thesis, Aalborg University, Department of Electronic Systems, Denmark, June 2010.
7. Tomasz Bujlow, Kartheepan Balachandran, Tahir Riaz, and Jens Myrup Pedersen. Volunteer-Based System for classification of traffic in computer networks. In *Proceedings of the 19th Telecommunications Forum TELFOR 2011*, pages 210–213. IEEE, Belgrade, Serbia, November 2011.
8. Tomasz Bujlow, Kartheepan Balachandran, Sara Ligaard Nørgaard Hald, Tahir Riaz, and Jens Myrup Pedersen. Volunteer-Based System for research on the Internet traffic. *TELFOR Journal*, 4(1):2–7, September 2012. Accessible: `http://journal.telfor.rs/Published/Vol4No1/Vol4No1.aspx`.
9. Tomasz Bujlow, Tahir Riaz, and Jens Myrup Pedersen. A method for Assessing Quality of Service in Broadband Networks. In *Proceedings of the 14th International Conference on Advanced Communication Technology (ICACT)*, pages 826–831. IEEE, Phoenix Park, PyeongChang, Korea, February 2012. Accessible: `http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6174795`.
10. Tomasz Bujlow, Tahir Riaz, and Jens Myrup Pedersen. A method for classification of network traffic based on C5.0 Machine Learning Algorithm. In *Proceedings of ICNC'12: 2012 International Conference on Computing, Networking and Communications (ICNC): Workshop on Computing, Networking and Communications*, pages 244–248. IEEE, Maui, Hawaii, USA, February 2012.
11. Tomasz Bujlow, Tahir Riaz, and Jens Myrup Pedersen. Classification of HTTP traffic based on C5.0 Machine Learning Algorithm. In *Proceedings of the Fourth IEEE International Workshop on Performance Evaluation of Communications in Distributed Systems and Web-based Service Architectures (PEDISWESA 2012)*, pages 882–887. IEEE, Cappadocia, Turkey, July 2012.
12. Volunteer-Based System for Research on the Internet, 2012. [Online]. Available: `http://vbsi.sourceforge.net/`.