Analysis of Elephant Users in Broadband Network Traffic

Péter Megyesi, Sándor Molnár

Analysis of Elephant Users in Broadband Network Traffic

Péter Megyesi and Sándor Molnár

High Speed Networks Laboratory,
Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
H-1117, Magyar tudósok körútja 2.,
Budapest, Hungary
{megyesi,molnar}@tmit.bme.hu

Abstract. *Elephant and mice* phenomena of network traffic flows have been an interesting research area in the past decade. Several operational broadband measurement results showed that the majority of the traffic is caused by a small percentage of large flows, called the *elephants*. In this paper, we investigate the same phenomenon in regards of users. Our results show that even though the packet level statistics of *elephant users* and *elephant flows* show similar characteristics, there is only a small overlap between the two phenomena.

Keywords: traffic measurement, traffic analysis, elephant flows, elephant users

1 Introduction

Traffic profiling is a crucial objective for network monitoring and management purposes. Flow characterization has been given a large attention by the research community in the past decade. Flows has been classified by their size of traffic (as *elephant and mice*) [1] [2] [3], by their duration in time (as *tortoise and dragonfly*) [4], by their rate (as *cheetah and snail*) [5] and by their burstiness (as *porcupine and stingray*) [5]. Several studies were written about the correlation between these flow behaviors [6] [7].

However, current literature lacks in profiling users in such regards. In this paper we investigate the *elephant and mice* phenomena regarding Internet users. We analyzed two recent measurements taken from high speed operational networks and found that *elephant users* show similar packet level characteristics to *elephant flows*. We also determined that there is a much smaller overlap between these two phenomena that one would expect. We found that only a small portion (10%-30%) of *elephant flows* are generated by *elephant users* and also the generation of *elephant flows* is not a necessary condition for being an *elephant user*. These results indicate that further investigation of user characterization could aid network operators in the future to apply different services or charging policies for different users.

The contribution of this paper is threefold. First, to our knowledge, our study is the first that presents the discussed characteristics of *elephant users*. Second, we point out that there is only a small overlap between *elephant flows* and *elephant users*. Finally, our measurements from recent networks show that the *elephant and mice* phenomena of flows and users are still present in todays networks.

This paper is organized as follows. Section 2 presents the related work. In Section 3 we give the definition of *elephants* and the properties of the two datasets we used for our research. Section 4 presents the results of our measurements. Finally, in Section 5 we conclude our work.

2 Related Work

There are several different definitions for *elephant flows* in the literature. In [2] authors propose two techniques to identify *elephants*. The first approach is based on the heavy-tail nature of the flow bandwidth distribution, and one can consider a flow as an *elephant* if it is located in this tail. The second approach is more simple, *elephants* are the smallest set of flows whose total traffic exceeds a given threshold. Estan and Varghese [3] used a different definition. They considered a flow as an *elephant* if its rate exceeds the 1% of the link utilization.

However, the definition given by Lan and Heidemann [5] become a rule of thumb in later literature (e.g. both [7] and [8] use this definition). They define *elephant flows* as flows with a size larger than the average plus three times the standard deviation of all flows. They use the same idea for categorize flows by their duration, rate and burstiness as *tortoise, cheetah* and *porcupine*, respectively. [5] was also the first study that presented the *cheetah and snail* and the *porcupine and stingray* classifications. *Tortoise and dragonfly* properties of traffic flows were first investigated in [4]. Here, the authors considered a flow as *tortoise* simply if its duration was lager than 15 minutes. Given the generality and the rule of thumb nature of the definition by Lan and Heidemann [5] we will use the same definition for *elephants* later in this paper.

In [9] Sarvotham et al. present a comprehensive study that traffic bursts are usually caused by only few number of high bandwidth connections. They separate the aggregated traffic into two components, *alpha* and *beta* by their rate in every 500ms time window. If the rate of the flow is greater than a given threshold (mean plus three standard deviations) than the traffic is *alpha*, otherwise it is *beta*. Authors determine that while the *alpha* component is responsible for the traffic bursts, the *beta* component has similar second order characteristics to the original aggregate.

The term *elephant user* appears in [10] where the authors calculate the Gini coefficient for the user distribution. The Gini coefficient is usually used in economics for measuring statistical dispersion of a distribution. They calculate the value of the Gini coefficient for the distribution of the number of bytes generated by the users as 0.7895 but no further discussion is presented.

In [11] authors investigate application penetration in residential broadband traffic. They calculate the results separately for the top 10 *heavy-hitters* (the top 10 users that generated the most traffic) in their measurement data. Besides pointing out the fact that the majority of the data is generated by a small group of users the paper does not tackle any further issues about *elephant users*.

3 Methodology

In this section we present the source of the two network traces we used in this study. We also give the definition of *elephants* and the metrics we used to analyze them.

3.1 Datasets

The first trace was measured by the The Cooperative Association for Internet Data Analysis (CAIDA) [13] in a 10 Gbit/s backbone link between Chicago and San Jose. They periodically take measurements on this link and make them available for the research community upon request in an anonymous format (removed payload and hashed IP addresses). We analyzed multiple subsets of these data and since we found similar result we chose one given time period to present our findings. This trace was recorded on 13:15 (UTC), 20th of December 2012 and contains four minutes of network traffic. Furthermore, we refer to this measurement as *CAIDA Trace*.

The second measurement was taken in the campus network of the Budapest University of Technology and Economics (BME) on 16:31 (CET), 18th of December 2012 and contains six minutes of traffic. The measured link was a 10Gigabit Ethernet port of a Cisco 6500 Layer-3 switch which transfers the traffic of two buildings on the campus site to the core layer of the university network. This measurement is not available to the public. However, we consider the results relevant to present since our findings are similar to the *CAIDA Trace* even though the nature of network is different. We refer to this measurement as *BME Trace*.

Table 1 presents the basic statistics of the two traces. Generally, the *CAIDA Trace* contains more data than the *BME Trace* by one order of magnitude.

3.2 Measuring Elephants

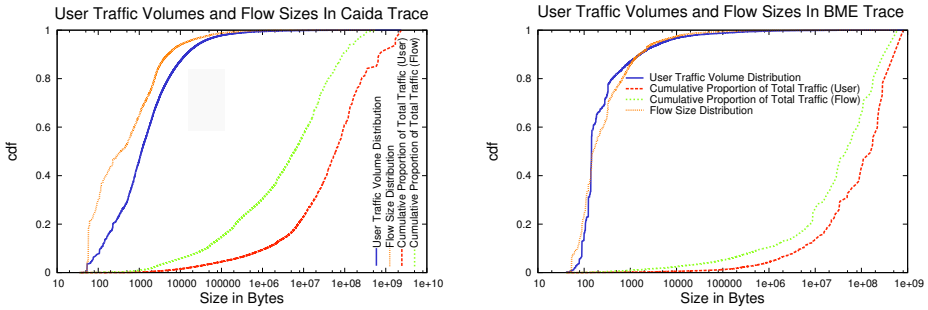
During the identification of *elephant users and flows* we use the definition presented in [5]: a user or a flow is considered an *elephant* if its flow size or traffic volume is grater than the average plus three times the standard deviation of all the flow sizes or traffic volumes of flows and users, respectively. Table 1 presents the values of these threshold for the two traces. The *elephant and mice* phenomena clearly exist: less than a thousandth of the users and flows are responsible for roughly 60%-80% of the total traffic.

In the next section we firstly show that the *elephant* phenomenon also exist with different threshold levels by plotting the cumulative distribution of user and

Table 1: Statistics of the two traces used for analysis

	CAIDA Trace	BME Trace
Number of packets	105444780	6804958
Number of users	680300	63668
Number of flows	3876982	264117
Total traffic	65.6 Gbyte	5.66 Gbyte
Elephant user threshold	15.9 Mbyte	13.7 Mbyte
Number of elephant users	661	56
Proportion of elephant users	0.097%	0.088%
Total traffic of elephant users	71.5%	84.5%
Elephant flow threshold	2.3 Mbyte	4.96 Mbyte
Number of elephant flows	2714	151
Proportion of elephant flows	0.07%	0.057%
Total traffic of elephant flows	61.7%	83.41%

flow sizes against theirs cumulative proportion of the total traffic. Furthermore, we present the comparison of the following three packet level metrics, (1) byte and packet throughput, (2) packet size distribution and (3) inter packet time distribution. We chose these metrics because they are the most frequently used packet level characteristics for comparing traffic traces [12]. Additionally, we investigate presence of both *elephant* and *non-elephant flows* in the traffic of *elephant users*.



A) CAIDA Trace B) BME Trace

Fig. 1: The *elephant and mice* phenomena presented by cumulative distribution of user traffic volume and flow sizes and their cumulative proportion of the total traffic

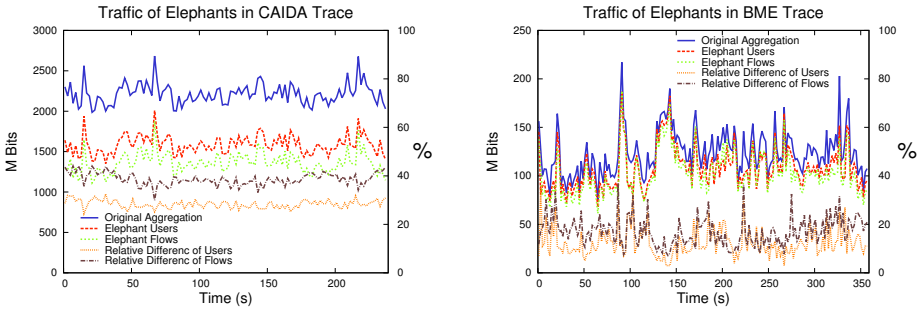
Table 2: Proportion of *elephants* with different thresholds

Threshold in Mbyte	CAIDA Trace				BME Trace			
	Users		Flows		Users		Flows	
	Ratio	Traffic	Ratio	Traffic	Ratio	Traffic	Ratio	Traffic
0.1	2.17%	95.48%	1.21%	85.29%	1.21%	98.75%	0.53%	94.84%
0.5	0.8%	92.44%	0.29%	74.21%	0.53%	96.97%	0.16%	91.08%
1	0.54%	90.55%	0.18%	69.56%	0.37%	95.66%	0.11%	89.46%
2	0.37%	88.17%	0.11%	63.71%	0.23%	93.45%	0.08%	87.04%
5	0.23%	83.48%	0.04%	52.03%	0.14%	89.96%	0.06%	83.23%
10	0.13%	76.92%	0.02%	42.87%	0.1%	86.35%	0.03%	73.56%
15	0.1%	72.2%	0.017%	37.25%	0.08%	83.96%	0.025%	70.65%
20	0.08%	68.78%	0.011%	32.09%	0.06%	80.91%	0.02%	66.67%
50	0.03%	53.55%	0.003%	17.88%	0.03%	68.12%	0.007%	47.93%

4 Measurement Results

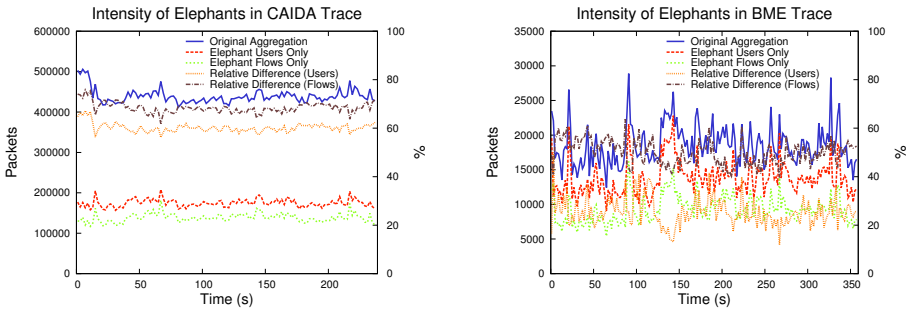
4.1 User and Flow Sizes

In Figure 1 one can investigate the *elephant and mice* phenomena for both traces. Here we plotted the cumulative distribution of user traffic volumes and flow sizes against theirs cumulative proportion of the total traffic. In Table 2 we collected the complementary values in percentage (1 minus the actual value) of the curves in Figure 1 for different thresholds. *Ratio* presents the proportion of users and flows whose traffic was larger than the *Threshold* value and *Traffic* represents their total share from the aggregated traffic.



A) CAIDA Trace B) BME Trace

Fig. 2: Traffic of *elephants*



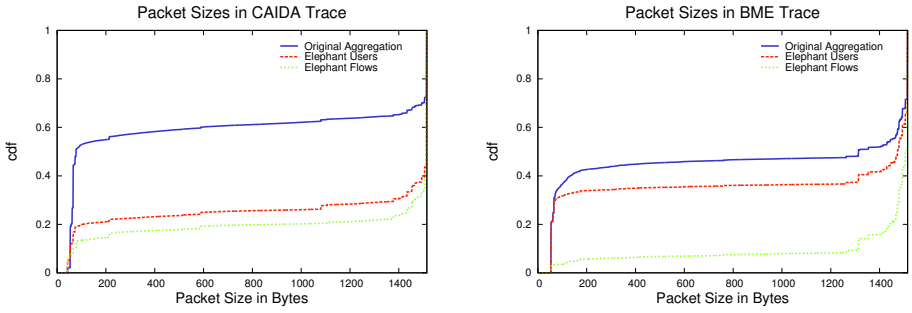
A) CAIDA Trace B) BME Trace

Fig. 3: Intensity of *elephants*

4.2 Byte and Packet Throughput

In Figure 2 the traffic of *elephant users* and *elephant flows* are plotted against the original traffic. The relative difference are also presented. In case of the *BME Trace* the *elephants* are responsible for sufficient amount of the total traffic (80%-85%), while in the *CAIDA Trace* this ration is a bit smaller (60%-70%). Since the traffic of *elephants* seems to follow the bursts in the original traffic (the relative differences are also smaller at these peaks), the results suggests that *elephant users* are main cause for traffic burstiness.

Figure 3 present the number of packets in every one second time interval. Here, the relative difference is much higher than in case of the byte throughput. In the *CAIDA Trace elephants* are responsible for only roughly 30%-40% of the total packets, while in case of the *BME Trace* this number is ratio is 50%-70%. Intensity of *elephants* are also following the packet burst of the original aggreagte since the relative difference is smaller in traffic peaks.



A) CAIDA Trace B) BME Trace

Fig. 4: Packet size distributions of *elephants*

Table 3: Packet proportions under different conditions

Condition	CAIDA Trace			BME Trace		
	Original Aggregate	Elephant Users	Elephant Flows	Original Aggregate	Elephant Users	Elephant Flows
PS <= 54 Byte	18.9%	11.9%	7.4%	20.9%	21.0%	2.9%
PS <= 66 Byte	44.3%	16.7%	9.9%	30.1%	29.4%	3.3%
PS >= 1450 Byte	32.8%	66.3%	72.8%	44.8%	54.6%	79.1%
PS = 1514 Byte	27.6%	54.2%	61.4%	28.5%	33.9%	49.7%

4.3 Packet Sizes

Packet size distributions of the two measurements is given in Figure 4. The joint property in both traces is that ratio of maximum and minimum sized packets is larger in *elephants* than in the original aggregate. Packets with intermediate size share similar proportion. We collected a few numerical example to Table 3 to present this phenomenon.

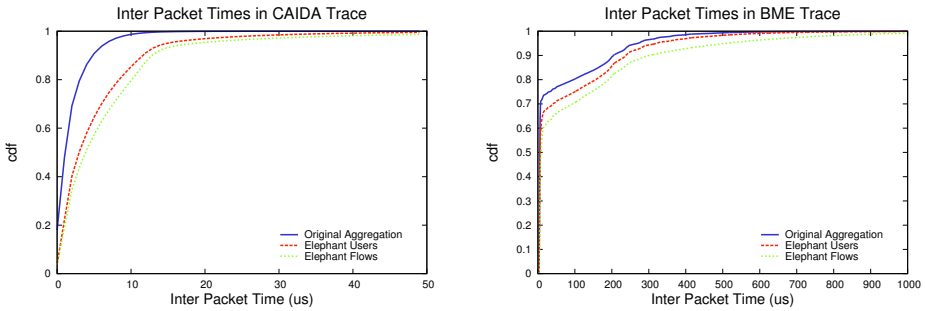
Table 4 present the ratio of number of packet in *elephants* compared to the number in the original aggregate under different conditions. It is clear from the values that *elephants* contains the majority of maximum sized packet and *elephant flows* exclude the majority of minimum sized packets. The ratio of minimum sized packets in *elephant users* shows different behavoir in the two measurements.

4.4 Inter Packet Times

Inter arrival time between consecutive packets corresponding the *elephant users or flows* are presented in Figure 5. The curves show similar characteristics for *elephant users* and *elephant flows*. The *cdf* curves of *elephants* are increasing slower than the original aggregate's which is an expected behavior since traffic of *elephants* are the rarefaction of the original packet stream.

Table 4: Ratio of number of packets in *elephants* compared to the original

Condition	CAIDA Trace		BME Trace	
	Elephant Users	Elephant Flows	Elephant Users	Elephant Flows
PS <= 54 Byte	25.0%	12.3%	74.7%	6.8%
PS <= 66 Byte	15.0%	7.0%	71.0%	5.3%
PS >= 1450 Byte	80.6%	70.2%	90.2%	88.6%
PS = 1514 Byte	81.2%	70.3%	88.4%	87.8%



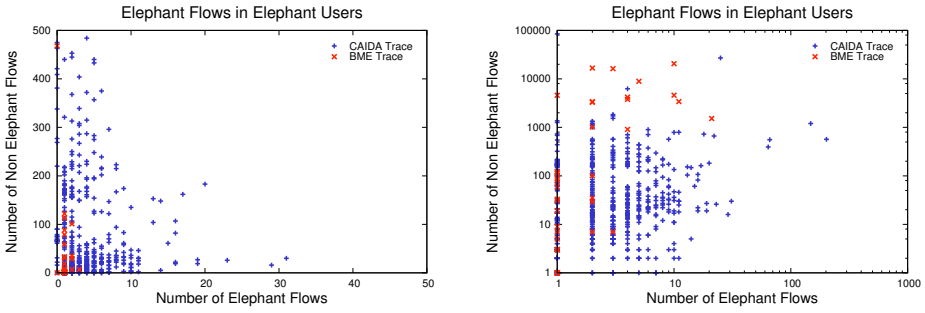
A) CAIDA Trace B) BME Trace

Fig. 5: Inter Packet Time distributions of *elephants*

4.5 Elephant and Non-Elephant Flows in Elephant Users

In Figure 6 every dot represents an *elephant user* according to the generated number of *elephant flows* and *mice flows*. These results indicate that there is no correlation between the number of *elephant flows* and *mice flows* generated by an *elephant users*. Furthermore, a user can be an *elephant* without generating any *elephant flows*. There was 53 *elephant users* in the *CAIDA Trace* who did not generated any *elephant flow*. They account for 8% of all *elephant users* in the *CAIDA Trace*. In the *BME Trace* this number is only 3, but since there were only 56 *elephant users* in that measurement their share is 5%.

Another interesting result is that in case of the *CAIDA Trace* only the 9.13% of *elephant flows* were generated by *elephant users*. In case of the *BME Trace* this value is higher, namely 37.85%. These result clearly indicate that the overlap between the *elephant user* and *elephant flow* phenomenons could be much smaller in some cases that one would expect.



A) Linear scale B) Logarithmic scale

Fig. 6: The number of *elephant* and *non-elephant flows* generated by *elephant users*

5 Conclusion

In this paper we investigated the *elephant and mice* phenomena of Internet users in recent broadband network measurements. We found that *elephant users* show similar packet-level characteristics to the well-investigated *elephant flow* phenomenon. However, we pointed out that only a small portion (10%-30%) of *elephant flows* were generated by *elephant users*. We also found that the generation of *elephant flows* by a user is not a necessary condition for being an *elephant user*.

As future work we would like to further analyze the *elephant user* phenomenon in the same way that *elephant flows* were analyzed in [5] [9]. Such study would aid us in the understanding of how particular users are affecting the second-order characteristics of network traffic.

Acknowledgments. The authors would like to thank CAIDA for granting access to the measurement data that was analyzed in this paper. The research was supported by OTKA-KTIA grant CNK77802.

References

1. Thompson, K., Miller, G., Wilder, R.: Wide Area Internet Traffic Patterns and Characteristics. IEEE Network Magazine, Vol. 11, No. 6, pp. 10-23 (1997)
2. Papagiannaki, K., Taft, N., Bhattacharyya, S., Thiran, P., Salamatian, K., Diot, C.: A pragmatic definition of elephants in internet backbone traffic. In Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement (IMW '02), pp. 175–176. Marseille, France (2002)
3. Estan, C., Varghese, G.: New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice. In ACM Transactions on Computer Systems, Vol. 21, No. 3, pp. 270–313. ACM, New York, NY, USA (2003)
4. Brownlee, N., Claffy, K.: Understanding Internet traffic streams: dragonflies and tortoises. IEEE Communications Magazine, Vol. 40, No. 10, pp. 110–117 (2002)
5. Lan, K., Heidemann, J.: A measurement study of correlations of Internet flow characteristics. In Computer Networks, Volume 50, Issue 1, pp. 46–62 (2006)
6. Markovich, N., Kilpi, J.: Bivariate statistical analysis of TCP-flow sizes anddurations. In Annals of Operations Research, Vol. 170, No. 1, pp. 199–216. Springer US (2009)
7. Molnár, S., Móczar, Z.: Three-Dimensional Characterization of Internet Flows. In Proceedings of IEEE International Conference on Communications (ICC '11). Kyoto, Japan (2011)
8. Callado, A., Kamienski, C., Szabo, G., Gero, B., Kelner, J., Fernandes, S., Sadok, D.: A Survey on Internet Traffic Identification. IEEE Communications Surveys & Tutorials, Vol. 11, No. 3, pp. 37-52, IEEE (2009)
9. Sarvotham, S., Riedi, R., Baraniuk, R.: Connection-level analysis and modeling of network traffic. In Proceedings of the 1st ACM SIGCOMM Internet Measurement Workshop (IMW '01), pp. 99–103. San Francisco Bay Area, CA, USA (2001)
10. Liu, P., Liu, F., Lei, Z.: Model of Network Traffic Based on Network Applications and Network Users. In International Symposium on Computer Science and Computational Technology (ISCSCT '08), pp. 171–174. Shanghai, China (2008)

11. Pietrzyk, M., Plissonneau, L., Urvoy-Keller, G., En-Najjary, T.: On profiling residential customers. In Proceedings of the Third International Workshop on Traffic Monitoring and Analysis (TMA '11), pp. 1–14. Vienna, Austria (2011)
12. Molnár, S., Megyesi, P., Szabó, G.: How to Validate Traffic Generators? In Proceedings of the 1st IEEE Workshop on Traffic Identification and Classification for Advanced Network Services and Scenarios (TRICANS 2013). Budapest, Hungary (2013)
13. The Cooperative Association for Internet Data Analysis, `http://www.caida.org`