Three-dimensional Characterization of Internet Flows

Sándor Molnár, Zoltán Móczár
High Speed Networks Laboratory, Dept. of Telecommunications and Media Informatics
Budapest Univ. of Technology and Economics, H–1117, Magyar tudósok krt. 2., Budapest, Hungary
E-mail: molnar@tmit.bme.hu, moczar.zoltan@gmail.com

***Abstract*—In this paper a three-dimensional (size, duration and rate) flow traffic characterization framework is proposed. Detailed investigation of application clusters in the characterization space is discussed and examples from specific applications (P2P, gaming, social networking, video playback) are presented. The framework provides an in-depth understanding of Internet traffic as demonstrated on measured traffic from a commercial network.**

I. Introduction

It is vital to understand the nature of Internet traffic flows and to characterize them for proper traffic monitoring, modeling and management purposes.

There are a number of prior studies that characterized the flows using different classification schemes [1], [2], [3], [4], [5], [6], [7]. One of the aspects, which is frequently investigated is the *size* of traffic flows [1], [2], [4], [7], [8]. These studies revealed a number of important properties of Internet traffic, for example, one of the well-known findings is that only a small percentage of flows carry the majority of the bytes. This research yielded to the elephant–mice view of the flow size characterization. Another aspect of the investigations is the *duration* of flows and related studies resulted in a tortoise–dragonfly view of the flow duration characterization [9]. The important finding is that most flows (about 45%) are short (dragonflies), lasting less than 2 seconds, and only a small number of flows are long with duration of hours to days, but they carry a high proportion (50–60%) of the total bytes. Beyond these characteristics other aspects of traffic flows like *rate*, *burstiness*, etc. are also studied. For example, burstiness is intensively investigated in [10]. Several authors used extreme value theory to study the relation between flow rate and duration [5], [6]. Some of them successfully obtained regression equations among flow characteristics providing further insights on the dependencies of these characteristics [7].

The previous studies provided excellent descriptions about Internet traffic, but most of them focused only on one aspect of the flow characteristics or on a special region. However, the correlations and dependencies among the different characteristics in a general view are crucial and can yield to a better understanding. For example, Zhang et al. [11] showed that for large flows the size and rate are highly correlated. Brownlee et al. [9] also pointed out that flows can be classified not only by their sizes (elephants and mice), but also by their lifetimes (tortoises and dragonflies). Lan et al. [12] investigated the size, rate, duration and burstiness and provided a correlation study among these parameters.

In this paper we consider three main flow characteristics (size, duration and rate) and propose a *three-dimensional flow characterization*. In addition to this characterization we investigate the main categories (elephant–mice, tortoise–dragonfly and cheetah–snail) of each dimension and the relationships among these categories. We advocate that these three characteristics together yield to a better understanding of Internet traffic than investigating them separately and show that we can reveal important characteristics about applications by identifying *application clusters* in this three-dimensional characterization space. In order to justify these statements we carried out a comprehensive and detailed traffic analysis study on actual measured Internet traffic taken from a commercial network.

The main contributions of the paper are the followings: (1) we propose a three-dimensional flow traffic characterization with identifying dominant categories, (2) we suggest application cluster identification, (3) we demonstrate this characterization on measured Internet traffic, (4) we present a number of interesting findings based on this characterization.

The paper treats these topics in the following order. In Section II we overview the details of measurements including the investigated network architecture and analysis tools. Section III presents the three-dimensional characterization concept and the related results with discussions. Finally, Section IV concludes the paper with our main results.

II. Traffic measurements

Measurements were taken from one of the commercial networks in Stockholm, Sweden. This company maintains network infrastructure for several service providers. These ISPs provide many different services such as Internet access, IP telephony or IPTV for both residential and business users. During the measurement period more than 1800 customers used the network for their own purposes. The network infrastructure of Swedish backbone network and the related residential network are shown in Fig. 1.

The backbone network consists of three core routers linked to each other with 3 Gbps optical fibres. The subscribers are connected to the area switches, and their traffic is aggregated in a migration switch through 100 Mbps links. The migration

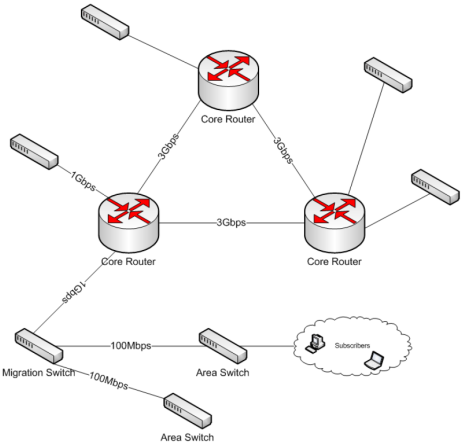


Fig. 1. The architecture of the measured network

TABLE I
BASIC DESCRIPTION OF MEASUREMENTS

Trace	Measurement period Oct., 2008 [(day) hour:min]	Duration [hour:min]	Flows [million]	Packets [million]
FL-1	(7) 08:19 – (8) 08:51	24:33	37.6	4627
FL-2	(7) 11:18 – (8) 22:16	34:59	59.1	3892
FL-3	(7) 15:00 – (7) 15:59	01:00	1.53	69.96

switch is linked to one of the core routers with a 1 Gbps capacity link.

The workstation responsible for data capturing was connected to one of the core routers with a 1 Gbps capacity fibre. The router mirrored its traffic to the workstation, which let the capturing device store and dump the data on its hard drives. Only the packet headers were captured to get information for the analysis such as protocol, size and direction.

Traffic identification was done with a tool developed by Ericsson Hungary. This software uses various techniques to identify the traffic such as port-based, signature-based and heuristic-based approaches, but the algorithm is not public.

Table I describes the main parameters of the investigated traces. After the preprocessing phase the cleaned data were loaded into database tables using Microsoft SQL Server 2005. Data retrieving was performed by SQL queries, and the results were processed by Matlab routines. Moreover, Matlab was also used for visualizing and creating charts.

III. FLOW CHARACTERIZATION

In this section we provide our three-dimensional characterization framework and demonstrate its use on actual measured traffic. We characterize Internet flows in three different dimensions, namely size (elephant and mice), duration (tortoise and dragonfly) and rate (cheetah and snail). First, we give definitions of these dimensions and categories. After that, we demonstrate how flows are located in the characterization space and present the ratio of top applications among these dimensions. Some of our results (distributions of the main traffic parameters, application clusters, etc.) from our detailed analysis are also provided and discussed.

We define flow as a bidirectional series of IP packets with the same source and destination addresses as well as port numbers (and vice versa for the backward direction). In other words, in our interpretation a flow contains both incoming and outgoing packets. For the analyzed dimensions the following definitions were used:

- *size:* the number of bytes transferred,
- *duration:* the time elapsed between start and end times,
- *rate:* the size divided by the duration.[1]

A. Categories

The category of a flow was determined by a threshold-based method scheme described in the following subsections where the exact formulas used to compute these values are also given. The threshold was defined as the mean plus the treble of the standard deviation of the sampled data in all cases following [12].

1) Elephant and mice: Prior studies tried to define elephant flows in different ways. We used the following definition to identify them. Let n be the number of flows, b_j the bytes carried by the flow j in both directions $(j = 1, 2, \ldots, n)$ and $\overline{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$ the mean flow size. Then the flow j is called elephant if

$$b_j > \overline{b} + 3\sqrt{\frac{\sum_{i=1}^{n}(b_i - \overline{b})^2}{n-1}}$$

otherwise the flow is defined as mice.

2) Tortoise and dragonfly: The separation of flows into tortoises and dragonflies is based on the flow duration. Similar to the previous definition, most of the flows are short, but the huge traffic is generated by the long ones. Let n be the number of flows, d_j the duration of the flow j $(j = 1, 2, \ldots, n)$ and $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ the mean flow duration. Then the flow j is defined as tortoise if

$$d_j > \overline{d} + 3\sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}$$

otherwise the flow is called dragonfly.

3) Cheetah and snail: This separation is taken by using the flow rate. Like the previous rules, a flow is related to the group of cheetahs if

$$r_j > \overline{r} + 3\sqrt{\frac{\sum_{i=1}^{n}(r_i - \overline{r})^2}{n-1}}$$

where n is the number of flows, r_j is the rate of the flow j $(j = 1, 2, \ldots, n)$ and $\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$ is the mean flow rate. If the above condition is not true, the flow is defined as snail.

B. Flow sizes and durations in details

This subsection investigates the relationships between some parameters of Internet traffic by a detailed analysis at the flow level. These parameters are the number of bytes carried by the flows, the flow duration and the number of packets.

[1]Flows with duration equal to zero were excluded.

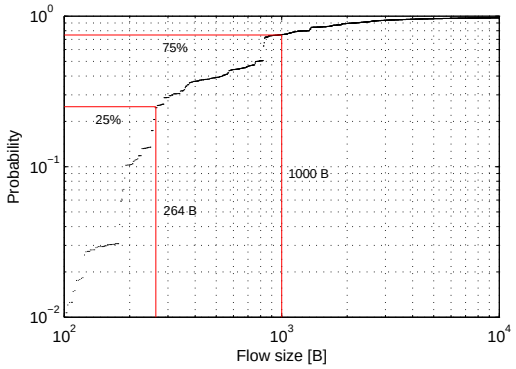


Fig. 2. CDF of the bytes transferred by the flows (FL-1)

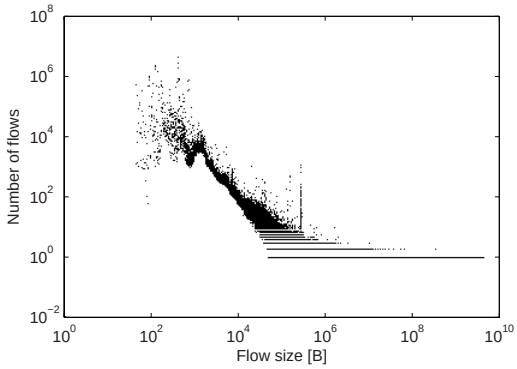


Fig. 3. Histogram of the bytes transferred by the flows (FL-1)

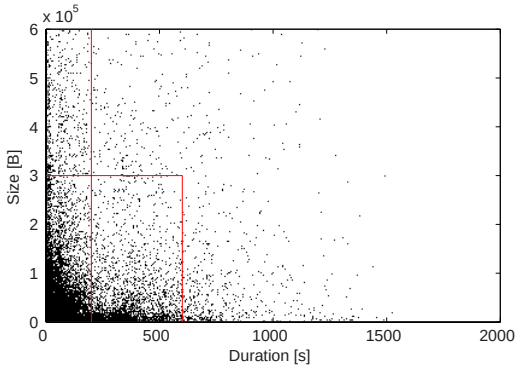


Fig. 4. Relationship between flow size and duration (FL-1)

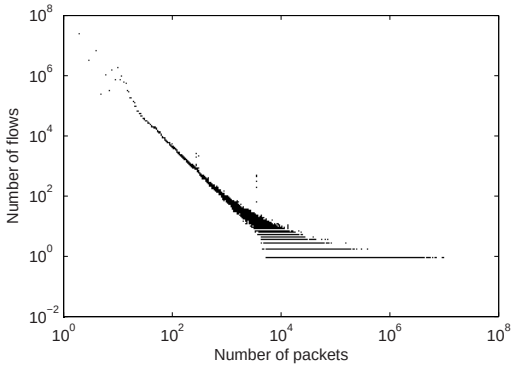


Fig. 5. Relationship between number of flows and packets (FL-1)

Fig. 2 shows the cumulative distribution function (CDF) of the flow size. Some breakpoints can be seen in this graph, which show that 75% of flows carry less than 1 kB, and 25% of flows are smaller than 264 B. Accordingly, it can be stated that most of the flows do not transfer large volume of data.

Fig. 3 depicts the histogram of the flow size. It can be observed that the number of packets has a quasi-uniform distribution with high variance in the interval 100–1000 B, while it shows a heavy-tailed decrease at flow size greater than 1 kB. The figure indicates that flows with several MB are quite rare. These flows are primarily originated from peer-to-peer applications (e.g. BitTorrent) and video playback (e.g. YouTube).

Fig. 4 illustrates the relationship between flow size and duration. Every point in the plots represents exactly one flow. As shown in the figure the bytes carried by the flows are almost independent of the flow duration. Measurements showed that approximately 99.7% of flows have a duration less than 600 seconds and they smaller than 300 kB. Furthermore, 99.3% of them are shorter than 200 seconds regardless of the flow size.

Fig. 5 shows the connection between number of flows and packets. We can see a cluster having heavy-tailed decay between 10 and 10^4 packets.

C. Applications in flow categories

First of all, we analyzed the traces to determine the distribution of the network traffic over different flow categories.

As shown in Table II a very small percentage (less than 0.1%) of flows called elephants are responsible for more than three quarters of the generated traffic and approximately one third of the traffic is related to the long flows (tortoises). Table II also indicates that in some cases big flows can be short. Since the FL-3 trace contains packets from a busy hour, it can be seen that the ratio of long flows in this hour is more than in other periods of the day.

Table III summarizes the average transfer rates for the whole

TABLE II
FRACTION OF INTERNET TRAFFIC FOR EACH CATEGORY OF FLOWS

Trace	Category	Ratio of no. of bytes	Ratio of no. of flows
FL-2	Elephant	77.18%	0.07%
	Tortoise	33.57%	0.08%
	Cheetah	0.04%	0.02%
FL-3	Elephant	75.12%	0.09%
	Tortoise	68.21%	0.89%
	Cheetah	0.006%	0.02%

TABLE III
AVERAGE TRANSFER RATES (FL-2)

Category	Transfer rate
Elephant	251 kbps
Tortoise	14 kbps
Cheetah	55 Mbps
All	19 kbps

TABLE IV
TOP FIVE APPLICATIONS IN DIFFERENT CATEGORIES

	Elephant	Tortoise	Cheetah
1	BT (69.64%)	Unid. (48.98%)	DC (48.56%)
2	Unid. (9.56%)	BT (38.46%)	Unid. (47.33%)
3	HTTP (3.68%)	MSN (3.77%)	BT (2.23%)
4	YT (3.35%)	SSL (2.03%)	HTTP (0.45%)
5	DC (2.43%)	HTTP (1.17%)	DNS (0.44%)

(a) Trace: FL-2

	Elephant	Tortoise	Cheetah
1	BT (43.15%)	BT (59.87%)	Unid. (54.04%)
2	YT (16.74%)	Unid. (28.96%)	DC (41.81%)
3	Unid. (15.18%)	HTTP (3.71%)	Browsing (3.68%)
4	HTTP (8.33%)	MSN (1.76%)	BT (1.10%)
5	Video pb. (5.51%)	SSL (0.77%)	HTTP (0.37%)

(b) Trace: FL-3

BT: BitTorrent, DC: DirectConnect, YT: YouTube
Unid.: Unidentified, Video pb.: Video playback

measurement period. It clearly shows that cheetahs are the fastest and tortoises are the slowest flows.

Table IV shows that BitTorrent contributes more than half of flows in case of elephants. Furthermore, YouTube is the second most popular application for the users in busy hours. It is worth mentioning that in case of cheetahs a relatively large portion of flows cannot be identified. These flows are probably originated from P2P applications, which are not easy to identify because they generally use arbitrary ports and encrypted method for the communication. Accordingly, most of the fast flows are related to P2P applications such as BitTorrent, DirectConnect, Gnutella and Skype. Similarly, a part of the tortoise flows are unidentified, thus it can be stated that long flows can also be generated by P2P applications.

As shown in Table V there are relationships between different characterizations where E and G denote the expected and the given category, respectively. For example, approximately 23% of elephants are also tortoises and 19% of tortoises are also elephants in case of FL-2 trace (see Table Va), which implies that big flows are tend to be short. At the same time, it is astonishing that in busy hours (Table Vb) almost 64% of elephants are tortoises what is nearly three times greater than the ratio calculated in case of FL-2 trace. Table V also indicates that there is no fast flow (cheetah) among big flows (elephants) and long flows (tortoises).

D. Cumulative distribution functions

This subsection presents the distribution functions of flow sizes, durations and rates for different classifications.

Fig. 6 shows the distributions of flow sizes. We can see that 70% of fast flows (cheetahs) are smaller than 1 kB as shown in Fig. 6a. This means that most of cheetahs are small bursts with only a few packets. Furthermore, Fig. 6b shows that half of tortoise flows are smaller than 10 kB in busy hours while this value is more than 100 kB for the whole measurement period. Consequently, in busy hours the number of bytes carried by

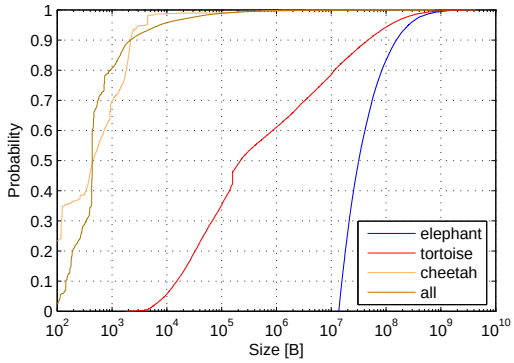
TABLE V
RELATIONSHIPS BETWEEN DIFFERENT CATEGORIES

E (↓) \ G (→)	Elephant	Tortoise	Cheetah
Elephant	100%	18.58%	0.047%
Tortoise	23.19%	100%	0%
Cheetah	0.013%	0%	100%

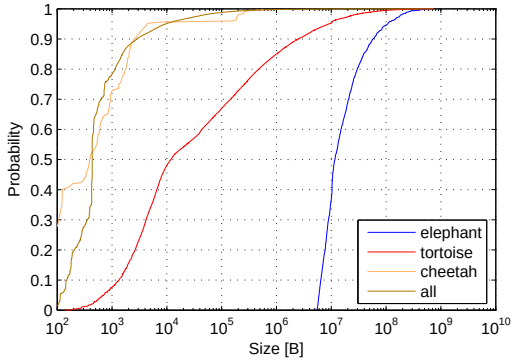
(a) Trace: FL-2

E (↓) \ G (→)	Elephant	Tortoise	Cheetah
Elephant	100%	6.66%	0%
Tortoise	63.84%	100%	0%
Cheetah	0%	0%	100%

(b) Trace: FL-3



(a) Trace: FL-2

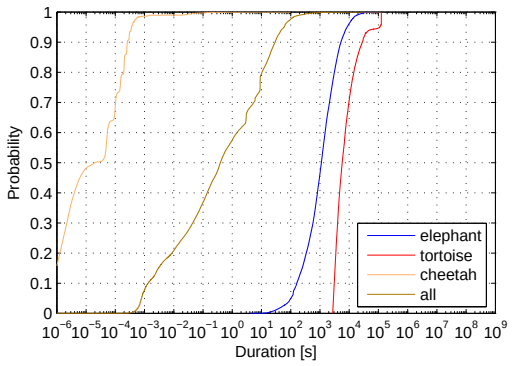


(b) Trace: FL-3

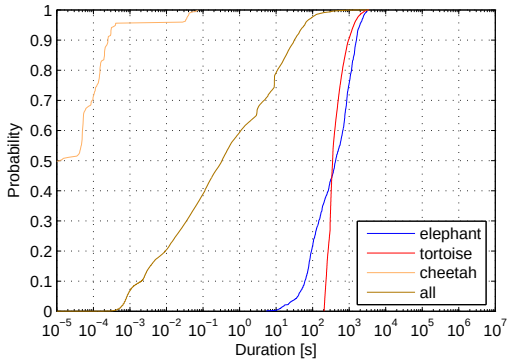
Fig. 6. Distributions of flow sizes

long flows is less than in other periods of the day. It can also be seen in Fig. 6a that about 15% of elephant flows transfer more than 100 MB of data and none of the elephants generate traffic less than 10 MB. Thus, we can conclude that mostly elephants are responsible for the majority of traffic.

Fig. 7 shows the distributions of flow durations for different types of flows. Fig. 7a clearly indicates that 80% of all flows are shorter than 10 seconds, so it can be stated that most of the Internet flows are short. About half of elephant flows are longer than 1000 seconds ($\approx$ 17 minutes) and 4% of them have a duration more than 10000 seconds (almost 3 hours), which implies that most elephant flows are long. We can see that tortoises also have very long durations. Moreover, 95% of

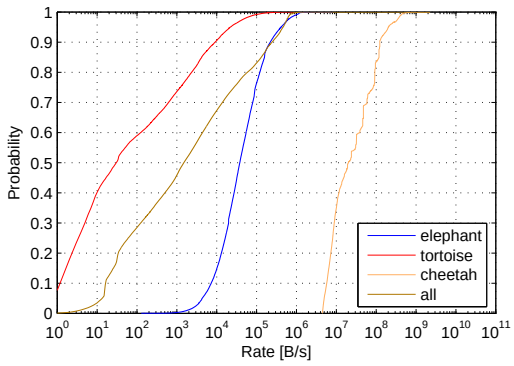


(a) Trace: FL-2

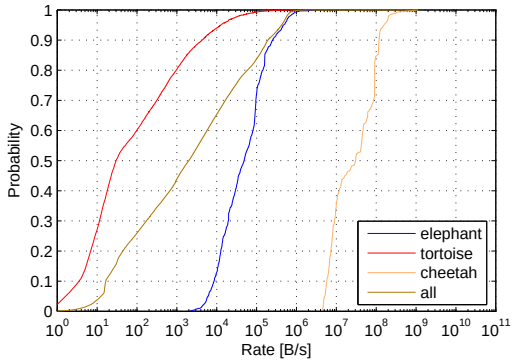


(b) Trace: FL-3

Fig. 7. Distributions of flow durations



(a) Trace: FL-2



(b) Trace: FL-3

Fig. 8. Distributions of flow rates

cheetahs are very short showing their bursty nature. In contrast, while none of the tortoises are shorter than 3000 seconds as shown in Fig. 7a, almost all of them have a duration less than 3000 seconds in busy hours as seen in Fig. 7b.

Fig. 8 shows the distributions of flow rates for each category. It can be observed that there is no significant difference in distributions between busy and non-busy hours. However, while in busy hours almost all of the tortoises are faster than 1 B/s as shown in Fig. 8b, in other periods of the day this ratio is about 90% (instead of almost 100%). Furthermore, 80% of tortoises are slower than 1 kB/s, which means that long flows are very slow. Fig. 8 indicates that approximately 70% of elephants have a rate between 1 kB/s and 100 kB/s, hence it can be stated that elephants are not so fast. Cheetahs are the fastest flows, because all of them have a rate greater than 4 MB/s, moreover, 20% of them are faster than 400 MB/s. In addition, we can observe that about 30% of all flows are faster than 10 kB/s, thus it can be concluded that in general, flows are relatively slow.

E. Application clusters

This subsection focuses on investigating the applications in the three-dimensional flow characterization space. Since the *peer-to-peer* traffic is dominant, moreover, *gaming, social networking* and *video playback* applications are also popular, we show results concerning these applications for the FL-3 trace.

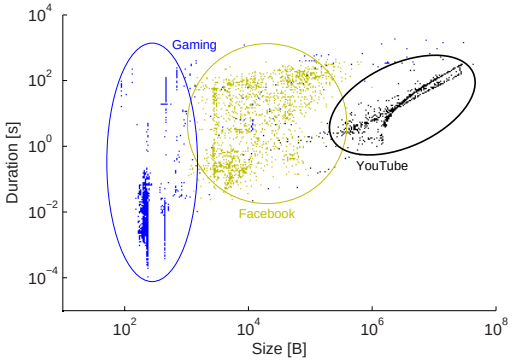
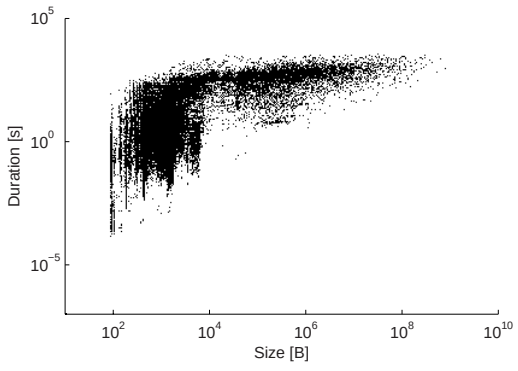
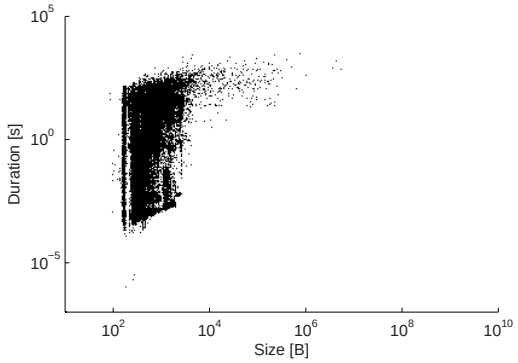


Fig. 9. Relationship between size and duration of gaming, Facebook and YouTube flows (FL-3)

Fig. 9 illustrates the relationship between flow size and duration for the application categories of gaming, social networking and video playback. The gaming has only 0.2% of the total traffic and it includes a number of different games like World of Warcraft, Counter-Strike, Xbox, etc. The social networking represents Facebook and the video playback is identical to YouTube. Facebook takes only 0.1% and YouTube contributes 7.4% of the total traffic, but both of them have increasing popularity. We can identify three different clusters in Fig. 9 which are related to these application categories. The gaming flows display clusters around horizontal lines concentrated in a small region between 50 B and 1 kB. On the



(a) TCP flows



(b) UDP flows

Fig. 10. Relationship between size and duration of BitTorrent flows (FL-3)

other hand, the duration of these flows varies in a large scale between 100 μs and 250 s. The cluster of Facebook flows covers a broader scale in size between 1 kB and 350 kB, but it shows a smaller region concerning duration between 20 ms and 350 s. YouTube flows present a cluster around a diagonal line indicating that duration is proportionally increasing to size at log-log scale due to positive correlation. While the sizes of YouTube flows are between 320 kB and 26 MB, the durations are in the interval 1 s and 500 s. Concerning the flow rate we found that gaming and Facebook flows have 360 kb/s and 96 kb/s average flow rates. In contrast, the average YouTube flow rate is 1.4 Mb/s. The advantage of this representation of flow characteristics is that we can identify application dependent clusters with representative characteristics.

Fig. 10 shows the relationship between flow size and duration of BitTorrent for TCP and UDP flows, respectively. BitTorrent is the most dominant P2P application in FL-3 contributing 46.8% of the total traffic. We can observe a peculiar difference between TCP and UDP flows in this figure. In case of TCP flows the graph is composed of two clusters: a cluster mainly consisting of horizontal lines and filling up almost completely the region between size 90 B and 6 kB and duration 100 μs and 320 s, and another cluster covering a large range in size between 1.5 kB and 580 MB, but a small range in duration between 5 s and 50 min. In contrast to TCP we can only find a single cluster in case of UDP flows, which is similar to the related cluster concentrated around the horizontal lines in Fig. 10a. Although, only 16% of BitTorrent flows are transferred over TCP and 84% of them are sent over UDP, TCP flows carry almost 99% of the total bytes.

IV. Conclusion

In this paper we advocated that a more solid understanding of Internet traffic can be obtained by analyzing properly chosen flow characteristics jointly and we proposed a three-dimensional (size, duration and rate) flow traffic characterization framework. The eligibility of this approach was shown by applying it for actual Internet traffic taken from a commercial network. We demonstrated that in these dimensions meaningful cluster categories can be identified and we also showed the dependencies among them. Moreover, we investigated how applications are allocated in the characterization space showing examples from the application categories of P2P, gaming, social networking and video playback.

Acknowledgement

We thank Sollentuna Energi AB for the measurements, and Ericsson Sweden and Ericsson Hungary for the cooperation. The research was supported by NKTH-OTKA grant CNK77802 and S. Molnár was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[1] K. Thompson, G. Miller, R. Wilder, "Wide Area Internet Traffic Patterns and Characteristics", IEEE Network Magazine, 11(6):10–23, November 1997.

[2] K. C. Claffy, H.-W. Braun, G. C. Polyzos, "A Parameterizable Methodology for Internet Traffic Flow Profiling", IEEE Journal of Selected Areas in Communications, 13(8):1481–1494, 1995.

[3] W. Fang, L. Peterson, "Inter-AS Traffic Patterns and Their Implications", Proceedings of IEEE GLOBECOM '99, pp. 1859–1868, Rio de Janeiro, Brazil, 1999.

[4] D. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, C. Diot, "A Pragmatic Definition of Elephants in Internet Backbone Traffic", Proceedings of ACM SIGCOMM Internet Measurement Workshop 2002, pp. 175–176, Marseille, France, November 2002.

[5] F. Hernandez-Campos, J. S. Marron, S. I. Resnick, C. Park, K. Jeffay, "Extremal Dependence: Internet Traffic Applications", Stochastic Models, 22(1):1–35, 2005.

[6] J. Kilpi, N. M. Markovich, "Bivariate Statistical Analysis of TCP-flow Sizes and Durations", Euro-FGI Deliverable D.WP.JRA.5.1.1, New Mathematical Methods, Algorithms and Tools for Measurement, IP Traffic Characterization and Classification, 2006.

[7] M. R. de Oliveira, A. Pacheco, C. Pascoal, R. Valadas, P. Salvador, "On the Dependencies between Internet Flow Characteristics", Lecture Notes in Computer Science, Volume 5464/2009, pp. 68–80, 2009.

[8] S. Bhattacharyya, C. Diot, J. Jetcheva, N. Taft, "Pop-level and Access-link-level Traffic Dynamics in a Tier-1 POP", Proceedings of ACM SIGCOMM Internet Measurement Workshop 2001, pp. 39–54, San Francisco Bay Area, November 2001.

[9] N. Brownlee, K. C. Claffy, "Understanding Internet Traffic Streams: Dragonflies and Tortoises", IEEE Communications Magazine, 2002.

[10] S. Molnár, G. Miklós, "On Burst and Correlation Structure of Teletraffic Models", 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, West Yorkshire, U.K., July 1997.

[11] Y. Zhang, L. Breslau, V. Paxson, S. Shenker, "On the Characteristics and Origins of Internet Flow Rates", In SIGCOMM, Pittsburgh, PA, USA, August 2002.

[12] K.-C. Lan, J. Heidemann, "A Measurement Study of Correlations of Internet Flows Characteristics", Computer Networks: The International Journal of Computer and Telecommunications Networking, 50(1):46–62, 2006.