Telecommun Syst
DOI 10.1007/s11235-010-9324-0

Traffic characterization of a residential wireless Internet access

Florian Wamser · Rastin Pries · Dirk Staehle ·
Klaus Heck · Phuoc Tran-Gia

Abstract Traffic characterization is an important means for Internet Service Providers (ISPs) to adapt and to optimize their networks to the requirements of the customers. Most network measurements are performed in the backbone of these ISPs, showing both, residential and business Internet traffic. However, the traffic characteristics of business and home users differ significantly. Therefore, we have performed measurements of home users at a broadband wireless access service provider in order to reflect only home user traffic characteristics.

In this paper, we present the results of these measurements, showing daily traffic fluctuations, flow statistics as well as application distributions. The results show a difference to backbone traffic characteristics. Furthermore, we observed a shift from web and Peer-to-Peer (P2P) file sharing traffic to streaming applications.

Keywords Traffic measurements · Traffic classification · Broadband wireless access

F. Wamser (✉) · R. Pries · D. Staehle · P. Tran-Gia
University of Würzburg, Institute of Computer Science,
Würzburg, Germany
e-mail: florian.wamser@informatik.uni-wuerzburg.de

R. Pries
e-mail: pries@informatik.uni-wuerzburg.de

D. Staehle
e-mail: dstaehle@informatik.uni-wuerzburg.de

P. Tran-Gia
e-mail: trangia@informatik.uni-wuerzburg.de

K. Heck
Hotzone GmbH, Berlin, Germany
e-mail: heck@hotzone.de

Published online: 29 May 2010

1 Introduction

During the last years, the Internet has emerged as the key component for business and personal communication which is reflected by the exponential traffic increase. The German Commercial Internet Exchange (DE-CIX) point has had for example a peak traffic rate of 803.7 Gbps on September 6th, 2009 [1] whereas the peak traffic rate was approximately 400 Gbps 12 months before. According to Cisco Systems [2, 3] this trend will continue. Over two third or 10,500 Petabytes (PB) of the monthly traffic is generated by consumers. The large bandwidth demands are caused by the fast changing application requirements. The applications range from low bandwidth email traffic over web browsing and P2P file sharing traffic to high bandwidth multimedia streaming. YouTube, as an example for video streaming, generated approximately 45 PB per month in the US in 2008 which is 1.75% of the complete Internet traffic [2]. Cisco Systems claims that 34% of the consumer Internet traffic is generated by streaming traffic in 2009 and expects that it increases to 55% in 2013 [2].

The total traffic increase and especially the increase of real-time applications require a careful network planning and optimization. This applies for fixed-line as well as for wireless providers. Traffic measurements are one essential part for the ISP to optimize their network. According to the measurement results, the ISP can adapt its prioritization strategies in order to guarantee a good perceived quality for the end user. However, most public available measurement data was gathered in the backbone and show the global traffic characteristics but do not reveal the user and application demands. According to Fukuda [4] there is a significant different traffic usage pattern in residential broadband traffic. Therefore, we performed the measurements close to the end user, namely at an ISP for home users who provides a broadband wireless Internet access.

The measurements were performed in summer 2008 and reflect the Internet usage of 250 households. Afterwards, the measurement data was classified using a combination of payload-based classification and host behavior. This paper shows the results of these measurements like daily traffic fluctuations, packet size distributions, flow and session statistics as well as application distributions. In contrast to our previous publication which is based on measurement data from 2007 [5], we have seen an immense growth of streaming traffic which is also underlined by Cisco Systems [3].

The remainder of the paper is organized as follows. Section 2 gives an overview of traffic measurements and its classification together with the related work. This is followed by Sect. 3 introducing our measurement scenario and methodology. Section 4 shows the results of the measurements and finally, conclusions are drawn in Sect. 5.

2 Background and related work

It has been a challenge for years to structure a reliable and feasible measurement architecture. First, a measurement has to generate detailed traffic characteristics, including global and special statistics, like application-based or user-based ones. Second, every measurement affects the measured data. If you meter an attribute, you have to take part in the system which influences the behavior of the system.

2.1 Traffic measurements

Commonly, there are two different approaches to measure a network: active probing and passive monitoring [6, 7]. The measuring process of the active measurements generate new traffic and inject it into the network, while passive measurements monitor and capture the network traffic. Latter systems use the recorded traffic to produce several statistics with the help of analysis software. The following monitoring systems use the passive approach.

Brownlee et al. [8] use RTFM [9], an Internet standard real-time flow measurement system with its open source implementation NeTraMet. It is a versatile and very general system for collecting flow data and includes a high level language for filtering, managing, and aggregating observed packets into flows. However, due to the fact that it needs to see headers for every packet through a device, it is not easy to implement in a switch or router.

Fraleigh et al. [10] designed a passive monitoring system to capture packet level traffic measurements on various ATM and SONET links. It is called IPMON and is inspired by the well-known OC3MON architecture by MCI [11] that is used by Thompson et al. [12] and McCreary et al. [13] to monitor optical ATM OC-3 links. IPMON has the capability to collect packet traces of up to OC-48 link speeds (2.4 Gbps) for a period of at least several hours. In addition, it uses GPS for synchronization. The CoralReef suite [14], developed by CAIDA, is originally based on the OC3MON, too. It is similar to IPMON, but does not support GPS timing and allows only link speeds of up to OC-12 (622 Mbps). Tools like CoralReef provide network card drivers, various programming APIs, and applications for capturing and analysis. A popular application programming interface for capturing network traffic is libpcap. Compared to the solutions above, it is only a computer library on top of network drivers and not a whole architecture. Shannon et al. [15] used libpcap to capture network traffic for further analysis.

Commercial solutions are available from Endace. First developed at the University of Waikato in the DAG project, the measurement cards are now able to capture Ethernet and optical links of up to OC-192 or 10 Gigabit Ethernet link speeds. Karagiannis et al. [16, 17] and John et al. [18] used DAG cards for their measurements.

Finally, some routers have the ability to export global per-flow summaries including start time, flow duration, byte and packet volume, IP addresses, and port numbers. In Cisco routers the tool for this purpose is called Netflow [19, 20]. It is embedded within the Cisco IOS software and is widely used to collect IP traffic information. Even though initially implemented by Cisco, Netflow is standardized by the IETF as Internet Protocol Flow Information Export (IPFIX) in the Request for Comments (RFCs) 5101 [21] and 5102 [22]. Juniper Networks, Nortel Networks, and Huawei Technology provide similar features within their routers.

2.2 Traffic classification

After collecting the data, the services have to be classified. Service classification has its own research group and with the emergence of new services like P2P, it is getting more and more difficult to identify packets [17]. At the network link an unordered mix of packets is collected that should be first grouped into connections and afterwards classified connection-wise. Along with port-based classification, several techniques and methods exist to classify packets:

2.2.1 Port-based classification

The assignment of port number to application type is used as defined by the Internet Assigned Numbers Authority (IANA). It is the simplest and most traditional method, but has several drawbacks. The port numbers are not defined for all applications. Especially some applications use port ranges or they even assign the ports dynamically so that the mapping of the ports and the applications can not be trusted. Applications like Skype vary the ports and even use port 80 to get through firewalls. Hence, a detection with this method is not possible. Thompson et al. [12], McCreary et al. [13], and Shannon et al. [15] used port-based classification and mapped each IP packet to a named application by choosing the first matching rule from an ordered collection of protocol/port patterns.

2.2.2 *Payload-based classification*

It is also known as content-based method. Payload-based classification is a syntactic analysis of the applicative layers of a packet. The classification entity is seeking deterministic character strings in the IP packet payload with fast regular expressions. The problem is that a detailed knowledge of the application as well as the format of its packets are needed. Some disadvantages are known: Character strings are not always available or the payload may be encrypted. However, this method only depends on a few characteristic packets. Karagiannis et al. [16, 23, 24] developed a heuristic for transport layer identification of P2P traffic which includes payload-based methods. A Wiki devoted to the identification of network protocols is used by the Application Layer Packet Classifier for Linux (L7-filter) [25] to allow a real-time classification.

2.2.3 *Host behavior classification*

Due to the limitation above, Karagiannis et al. [26] proposed another approach for traffic classification. They try to classify the popularity and the transport layer interactions with the help of inherent host behavior. The focus is shifted from classifying flows to associating hosts with applications. The flows are then classified accordingly. With this method, Karagiannis was able to present some heuristics to detect malware, P2P, web, chat, FTP, game, and streaming traffic.

2.2.4 *Statistical classification*

This is a recent method that uses statistical descriptions of the traffic with supervised learners. A statistical parameter can be the packet size or the inter-arrival time. First order Markov chains or k-Nearest Neighbors, Linear or Quadratic Discriminant Analysis are proposed by [27, 28] to calculate the probability of a packet to the statistical data model of an application. The statistical method is also able to detect tunneled or encrypted traffic.

3 Measurement scenario and methodology

In this paper, we focus on traffic characteristics of home users in a wireless network. The measurements have been performed at a Germany-wide wireless access provider who offers, along with business network access, private Internet access in large housing estates. The measurement and the classification is done according to proposals and papers introduced in the related work section.

3.1 Measurement setup

The measurements were performed at an ISP switching center which provides access for 250 households. The customers have access over Wireless LAN at several access points before the traffic is multiplexed at an IEEE 802.11a radio link. The dimensioning of the radio link is done by the provider according to the upcoming traffic of the users. Measurements of the provider confirmed that the link almost never operates at full capacity.

The measuring unit is set up right after the access points in the wired network. The monitoring point for the measurement is shown in Fig. 1. We measured both directions with the help of a receive-only network tap which ensures that the productive network is not interfered by our measurement. Our meter runs on a Linux system. It observes packet headers using two commodity 100Base-T Ethernet cards via libpcap.

The measurement process basically consists of five steps. First, raw traces are captured in pcap packet capture files. Additionally, the real-time classification entity described in the next paragraph stores detection data in log files. Second, the traffic traces are filtered to suppress or to make sensitive information anonymous. The anonymization module

Fig. 1 Measurement setup

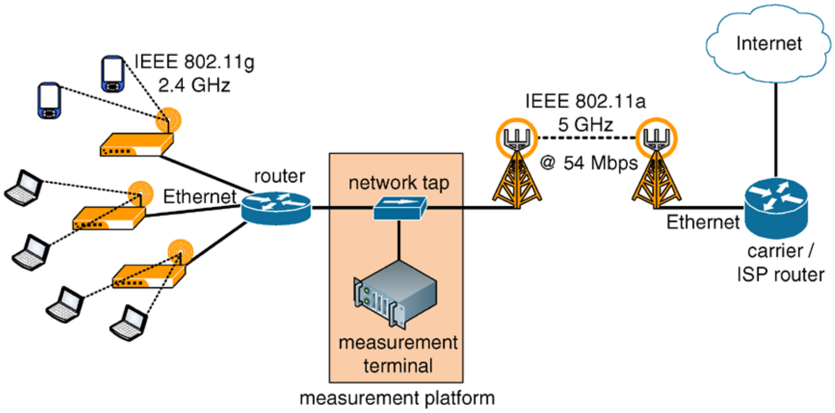


Fig. 2 Realtime classification

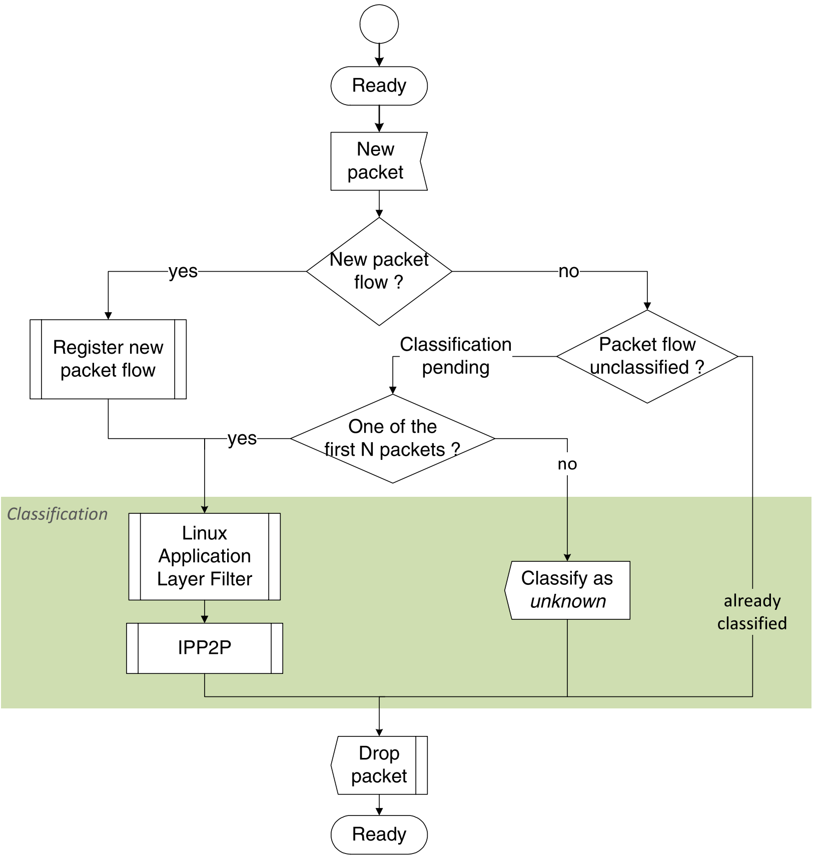


Table 1 Linux application layer filter sequence

	Protocol	Classification
1	Remote traffic	payload-based
2	Operating system tools or services	payload-based, port-based
3	P2P traffic	payload-based
4	Gaming traffic	payload-based
5	HTTP-based protocols	payload-based
6	Streaming	payload-based
7	Mail, instant messenger	payload-based
8	Web-traffic	payload-based
9	FTP, H323	special detection with kernel modules
10	VoIP, data transfer	payload-based

scrambles data in order to raise effort needed to obtain sensitive information about the internals of an operational network. Afterwards, the filtered traces are checked for errors and submitted in a database-driven repository. The last step is the analysis of the traces which is performed offline at external computers. All further work is done either within the database itself with the help of database languages or by querying the database.

3.2 Service classification

Our classification involves two levels of detection. On the one hand we use a payload-based detection with the Application Layer Packet Classifier for Linux (l7-filter) [25] and IPP2P [29]. However, this method requires the payload of the packets, which we are not allowed to store in capture files because of privacy concerns.

The payload-based classification is done in the following way: First, in real-time, a connection tracking assigns the packets to flows. If a new flow is detected, the classification scans up to $N = 20$ packets of this flow until the flow is classified, see Fig. 2. This is done online before capturing the data. It is scanned for well-known common applications. Thereby, it is important to use a good scanning sequence in order to avoid false detections. Table 1 shows the used order. First, the application tries to filter remote traffic, before operating system tools or services are filtered. Thus, the last classification rules applied are those for VoIP and data transfer. If the payload does not match at all, the packet is classified as "unknown". Especially all encrypted and new protocols cannot be detected by the Linux Application Layer Filter and are thus classified as unknown. Afterwards, the traffic is checked for P2P file sharing data (IPP2P) because it may use arbitrary ports.

Our second classification method is a host behavior analysis similar to the proposed one by Karagiannis [26]. The connections of a host are investigated as in the functional level approach. We record the usage of ports and IP addresses per host and compare the results of unknown hosts to already classified hosts. Thus, we are able to distinguish between P2P file sharing, web, and streaming traffic. The host behavior classification is done at the data repository after the packet capturing. The major advantage is that it is also capable to detect encrypted traffic. However, a detection of certain applications is in turn not possible. Therefore, a traffic class called “unclassified P2P” is shown in Sect. 4 which is P2P file sharing traffic of an unrecognized application.

3.3 Limitations

The monitoring and classification of unknown traffic has always some difficulties and limitations which have to be taken into account. Several issues occurred during the measurement which are enumerated below for completeness.

3.3.1 Classification payload patterns

The traffic patterns tend to underestimate or overestimate the traffic. It is difficult to find reliable packet signatures that match only the intended protocol. In all cases, a random encrypted stream may fit to several patterns. The other way round, some patterns are only able to match a part of the whole desired traffic. Namely, in our case the Skype pattern is one of the patterns that tend to overestimate and therefore added to the unknown traffic. Furthermore, some badly designed unimportant application patterns are simply left out in our analysis.

3.3.2 Anonymization, packet capture length

During the capturing of packets, the capture length is set to 96 bytes to make sure that the whole header is included in the traces. Due to privacy issues, the IP and payload anonymization cleared the rest of the payload in such a way that only the packet headers remained in the trace files. Consequently, we have no usable information about the payload during the offline analysis.

3.3.3 Diverse HTTP usage

During the measurements, we noticed that the HTTP usage statistics are varying. Some customers use extreme HTTP downloads from large file-hosting sites like Rapidshare [30]. Although these downloads do not represent the typical web browsing behavior, they are included in the web traffic statistics. Although HTTP video and HTTP audio traffic might be counted to web traffic, it is added to the streaming traffic class in Sect. 4.

3.3.4 Traffic shaping

The wireless access provider uses traffic shaping to control the Internet traffic. Due to the fact that the Cisco router is configured to prefer web and real-time traffic, P2P traffic might be underestimated in the following results. Furthermore, the Cisco router blocks identified P2P traffic if its bandwidth threshold (3 Mbps) is exceeded.

3.4 Trace description

The measurements were performed from July 11th, 2008 until July 29th, 2008. The whole measurement last 19 days and about 400 GB measurement data was collected. Further on, the Internet service provider gave us Cisco Netflow statistics, which prove our measurements in data volume and packet count. The billing system of the ISP is flat rate. Moreover, the packet loss during the capturing of packets in trace files is negligible and sums up to 0.18% in downlink direction and 0.09% in uplink direction.

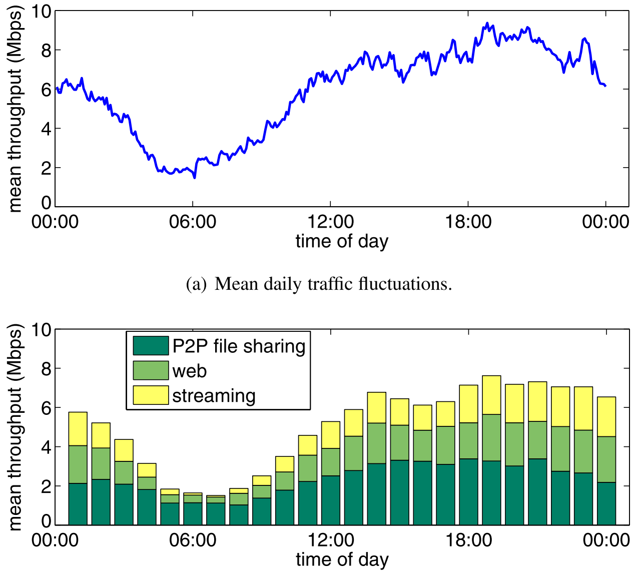
4 Measurement results

This section presents the results of the traffic measurements at the broadband wireless Internet access. The general daily traffic fluctuations and packet size distributions are included in the first part, the second part deals with flow and session statistics of the users, and the last part shows a detailed traffic classification.

4.1 Daily traffic fluctuations

First of all, we take a look at the mean throughput variations during a day. The mean throughput is calculated by first dividing all measurement data into days, then splitting each day into 5 min samples, and finally calculating the mean of all nineteen 5 min samples. The throughput fluctuations during the day are shown in Fig. 3(a). The x-axis shows the time of the day while the y-axis shows the mean throughput. It is obvious that the throughput decreases after 1:00 o’clock down to a minimum at 6:00 o’clock. Afterwards, the throughput increases with a maximum throughput at 19:00 o’clock. Similar daily traffic fluctuations can be found in [4, 31].

Since the traffic is varying over the day, we distinguish between constant and fluctuating traffic in Fig. 3(b). The figure shows the daily traffic statistics according to the three main applications P2P file sharing, web, and streaming traffic. Comparing the three different application categories, we can see that P2P file sharing traffic is still the dominating application during the whole day with the largest percentile between 5:00 and 8:00 o’clock. As P2P file sharing traffic does not require a user interaction like web traffic, this



(a) Mean daily traffic fluctuations.

(b) Mean daily application distribution.

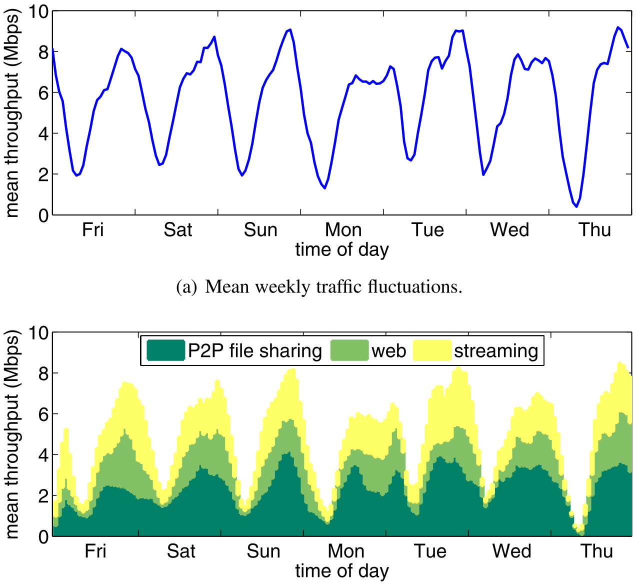
Fig. 3 Mean throughput and application distribution

high percentile at night show users running their computers 24 hours, 7 days a week.

Web and streaming traffic constitute almost the same amount of traffic but the total throughput varies during the day. At night, almost no web and streaming traffic is present in contrast to P2P file sharing traffic. The high streaming throughput is rather surprising compared to Ploumidis et al. [32] with only 0.177% of streaming traffic measured in 2005. The reason for this streaming traffic increase is the popularity of new video streaming services. YouTube for example generated 1.75% [2] of the complete Internet traffic in the US in 2008 and the platform was set up at the end of 2005.

The traffic fluctuations for a complete week are shown in Fig. 4. The plots are generated on an hourly basis and in addition, a moving average of size 4 is applied to smooth the curves. In contrast to our previous publication [5], the traffic fluctuation does not differ between weekdays and weekends. All days of the week have a similar distribution. A lower traffic volume can only be seen for Monday evening. As the traffic is averaged over all Mondays of the measurements, this is surprising and cannot be explained. Looking at Fig. 4(b) showing the application distribution, it can be seen that the high percentile of P2P file sharing traffic at night is not an exception of one night at the week.

Summarizing, we see that 1.6 Mbps is constantly used by P2P file sharing traffic over the day. Web and streaming traffic are varying over the day because they need user interaction. Further investigations on the traffic classes are shown in Sect. 4.5.



(a) Mean weekly traffic fluctuations.

(b) Mean weekly application distribution.

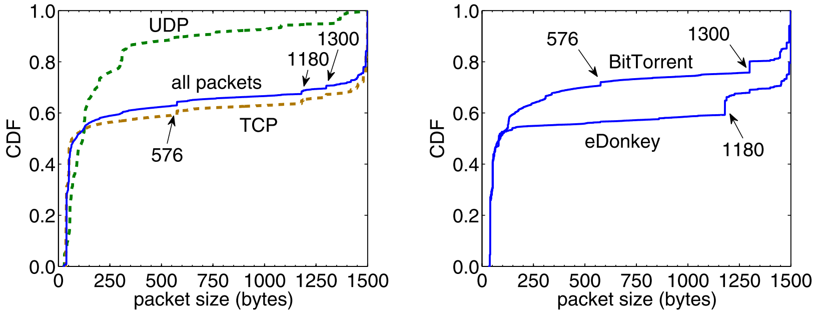
Fig. 4 Mean throughput and application distribution during one week

4.2 Packet size distributions

After having shown the daily and weekly traffic and application fluctuations, we now want to evaluate the packet size distributions and compare them to traffic measurements in the backbone. Thompson et al. [12] show a trimodal packet size distribution where nearly half of the packets are 40 to 44 bytes, 20% are 576 bytes, and 10% are 1500 bytes in length. Sean McCreary and Claffy [13] observed that about 80% of the packets are smaller than 600 bytes but have seen the same trimodal packet size distribution as Thompson. The newest backbone traffic packet distribution we found is presented by John and Tafvelin [18] in 2007. In contrast to the previous two papers, they show a bimodal traffic distribution where 40% are of size smaller than 44 bytes and another 40% of the packets are between 1400 bytes and 1500 bytes. Their results are similar to our measurements results shown in Fig. 5(a).

The figure displays the packet size on the x-axis and its Cumulative Distribution Function (CDF) on the y-axis. From the figure, we can observe several things. First, 90% of the UDP packets are smaller than 500 bytes. This might be P2P control or real-time streaming traffic. Second, looking at the curve for all packets, we observe a bimodal packet size distribution. The first peak occurs at around 40 bytes and the second step at 1500 bytes. This shows that most packets are transmitted via TCP, with the 40 bytes Acknowledgments and the 1500 bytes Ethernet Maximum Transfer Unit (MTU), which is also underlined with the TCP packet size distribution curve. However, we can also observe three small steps at 576 bytes, 1180 bytes, and 1300 bytes. These

Fig. 5 Cumulative IP packet size distribution



(a) TCP and UDP packet size distribution.

(b) eDonkey and BitTorrent packet size distribution.

Table 2 Flow statistics

	Mean	Max.	Percentile		
			50th	90th	99th
duration	19.12 s	24 h	0 s	17.38 s	338.67 s
packets	26.93	1.57 Mio	1	10	139
size	16.98 KB	1.01 GB	144 B	1481 B	74.24 KB

packet sizes are used by P2P file sharing protocols as shown in Fig. 5(b).

Packets of size 1180 bytes are only used by the eDonkey protocol which was also observed by Karagiannis et al. [16]. Furthermore, they have shown a similar packet size distribution for BitTorrent. 1300 bytes is the MTU recommended by some ISPs for DSL connections. Therefore, we think that these packet sizes result from downloads from clients of such ISPs. Finally, we take a look at the protocol distribution on the transport layer. Almost 88% of all measured packets are transmitted via TCP, only 11.6% via UDP, and less than one percent is used for ICMP control traffic. Considering the total throughput in bytes, 95% of the complete data is transmitted via TCP.

4.3 Flow statistics

The observed packet size distributions are very similar to the latest found publication. Let us now take a look at flow statistics. In order to assign packets to flows, we use the crl_flow tool from the CoralReef suite of Caida. The standard expiry timeout of a flow is thereby set to the default 64 seconds. In total, 73.4 Mio packet flows were identified which have caused a complete traffic volume of 1.25 TB. Table 2 shows the detailed statistics of the duration, number of packets, and size of the flows.

The maximum duration of a flow is 24 h due to the fact that each measurement run lasts 1 day. Comparing the results to Brownlee and Claffy [8], the percentage of dragonflies, flows lasting less than 2 seconds, is with 70.9% much higher than the 45% shown in the paper. According to them, 1.5% of the flows are longer than 15 min, called tortoise, but carry 50% to 60% of the complete traffic. In our measurements, only 0.29% of the flows last longer than 15 min but they carry 720 GB or 57.74% of the whole traffic.

These statistics reveal that by far the largest number of flows carry a small amount of data (mice) and last only short (dragonflies). This has to be taken into account by the ISPs, especially when performing traffic management on a per flow basis.

4.4 Session statistics

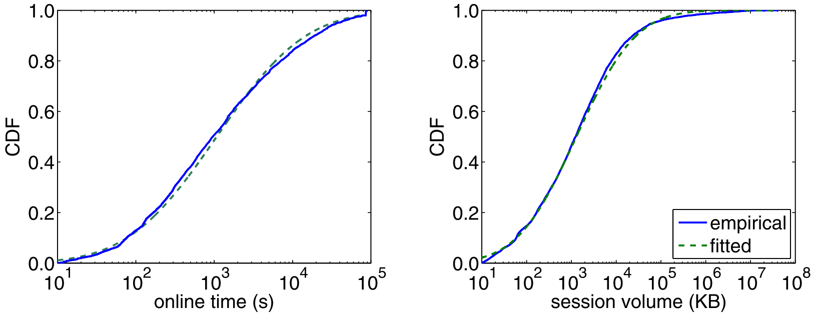
In order to get rid of the large number of the dragonfly flows and to show a more general user behavior, we set up the definition of a session. The parameters of a sessions are based on experimentation with different parameters. It is defined as follows. Several flows of one user regardless of the application with an inter-flow-time lower than 5 min (timeout: 5 min) belong to one session. Furthermore, a session has to last longer than 10 s and needs a minimum session volume of 10 KB in order to distinguish between periodic signaling and normal traffic.

4590 sessions are identified during the whole measurement and the maximum online time is 24 h for the same reason as the longest flow duration. The mean number of sessions per user is 2.1 and the maximum number of sessions 74. Further session statistics are shown in Table 3.

The weekend data is gathered at 2 weekends and the mean session duration during the weekend is 167 min compared to 129 min during the week. This is not surprising as most home users spend more time in front of their computers during the weekend. However, what is surprising is the median of the session volume. 1.24 MB in 24.6 min seems to be a very low amount of data. We think that the reason for this

Table 3 Online session duration and session volume

	Mean	Max.	Percentile		
			25th	50th	90th
Session duration					
All-day	136 min	24 h	6.7 min	24.6 min	395.2 min
Weekend	167 min	24 h	8.3 min	33.4 min	465.0 min
Weekday	129 min	24 h	6.5 min	23.7 min	362.9 min
Session volume					
All-day	80 MB	42 GB	246 KB	1.24 MB	23.52 MB
Weekend	98 MB	22 GB	294 KB	1.74 MB	33.55 MB
Weekday	76 MB	42 GB	238 KB	1.15 MB	22.29 MB



(a) Session duration.

(b) Session volume.

Fig. 6 Empirical cumulative distributions and lognormal distributions fitting the empirical functions

lies in instant messaging services and periodic email checking. Although only 3% of the complete data volume belong to messaging services, 10% of all traffic flows belong to this class.

Finally, we can see a large gap between the 50% quantile and the maximum session volume (1.24 MB to 42 GB). The few large sessions belong to P2P and web file downloads. In order to further analyze the session duration and session volume, the CDFs of both statistics are plotted in Fig. 6.

The x-axis in Fig. 6(a) shows the logarithmic scale of the session duration. Unfortunately, we gathered the measurements on a daily basis and therefore it is not possible to identify session longer than one day. However, we can see that 3% of the sessions last at least one day and these sessions belong to P2P file sharing traffic. The curve can be best fitted by a lognormal distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \tag{1}$$

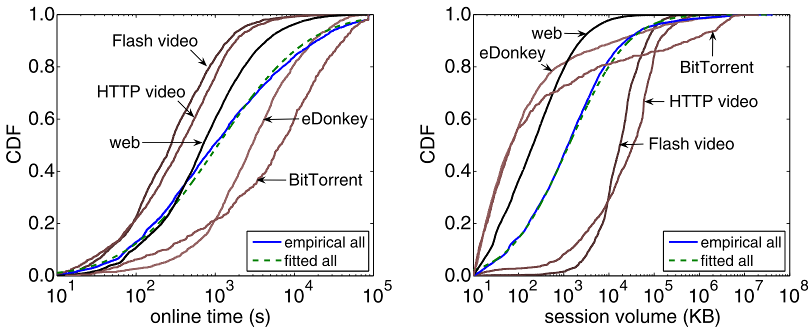
with $\mu = 2.8879$ and $\sigma = 2.0577$.

Looking at the CDF of the session volume in Fig. 6(b), we can see a larger heterogeneity compared to the session duration. About 3% of the sessions have a volume larger than 1 GB whereas 90% of the sessions have a volume smaller than 23.52 MB. This curve can also be well fitted by a lognormal distribution with $\mu = 7.1650$ and $\sigma = 2.4066$.

Chlebus and Divgi presented session statistic of a Wireless LAN in [33, 34]. Their definition of a session slightly differs from ours. It is created when a user logs into the network and ends when the user logs out or is timed out of the network. Unfortunately, the length of the timeout is not defined. According to their statistics, a user has on average 2.16 sessions per day, consuming a mean of 12.24 MB in about one hour. The maximum session duration was 34 hours consuming 1.5 GB of data and the maximum number of sessions per user was measured 37. They fitted the curves with a truncated Pareto distribution. Although their results differ from ours, the general distribution of the session duration and the session volume are similar.

In order to see the impact of each application on the session statistics, we split the session statistics by application. Figure 7(a) shows the online times of each application. It is clearly visible that the overall session times are prolonged by P2P traffic. 50% of all eDonkey sessions last longer than 50 minutes and 50% of the BitTorrent session are longer than 110 min. This is by far longer than a video streaming session. However, what is surprising is the fact that the session volume of P2P traffic is smaller compared to video streaming sessions, cf. Fig. 7(b). The reason for this might be traffic shaping of the service provider. While streaming traffic is prioritized, P2P file sharing applications are regarded as best effort traffic and blocked if they exceed a certain bandwidth. Looking at the BitTorrent curves, we can see that the number of BitTorrent sessions is lower compared to other sessions. In total, about 1000 session were identified.

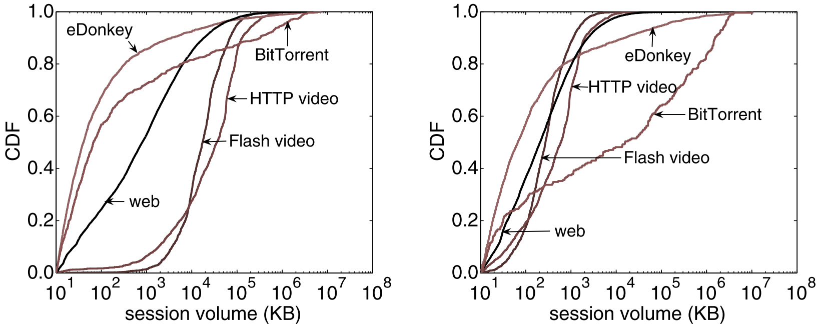
Fig. 7 Session duration and volume split by application



(a) Session duration.

(b) Session volume.

Fig. 8 Session volume separated between downlink and uplink traffic



(a) Session volume on the downlink.

(b) Session volume on the uplink.

The duration of a web surfing session is on average longer than a video streaming session, meaning that a user normally browses through webpages longer than watching YouTube. In addition, the traffic volume of a web session is in general smaller than a video streaming application. Just a few web session can be observed with a traffic volume of several hundred MB up to 2.9 GB. These sessions belong to HTTP downloads like Rapidshare.

In Fig. 8, we differentiated for the session volume between downlink and uplink. As expected, the ratio between downlink and uplink traffic volume for P2P traffic is almost 1:1, caused by the tit for tat strategy of P2P file sharing. In contrast, streaming applications like HTTP video and Flash video transmit mainly on the downlink and the uplink statistics only reflect the TCP acknowledgments. The figures furthermore reveal that the average Flash video size is 17 MB, which can also be find in YouTube statistics.

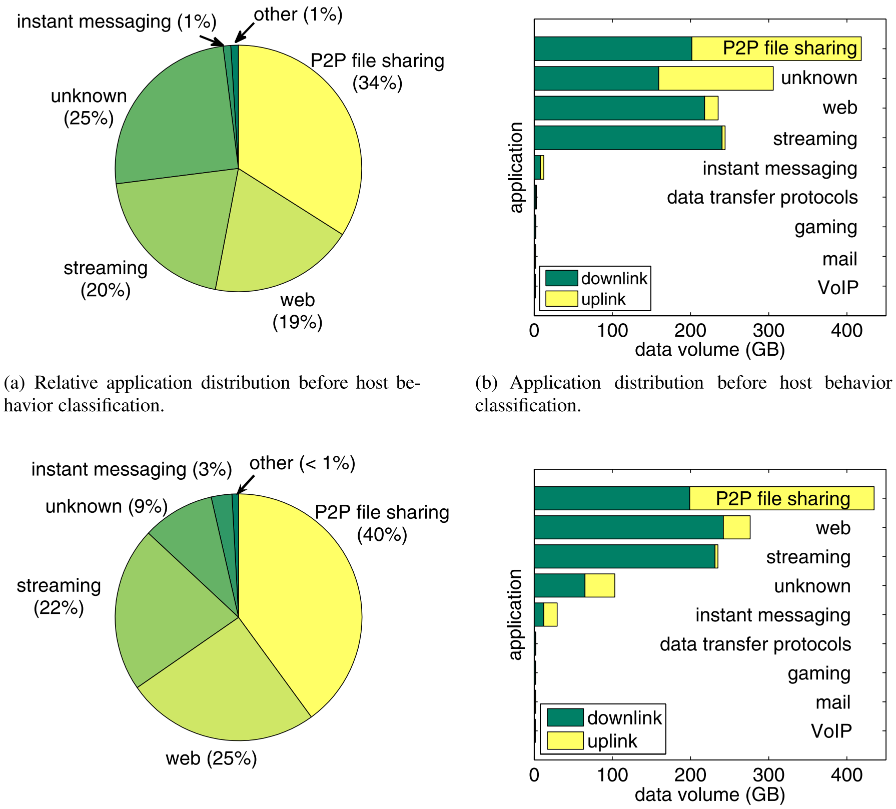
4.5 Traffic classification

After we evaluated the flow and the session statistics and compared them to the related work, we want to evaluate if the application distribution differs compared to fixed-line networks. Towards the end of 2005, P2P file-exchange applications overtook web traffic as the major contributor of traffic on the Internet. P2P traffic was measured at 60% to 80% of the total broadband traffic [31]. Cisco Systems states in their annual report that 60% or 1358 PB per month belong to P2P traffic at the end of 2006 [35]. However, this percentage decreases and Cisco predicts a P2P traffic percentage of 40% (5192 PB per month) at the end of 2010 and an increase of streaming traffic to 43% (5469 PB per month).

Figure 9 shows the application distribution after the payload-based classification and after the host behavior analysis. The figures illustrate that a payload-based classification alone leads to a quarter of unknown traffic. After applying the host behavior, which tries to classify the popularity of the transport layer interactions, the percentage of unknown traffic was decreased down to 9 percent. This shows that although it is not clear how trustworthy the host behavior analysis is, it helps to reduce the percentage of unknown traffic. However, both, the results after the payload-based classification and the final application distribution after the host behavior analysis underline the statements from Cisco Systems.

Looking now at Fig. 9(c), 40% of the complete measured traffic belong to P2P file sharing applications. This relatively low percentage of P2P is only achieved with the traffic shaping of the ISP which is essential in order to perceive an acceptable streaming and web quality. This shaping becomes obvious when looking at Fig. 9(d). P2P file sharing uses 71% of the uplink bandwidth whereas only 26% is used on the downlink. If no shaping would be performed by the ISP, P2P file sharing traffic would use around 60% [5] in total, which is a higher value as in backbone measurements. This higher percentage of P2P file sharing traffic clearly results from the

Fig. 9 Application distribution



(a) Relative application distribution before host behavior classification.

(b) Application distribution before host behavior classification.

(c) Relative application distribution after host behavior classification.

(d) Application distribution after host behavior classification.

measurements in a residential network. Mainly, this is especially interesting for home network service providers to optimize their services.

Web traffic was measured with up to 50% in the core [16]. In our environment only 25% web traffic was detected. However, our web traffic fraction includes browsing and file downloads with HTTP but not streaming over HTTP which belong to a separate streaming traffic category. Surprisingly, we notice a new user download behavior. Some customers use extreme HTTP downloads from large file-hosting sites as an alternative to P2P file sharing. Most notably, during the prioritizing and the shaping of the traffic, this is detected as a problem. HTTP proxies may help here to limit the outbound traffic. Figure 9(d) shows the exact data volume of the traffic categories and further distinguishes between downlink and uplink volume.

Although VoIP and FTP (data transfer protocol) are prioritized, the usage is very low. In case of VoIP this has several reasons. First, the network can not meet the user expectations (relatively high delays and jitter) and second, IP phones and VoIP devices mainly provide wired interfaces. Besides the low usage of VoIP and FTP, we have also seen only a few gaming traffic. One can think that this might result from the fact that gamers normally use a DSL connection with smaller delays compared to the measured multi-hop broadband wireless Internet access, but Cisco Systems has also measured less than 1% of gaming traffic in the year 2008 [2]. The low usage of VoIP and Internet games is seen as characteristic for a wireless broadband access network at the moment.

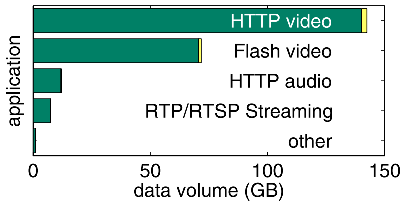
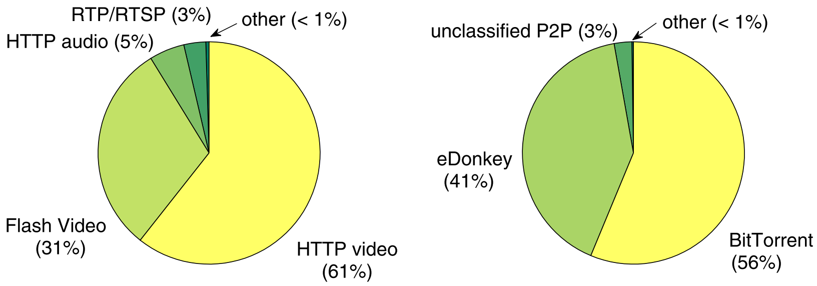


Fig. 10 Application distribution of streaming media

In contrast, streaming traffic with about 22% of the whole traffic is now besides web and P2P file sharing traffic one of the main traffic categories used in home environments. On the one hand this is surprising when comparing it with previous publications from Ploumidis et al. [32] with 0.177% and Pries et al. [5] with 4% of streaming traffic. On the other hand, this result is conform with the values predicted by Cisco Systems [35]. The exact distribution of the streaming traffic is shown in Fig. 10.

It is rather complicated to assign specific media players to the different protocols since most players are able to handle

Fig. 11 Subcategory application distribution



(a) Streaming media.

(b) P2P traffic.

several protocols. The biggest portion, HTTP video are used by Quicktime, Real Player, and the Windows Mediaplayer. However, all players support RTP/RTSP streaming as well. The only difference between these two groups is the way the connection is established.

If the player is called using "rt[s]p://", the l7-filter assigns the connection to the RTSP class and if the connection is established using "http://" the connection belongs to the HTTP video class. However, the Real Player normally uses RTSP for streaming. Besides these two classes, another similar streaming protocol can be used called Microsoft Media Server Protocol (MMS), which was not detected during our measurements.

Figure 11(a) shows the percentage of the streaming traffic. Similar to VoIP traffic, real-time streaming traffic has higher QoS requirements. Consequently, it is not surprising that the fraction of non live streaming as Flash Videos is measured with 31% of the whole streaming traffic.

Finally, the P2P differentiation is shown in Fig. 11(b). This statistics differs from the latest ipoque statistics [36]. About 71% of their complete P2P traffic belongs to BitTorrent and only 24% to eDonkey, whereas we measured 56% BitTorrent P2P traffic and 41% eDonkey traffic. The difference might be caused by the location of the measurement. Schulze and Mochalski measured at 3 different universities and we measured in 250 households. The 3% unclassified P2P file sharing traffic shown in Fig. 11(b) has been detected by the P2P host behavior statistics as P2P traffic but the filter was unable to assign the traffic to eDonkey or BitTorrent.

5 Conclusion

This paper presents the results of our Internet traffic measurements in a commercial broadband wireless access network for residential Internet users. The presented results are divided into general traffic statistics and application distributions. The findings of the daily and weekly traffic fluctuations show a similar behavior compared to the statistics from the German Internet point DE-CIX [1]. A breakdown of the application distribution during the day shows that P2P file sharing traffic is used all day long whereas the amount of web and streaming traffic increases in the evening hours with a peak at 19:00 o'clock. The reason is that streaming traffic and especially web traffic requires user interaction which is not the case for P2P file sharing traffic which means that the file sharing applications run 24 hours, 7 days a week. The packet size distribution of all packets is similar to the latest backbone measurements [18] and follows a bimodal distribution. 43% of the packets have a length of 40 bytes and 30% of the packets contain 1500 bytes of information. This results from the 88% measured TCP packets, containing 95% of the complete measured traffic.

Brownlee and Claffy [8] observed a large number of short flows carrying a small amount of data. This was also monitored in our measurements and the percentage of these so called dragonflies was with 70.9% compared to 45% much higher. In addition, the number of flows lasting more than 15 min was with 0.29% much smaller compared to 1.5%. These few flows carry however 57.74% of the whole traffic. The following session statistics reveal a similar behavior compared to the flow statistics and although they differ from Chlebus and Divgi [33, 34], the general distribution of the session volume and session duration are similar.

Our traffic classification statistics showed that a combination of payload-based and host behavior classification is a good means to perform traffic characterization. The results affirm the predicted trends of P2P, web, and streaming traffic with empirically determined values. The percentage of P2P file sharing traffic is with 40% lower compared to 62% measured in 2007 [5]. The decrease is caused by the increase of streaming traffic to 22%. Within the streaming traffic class Flash Video increases to 31%. Furthermore, a second reason for the decrease might be a change in the download behavior of some customers. They use extensive HTTP downloads as alternative to P2P and FTP file sharing. A breakdown of the P2P file sharing traffic shows that BitTorrent is with 56% responsible for the largest portion of the P2P traffic followed by eDonkey with 41%.

The low fraction of VoIP and gaming traffic in our measurements is seen as characteristic for broadband wireless access networks and isolates them from other access technologies. In case of VoIP it is on the one hand caused by the network and on the other hand by the lack of wireless IP phones and wireless capable VoIP devices.

Summarizing we want to point out that the general traffic fluctuations remain similar to previous measurements whereas the application distribution differs. Streaming applications become more and more important and are now responsible for one fourth of the complete traffic. With radios connected over Wireless LAN to the Internet and new television screens with Internet connection, it is expected that the percentage of streaming traffic will soon overtake P2P file sharing traffic.

References

1. DE-CIX German Internet Exchange. (2010). http://www.de-cix.net/.
2. Cisco Systems Inc. (2009). Cisco visual networking index—forecast and methodology, 2008–2013. White Paper.
3. Cisco Systems Inc. (2009). Hyperconnectivity and the approaching zettabyte era. White Paper.
4. Fukuda, K., Cho, K., & Esaki, H. (2005). The impact of residential broadband traffic on Japanese ISP backbones. *SIGCOMM Computer Communication Review*, *35*(1), 15–22.
5. Pries, R., Wamser, F., Staehle, D., Heck, K., & Tran-Gia, P. (2009). Traffic measurement and analysis of a broadband wireless Internet access. In *IEEE VTC spring 09*, Barcelona, Spain.
6. Paxson, V., Mahdavi, J., Adams, A., & Mathis, M. (1998). An architecture for large-scale Internet measurement. *IEEE Communications*, *36*(8), 48–54.
7. Fraleigh, C., Moon, S., Lyles, B., Cotton, C., Khan, M., Moll, D., Rockell, R., Seely, T., & Diot, C. (2003). Packet-level traffic measurements from the sprint IP backbone. *IEEE Network*, *17*(6), 6–16.
8. Brownlee, N., & Claffy, K. C. (2002). Understanding Internet traffic streams: dragonflies and tortoises. *IEEE Communications Magazine*, *40*(10), 110–117.
9. Brownlee, N., Mills, C., & Ruth, G. (1999). Traffic flow measurement: architecture.
10. Fraleigh, C., Diot, C., Lyles, B., Moon, S. B., Owezarski, P., Papagiannaki, D., & Tobagi, F. A. (2001). Design and deployment of a passive monitoring infrastructure. In *IWDC '01: proceedings of the thyrrhenian international workshop on digital communications*, Taormina, Italy (pp. 556–575).
11. Apisdorf, J., Claffy, K. C., Thompson, K., & Wilder, R. (1997). OC3MON: flexible, affordable, high performance statistics collection. In *Proc. of INET 97*.
12. Thompson, K., Miller, G. J., & Wilder, R. (1997). Wide-area Internet traffic patterns and characteristics (extended version). *IEEE Network*, *11*(4), 10–23.
13. McCreary, S., & Claffy, K. C. (2000). Trends in wide area IP traffic patterns—a view from ames Internet exchange. In *Proceedings of the 13th ITC specialist seminar on Internet traffic measurement and modelling*, Monterey, CA.
14. Keys, K., Moore, D., Koga, R., Lagache, E., Tesch, M., & Claffy, K. C. (2001). The architecture of CoralReef: an Internet traffic monitoring software suite. In *PAM2001—a workshop on passive and active measurements*.
15. Shannon, C., Moore, D., & Claffy, K. C. (2002). Beyond folklore: observations on fragmented traffic. *IEEE/ACM Transactions on Networking (TON)*, *10*(6), 709–720.
16. Karagiannis, T., Broido, A., Brownlee, N., Claffy, K. C., & Faloutsos, M. (2003). *File-sharing in the Internet: a characterization of P2P traffic in the backbone* (Tech. rep.). University of California, Riverside, University of California, Riverside Department of Computer Science, Surge Building, Riverside, CA 92521.
17. Karagiannis, T., Faloutsos, M., Broido, A., Brownlee, N., & Claffy, K. C. (2004). Is P2P dying or just hiding? In *IEEE global telecommunications conference*, 2004, GLOBECOM'04 (pp. 1532–1538).
18. John, W., & Tafvelin, S. (2007). Analysis of internet backbone traffic and header anomalies observed. In *IMC'07: proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA (pp. 111–116).
19. Caceres, R., Duffield, N. G., Feldmann, A., Friedmann, J., Greenberg, A., Greer, R., Johnson, T., Kalmanek, C., Krishnamurthy, B., Lavelle, D., Mishra, P. P., Ramakrishnan, K. K., Rexford, J., True, F., & van der Merwe, J. E. (2000). Measurement and analysis of IP network usage and behavior. *IEEE Communications Magazine*, *38*(5), 144–151.
20. Cisco Systems Inc. (2007). Cisco IOS NetFlow. White Paper.
21. Claise, B. (2008). Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information. RFC 5101. http://www.ietf.org/rfc/rfc5101.txt.
22. Quittek, J., Bryant, S., Claise, B., Aitken, P., & Meyer, J. (2008). Information model for IP flow information export. RFC 5102. http://www.ietf.org/rfc/rfc5102.txt.
23. Karagiannis, T., Broido, A., Faloutsos, M., & Claffy, K. C. (2004). Transport layer identification of P2P traffic. In *IMC '04: proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA (pp. 121–134).
24. Karagiannis, T., Molle, M., & Faloutsos, M. (2004). Long-range dependence: ten years of Internet traffic modeling. *IEEE Internet Computing 8*(5), 57–64.
25. Application Layer Packet Classifier for Linux (L7-filter) (2010). URL http://l7-filter.sourceforge.net/.
26. Karagiannis, T., Papagiannaki, K., & Faloutsos, M. (2005). BLINC: multilevel traffic classification in the dark. In *SIGCOMM '05: proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications*, New York, NY, USA (pp. 229–240).
27. Dahmouni, H., Vaton, S., & Rosse, D. (2007). A Markovian signature-based approach to IP traffic classification. In *MineNet '07: proceedings of the 3rd annual ACM workshop on mining network data*, New York, NY, USA (pp. 29–34).
28. Bernaille, L., Teixeira, R., & Salamatian, K. (2006). Early application identification. In *CoNEXT '06: proceedings of the 2006 ACM CoNEXT conference*, New York, NY, USA.
29. IPP2P (2010). http://www.ipp2p.org.
30. Rapidshare—easy filehosting (2010). URL http://www.rapidshare.com.
31. Perenyi, M., Dang, T. D., Gefferth, A., & Molnar, S. (2006). Identification and analysis of peer-to-peer traffic. *Journal of Communications (JCM)*, *1*(7), 36–46.
32. Ploumidis, M., Papadapouli, M., & Karagiannis, T. (2007). Multilevel application-based traffic characterization in a large-scale wireless network. In *International symposium on a world of wireless, mobile and multimedia networks (WoWMoM)*, Helsinki, Finland.
33. Chlebus, E., & Divgi, G. (2007). The Pareto or truncated Pareto distribution? Measurement-based modeling of session traffic for Wi-Fi wireless Internet access. In *IEEE wireless communications and networking conference, 2007, WCNC 2007*, Hong Kong, China.
34. Divgi, G., & Chlebus, E. (2007). User and traffic characteristics of a commercial nationwide Wi-Fi hotspot network. In *IEEE 18th international symposium on personal, indoor and mobile radio communications, 2007, PIMRC 2007*, Athens, Greece.

35. Cisco Systems Inc. (2008). Cisco visual networking index—forecast and methodology, 2007–2012. White Paper.
36. Schulze, H., & Mochalski, K. (2009). Internet study 2008/2009. http://www.ipoque.com/resources/internet-studies/.

Florian Wamser studied in Wuerzburg, Germany and at the Helsinki University of Technology, Finland. He received his diploma degree in computer science from the Department of Computer Science of the University of Wuerzburg in 2009 where he is currently a research assistant. He is interested in wireless broadband access networks and related fields as well as cellular communication.

Rastin Pries graduated in computer science at the University of Wuerzburg, Germany. Since 2004 he is a research fellow at the chair of Prof. Phuoc Tran-Gia, working towards his Ph.D. Previously he was at academia in Wedel and Wuerzburg (Germany) as well as industries at Computer Partner (Hamburg, Germany) and Infosim (Dallas, USA). His research interests are performance analysis and optimization of broadband wireless access networks. Rastin Pries is involved in several industry projects and takes an active part in European Cost, Network of Excellences, and BMBF projects. He is currently coordinating the BMBF project "G-Lab" together with Prof. Phuoc Tran-Gia.

Dirk Staehle is Assistant Professor at the Chair of Distributed Systems at the University of Würzburg, Germany. Dirk Staehle has lead multiple industry co-operations in the field of GPRS, UMTS, and HSPA radio network planning with T-Mobile International, France Telecom R&D, and Vodafone Netherlands. His research interests include analytic modeling and simulation of wireless networks; radio network planning; (application layer aware) radio resource management; and source traffic modeling of wireless applications. He is currently working on cellular HSPA and OFDMA networks, as well as on mesh and sensor networks.

Klaus Heck is founder and CEO of the Hotzone GmbH, a (Wireless) Internet Service Provider located in Wuerzburg and Berlin, Germany. The Hotzone uses Wireless LAN, LAN and DSL as their access technology. Where applicable, full triple-play service is offered. Previously, Dr. Klaus Heck was a research fellow at the chair of Prof. Phuoc Tran-Gia at the University of Wuerzburg and has lead different industry cooperations. He worked as an external consultant at InfoSim Germany in Wuerzburg, Germany and as a consultant at InfoSim Inc., Dallas, TX.

Phuoc Tran-Gia is professor and director of the Institute of Computer Science at the University of Wuerzburg, Germany. Previously he was at academia in Stuttgart, Siegen (Germany) as well as industry at Alcatel (software development System 12), IBM Zurich Research Laboratory (Zurich, Switzerland, architecture and performance evaluation of communication networks). He is consultant and cooperation project leader with Siemens (ICN Board, Munich, ICM Berlin), Nortel (Texas), T-Mobile International (Bonn), France Telecom (Belfort), European Union (European Science Foundation, Brussels), and is coordinating the G-Lab project "National Platform for Future Internet Studies".