On traffic modelling in GPRS networks

Madsen, Tatiana Kozlova; Schwefel, Hans-Peter; Prasad, Ramjee; Bøgh, Jan; Hansen, Martin Bøgsted

Published in:
Proceedings of the Eight International Symposium on Wireless Personal Multimedia Communications (WPMC ' 05)

Publication date:
2005

Document Version
Publisher's PDF, also known as Version of record

On Traffic Modelling in GPRS Networks

Tatiana K. Madsen[†], Hans Peter Schwefel[†], Martin B. Hansen[‡], Jan Bøgh[♮], Ramjee Prasad[†]
[†]Dept. of Communication Technology, Center for TeleInfrastructure, Aalborg University, Denmark
Email: [tatiana—hps—prasad]@kom.aau.dk
[‡]Dept. of Mathematical Science, Aalborg University, Denmark
E-mail: [mbh@math.aau.dk]
[♮]Sonofon A/S, Denmark

***Abstract*— Optimal design and dimensioning of wireless data networks, such as GPRS, requires the knowledge of traffic characteristics of different data services. This paper presents an in-detail analysis of an IP-level traffic measurements taken in an operational GPRS network. The data measurements reported here are done at the Gi interface. The aim of this paper is to reveal some key statistics of GPRS data applications and to validate if the existing traffic models can adequately describe traffic volume and inter-arrival time distribution for different services. Additionally, we present a method of user session identification in case when only measurements on IP level are available.**

I. Introduction

The recent development in telecommunication networks has revealed two important tendencies. The first one is the increase in Internet traffic. The second one is the demand for wireless delivery of data services. The introduction of the General Packet Radio Service (GPRS) was an important step in converging wired and wireless communication. GPRS provides a packet-switched access service for IP traffic, integrated into the GSM architecture [1].

The performance analysis of new wireless data technologies, their optimization and network planning require information about the data traffic profile. While much measurement and analysis has been done for wireline networks [2], there is still very limited material available about wireless traffic models. It is questionable whether for wireless traffic modelling the existing models can be used and extrapolated. The use of wireless Internet services has different characteristics comparing with the traditional Internet since it is greatly influenced by mobility of a user, the access speed, the access client and the pricing. It is expected that the traffic characteristics (the parameter settings, but likely also distribution types) are strongly dependent on the used wireless access technologies. Different wireless networks will dictate different user behavior: e.g. measurements of WLAN traffic will reflect the specifics of this particular access technology, but it can not be used to model traffic in a wireless network of other kind. Therefore, a technology-specific measurements should be carried out.

This paper is dealing with the analysis of measured data traffic of a live GPRS network. GPRS networks are now fully in operation and measurements can be used to validate and update the existing traffic models. Even though the first GPRS network was deployed in 2000, there is still very limited measured data analysis reported in the literature. Nowadays these networks are mature and they carry traffic levels that allow reliable analysis of the measured data. One should also note that with time, data services become more and more popular, resulting in the changes in traffic characteristics. Thus, the analysis based on recent series of measurements is needed. Identification of traffic changes over long period of time based on the set of three independent measurements performed from 2002 to 2004 is presented in [3]. However, that work is considering only the aggregate workload statistics. GPRS service usage and traffic volumes are investigated in [4] based on measurements taken in two live GPRS networks. There are a few studies available describing particular services (e.g. WAP [5]). The aim of this paper is to investigate the application level activity of the subscribers and to identify models that suitably describe the statistics of different applications (including such dominant applications as WWW and e-mail).

Since only traces at the Gi interface were at our disposal, some assumptions should be made about user sessions taking into account the total amount of IP addresses and distribution of inactive periods. We introduce an approach to identify a user session based on the idle period threshold. The impact of the threshold choice on the data statistics is investigated.

The paper is organized as follows. Section II describes the data collection process. Section III elaborates on how to identify a user session. Sections IV and V present flow and service-specific analysis. Section VI offers some concluding remarks and outline for the future work.

II. Data Collection

Traffic traces can be obtained at appropriate network aggregation points, over a substantially long period of time. Arrival and departure process statistics can be deduced from the traffic traces, e.g. packet length and packet inter-arrival time distributions.

In GPRS networks traffic measurements can be either done at IP level (revealing information about application types and packet characteristics) or at the logical link-control (LLC) and radio link control (RLC) level (allowing analysis of mobility events). The measurements presented in this paper are done at the IP level. We captured IP packets at the Gi interface. The Gi interface connects the GPRS network to external IP based networks, such as the Internet or the operator's service network. The base for analysis is firewall data. Table I shows a sample of a firewall log. One row presents an information

for one firewall session for the same source- destination pair, containing information about time when the session began and session duration, what service, transfer protocol and source port were used, the IP address of the source, how many bytes were transferred, how many bytes were uploaded (client inbound bytes, CIB) and downloaded (client outbound bytes, COB) and how many packets were involved in the transmission. To preserve subscriber privacy the destination IP address was not available in the measurement data.

The data available for analysis on this paper consists of approx. 90 min firewall traces collected in March 2004.

III. User Session

A. Identification of user sessions

Since IP addresses are assigned to the subscribers temporally and the same IP address may be reused later (even within a few seconds), the IP address does not identify the user. Some additional information (the relevant GPRS signalling) is needed to sort out packets to separate Packet Data Protocol (PDP) contexts. In what follows, we call an *IP session* a period from the moment we have observed the first usage of a particular IP address until the last registered activity period with this IP address (during the whole interval of observations). Since one IP session may comprise of multiple activity periods of the same user or even of different users, we introduce a term *user session* to isolate different users and activity periods of one user.

It was observed that approx. half of the IP addresses from the available pool of the addresses appeared in the traces. Since nothing was known about the address assignment strategy and taking into account the fact that the used addresses were randomly distributed over the whole address interval, we could not conclude that one observed IP session would belong only to one user.

Each of the IP sessions consists of a number of activity periods, that we refer to as *flows*. Different flows that belong to one IP session can in principle overlap, this will correspond the case when different applications are active at the same time or when the same application initiates several data streams. At the same time, it was observed that 170 IP sessions consist of a single flow.

Without having a related measurements of a PDP context, we have decided to identify user sessions among sessions with the same IP address based on the distribution of inactivity periods. Figure 1 presents a histogram of duration of idle periods within IP sessions. It can be seen that even though the majority of idle periods are short, that is less than 10 min, a large number of IP sessions contains inactivity periods that are 30 min long and up to 1 hour long. Considering these long idle periods, it makes sense to treat a bursts of flows appearing after these idle times as a separate user session. Even if there was no physical change of user, the activities of a user with 30 min gap will be uncorrelated and they can be considered as two independent sessions. Therefore, a threshold value, T, should be found such that the activity periods of the same

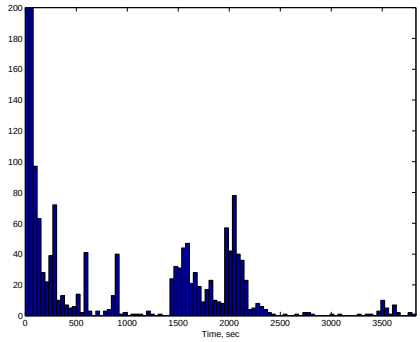

Fig. 1. Histogram of inactivity periods in IP sessions.

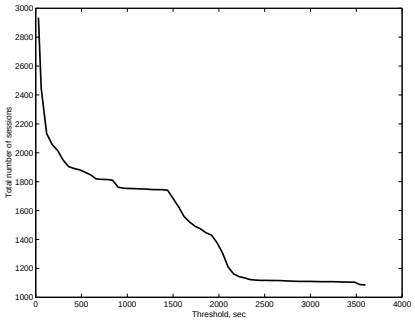

Fig. 2. Total number of user sessions as a function of the threshold value.

IP session are regarded as different user sessions if the idle period between them is larger than T.

The choice of the threshold T has a big influence on the user sessions' statistics. Figure 2 and 3 shows how the total number of sessions and the average number of flows per session depend on the threshold value. If the threshold is chosen small, then the statistics of the system is very sensitive to any change in this value. At the same time, there is a range of values (intervals $[600, 1500]$ and $[2200, 3500]$) where the lines on Figures 2 and 3 become almost constant. It is advisable to choose T exactly within those "unsensitive" intervals. In what follows, we have fixed the value of the threshold equal to 1200 sec.

To be sure that we do not split too many sessions by introducing a too small threshold, we perform the following

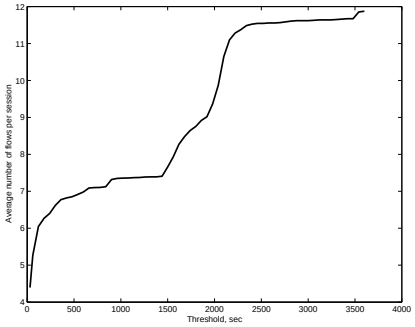

Fig. 3. Average number of flows per session as a function of the threshold value.

Time	Service	Source	Protocol	Source port	Elapsed	Bytes	CIB	COB	Packets
09:25:32	dom.udp	212.88.73.106	UDP	49192	00:00:00	138	61	77	2
09:25:32	pop-3	212.88.73.106	TCP	1179	00:01:33	1048	439	609	19
09:51:40	http	212.88.73.107	TCP	1036	00:00:04	4464	3528	936	11

TABLE I

A SAMPLE FROM A FIREWALL-LOG

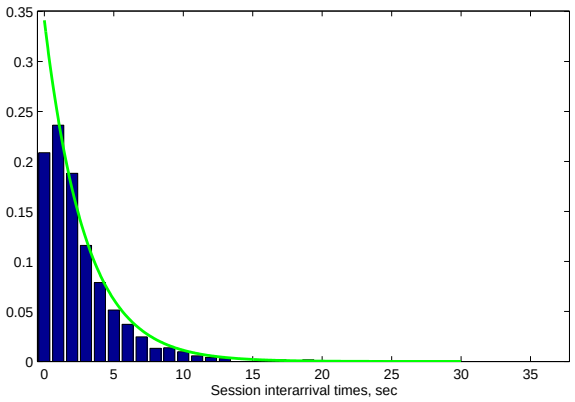

Fig. 4. Histogram of interarrival times of user sessions (smaller than 40 sec) with exponential fit.

test. We verify if the arrival time process for the obtained user sessions resembles a Poisson process as we would expect for a large number of independent events. Figure 4 presents a histogram of the corresponding interarrival process with the exponential fit.

It should be noted that there exist large interarrival times, up to 10 min (not shown on Figure 4), something we would not expect if distribution follows strictly the exponential distribution. These big interarrival times do not disappear even if we will decrease the value for the threshold. Further investigations are required.

B. Window censoring

Due to the data collection mechanism, the following problem appears regarding IP sessions start and end time. When we observe a flow with a particular IP address, we can not say if it is the first flow of this IP session or if the session has started in the past and this is a next activity period after an idle period. Since none of the flows that have started before our initial observation point is recorded in our data, the flow interarrival times can not be estimated. Introducing the threshold approach does not help to combat this problem. The problem stays unless a user session starts later than the threshold value T after we begin to record the data. The similar phenomena appears also at the end of the data record: if the last time interval between the last flow and the end of the study is less than the threshold value, we do not know if this bit belongs to the last IP-session or a new IP-idle time. The above mentioned problem with left- and right truncation of the data is known in the literature as *window censoring*.

Let consider IP-sessions as a process that jumps back and forth between two states of being busy (transmitting flows) and idle (waiting for a new IP assignment) and suppose the durations of subsequent busy and idle times are i.i.d. and that the process has started in the far past, so it has achieved stationarity. Such alternating renewal processes (ARP) have been taken as models in a variety of contexts such as systems reliability in engineering or the behaviour of healthy-sick cycles in actuarial and insurance mathematics.

For ARPs window censoring problem is often present in one or another form. Efficient methods for dealing with censoring have only recently been proposed. A study which resembles the present problem can be found in [6]. It tackles a similar problem but this problem does not match exactly the phenomenon we have observed. In that paper the status of the first waiting and last waiting times are know as opposite to our case, but it is possible to modify the procedure for the problem at hand. This and other related issues will be touched upon in a separate paper.

IV. FLOW ANALYSIS

In the previous section we have explained that due to specifics of data collection process a window censoring problem occurs when we consider IP session process. At the same time, if we consider an aggregation of all flows, we are avoiding this problem. Since we are recording all flows that have started within a specified interval, the flows that started before the first observation point, t_{start}, will not be collected even if they ended after t_{start}; but all flows with the starting point before the last observation moment t_{end} are recorded regardless how long they last.

In this section we are looking at the flow properties, especially at such phenomena as heavy tails. Long range dependence and heavy tails cannot be ignored if estimation of network capacity is the goal [7]. These phenomena have received some attention (see e.g. [8]) and it is expected that heavy tails will be of crucial importance for a wide range of network engineering problems.

To help with assessing whether heavy tails are present and to estimate the index α (loosely speaking, the reliability function drops off as a power of x, $R(x) \sim c/x^{-\alpha}$ for large x), various exploratory plotting techniques are available. Following one of the approaches, we plot the reliability function of flow volumes on log-log scale (see Figure 5). The tail shows the straight-line behavior with negative slope $\alpha \approx 1.1$. The same type of behavior can be observed for flow durations (Figure 6). In this case the slope is approx. 1.2. The small difference in parameter α for these two distribution can be explained by

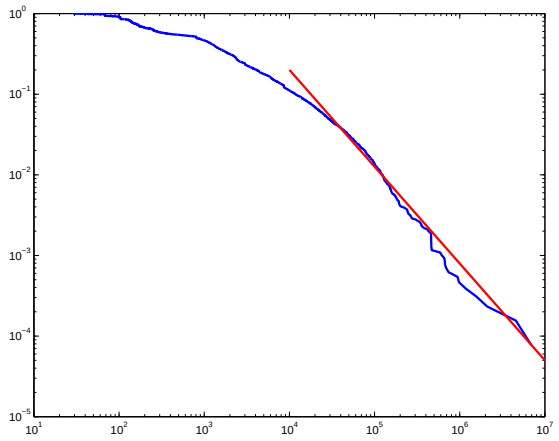

Fig. 5. Indication for heavy-tail distribution of flow volumes.

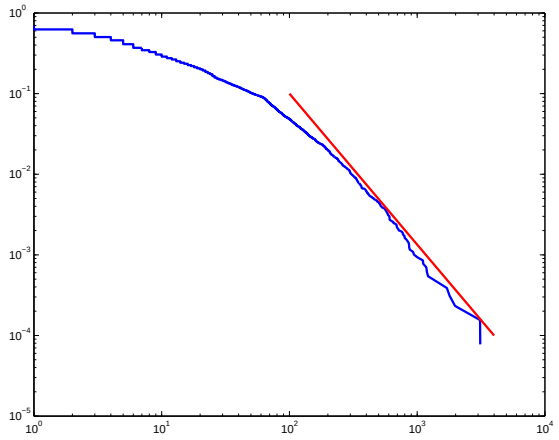

Fig. 6. Indication for heavy-tail distribution of flow duration.

the difference of the connection speed for different users. This can lead to differences in flow duration even for the same flow volumes. What is more, usually users with a good connection tend to download bigger files, but this will not take longer time. Therefore, in the collected data we observe bigger number of ”heavy” volumes compared with ”heavy” durations.

We would like to mention that in [3] it was concluded that the Poisson Pareto Burst Process (PPBP) model provides an accurate model of the aggregated GPRS traffic. It is in correspondence with our observations of heavy tail behavior of flow volumes and durations.

V. Service-specific analysis

This section presents in-detailed analysis of different applications usage in a GPRS network. We start with providing an overall picture of network services and later concentrate on two dominant applications, namely web browsing and email service.

A. Distribution of Services

Table II summarizes the information about different service types observed in the measured data. The total of 18 services were counted. Web browsing and email service constitute a large proportion of both the total number of flows and the total traffic volume.

Service type	No of flow ratio, %	Volume ratio, %
http	31	51
https	10	15
e-mail (pop-3, imap, smtp)	11	19
WAP	6	9
ftp	1	2
DNS	33	1
other	8	19

TABLE II
SERVICE TYPES DISTRIBUTION

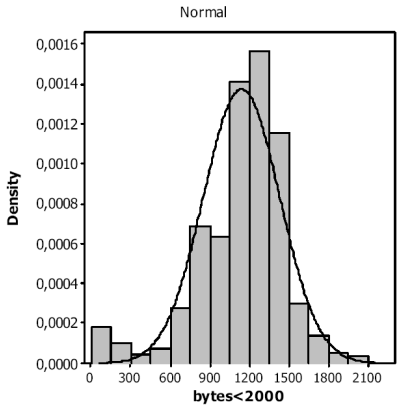

Fig. 7. Histogram for transfer size of email service smaller than 2000 bytes with a normal distribution fit.

As expected from the properties of Domain Name System (DNS), DNS request and response occurs frequently, thus it has a large share in the total number of flows. But since only files of small sizes are transferred, this kind of service occupies only 1% of the total traffic volume.

B. Email service

Table III shows summary statistics for e-mail service. The data is investigated as two distributions, small transfer size (< 2000 bytes) and large transfer size (> 2000). A histogram for small transfer size is fitted with a normal distribution (see Figure 7).

Observing interarrival times of different sessions, it was concluded that they are independent: from Figure 8 we can see that the autocorrelations are near zero for all time-lag separations and no one is significantly non-zero. Exponential

Mean	17331.58
Standard deviation	152918.7
Sample variance	2,34E+10
Minimum	44
Maximum	4571428
Number of sessions	1288

TABLE III
E-MAIL SERVICE (BYTES)

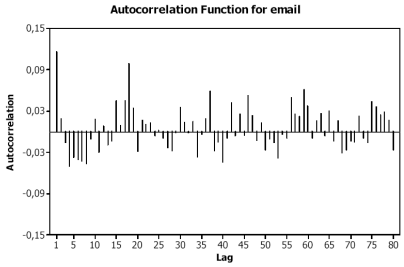

Fig. 8. Autocorrelation function for interarrival time of email sessions.

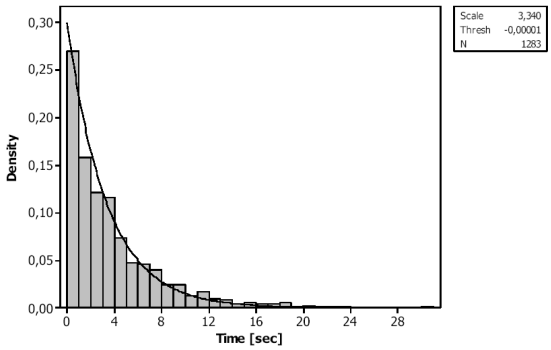

Fig. 9. Histogram of e-mail session interarrival times with the exponential distribution fit.

distribution provides a good fit for the histogram (see Figure 9) and Poisson distribution for counts.

C. WWW service

In this section we present analysis of flows that contain only http or https traffic. The distribution of www volumes can be approximated by exponential distribution (see Figure 10). Table IV presents the summary of the www flow volume statistics. As for the case of email traffic, interarrival times can be approximated by the exponential distribution.

The complete histograms of transfer sizes for both www and email gives an indication of the presence of heavy tails. This is in correspondence with our findings of heavy-tailed behavior of the aggregated traffic volumes (Section III.C).

VI. Conclusion and Outlook

This paper presents an analysis of an IP-level traffic measurements taken in an operational GPRS network. We address the issue of a user session identification when only traces at the IP level are at the disposal and no correlated measurements of PDP contexts are available. A discussion on heavy-tailed traffic characteristics for flow volumes and duration is provided. As a dominant applications, examples of www and email traffic are considered. A visual fit of known distributions to the measured data is presented.

Mean	15368.02
Standard deviation	1.785E+05
Sample variance	3.186E+10
Minimum	40
Maximum	10147134
Number of sessions	4938

TABLE IV
WWW SERVICE (BYTES)

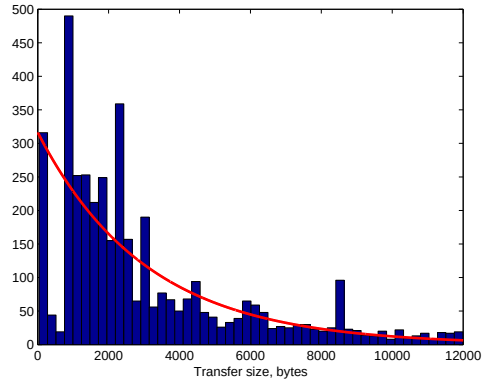

Fig. 10. Histogram for transfer size of www traffic smaller than 12000 bytes with an exponential distribution fit.

Furthermore, discussion about window censoring problem is provided. This problem occurs due to the data collection method and some corrections should be introduced in the estimation of distribution parameters. How to find the necessary corrections will be presented in our future work. It is our future plans to make correlated measurements on RLC/LLC layer.

VII. Acknowledgements

This work has been supported by Center for Network and Service Convergence, Denmark. The authors acknowledge Sonofon A/S for permission to publish this paper.

The authors would like to thank Aki Sigurdsson, Hreidar Joelsson, Hui Zhao, Kjartan Jonsson, and Valur Porsson for their help with the initial data analysis.

References

[1] 3GPP, General Packet Radio Service Description, TS 03.60 V7.8.0, January 2002.
[2] Leland W.E., Taqqu M.S., Willinger W. and Wilson D.V., "On the Self-similar Nature of Ethernet Traffic", IEEE/ACM Transactions on Networking, Vol.2, No.1, February 1994.
[3] Ivanovich M., Li. J., Neame T. and Fitzpatrick P., "Modelling GPRS Data Traffic", In *Proceeding of Globecom*, 2004.
[4] Kalden R., Varga T., Wouters B. and Sanders B., "Wireless Sevice Usage and Traffic Characteristics in GPRS Networks", In *Proceeding of the 18th International Teletraffic Congress*, 2003.
[5] Nieminen T. and Halme S.J., "An Analysis of WAP Packet Traffic Measurenments", In *Proceeding of the 16th Nordic Teletraffic Seminar*, 2002.
[6] Enrique E. Alvarez, E.E. 2005. Asymptotic Efficiency in Censored Alternating Renewal Processes. *Proceedings of ASMDA 2005*, Brest, France.
[7] H.-P. Schwefel, L. Lipsky, Impact of aggregated, self-similar ON/OFF traffic on delay in stationary queueing models, Performance Evaluation 43, pp. 203-221, 2001.
[8] S. Resnick, Heavy Tails Modeling and Teletraffic Data, The Annals of Statistics, Vol. 25, No. 5, 1997, pp. 1805-1869.