© Urban & Fischer Verlag
http://www.urbanfischer.de/journals/aeue

AEÜ International Journal of Electronics and Communications

Source Traffic Modeling of Wireless Applications

Phuoc Tran-Gia, Dirk Staehle and Kenji Leibnitz

Herrn Professor Paul J. Kühn zum 60. Geburtstag gewidmet

Abstract **Data applications are expected to play an increasingly important role in the next generation mobile communication services. To plan these networks, detailed models of data users with diverse applications are required. The purpose of this study is to provide practically usable traffic models for data users in wireless networks, e.g. GPRS. We start with a discussion of well-known models in the literature, where the focus is on their applicability in mobile networks. A practical model of mobile HTTP users, based on an extrapolation of WWW users in current wire-line networks, is presented; this model is expected to be accurate enough to be used in simulation studies for network planning purposes. Finally, models for other emerging data applications are also taken into account.**

Keywords **Source modeling, Wireless networks, WWW, GPRS, Internet access**

1. Introduction

Selecting the appropriate source traffic model to reflect the behavior of the users in a telecommunication system is an important issue in order to perform a successful design of new networks. Especially networks providing services for multiple classes of users, e.g. the General Packet Radio Service (GPRS) [1] for the Global System of Mobile Communications (GSM), require a detailed characterization of the carried traffic to operate efficiently since the underlying radio frequencies are scarce resources. Unlike existing GSM systems that primarily serve voice users and to some extent simple facsimile or short message services (SMS), GPRS supports a wide range of variable-bit-rate applications with high bandwidth efficiency.

Teletraffic engineering has been used for a long time to dimension telephone networks. Similar to the more sophisticated models described later, the voice users can be characterized in a layered structure. The most macroscopic behavior describes the activity of the user from the time of a connection setup until call termination and consists of the mean interarrival time between connection setups, together with the mean connection duration. In conventional systems and classical telephone networks, it has been widely accepted to use exponential distributions for both parameters resulting in a Poisson process description for voice users. Within a call, the user will create talk spurts corresponding to the time when he actively talks followed by periods of inactivity while listening to his counterpart. Since the activity phases continuously occupy the channel, the traffic generated by this type of users is usually characterized by an ON/OFF Process. Parameters for this type of traffic can be obtained from empirical measurements.

While the characterization of voice users is straightforward, the traffic generated by data users is highly dependent on the application and has a high burstiness, i.e. the variance of the interarrival times between data packets as well as the variance of the packet length can be very high. Therefore, due to the mostly feedback-oriented nature of packet switched data applications the simple Poisson model is no longer sufficient. In [2] measurements in an office automation environment of an ISDN network were conducted and a canonical model of the packet arrival process by a single user was constructed. This model captured the correlation in the interarrival time distribution of the packet streams. Furthermore, this study showed that for applications requiring interactions (e.g. waiting time for host responses, thinking time, etc.) long range dependence is observed and could be well fitted by a Pareto distribution. However, since the Pareto distribution can have an infinite mean it also shows that the standard summary statistics such as the sample mean and standard deviation are often unusable in such cases.

Recently, a lot of attention has been paid to long range dependent or self-similar models of traffic streams in telecommunication networks [3, 4]. Many of such models have been used to investigate variable-bit-rate video sources [5] with a statistical analysis of empirical sequences and estimation of the grade of self-similarity in terms of the Hurst parameter. Since MPEG video traffic consists of a highly correlated sequence of images due to its encoding, the correct modeling of the correlation structure of the video streams is essential.

For Internet traffic, analytical models were derived in [6] for telnet, NNTP, SMTP, and FTP from empirical traces. It was found that wide-area traffic is difficult to model exactly in a statistical sense, but simple and accurate analytical models can be constructed that serve as good approximations. An extension to this work is found in [7]. Here, it is shown that user-initiated TCP connection arrivals (e.g. remote login, file transfers) can be modeled as Poisson; however, packet interarrivals within such sessions are significantly more bursty. This leads to an underestimation of performance measures, like average packet delay or maximum queue size.

Received October 1, 2000. Revised November 30, 2000.

Phuoc Tran-Gia, Dirk Staehle and Kenji Leibnitz, Lehrstuhl für Informatik III, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany.
E-mail: {trangia|staehle|leibnitz}@informatik.uni-wuerzburg.de

A more recent study was done in [8], where also a self-similar relationship was observed for local and wide area networks. The heavy-tailed file size distributions and the operation of the reliable transmission and flow control mechanism of TCP cause the long range dependence structure. In contrast, non-flow-controlled and unreliable UDP streams show very little long-range dependence. It was also shown that the performance in terms of queuing delay is influenced by the self-similarity, while throughput-related measures such as packet loss and retransmission rate are left mainly unaffected, provided that a reliable, flow-controlled transport protocol is used. Apart from the application, the user behavior is also influenced by external sources, such as the tariff structure [9] within the network. If the service is expensive fewer customers will be ready to pay for the service or limit their access to a shorter connection duration. Additionally, due to the rate adaptation mechanism of TCP, the transmitted traffic amount also depends on the load of the network and the processing speed of the contacted server. Thus, the behavior of the customer is feedback-oriented and is not only determined by his application but also depends on external influences that are difficult to foresee.

In this paper we will provide practically usable traffic models for GPRS data users which will probably be the most frequent type of wireless data user in the near future. We will start with a discussion of well-known models in the literature that are based on wire-line HTTP users. By observing the commonalities of the proposed models we will give a general description of a data user and extrapolate this behavior to the operation of a WWW browser in a wireless environment. Due to the flexibility of the recommended model, it is easy to obtain the parameters for existing applications by fitting them to measurements. This also facilitates the extension of the model to emerging data applications.

The remainder of this paper is organized as follows. In Section 2, we investigate the literature on source traffic models of HTTP users. We will review the proposed models briefly and particularly the work in [10] in more detail, which we will use as the basis for our proposed model. Furthermore, we will describe some references on characterization of some important applications (E-Mail, FTP, WAP) expected to be widely used in GPRS systems. Additionally to presenting published work, we will also compare the values to those from measurements in our network environment. This is followed in Section 3 by the presentation of our recommended models and its parameters. The paper is concluded in Section 4, followed by a detailed list of all considered references.

2. Source traffic models of major applications

2.1 General remarks

In the introduction the former attempts to characterize TCP traffic as one common traffic stream of different applications were mentioned. However, if we look at different applications like WWW or E-Mail, traffic streams with distinct characteristics are generated. Vicari [9] investigated the behavior of dial-in users at the University of Würzburg. Figure 1 shows the traffic share of the most important applications for access speeds between 9.6 kbps and 64 kbps.

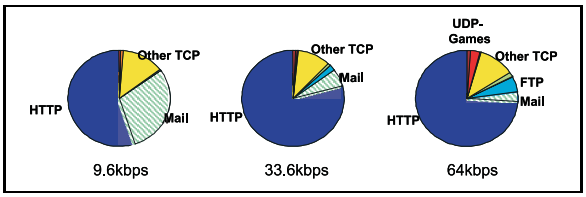

Fig. 1. Traffic mixture of dial-in users with different access speeds.

To predict the future traffic mixture of GPRS certain aspects should be considered:

1. A GPRS user will have a load-dependent bandwidth. During the high-load period he will have an access speed in the range of a 9.6 kbps modem or even worse. In the low-load period the access speed could approach that of a 64 kbps ISDN modem.
2. The billing in GPRS will influence the length and frequency of user activity.
3. The applications will be different for a user working with a notebook in contrast to one with a WAP handset.

All these points will influence the composition of the applications of a future GPRS user. Due to lack of better information the average traffic mixture of a GPRS user is supposed to be similar to that of a 33.6 kbps user. The main part of roughly 75 % belongs to web traffic. With GPRS and the introduction of WAP-handsets a part of this traffic will change towards WAP. The other important application is E-Mail with a share of about 10 %. The last service with remarkable volume is FTP. The part labeled "Other TCP" indicates different services like games, telnet, etc. with a share of less than one percent each. However, the usage of different games might increase in future while neither the amount nor the characteristics of the traffic they produce is predictable. The proportion of the other minor services is assumed to diminish in future. Thus, this paper is focused on the modeling of WWW and E-Mail traffic sources. Furthermore, a presumption about the behavior of FTP and WAP users is made.

A source traffic model consists of two parts: the arrival process for user activities and the process describing the activity phase. The models for the activity phase will be described in the following sections individually for the different services. The modeling approaches for the arrival process are outlined here since they are identical for the various applications. The arrival process determines the instants when a user starts his activity. Such instants are denoted as arrival times. If WWW users are considered, the instant a user begins his web browsing session is the arrival time of the HTTP application. The arrival times can be described either by the distribution of the interar-

rival time, i.e. the time between two consecutive arrivals, or by the distribution of the length of the OFF phase, i.e. the time between the end of a user's activity and the start of the next activity phase of the same user. The approach using the interarrival time is suitable if a large number of users can be assumed and the length of the activity phase forms only a fraction of the length of the OFF phase. The most frequently used arrival process is the Poisson process with an exponentially distributed interarrival time which was already mentioned in the introduction with regards to the arrivals of voice users. In the literature this approach is used e.g. in [11–13]. If the arrival times are determined by the length of the OFF phase of a single user the whole model is referred to as an ON/OFF model. Among others [10, 14, 15] used an ON/OFF model to describe WWW traffic.

2.2 Web traffic

Web traffic is nowadays the most important application used by the Internet community. The term web traffic comprises all HTTP traffic generated during a session with a typical web browser like the Netscape Navigator or the Microsoft Internet Explorer. However, not all of this traffic is transported in the same way. There are differences both between the different versions of HTTP and between the different browsers. These versions affect the way a web page is downloaded.

In order to understand the different methods for page transmissions the basic method and the structure of a web page have to be explained. A web page consists of ASCII text, the HTML code. This part of the page is referred to as the main object in the remainder of the paper. In the HTML code, images or other objects like Java classes may be embedded with a reference to a separate file on the server. We will use the term inline objects for them, henceforth. The downloading procedure of a WWW page starts when the user clicks on a link referencing this page. Then a GetRequest is sent to the URL (Universal Resource Locator) indicated by the link. To transmit this request, a TCP connection has to be established between the user and the web server. After receiving the request at the web server, the main object is sent back to the browser. There, the HTML code is parsed for the occurrence of inline objects which have not yet been transmitted.

At this point the different versions of HTTP come into account. The first and least efficient method is to establish an own TCP connection for each inline object and to transmit all objects successively. Later versions improved the performance by downloading several inline objects in the same TCP connection and through the establishment of several TCP connections in parallel. Furthermore, it is possible to keep a TCP connection alive even after the requested web page is completely transmitted. This is reasonable as the next requested page might reside at the same server. So for the GetRequest no new TCP connection has to be established, but the existing one is used. Due to the Slow Start mechanism of TCP a significantly better performance is achieved. A more detailed description of the functionality of the different HTTP versions can be found in [16–18].

The complex functionality of HTTP requires to model a WWW page in more detail than only by the distribution of its size. For an adequate simulation the structure of a page has to be known as well as whether the page is located on the same web server as the previous one.

2.2.1 Overview of existing web traffic models

The problem to measure HTTP traffic and to design an appropriate model has been a challenge for several years. Two methods are popular to measure the web traffic directly, server logs and client logs. Another method is to measure all traffic flowing through a LAN and to extract the WWW traffic.

The first method, the server logs, is based on the property of the web servers to keep track of the files they served. With this information it is possible to create a workload model [19] of a server. However, the behavior of a single user is not seen and consequently from such data no appropriate model can be derived for his behavior.

The second technique, the client logs, relies on data gathered by using the NCSA Mosaic browser to keep track of all retrievals made during web user sessions. Such measurements were conducted by Cunha [20] and Catledge [21]. These studies investigated the characteristics of web accesses. With this data a corresponding model can be constructed to describe the behavior of a WWW user. However, these studies were all performed around 1995 and due to the fast changes in the Internet, they are not up to date any more. The difficulty to perform such a measurement nowadays is that the source code of present browsers is proprietary. Another problem is to equip a sufficiently large number of "typical" WWW users with a modified browser.

The third strategy, the packet traces, is the most popular one in the last years. The measurements are performed by taking packet traces from a subnet carrying HTTP traffic, typically an Ethernet or other shared-media LAN. From this data and under consideration of higher-layer protocols the traffic analysis can produce a model of the original application. This approach was used e.g. in [13, 14, 22, 23]. Vicari [9] and Färber [24] collected packet traces from dial-in routers. This third approach eliminates the principal disadvantages of the two previous methods, so a detailed study of models corresponding to measurements performed with the third method is given and only one model derived from client traces is discussed. Since this paper focuses on user behavior models measurements of server logs are neglected.

The model resulting from client logs will be described first. It is based on the measurements performed by Cunha [20]. Barford and Crovella used the results of these client traces to implement a workload generator called SURGE [15]. The model takes into account a number of rele-

vant traffic parameters: i) size of objects on the server, ii) size of requested objects and their popularity, i.e. hit rate, iii) number of inline objects and locality, and iv) viewing time. The browser considered in this model was a modified version of the "Mosaic" browser, which doesn't play an important role in the current web landscape. The statistics are based on a population of 37 clients, showing interesting evidence concerning the self-similarity of traffic; however, the observed population is rather small and the modification of the browser causes difficulties to generalize the results to model present web traffic users.

From the models derived from packet trace measurements we will concentrate on Deng [14], Mah [22], Vicari [23], Färber [24], Reyes-Lecuona [13], and Choi [10]. These papers can be distinguished in several ways:

- Färber measured the behavior of dial-in users whereas the other papers rely on packet traces taken in a LAN.
- Vicari, Färber, and Reyes-Lecuona distinguished their traces into sessions/subsessions and proposed distributions for both the interarrival times of sessions and the duration or "size" of a session. Deng, Mah, and Choi either did not split their traces into sessions or gave no distribution describing the session level, but only the composition of an "unlimited" session.
- Furthermore, the papers differ in the way the start and the end of a web page was detected inside the packet stream. In all papers except [10] the start of a new page was assumed after a certain gap, i.e. time without packets before the start of a new TCP connection. Choi, however, discerned a new web page via the extension of the requested file. Thus, Choi was the only one to distinguish between two consecutive pages when the download process was interrupted by the request of a new page.

A detailed description of the models and distributions of the three papers dealing with the session level is outlined first.

In [23] the term subsession is introduced. It describes the time period during which a user does actively serve in the Internet. Vicari found that such a subsession comprises 19.6 web pages on average and each of these pages consists of four separate files on average, i.e. the main object and three inline objects. A distribution of the object size is not given. The size of the complete web document and the time between web requests are modeled with normalized Pareto distributions. Färber et al. [24] suggested a different model. They assumed aggregated traffic, i.e. superposition of the traffic produced by many users, and compared both the measured time interval between session starts and the measured duration of a session with selected distributions.

Reyes-Lecuona et al. [13] provided a more detailed model consisting of three levels, the session level, the page level, and the packet level. The session level was described by the session interarrival time, modeled by a Poisson process, and the number of pages downloaded per session which was found to be Lognormal distributed. The model at page level was defined by the distributions

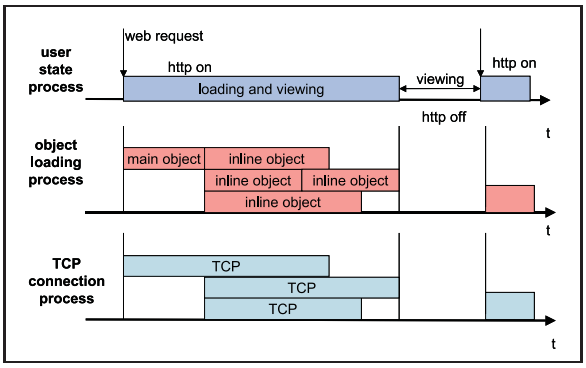

Fig. 2. Behavioral user model of Choi.

for two parameters, the page size and the time between consecutive pages. In their model the structure of a WWW page was not considered in more detail. The packet level contains two parameters, the packet interarrival time and the packet size.

In the other three papers [10, 14, 22] an ON/OFF model is used. The ON phase is the time period during which a page is downloaded. The OFF phase is the time between the end of a download process of one page and the start of the download of the following page.

Deng [14] developed the simplest ON/OFF model. The ON phase is the time during the download process of a web document. The OFF phase is the time between two download processes while the user views the page. Deng proposed distributions for three parameters: the duration of the ON period was found to be Pareto distributed, the duration of the OFF period Weibull distributed as well as the interarrival time of web requests during the ON phase. The values regarding the size of a document were taken from [25]. There are two major drawbacks in Deng's model. First, the network load is already included in the distribution of the ON phase and second, the document structure was not considered in the model. So, the behavior of a user with a certain HTTP version can not be accurately simulated.

The model of Mah [22] is more sophisticated. Like Deng he describes an ON and an OFF phase. However, in contrast to Deng, the OFF phase is not only the time a user views the downloaded page, but also the time between two sessions. The ON phase corresponds to the downloading time of a web document. Mah proposes distributions for the size of a web document as well as for the size of main objects and inline objects. However, the number of files per web document is only given by the mean and the median. Furthermore, Mah measured the size of GetRequests which are matched best by a bimodal distribution. An interesting feature investigated by Mah is the number of consecutive documents requested from the same server. Beside the absence of a distribution for the number of inline objects two major drawbacks regarding Mah's model exist.

In his measurement he did not consider interruptions of downloads with new page requests. The request for

a new page would be taken into account as an inline object of the previous, interrupted page. Thus, both the user think time and the inline object size distribution show a certain deficit. Furthermore, Mah's paper is not intended to give distributions for the different parameters of his model. Instead, the probability distributions are represented by their measured CDFs, which can also be used in simulation.

In [10], Choi and Limb presented a comprehensive model for web traffic, called the "behavioral model". The model is of hierarchical nature like ATM source traffic models presented numerously in the high-speed network literature. Since a slightly modified model arising out of Choi's approach will play a major role in our proposed model, we will describe the behavioral model in more detail. The model is shown in Figure 2. Basically, the activity of a user browsing the web is split into the following traffic layers:

Web request: the arrival process of web requests is observed on this layer. The model of a web user is based on an ON/OFF type process, where the ON state represents the activity after accepting a web request and the OFF state represents a silence period after all objects in a web request are retrieved. The duration of the ON state (HTTP ON) and the OFF state (HTTP OFF) correspond to the page loading time and viewing time, respectively. The web user can thus be in the following states:

- Loading and Viewing (HTTP ON): Time for fetching all objects belonging to one web request. The user is waiting for objects to be loaded; he can start viewing the objects while waiting. Alternatively, the user can interrupt the loading process (e.g. by clicking on appearing links) and thus start another web request.
- Viewing (HTTP OFF): Viewing time is the inactive interval between two web requests. A new web request is considered to be generated immediately after the expiration of the viewing period.

Objects to load: In this layer a detailed model for the web document structure is given similar to the one contained in Mah's work. The size of both the main objects and the inline objects was found to be Lognormal distributed. The improvement in contrast to Mah's models is that additionally a distribution of the number of inline objects per web page is derived. Furthermore, Choi characterized the size of GetRequests with a Lognormal distribution.

TCP connection activation and deactivation: In the ON state several TCP connections are activated and used to deliver objects. Depending on the browser version they can be activated in serial or in parallel. Parallel TCP connections for inline objects are opened consecutively after the single connection for the main object. The number of TCP connections activated depends on the version of HTTP considered, of which four are taken into account. In the early version 1.0 objects are downloaded back-to-back; each corresponds to one TCP connection. In the modified version 1.0 with multiple connections, the browser opens multiple TCP connections to download objects in parallel. The number of TCP connections which can be simultaneously activated is however limited. In version 1.0 with keep-alive, the deactivation of connections is not done immediately; the connection is kept alive for a short period to wait for possible connection requests. In the recent version 1.1, persistent connections are allowed and requests can be pipelined.

Table 1. Overview of WWW models

literature source	main object size	inline object size	number of inline objects
Choi [10]	Lognormal mean: 10 kB med.: 6 kB $\sigma = 25$ kB	Lognormal mean: 7.7 kB med.: 2 kB $\sigma = 126$ kB	Gamma mean: 5.5 med.: 2 $\sigma = 11.4$
Mah [22]	Pareto $\alpha = 0.85$–0.97 med.: 2–2.4 kB	Pareto $\alpha = 1.12$–1.39 med.: 1.2–2 kB	— mean: 2.8–3.2 med.: 1
Barford [15]	Pareto $\alpha = 1$ $k = 1000$		Pareto $\alpha = 2.43$ $k = 1$
Reyes-Lecuona [13]	page size: Pareto $\alpha = 1.5$–1.7 $\beta = 30000$–31000		
cdma2000 [12]	packets per page: Pareto $\alpha = 1.1$, $k = 2.27$		
UMTS [11]	page size: Pareto (truncated at 66kB) $\alpha = 1.1$, $k = 81.5$		

literature source	viewing time	inter session/ request time	pages per session
Choi [10]	Weibull mean: 39.5 s med.: 11 s $\sigma = 92.6$ s	—	—
Mah [22]	mean:1000–1900 s med.: 15 s		—
Barford [15]	Weibull $\alpha = 1.46$ $\beta = 0.38$	Pareto $\alpha = 1.5$ $k = 1$	—
Reyes-Lecuona [13]	Gamma mean: 25–35 s $\sigma = 133$–147 s	Poisson Arrival process	Lognormal mean: 22–25 $\sigma = 78$–166
cdma2000 [12]	geometric mean: 120 s	Poisson arrival	geometric mean: 5
UMTS [11]	geometric mean: 412 s	Poisson arrival	geometric mean: 5

The section on web traffic models is completed by two tables summarizing the results of the presented publications. Additionally, the WWW part of the models recommended by the ETSI and the ITU further described in Section 2.6 are also included. Table 1 contains both the distributions defining the structure of the web page and the distributions concerning web sessions.

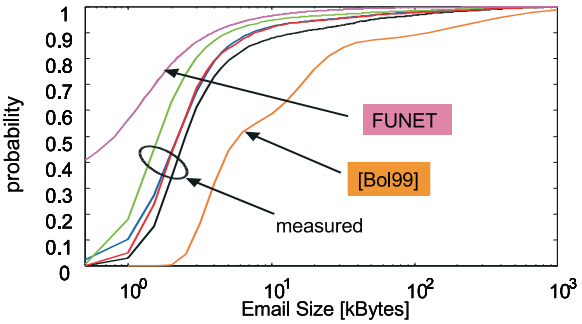

Fig. 3. Cumulative distributions of E-Mail sizes.

Table 2. Values of E-Mail size measurements

Source	number of samples	mean	standard deviation
user1	12110	17.6 kB	174.8 kB
user2	7026	28.2 kB	234.9 kB
user3	6787	30.1 kB	209.6 kB
all users	34670	22.7 kB	200.3 kB
FUNET	—	2.7 kB	17.7 kB
Bolotin	1760	77.9 kB	337.6 kB

2.3 E-Mail

Beside HTTP and future WAP, the most important application will probably be E-Mail. This is also confirmed by the measurements of Vicari shown in Figure 1. In literature, models of E-Mail usage are not studied as extensively as the behavior of WWW users. Bolotin et al. [26] provided the distribution of the E-Mail size. However, there is no model given for the frequency of E-Mail usage. Another characterization of the E-Mail size is given in the FUNET model [1] which was derived from E-Mail measurements performed at the University of Helsinki. In this model the E-Mail size is approximated by a Cauchy distribution function with $\alpha = 0.8$ and $\beta = 1.0$. It is defined by the density function $f(x) = 1/(\pi\beta(1+(\frac{x-\alpha}{\beta})^2))$.

The number of references regarding E-Mail traffic is quite small and additionally the given distributions differ considerably. Therefore, the E-Mail collection of sample users at the University of Würzburg was also evaluated. Figure 3 shows the CDFs (cumulative distribution functions) obtained by the measurements together with the CDFs corresponding to the above sources. The three grey curves belong to the E-Mail archive of sample users with more than 5000 E-Mails each. The detailed numbers are given in Table 2. The mean values and the standard deviations of the plotted curves are also listed there. The black graph in Figure 3 is the CDF of a collection of about 34000 E-Mails at the University of Würzburg including those belonging to the three other CDFs. The CDFs gained from measurements are located between the curves of the FUNET model and the model of Bolotin. For larger E-Mails the Cauchy model approaches the measured values, whereas for small E-Mails the probabilities are overvalued. Obviously an E-Mail has a minimum size of some hundred bytes containing the header. The CDF belonging to Bolotin's model overestimates the weight of large E-Mails. All CDFs of the Würzburg measurement show a similar shape while the averages and standard deviations are different. However, this feature was expected due to the different behavior of various users.

2.4 FTP

As indicated in the diagram shown in Figure 1 we notice that another important service in wire-line environments is the file transfer protocol (FTP). However, in recent years the importance of FTP is decreasing with the upcoming of HTTP downloads. Particularly, in networks with wireless access the usage of FTP is expected to diminish.

For this reason the references in the literature regarding FTP user models are all rather old. Thus, only the three most important papers describing those models are mentioned here. In [6, 27, 28] distributions for the size of FTP bulk transfers are given either graphically or with detailed parameters.

2.5 WAP

The Wireless Application Protocol (WAP) is a new specification to merge wireless data and Internet technologies [29]. Contrary to most technology developed for the Internet which assumes large powerful computers, WAP aims at extending WWW standards for optimised use in wireless data networks. The major limitations are here:

- Less computing power and memory.
- Smaller displays and limited input facilities.
- Narrowband and unpredictable network connections.

The transmission between a Wireless Application Environment (WAE) user agent and the origin server is performed via WAP proxies. The server can be a WWW server that provides HTML information or a specific WAP server supplying WML pages. At the WAP proxies the information is translated from its textual form to a tokenized binary WML encoding, which can contain the data of the WML/HTML pages and compiled programs in WMLScript (similar to JavaScript). The information is then displayed according to the capabilities of the user agent. If the browser supports images, they will be downloaded, otherwise an alternative representation is used.

The functional areas of WAP are either text presentations with basic formatting and layout or more commonly the deck/card mechanism. The deck can be considered as a web page which contains several cards that can perform certain actions, like a choice menu, navigation and linking between cards, etc. In both cases, the behavior will be similar to that of an HTTP session. The main difference

will probably be the structure and size of a viewed page. The size of a WAP page will be smaller than the size of a normal web page due to its binary encoding and additionally either no inline objects or less inline objects with a smaller size will be included.

2.6 Models recommended in standards

In recent standards, recommendations towards source traffic models can be found. In the UMTS [11] and the cdma2000 [12] standards, models for a wireless data user are provided which are recommended to be used in performance analyses. In the cdma2000 model the traffic is distinguished both in uplink or downlink traffic and in interactive or download traffic. The proposed model has a multilayer structure. On the session level a Poisson arrival process is assumed and a session consists of a geometrically distributed number of packet calls (WWW pages in the case of HTTP). The time between two packet calls is also assumed to follow a geometric distribution as does the interarrival time between two packets. The number of packets in a packet call is Pareto distributed for downloads and geometrically distributed for the reverse link. The packet size is assumed to be constant with different values for the different traffic types. The parameters for the distributions also depend on the traffic type.

The UMTS traffic model is similar, however, with different distributions and parameters. The problem of these models is that they are not based on measurements of existing traffic. In [12] no references are given, the UMTS standard does reference [30] and [31], but these papers do not contain any parameters which are derived from traffic measurements.

3. Recommended model

In this section we will present the model which we suggest to use for a wireless environment performance analysis. We consider a number m of GPRS users active in a coverage area. Each of the users generates different traffic streams related to different applications. Since most of the applications of GPRS users can be activated in parallel (E-Mail while browsing WWW, etc.), we assume in our model separate traffic processes according to the observed applications. For the arrival process of the different applications for a single user the following process descriptions are used:

ON/OFF-process: the ON-state stands for the activity phase while using the application, the OFF-state is the idle period. ON- and OFF-state are described by random variables and distribution functions for their durations. An example of the parameterization of HTTP users will be given in the following.

Renewal process: when the ON-state is short compared to the ON/OFF-cycle length, a renewal process will be employed to describe the traffic. An example of a renewal process for E-Mail traffic is presented below.

3.1 E-Mail model

At the base station a merging process of m individual E-Mail processes can be observed. Given the user i to generate an input process of E-Mails with rate $\lambda_{i,Email}$, the resulting E-Mail process offered to the base station will be described with the following parameters

A_{Email}: random variable (r.v.) for the inter-email time of all users with $E[A_{Email}] = 1/\sum_{i=1}^{m} \lambda_{i,Email}$.

S_{Email}: r.v. for the size of an E-Mail.

For a model of an activity phase it has to be clarified in which way E-Mail is used. Basically, the occurrence of two different mail scenarios has to be concerned:

1. User initiated E-Mail:
   (a) User sends one or more E-Mails.
   (b) User checks at the server for new E-Mails and should the occasion arise downloads them.
2. "Network" initiated E-Mail:
   (a) The server notifies the user about new mails.
   (b) Periodical events: the mail client checks the server for new messages periodically.

All these scenarios follow different arrival processes. What all scenarios have in common is the distribution of the size of an E-Mail.

In the scenarios of item 1(b) and item 2(b) two possibilities exist. Either new E-Mails are waiting on the server or not. Since the messages for checking the E-Mail are rather small they are neglected in our model. Therefore, we only consider events when E-Mails are transmitted either from the server to the mobile or in the reverse direction. It is assumed that the number of E-Mails in both directions is equally distributed. For the arrival of new E-Mails the Poisson process described above is used. The parameter $\lambda_{i,Email}$ depends on the behavior of a user and is chosen as 10^{-4} E-Mails per second.

The other distribution required to describe the E-Mail model defines the size of an E-Mail. In Section 2.3 was outlined that the distributions in literature differ considerably from the distributions evaluated at the University of Würzburg. Thus, the measured CDFs are used for characterizing the E-Mail size. The respective moments are given in Table 2.

3.2 HTTP model

The resulting WWW-related traffic process at the base station is the superposition of individual user WWW traffic streams. Taking into account a number m of GPRS users actively using HTTP in a base station area, each of them can be in an activity or idle phase . For the WWW activity phase we employ the basic concept presented in several papers, e.g. [10, 22, 23]. The model is shown in Figure 4. Basically, the WWW activity phase consists of a sequence of HTTP-ON and HTTP-OFF phases.

An HTTP-ON phase starts after the arrival of a web request and it models the activity after accepting the request. The HTTP-OFF phase represents a silence period after all

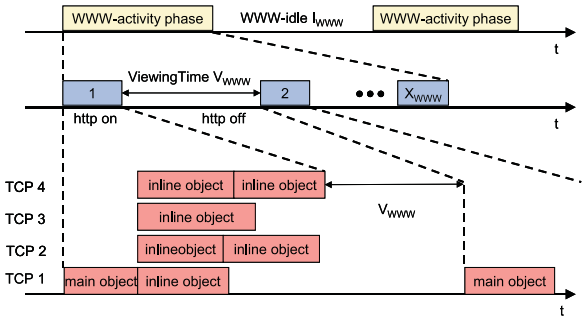

Fig. 4. WWW activity phase.

objects in a web request have been retrieved. Thus, the duration of the HTTP-ON and HTTP-OFF phase corresponds to the page loading time and viewing time, respectively, where the viewing time can be zero, reflecting the case that a user starts a web request during the loading time of an object.

During the HTTP-ON phase, objects forming the requested page are loaded. As mentioned before, two types of objects are distinguished: main object (main file containing an HTML document) and inline object (object linked from the Hypertext document). Depending on the browser version in use several TCP connections can be activated to load objects. The activation and deactivation of TCP connections can be partly in serial or in parallel. Parallel TCP connections, e.g. for inline objects, are opened consecutively after the single connection for the main object.

In the following, a short summary of the recommended HTTP model for a single user and the respective distributions with parameters are outlined. The model consists of two major parts: i) the first set of parameters describes the session level with the time between two sessions and the number of web requests per session and ii) the second parameter set describes the composition of a web request with the size of the main object and the number of inline objects and their size, the size of the GetRequest and the length of the viewing time for the web page. This model contains the following random variables:

I_{WWW}: time between two WWW sessions
X_{WWW}: number of web requests per WWW session
V_{WWW}: web request viewing time
M_{WWW}: size of main object
N_{WWW}: number of inline objects
O_{WWW}: size of inline object
R_{WWW}: size of GetRequest

The distribution of the random variables concerning the composition of a web request are taken from Choi [10]. This paper is the most recent one and has the most detailed model for a web request. Choi's measurements yielded the distributions shown in Table 3.

The parameters above result in an average web request size of about 50 kB. This value is reasonable as it corresponds to the values given in recent publications, e.g.

Table 3. Results of Choi

r.v.	Distribution	mean	variance
V_{WWW}	Weibull	39.5 s	92.6 s
N_{WWW}	Gamma	5.55	11.4
M_{WWW}	Lognormal	10 kB	25 kB
O_{WWW}	Lognormal	7.7 kB	126 kB
R_{WWW}	Lognormal	360 B	106 B

[13, 23]. The moments of the viewing time are also similar to those observed in [13].

The distributions regarding the session layer are not available in a single paper. The average number of web requests during a session in [13] is consistent with the value given in [23]. Thus, for X_{WWW} a Lognormal distribution with $\alpha = 1.8$ and $\beta = 1.68$ is chosen resulting in a mean of 25 pages and a standard deviation of 100 pages. The length of the time between two sessions can only be determined from sufficiently long measurements and each session has to be assigned to a specific user. The only one to perform a measurement of this kind was Mah. Consequently, the distribution was taken from his paper. However, he did not distinguish between the viewing time and the time between two sessions. Since Mah did not fit any distribution the CDFs of Mah's homepage are used (http://http.cs.berkeley.edu/~bmah/Software/HttpModel). Values below 900 s are truncated since Mah did not distinguish viewing and intersession time. The truncated distribution yields a mean time for I_{WWW} of 14 300 s and a standard deviation of 20 600 s.

3.3 FTP and WAP

The importance of the FTP service in a wireless environment is expected to be less than in the wire-line Internet. Thus, for the model of a GPRS user it may be negligible. However, due to reasons of completeness an FTP model is recommended as well. Like in the reference model of UMTS [11] the FTP service can be included by describing an FTP file transfer like a web session with only one web request. The arrival process of FTP sessions is chosen as Poisson with a rate of $6 \cdot 10^{-6}$ sessions per second. The structure of the WAP model is similar to that of the HTTP model except that the web request size will be different. In WAP, a web page will only comprise inline objects if the capability of the handset is appropriate. The type of the handset, therefore, influences the distribution of the number of inline objects. Of course, if the mobile can not handle images there will be no inline objects, otherwise their distribution is equivalent to the HTTP case. The document size will be smaller due to the tag compression. The shrinking factor will be around 40%. The distribution of the size of inline objects depends on the resolution which is possible in the handset. Thus, like in the HTTP model the size is Lognormal distributed. For each user a fixed shrinking factor dependent on the handset resolution has to be introduced to adopt the distribution appropriately.

4. Conclusions

In this paper we presented a survey of common models for packet data users found in the literature. While most of the applications have been examined in the past for wire-line systems, our focus lies on users operating in a wireless network environment. However, the basic source traffic behavior will be the same in both cases as this is determined mainly by the application and the user and only to some extent by the bearer system.

Due to the enormous increase in WWW traffic, it is not surprising that most of the more recent publications deal primarily with HTTP modeling. It could be seen that different approaches were taken to measure and characterize the user behavior. While some of the authors split the traces into sessions, others consider the whole measurement phase as one session and describe the more microscopic level of web requests. Additionally, there are differences in determining the end of a web page. All of this leads to slightly differing models and their appropriate parameters. Combining the advantages of the various approaches, we propose a recommended model for WWW traffic which takes all of the described features into account.

In a similar fashion to HTTP, we also investigated the literature on models covering E-Mail and FTP traffic. Apart from these rather "traditional" services, new services like WAP need to be considered. However, as users for such systems do not exist at the time this report is made, we can only give an estimate on the actual behavior of such users. Our model can be enhanced when measurements on WAP users are available.

References

[1] Brasche, G.; Walke, B.: Concepts, services, and protocols of the new GSM phase 2+ general racket radio service. IEEE Communications Magazine (August 1997).

[2] Meier-Hellstern, K.; Wirth, P.; Yan, Y.-L.; Hoeflin, D. A.: Traffic models for ISDN data users: Office automation application. Proc. Of International Teletraffic Congress (ITC-13), A. Jensen and V. B. Iversen (Eds.), 1991.

[3] Leland, W.; Taqqu, M.; Willinger, W.; Wilson, D.: On the self-similar nature of ethernet traffic. IEEE/ACM Transactions on Networking (1994), 2:1–15.

[4] Willinger, W.; Taqqu, M.; Sherman, R.; Wilson, D.: Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level. IEEE/ACM Transactions on Networking **5** (Feb. 1997), 71–86.

[5] Rose, O.; Frater, M.: A comparison of models for VBR video traffic sources in B-ISDN. IFIP transactions C-24: Broadband communications II, 1994. 275–287.

[6] Paxson, V.: Empirically-derived analytic models of wide-area TCP connections. IEEE/ACM Transactions on Networking (1994).

[7] Paxson, V.; Floyd, S.: Wide-area traffic: The failure of Poisson modeling. IEEE/ACM Transactions on Networking **3** (1995), 226–244.

[8] Park, K.; Kim, G.; Crovella, M.: On the relationship between file sizes, transport protocols, and self-similar network traffic. Proceedings of the Fourth International Conference on Network Protocols, Oct. 1996. 171–180.

[9] Vicari, N.; Koehler, S.: Measuring internet user traffic behavior dependent on access speed. Proc. of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management, Monterey, CA, Sept. 2000.

[10] Choi, H.; Limb, J.: A behavioral model of web traffic. International Conference of Networking Protocol 99' (ICNP 99'), Sep. 1999.

[11] ETSI: Universal mobile telecommunication system (UMTS); selection procedures for the choice of radio transmission technologies of the UMTS. Tech. Rept. TR 101 112 v3.2.0. ETSI, April 1998.

[12] Radio Communication Study Group: The radio cdma2000 RTT candidate submission. Tech. Rept. TR45-5. ITU: US TG 8/1, June 1998.

[13] Reyes-Lecuona, A.; González-Parada, E.; Casilari, E.; Díaz-Estrella, A.: A page-oriented WWW traffic model for wireless system simulations. Proceedings of the 16th International Teletraffic Congress (ITC'16), Edinburgh, UK, June 1999. 1271–1280.

[14] Deng, S.: Empirical model of WWW document arrivals at access link. Proceedings of ICC 96, June 1996.

[15] Barford, P.; Crovella, M.: Generating representative workloads for network and server performance evaluation. Proceedings of ACM SIGMETRICS 98, June 1998. 151–160.

[16] Nielsen, H. F.; Gettys, J.; Baird-Smith, A.; Prud'hommeaux, E.; Lie, H. W.; Lilley, C.: Network performance effects of HTTP/1.1, CSS1, and PNG. Proc. SIGCOMM '97, Sept. 1997.

[17] Berners-Lee, T.; Fielding, R.; Frystyk, H.: Hypertext transfer protocol HTTP/1.0. RFC1945, May 1996.

[18] Fielding, R.; Gettys, J.; Mogul, J.; Frystyk, H.; Masinter, L.; Leach, P.; Berners-Lee, T.: Hypertext transfer protocol – HTTP/1.1. RFC2616, June 1999.

[19] Arlitt, M.; Williamson, C.: Web server workload characterization: the search for invariants. ACM SIGMETRICS Conference, Philadelphia, PA, May 1996.

[20] Cunha, C.; Bestavros, A.; Crovella, M.: Characteristics of WWW client-based trace. Tech. Rept. TR-95-010, Boston University: Dept. of Computer Science, April 1995.

[21] Catledge, L. D.; Pitkow, J. E.: Characterizing browsing strategies in the world-wide web. Proc. of the Third World Wide Web Conference, April 1995.

[22] Mah, B.: An empirical model of HTTP network traffic. Proceedings of the IEEE INFOCOM 97, Kobe, Japan, April 1997. 592–600.

[23] Vicari, N.: Measurement and modeling of WWW-sessions. Tech. Rept. 184, University of Würzburg: Institute of Computer Science, 1997.

[24] Färber, J.; Bodamer, S.; Charzinski, J.: Statistical evaluation and modelling of internet dial-up traffic. Proceedings of the Conference on Performance and Control of Network Systems III, SPIE's International Symposium on Voice, Video, and Data Communications (SPIE PE99), Boston, Sept. 1999. 112–121.

[25] Crovella, M.; Bestavros, A.: Self-similarity in the world wide web traffic: Evidence and possible causes. Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, May 1996.

[26] Bolotin, V.; Levy, Y.; Liu, D.: Characterizing data connection and messages by mixtures of distributions on logarithmic scale. Proc. of ITC-16, Edinburgh, UK, June 1999.

[27] Cáceres, R.; Danzig, P. B.; Jamin, S.; Mitzel, D. J.: Characteristics of wide-area TCP/IP conversations. Proc. of SIGCOMM '91, 1991. 101–112.
[28] Danzig, P.; Jamin, S.: tcplib: A library of TCP internetwork traffic characteristics. Tech. Rept. CS-SYS-91-01, University of Southern California: Computer Science Department, 1991.
[29] WAP Forum: Wireless application protocol specifications. http://www.wapforum.org/, April 1998.
[30] Anderlind, E.; Zander, J.: A traffic model for non real-time data users in a wireless radio network. IEEE Communications Letters **1** (Mar. 1997).
[31] Anagnostou, M. et al.: A mulitservice user descriptive traffic source model. IEEE Transactions on Communications **44** (Oct. 1996).

Phuoc Tran-Gia Phuoc Tran-Gia received degrees in electrical engineering from Stuttgart University (Dipl.-Ing.) and University of Siegen (Dr.-Ing.), Germany. In 1977, he joined Standard Elektrik Lorenz (now Alcatel), Stuttgart, where he worked on the software development of digital switching systems. From 1979 to 1982, and from 1983 to 1986, he was with the University of Siegen and University of Stuttgart. In 1986, he joined IBM Zurich Research Laboratory, Zurich, Switzerland, where he worked on the architecture and performance evaluation of computer communication systems. Since 1988 he is professor in the Faculty of Mathematics and Computer Sciences, University of Würzburg, Germany. He is currently director of the institute of computer science. Professor Tran-Gia is chairperson of the management committee of the COST 257 action of the European Union entitled "Impact of new services on the performance and architecture of broadband networks". He is also founding director of the multi-university Nortel's "Center of Network Optimization". His current research areas include performance modeling and management of communication systems, and planning and optimization of mobile communication networks. Phuoc Tran-Gia is also founder and CEO of InfoSim Networking Solutions AG (Würzburg, Germany and Richardson/Dallas, Texas, USA), which is specialized in IP network management products and services.

Dirk Staehle received the Diploma degree in computer science in 1999 from the University of Würzburg, Germany. Thereafter he became member of the Scientific Staff of the Department of Distributed Systems at the University of Würzburg, where he is presently working towards his PhD degree. His main research interest lies in the performance evaluation of wireless Internet access using GPRS and UMTS.

Kenji Leibnitz received the Diploma degree in computer science in 1996 from the University of Würzburg, Germany. In November 1996 he joined the Scientific Staff of the Department of Distributed Systems, where he is also currently working towards the PhD degree. His main fields of research are performance evaluation and capacity planning of CDMA mobile communication systems, especially third generation UMTS networks.