When Do We Need Rate Control for Dedicated Channels in UMTS?

Tobias Hoßfeld, Andreas Mäder and Dirk Staehle
Department of Distributed Systems, University of Würzburg, Am Hubland, 97074 Würzburg, Germany
Tel.: +49-(0)-931 8886642, Fax: +49-(0)-931-8886632
{hossfeld,maeder,staehle}@informatik.uni-wuerzburg.de

***Abstract*— The Universal Mobile Telecommunication System (UMTS) offers rate controlled radio bearers for best effort traffic. The purpose of rate control is to maximize resource utilization and concurrently to provide best-effort users with an acceptable grade-of-service. Due to the user behaviour or the used application of a mobile subscriber we distinguish between two types of best-effort users. The time-based users stay for a certain time within the network and download an arbitrary amount of data. Volume-based users leave the network after they have downloaded a specific data volume. The contribution of this work is an evaluation of time-based and volume-based best-effort traffic over rate-controlled DCHs by analytic means and with a detailed packet-level simulation. The results of both approaches are qualitative equal: while the rate control mechanism works effectively for time-based users and allows any rate, the assigned rates for the volume-based users take values between two extremes, either the minimal or the maximal rate.**

I. Introduction

The purpose of the *Rate Control* mechanism in UMTS is to maximize the utilization of resources and concurrently to provide best-effort users with an acceptable grade-of-service. Users of the conversational and streaming classes (QoS users) have stringent resource requirements and thus have precedence before users of the interactive and best-effort classes (BE users). The rate control distributes the remaining capacity to the BE users by adapting the bit rate of each user.

Web browsing is a typical application for the best-effort service. On the downlink the web traffic is currently transmitted over dedicated channels (DCHs) which are typically slowly rate-controlled. In future best-effort traffic will be mainly exchanged via the high-speed downlink shared channel (HS-DSCH).

A web user is mostly described with a volume-based source traffic model. It is assumed that the user stops browsing in the web after successfully downloading the required information on different web pages, i.e. after downloading the corresponding data volume. However, if we consider measurements of Internet user traffic depending on the access speed [1], the observed traffic volumes were found to increase with the access speed and the session duration was almost independent of the access speed. This means that the users stay for a certain time within the network and download an arbitrary amount of data. Hence, a time-based source model might better fit the emerging traffic. Considering rate control, we therefore distinguish between two types of users, time-based and volume-based. For other applications, like data exchange via FTP or Voice over IP, the kind of application itself determines already the user behaviour and results either in a volume-based or a time-based source traffic model.

Kröner [2] gives an overview of some generic rate allocation principles for an UMTS network to assign the remaining power to the best effort users: equal rates, equal power, and Min/Max rate allocation. In [3], an analytic method is proposed to approximate the quality in terms of bandwidth that a best effort user experiences in an environment with heterogeneous traffic. We use the equal rate algorithm and calculate the optimal rate allocations as in [3].

Our contribution is a performance evaluation of volume and time-based services like web traffic over rate-controlled DCHs using an equal rate allocation strategy. We implemented a discrete-event simulation with TCP New Reno and rate and power control according to the 3GPP standard. The resulting distributions of the assigned bit rates and NodeB transmit powers show that the benefit of rate control depends strongly on the user behaviour. We conclude further that in some cases rate control is dispensable or even to the disadvantage for the grade-of-service.

The rest of this paper is organized as follows. In Section II, we illustrate the rate control mechanism and introduce the optimal rate control model. The effect of the different source traffic models on the performance and the achieved bit rates is analytically derived in Section III. Section IV shows the web traffic model and, in particular, how time-based and volume-based web users are modelled. The analytical observations are confirmed for web traffic with its complex request response pattern. The numerical results are obtained by a detailed discrete-event simulation. Finally, we conclude this work in Section V and give an outlook on future work.

II. Rate Control Model

The capacity limiting link due to the asymmetry of web traffic is the downlink on which the UMTS system capacity is limited by the NodeB transmit power. The maximal amount of the NodeB transmit power is $\hat{T}_{max}$. Every NodeB x tries to transmit with a target power $\hat{T}_C$ which is consumed by common channels with constant power $\hat{T}_{CCH}$ and by QoS users with power $\hat{T}_{x,QoS}$. Thus, the best effort users share the remaining power $\hat{T}_{x,BE} = \hat{T}_C - \hat{T}_{CCH} - \hat{T}_{x,QoS}$. The basic rate control principle is illustrated in Figure 1. The packet scheduler which is typically located in the Radio Network

Controller (RNC) implements the particular rate control mechanism. In our model, we consider the equal rate allocation strategy since it follows the same principle of fairness used for the QoS users, i.e. all web users are assigned the same rate irrespectively of the radio channel conditions and the locations of the users. The transmission rate is mainly determined by the users who experience bad radio channel conditions or are located at the edge of the cell. These users consume considerably more power than users near the NodeB.

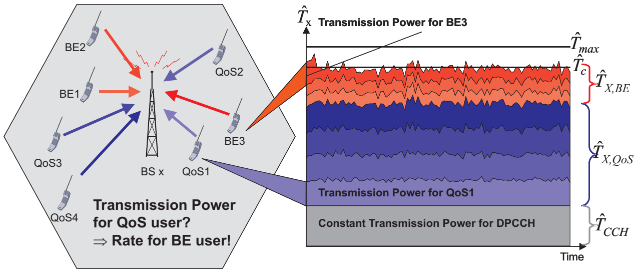


Fig. 1. Rate control illustration

In order to determine the optimal rate for the web users of a NodeB x at an arbitrary time instant, we use the *optimal rate control algorithm* [3] which utilizes knowledge of the current UMTS system state, the radio channel conditions for all mobile stations, and the NodeB transmit power for each NodeB of the network.

The required transmit power $\hat{S}_{x,k}$ of the dedicated channel from NodeB x to mobile k is

$$\hat{S}_{x,k} = \frac{R_k \hat{\varepsilon}_k}{W + R_k \hat{\varepsilon}_k} \left(\frac{W \cdot \hat{N}_0}{\hat{d}_{x,k}} + \sum\nolimits_{y \neq x} \hat{T}_y \frac{\hat{d}_{y,k}}{\hat{d}_{x,k}} + \alpha \hat{T}_x \right), \quad (1)$$

where R_k and $\hat{\varepsilon}_k$[1] are the bit rate and the target E_b/N_0 value of a mobile k that entirely specify a mobile's service. Accordingly, we introduce the variable

$$\omega_k = \frac{R_k \cdot \hat{\varepsilon}_k}{W + R_k \cdot \hat{\varepsilon}_k} \quad (2)$$

for the load of a mobile, where W is the system bandwidth of 3.84Mcps. Note, that the load ω_s of all mobiles with the same service s is equal. In particular, applying the equal rate allocation strategy leads to equal loads for all best effort users which we denote as ω_B. Furthermore, N_0 is the thermal noise spectral density of -174dBm/Hz, T_y is the total emitted power of a NodeB y including the power for common and dedicated channels, α is the orthogonality factor that specifies the part of the power from the own cell that is received as interference. And finally, $d_{x,k}$ is the propagation gain from NodeB x to mobile k. Now, consider the term in brackets in Eq. (1). It corresponds to the total interference at the mobile k divided by the propagation gain $\hat{d}_{x,k}$ from its serving NodeB and consequently depends on the transmit powers of its own NodeB x, other NodeBs y, and on the radio conditions to these NodeBs. If we assume optimal rate control, i.e. every NodeB manages to transmit always with its target power T_C, then we can replace T_x and T_y by T_c. Thus, this term depends only on the propagation gains to the various NodeBs, hence on its location. We define the variable

$$Z = E\left[\frac{W \cdot \hat{N}_0}{\hat{d}_{x,k}} + \sum\nolimits_{y \neq x} \hat{T}_C \cdot \frac{\hat{d}_{y,k}}{\hat{d}_{x,k}} + \alpha \hat{T}_C \right] \quad (3)$$

as the expectation of this term for a user with random location in the cell area and obtain the mean transmit power $\hat{S}_s$ for a mobile with service s as

$$E\left[\hat{S}_s\right] = \omega_s \cdot Z. \quad (4)$$

Now, consider again the total power of a NodeB. It comprises the power for common channels, the power for the n_s mobiles with a QoS service $s \in \mathbb{Q}o\mathbb{S}$, and the power for the n_B best effort mobiles:

$$\hat{T}_C = \hat{T}_{CCH} + Z \cdot \left(\sum\nolimits_{s \in \mathbb{Q}o\mathbb{S}} n_s \omega_s + n_B \omega_B \right). \quad (5)$$

Consequently, the maximum load for the best-effort mobiles is a function $\omega_B(n_s, n_B)$ of the number of mobiles:

$$\omega_B(n_s, n_B) = \frac{\hat{T}_C - \hat{T}_{CCH} - \sum_{s \in \mathbb{Q}o\mathbb{S}} n_s \cdot \omega_s \cdot Z}{n_B \cdot Z}. \quad (6)$$

Assume further, that the best effort mobiles can adapt only a set of predefined rates $\mathfrak{R}_B = \{R_1, ..., R_M\}$ with corresponding target E_b/N_0 values $\hat{\varepsilon}_i^*$. Then, the load of a best effort user can be mapped to the bit rate which is consequently also a function of the number of mobiles in the cell

$$R(n_s, n_B) = max\left\{ R_i \in \mathfrak{R}_B \middle| \frac{R_i \hat{\varepsilon}_i^*}{W + \alpha R_i \hat{\varepsilon}_i^*} \leq \omega_B(n_s, n_B) \right\}. \quad (7)$$

Note that in the following numerical studies we assume data rates which are multiples of 16kbps up to a maximum of 128kbps and a common target E_b/N_0 value of 3dB.

III. Analytical Approach for Fully Active Users

This section is intended to demonstrate the impact of the source traffic model – time-based or volume-based – on the performance of a UMTS cell with rate control. Therefore, we assume a very simple scenario where the NodeB serves only best effort users, i.e. there is no power spent for QoS users. The target power for the rate control $\hat{T}_C$ is set to 8W and the power for common channels $\hat{T}_{CCH}$ is 2W. The orthogonality factor α is 0.4. The best effort users arrive according to a Poisson process with rate λ. The departure process is different for the time-based model and the volume-based model. In the time-based model, the users leave the system after an exponentially distributed service time with mean $1/\mu$. Accordingly, the time a user stays in the system is independent of her data rate, but the data rate determines the volume she can transmit. In contrast to that, the volume-based model specifies that the user leaves the system just after she has transmitted a certain data volume which is exponentially distributed with mean $E[V]$. Thus, the time she stays in the system depends on the data rate she obtains.

In the following, we show how to compute the steady-state distribution for the time-based and the volume-based service

[1] Note that $\hat{a}$ marks a linear value and a is the corresponding decibel value.

times, and how to derive the distribution of the data rate seen by a random user. Since we consider only best effort users we denote the rate that a best effort obtains when n users are in the system by $R_B(n)$ omitting the number of QoS users compared to the notation in Eq. (7).

A. Time-Based Source Traffic Model

A UMTS cell is modelled as an M/M/n-0 queue where n is the maximum number of users that still allows the minimum rate $R_1 = 16kbps$. The steady-state probability is then given by

$$p(i) = \frac{a^i}{i!} \cdot \left(\sum_{j=0}^{n} \frac{a^j}{j!}\right)^{-1} \tag{8}$$

with $a = \lambda/\mu$. The distribution of the best effort rates at a random instance of time is

$$P\{R_{B,time} = R_k\} = \sum_{i=0}^{n} \delta(i,k) \cdot p(i), \tag{9}$$

with $\delta(i,k) = 1$ if $R_B(i) = R_k$ and $\delta(i,k) = 0$, otherwise. The distribution of the best effort rates for a random user is

$$P\{R_{B,user} = R_k\} = \frac{\sum_{i=1}^{n} i \cdot \delta(i,k) \cdot p(i)}{\sum_{i=1}^{n} i \cdot p(i)}. \tag{10}$$

B. Volume-Based Source Traffic Model

For the volume-based source traffic model the departure rate for a single user is state dependent. Accordingly, the transition rate $q(i, i-1)$ from state (i) with i users to state $(i-1)$ with $i-1$ users is given as

$$q(i, i-1) = i \cdot \frac{R_B(i)}{E[V]}. \tag{11}$$

The transition for arrivals does not change, i.e. $q(i, i+1) = \lambda$. Consequently, we obtain the steady state distribution as

$$p(i) = \frac{\prod_{i=1}^{n} \frac{\lambda}{q(i,i-1)}}{1 + \sum_{j=1}^{n} \frac{\lambda}{q(i,i-1)}} \tag{12}$$

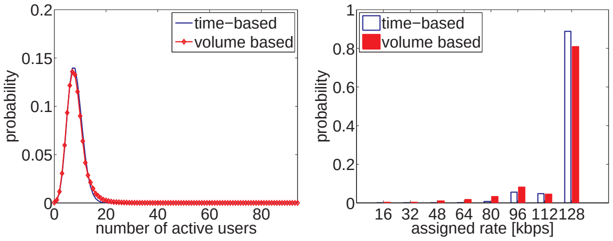
The distribution of the best effort rates at a random instance of time or for a random user are derived analogous to Eq. (9) and Eq. (10).

C. Numerical Results of Analytic Approach

We now study the impact of the source-traffic model on the distribution of the number of users in the system and the resulting best-effort rates. In order to vary the system load, we keep the arrival rate constant at 1 user per second and choose mean services times $1/\mu$ of 8s, 9.5s, and 11s. For these values we determine the distribution of the data rates for a random user and determine the mean data volume $E[V_{user}]$ that a user is able to transmit during its life time:

$$E[V_{user}] = \frac{E[R_{B,user}]}{\mu} = \sum_{i=1}^{M} P\{R_{B,user} = R_i\} \frac{R_i}{\mu} \tag{13}$$

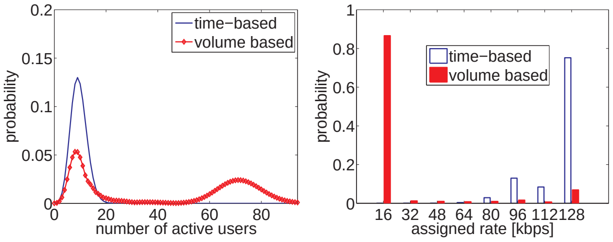
The resulting mean data volume for a random user according to the time-based model is then used to define the mean data volume that a user in the volume-based model has to transmit. Thus, the system load and the long-term throughput is the same for the time-based model and the volume-based model as long as no blocking occurs.



(a) number of active users

(b) assigned rate

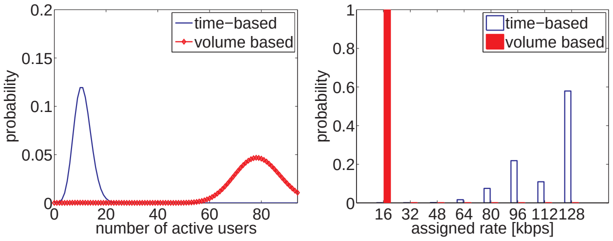
Fig. 2. For time-based $1/\mu = 8$s and for volume-based $E[V] = 1000$kbit



(a) number of active users

(b) assigned rate

Fig. 3. For time-based $1/\mu = 9.5$s and for volume-based $E[V] = 1148$kbit



(a) number of active users

(b) assigned rate

Fig. 4. For time-based $1/\mu = 11$s and for volume-based $E[V] = 1260$kbit

Figures 2, 3, and 4 show the steady-state distribution and the resulting distribution of the data rate for a random best-effort user in the three load scenarios. For the time-based source traffic model, the offered load determines the peak of the probability function $p(i)$ for the number of active users. Accordingly, this peak can be shifted to any possible number between 0 and n, i.e. between no users and the maximal number of users which are allowed to access the system due to admission control. As a result, for the time-based source traffic model the probability function of the assigned rate may take positive probabilities for any rates according to the offered load and the peak of this probability function is also shifted continuously over the set of available rates. This means that the rate control mechanisms works effectively for time-based users and allows any rate. In the three examples, mainly the highest data rates are assigned. Thus, we should expect a similar behaviour for the volume-based model as the rate control has only a small impact on the data rate.

For the lowest load resulting in a mean data volume of 1000kbit, we achieve the expected result: In Figure 2 the number of active users and thus the assigned bit rates are nearly identical for both models. If we consider, however, the scenarios with higher load resulting in mean data volumes of 1148kbit and 1260kbit, respectively, we obtain completely different results for time-based and volume-based users. In the volume-based scenarios, only extreme values for the active number of users and therefore for the assigned bit rate are obtained. The rate control mechanism tends towards assigning either the minimal or the maximal bit rate. The bit rates in between are not used. This is due to the fact that the probabilities to have a 'medium' number of active users in the system is close to zero. Either a small number of best-effort users is active which share the available transmit power and get the maximal rate, or the number of users is so high that only the minimal bit rate can be assigned, cf. Figure 3. With a slightly higher load, the blocking probability increases for the volume-based users, cf. Figure 4, while in the time-based scenario no user is blocked at all.

This effect can be explained by the fact that the volume-based users stay in the system until they have downloaded a certain amount of data. If the number of volume-based users increases, the assigned bit rate decreases which results in longer download times. Nevertheless, new users arrive with rate λ and thus the bit rate has to be decreased by the NodeB. In Fig. 3, the system becomes stable for the maximum rate of 128kbps and the minimum rate of 16kbps where we can observe Poisson distributions around the expected numbers of 8.97 and 71.75 users. The transition from the lowest rate to the highest rate occurs both rarely and fast.

IV. Web Traffic Simulation

In order to study the impact of the basic characteristic of the source traffic model in a more realistic scenario, we consider a hexagonal UMTS cell with three surrounding tiers. In the center cell two kinds of users are served: (1) QoS users on dedicated channels with a fixed rate of 64kbps, a target E_b/N_0 value of 4dB, and an exponentially distributed service time with mean $120s$, and (2) best-effort web users on rate controlled dedicated channels. The source traffic model of the web users is either volume-based or time-based. The packet-based discrete event simulation includes an exact implementation of the TCP stack with version New Reno, the UMTS power control, and the optimal rate control as defined in Section II. The location of the users is chosen randomly within the cell area and the users remain at that location, i.e. mobility is not simulated. Users in the surrounding cells are not simulated in detail. Instead, perfect rate control is assumed and the surrounding NodeBs transmit with constant power $\hat{T}_C = 8$W.

A. Web Traffic Model

The web traffic model for volume-based users is chosen according to an adapted version of [4]. It is structured into three layers: session, page, and TCP layer. On session layer, the users arrive according to a Poisson process. During a web session a user downloads and views a number of N web pages where N is lognormally distributed. A web page consists of the main object and a random number of optional inline objects. The main object is the HTML code and an inline object is a file or script and referenced in the main object. HTTP opens a TCP connection and sends a getRequest for the main object. After receiving the main object, the inline objects are requested from the server and sent back in up to four parallel TCP connections. The volume for both main and inline objects is also lognormally distributed. In between two web pages there is a viewing time during which the user is passive. For the exact values and distributions please refer to [4] or [5].

In order to specify an analogous time-based model for the web user we measured the session time in a simulation with volume-based web users and five QoS users on average, and fitted it by a Gamma distribution with a mean of 7.28 minutes and a standard deviation of 11.38 minutes. The simulation of the time-based web-users then generates the session duration at the arrival of a new session according to this distribution and the user downloads web pages until the end of the session.

B. Simulative Results

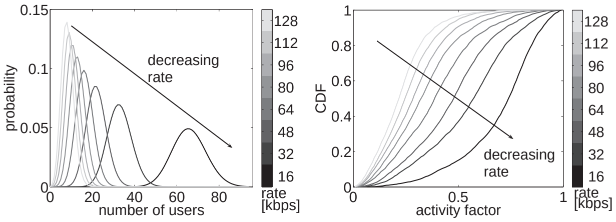
In Section III, the effect of time-based and volume-based source traffic models was demonstrated for fully active users under the assumption that the required service, either in terms of time or volume, has the Markov property. In this section, we consider web traffic which is transmitted over TCP/IP. The activity factor of a web user depends on several factors, like the amount of downloaded data volume (due to TCP's flow control mechanisms), the round trip time between the web server and the client, the packet loss probability on the link, the available bandwidth, and last but not least the ratio between the duration of a web page download and the viewing time. We assume the air interface to be the bottleneck in this scenario and hence the assigned bit rate by the NodeB determines the throughput.

If the web users transmit with a constant bit rate (e.g. via not rate controlled dedicated channels), the time-based and the volume-based source traffic model are equal and lead to the same results. Figure 5(a) shows the number of web users in the system for a given Poisson arrival rate $\lambda = 1/10$s and different fixed transmit rates. It has to be noted that the web users do not necessarily contribute to the required transmit power, since they are not active on the air interface for the whole session duration. This is expressed by the activity factor, which denotes the ratio of time being active on the air interface to the total time the web user stays in the system. The higher the bit rate is the smaller the activity factor is, see Figure 5(b).

We proceed to consider now the web traffic model for time-based and volume-based users in a UMTS system with rate-controlled channels. In order to vary the power available for the web users, circuit-switched (CS) QoS users arrive according to a Poisson process resulting in a load of 1,3,5, or 7 users. Figure 6(a) shows the distribution of the best effort rates for the time-based users. We can see that the rate control utilizes the full spectrum of possible rates. So in this case, the rate control works to the benefit of the users.

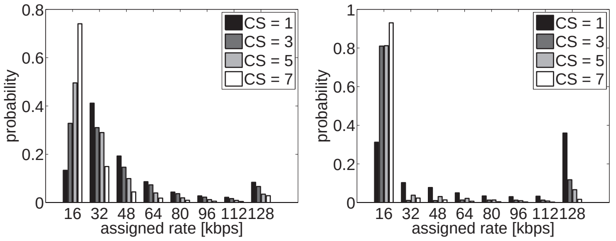
In contrast, the rate probabilities for the volume-based users as shown in Figure 6(b) are cumulated at the marginal values of the rate spectrum. Even in a medium loaded system, the assigned rates are either at the maximum of 128kbps or at the minimum of 16kbps. This effect results from the interaction between the assigned bit rate and the number of web users in the system as shown in Figure 7. The figure shows two interaction cycles. The assigned rate influences the session duration, which in turn determines the number of currently active users. So, in a highly loaded system with short session interarrival times, the left cycle leads to rate assignments which tend to the minimum rates. With lower load, the right cycle leading to high rate assignments is most probable. However, even if the system is medium loaded in terms of session arrivals, the instantaneous load situation leads either to the rate-increasing or rate-decreasing interaction cycle, keeping the probability for the rates in between very low. The effect that decreasing the data rate leads to an increased activity means that decreasing the data rate does not automatically reduce the load to the system. Consequently, for the volume-based web traffic model the tendency to assume only the extreme bit rates is even stronger than for the always active case in Section III.

In all scenarios, the effect of the rate control for the volume-based traffic model differs significantly from the effect for the time-based traffic model. Further, in case of volume-based user behaviour the rate control mechanism is not efficient, since it leads to very low bit rates or to very unequal distributed bitrates for high and medium loads. For lower loads, the rate control has no effect at all. As a result, it can be stated that for volume-based source traffic models only a degenerated rate control mechanism is required with only two possible rates, while for time-based users the whole spectrum of bit rates can be assigned.



(a) number of users in the system (b) activity on air interface

Fig. 5. WWW simulaton with constant bit rate



(a) time-based (b) volume-based

Fig. 6. WWW simulation with rate control

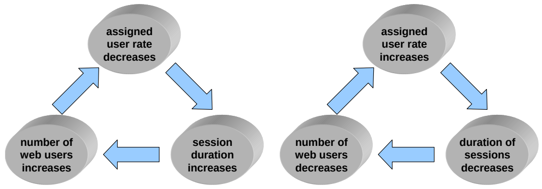

assigned user rate decreases
session duration increases
number of web users increases
assigned user rate increases
duration of sessions decreases
number of web users decreases

Fig. 7. Interaction of assigned rate, session duration, and number of users

V. Conclusions and Outlook

The goal of this work is the performance evaluation of time-based and volume-based web traffic over rate-controlled dedicated channels with the equal rate allocation strategy.

The numerical results suggest that the use of rate control should be planned with consideration of the user behaviour and with a proper designed admission control in order to soften the unwanted effects in high load situations. More specifically, in case of volume-based user behaviour, rate control degenerates in such a way that the rate control mechanism only switches between the minimal and the maximal possible rate.

We further conclude that the selection of the appropriate source model for investigating a system or mechanism, like the rate control in UMTS, is very important and has to be independent of the evaluated mechanism. Future work might focus on the development of different rate control strategies for time-based and volume-based web users. However, the real user behaviour lies in between those two extremes leading to another approach, the user-experienced WWW source traffic model. It does not determine the user behaviour a priori, but also considers the user perceived quality of the service: a user tends to stay longer in the system if it perceives a better quality. This model will be specified and investigated in future work.

Acknowledgments

The authors would like to thank Prof. Dr.-Ing. Phuoc Tran-Gia, Dr. Kenji Leibnitz from the University of Osaka, Japan, and Dr. Bernd Schröder from T-Mobile International for the fruitful discussions during the course of this work.

References

[1] N. Vicari and S. Köhler. Measuring internet user traffic behavior dependent on access speed. In *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, USA, 9 2000.

[2] H. Kröner. Radio resource allocation for data services in UMTS networks. *International Journal of Electronics and Communications*, 2001.

[3] D. Staehle, K. Leibnitz, K. Heck, P. Tran-Gia, B. Schröder, and A. Weller. Analytic approximation of the effective bandwidth for best-effort services in UMTS networks. Technical report, University of Würzburg, 2002.

[4] P. Tran-Gia, D. Staehle, and K. Leibnitz. Source traffic modeling of wireless applications. *International Journal of Electronics and Communications (AEÜ)*, 55, 2001.

[5] W. Choi and J.Y. Kim. Forward-link capacity of a DS/CDMA system with mixed multirate sources. *IEEE Trans. on Veh. Tech.*, 50(3):737–749, May 2001.