ANALYTICAL PERFORMANCE EVALUATION OF INTERNET ACCESS OVER GPRS AND ITS COMPARISON WITH SIMULATION RESULTS

Tim Irnich, Peter Stuckmann

Aachen University of Technology (RWTH), Chair of Communication Networks (ComNets)
Kopernikusstr. 16, D-52064 Aachen, Germany, {tim|pst}@comnets.rwth-aachen.de

***Abstract*— In this paper, we discuss the results of recently published analytical concepts for performance evaluation of packet-switched mobile radio networks. For comparison we regard the results of a simulation system (GPRSim), which emulates the GSM *General Packet Radio Service* (GPRS) in great detail. Our scope is to identify the possible reasons for deviations between analytical and simulation results in order to support the further development in this area. In the regarded analytical modelling framework, which is presented by U. Vornefeld [1, 2] source modelling is done by characterization of WWW traffic with means of marked Markovian arrival processes (MMAP), accordingly GPRS performance measures are derived by analysis of an MMAP/G/1 queue. By evaluation of WWW traffic generated during stochastic simulation we show that important characteristics of recent Internet traffic are depicted by Vornefeld's analytical source model. We find that the regarded analytical approach is well suited for understanding the general impact of various system parameters. Nevertheless, although the analytical model uses a very detailed characterization of WWW traffic, for quantitative performance analysis simulation results are still required.**

***Index Terms*— GPRS Analysis, Internet Applications, Dimensioning, MMAP, Validation, Event Driven Simulation.**

I. INTRODUCTION

In the context of the evolution towards 3rd Generation (3G) mobile radio networks, packet-switched data services like the General Packet Radio Service (GPRS) and the Enhanced GPRS (EGPRS) are presently introduced into GSM and IS-136 systems worldwide.

While for analytical dimensioning of circuit-switched networks the Erlang-B-Formula [3] has been successfully applied over decades, for packet-switched cellular radio networks such an applicable traffic engineering model is still missing.

With the change of the switching paradigm the focus in radio network dimensioning is shifted. While dimensioning of circuit switching systems focuses on user calls and determines call blocking probabilities and link utilization, packet switching dimensioning has to "zoom into" the user call, because *Quality of Service* (QoS) requirements of packet-switching services have to be defined on packet level. The corresponding measures of interest are packet delays, packet loss probability and required capacity.

Since it is well known that recent Internet traffic exhibits statistical effects like self-similarity and long range dependence, modelling of these effects under retainment of analytical tractability is a mandatory requirement. Besides that, in packet-based mobile radio networks the resulting service process is significantly influenced by characteristics of the radio link, packet scheduling strategies and protocol efficiency.

In his recent publications [1, 2] U. Vornefeld presents an analytical modelling framework for Web browsing users accessing the Internet using packet-switching mobile radio networks. He uses numerical algorithms to derive an analytically tractable representation of a well-known Web browsing application model [4] in terms of *Markovian Arrival Process with Marked Arrivals* (MMAP). He divides the Web model's behavior in active (on) and passive (off) phases and characterizes the sojourn times in these states by PH-type distributions. During the on phase, he assumes Poisson packet arrivals.

By comparison with traffic characteristics evaluated in a detailed GPRS simulation system (GPRSim) [5, 6], which is in fact an emulation of the GPRS system, we show that this analytical representation is able to depict the key aspects of internet traffic sources' behavior. In a second step we compare Vornefeld's results in [2] with *Quality of Service* (QoS) measures derived by means of stochastic simulation. In order to identify the crucial points in the analytical model and to guide the process of further development in the area of analytical modelling, we evaluate communities and differences in the results.

II. GENERAL PACKET RADIO SERVICE (GPRS)

GPRS has been standardized by the ETSI as part of the GSM *Phase 2+* development. It represents the first implementation of packet switching within GSM, which is essentially a circuit-switched technology.

Packet switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, the available radio resource can be concurrently shared between several users. Through multiplexing of several logical connections on one or more GSM physical channels, GPRS reaches a flexible use of channel capacity for applications with variable bit rates.

The GPRS provides logical channels (Packet Data Channel, PDCH) for transfer of user data. The available radio resources are shared between GPRS and GSM logical channels, which means that any time slot can be used for GSM or GPRS data transmission.

A PDCH's basic transmission unit is a *Radio Block* (see Fig. 1) that requires four time slots in four consecutive *Time Division Multiple Access* (TDMA) frames, called a *Radio Block Period*. A so-called GPRS *Multiframe* comprises 52 TDMA frames and every 13th burst is not used for transmission.

TABLE I
GPRS CODING SCHEMES (CS)

	CS 1	CS 2	CS 3	CS 4
Data rate of one PDCH [kbit/s]	9.05	13.4	15.6	21.4
MAC/RLC block size [Bit]	181	268	312	428
RLC information field size (payload) [bit (byte)]	152 (19)	232 (29)	280 (35)	392 (49)
Max. number of MAC/RLC blocks per IP packet	31	21	17	12

Thus, a multiframe comprises 12 Radio Block Periods. The length of a TDMA frame is 4.615 ms and four different Coding Schemes (CS) are defined, providing data rates from 9.05 kbit/s to 21.4 kbit/s per PDCH, see Tab. I. Since in GPRS the access of all eight slots of a TDMA frame is foreseen, data rates up to 160 kbit/s can be achieved. For a single mobile station its *Multi-Slot Capability* (MSC) defines how many consecutive slots within the TDMA frame may be used.

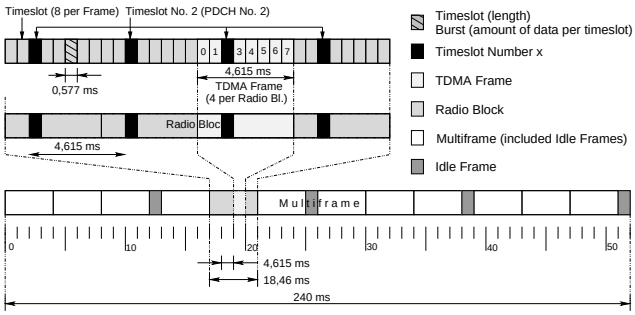


Fig. 1. GPRS TDMA structure

III. WWW TRAFFIC MODEL

Vornefeld uses a WWW traffic model recently proposed by Choi and Limb [4, 7] as a starting point for deriving an analytically tractable description of WWW traffic. The behavior of this model is divided into alternating phases of packet generation and silence. This type of behavior is usually called an on/off type of behavior. An on-phase starts after the arrival of a web request. During this phase, the model generates the packets corresponding to downloading the requested page. The off-phase represents a silence period after all objects have been retrieved. Thus, the on and off-phases equal the page loading times and page reading times, respectively.

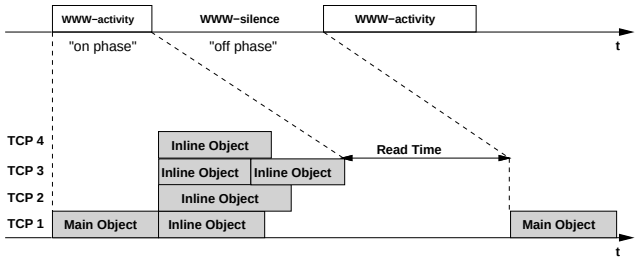


Fig. 2. WWW activity phases and object arrival vs. time

During the on-phase the page's objects are loaded, as shown in Fig. 2. Choi distinguishes two types of objects: the main object containing the document's HTML code and inline objects referred to by the page's HTML code, such as images or JAVA applets.

The random variables used by Choi to describe the object's sizes, the number of inline objects and the length of the viewing time are shown in Tab. II.

TABLE II
RANDOM VARIABLES DESCRIBING CHOI'S MODEL

	Random Variable Distribution	Mean	Standard Deviation
t_{View}	Viewing Time Weibull	39.5 s	92.6 s
n_{Inline}	No. of inline objects Gamma	5.55	11.4
s_{Main}	Size of main object Log-Normal	10 kB	25 kB
s_{Inline}	Size of inline objects Log-Normal	7.7 kB	126 kB

IV. SIMULATION ENVIRONMENT

The (E)GPRS Simulator GPRSim [5] is a pure software solution based on the programming language C++. Up to now models of Mobile Station (MS), Base Station (BS), Serving GPRS Support Node (SGSN), and Gateway GPRS Support Node (GGSN) have been implemented. The simulator offers interfaces to be upgraded by additional modules (see Fig. 3).

Different from usual approaches to establish a simulator, where abstractions of functions and protocols are being implemented, the approach of the GPRSim is based on the detailed implementation of the GSM and (E)GPRS protocols. This enables a realistic study of the behavior of EGPRS and GPRS. The real protocol stacks of (E)GPRS are used during system simulation and are statistically analyzed under a well-defined and reproducible traffic load.

TCP has been implemented based on the description in [8] including slow start and congestion avoidance algorithms. To be able to compare GPRSim results with the analysis of [1, 2], Choi's WWW model has been implemented and embedded into the GPRSim load generator module (see Fig. 3), where it was used as a replacement for the Mosaic WWW Model (please refer to [5] for a detailed description of the GPRSim load generator) used in most GPRSim publications.

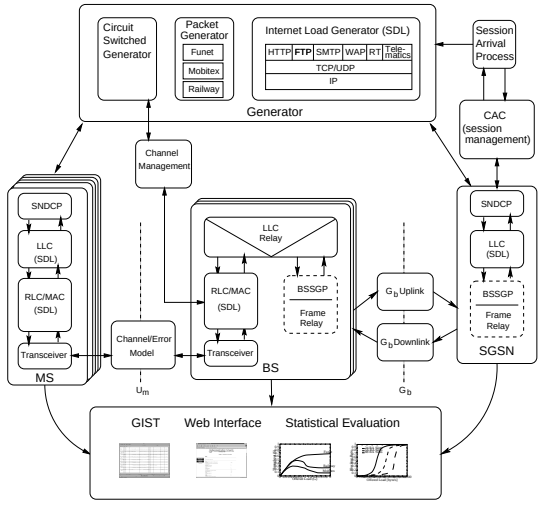


Fig. 3. Overview GPRSim structure

The Hypertext Transfer Protocol (HTTP) defines two ways setup and release of TCP connections can be controlled. There can be one TCP connection per requested Web object, or an existing connection can be "kept alive"for the request of the next Web object belonging to the same Web request. This behavior is called *keep-alive*. Simulations can be performed with keep-alive alternatively enabled or disabled. In our simulation results presented in this paper keep-alive was enabled, because a weaker influence of TCP flow control can be expected in this case.

Within the air interface transmission error model it is decided whether a received data or control block is error free or not. For this purpose a set of mapping curves is used gained from link level simulations that allow the mapping of a C/I value to the corresponding block error rate (BLER) of a radio block [9, 10]. Figure 4 shows the BLER versus C/I results gained from link level simulations. The TU3 (Typical Urban) channel model of GSM 05.05 was assumed there [11].

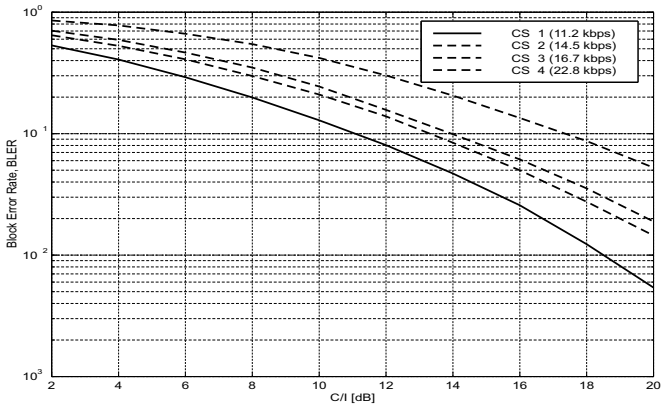


Fig. 4. C/I ratio vs. Block Error Probability (BLEP)

V. ANALYTICAL GPRS MODEL

To allow for analytical modeling, WWW traffic has to be characterized at the point of interest within the system by stochastic processes.

In his recent publications [1, 2] U. Vornefeld presents an analytical modelling framework for Web browsing users accessing the Internet using packet-switching mobile radio networks. He uses numerical algorithms to derive an analytically tractable representation of a well-known Web browsing application model [4] in terms of *Markovian Arrival Process with Marked Arrivals* (MMAP). More precisely, he divides the Web model's behavior in active (on) and passive (off) phases and characterizes the sojourn times in these states by PH-type distributions. During the on phase, he assumes packet arrivals to be generated by a Poisson arrival process.

The MMAP, introduced by He [12], is a stochastic process that is able to represent inherent correlation structures in combination with analytical tractability. It extends the concept of the *Markovian Arrival Process* [13] by allowing to assign an individual service time distribution to each arrival event, called *marking* an arrival. Thus, it is possible to account for individual conditions of radio link and resource allocation strategies regarding multiple users.

A. MMAP Representation of WWW Traffic

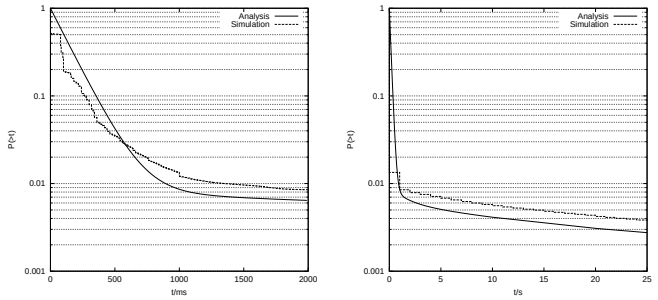
To be analytically tractable, a source description has to consist of exponential phases only. In order to derive such a description Vornefeld fits a PH-distributions to the simulated on phase distribution, using the EM algorithm of [14]. For generating the input curves of this algorithm, Vornefeld has implemented his own simulation model of Choi's application model. As the duration of the off phase of Choi's model is described by a heavy tailed Weibull distribution, which is available in closed form, the off phase duration's Weibull distribution can be approximated by a Hyper-exponential distribution. To achieve this, Vornefeld uses an extended version of the algorithm proposed by Feldmann and Whitt [15]. For a complete description of the procedure please refer to [1] and [2].

The arrival events of the resulting stochastic process are equivalent to an arrival of an IP packet with maximum size, which is assumed to be 536 bytes of IP payload and 40 bytes of TCP and IP header, 576 bytes in total.

Comparing the IP packet inter-arrival time resulting from Vornefeld's analytical representation of Choi's WWW model with the inter-arrival time evaluated during simulations with the GPRSim, see Fig. 5, reveals a good agreement between source characteristics in simulation and analysis, respectively. Therefore negligence of self-similar source characteristics, which is considered as one of the major sources of inaccuracy of many analytical modelling approaches can be considered as eliminated.

B. Service Process

Vornefeld uses link level simulations to determine the *Laplace-Stieltjes-Transform* (LST) of the service time distribution [2]. His simulation of the GPRS radio channel comprises



(a) Range of short inter-arrival time

(b) Tail of IP inter-arrival time CCDF

Fig. 5. Comparison of the analytically derived interarrival time CCDF of Choi's model with the CCDF obtained by simulation. Left: analysis see [2]

a non-frequency selective radio channel with slow and fast fading and Doppler shifts.

Taking the arrival of an IP packet as equivalent of 576 bytes of data at the top of the GPRS protocol stack, he models protocol overhead and segmentation into RLC Blocks. Thus, an IP packet arrival leads to a batch arrival of RLC Blocks (the RLC protocol's *Protocol Data Units* (PDUs)) on RLC level. The CS determines the size of these batches by the number of payload bytes an RLC Block can carry (see Tab. I).

Vornefeld determines the n-point distribution describing the number of transmission attempts needed for successful transmission of all Radio Blocks originating from one IP packet. In a next step he approximates the obtained distribution by a continuous PH-type distribution, using the EM algorithm.

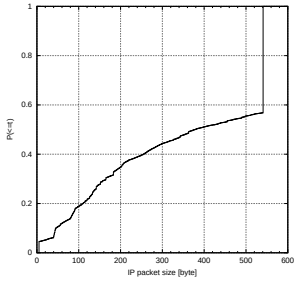


Fig. 6. IP packet size CDF, downlink transfer

In a real GPRS system the IP packet size of course is not always 576 bytes. In Fig. 6 the downlink IP packet size's *Cumulative Distribution Function* (CDF) is displayed for a scenario with 4 MS, 4 PDCHs, MSC 4 and CS-2. As the packet size distribution is independent of scenario parameters, this graph is representative for all simulations. We see that approximately 5% of all IP packets solely consist of TCP/IP header information and only 55% of all packets have the maximum size of 536 bytes. In Vornefeld's analytical model the negligence of smaller packets implies a tendency to overestimate the mean IP packet delay, because it is immediately clear that a smaller IP packet leads to a shorter service time.

VI. GPRS PERFORMANCE EVALUATION

In this section the results of [2] are compared with GPRSim simulation results. We select the IP packet delay as quantity of interest. The analytical results in this section are taken from [2], details on how these results are obtained can be found there. In Fig. 7 the mean IP packet delay vs. mean C/I ratio is displayed for a scenario comprising one WWW user employing CS-2 and using 2 PDCHs and 4 PDCHs, respectively. In both scenarios the MSC is set equal to the number of available PDCHs (MSC 2 resp. MSC 4). In the case of negligible

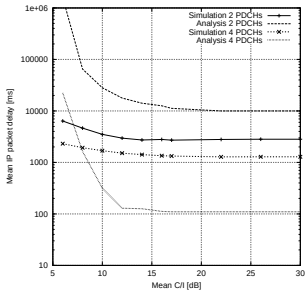


Fig. 7. Mean IP packet delay vs. mean C/I, 1 MS, CS-2, MSC equal to no. of PDCHs. Analysis see [2]

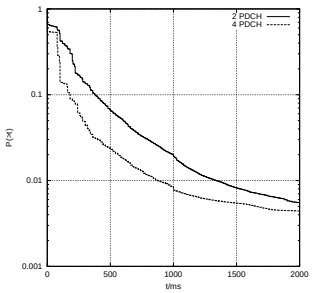


Fig. 8. IP packet inter-arrival time CCDF obtained by simulation, 1 MS, CS-2, 30 dB C/I

BLEP (represented by the high C/I values at the right margin of Fig. 7) we see that the impact of the system capacity is seen differently by simulation and analysis, respectively. While the analysis predicts the mean IP delay to rise two orders of magnitude when the system capacity is reduced from 4 to 2 PDCHs, simulation yields only a doubled mean IP delay.

This behavior is (at least partly) caused by the TCP protocol, which is present in the simulation model (see Fig. 3), but not in the analytical model. TCP performs flow control to avoid network congestion [8]. This means that on IP level the offered traffic is limited when the system capacity is reduced. As a result, in the simulation model we have what is usually called an elastic source behavior, while in the analysis we have an inelastic traffic source. This effect is visualized in Fig. 8. In the absence of packet loss at the air interface the reduction from 4 to 2 PDCHs causes higher IP packet inter-arrival times. Thus, in the same time interval the mean number of packet arrivals declines when the system capacity is reduced. It is immediately clear that a reduced offered traffic leads to a better delay performance and thus represents a counter-tendency to the primary effect of varying the system capacity.

In the analytical model the traffic is not elastic, since the traffic load is only determined by the poissonian packet arrival process in the on-state and not by the underlying system performance or the available capacity. This leads to the huge delay performance difference of the analytical model comparing the cell scenario with 4 PDCHs, where less traffic than the available channel data rate is offered and the cell scenario with 2 PDCHs that represents an overload condition during the on-phase (see Fig. 7). Since the simulation model represents TCP-based traffic that is bursty in nature due to the flow control mechanisms, the mean IP packet delay in the 4 PDCH scenario of the simulation model is higher than in the analytical

model. While the IP packets generated by a Poisson traffic source can be served immediately by the channel with relatively high capacity, batch arrivals generated by the TCP source lead to higher waiting times. Once the available channel capacity becomes smaller than the offered traffic during the on-phase like in the 2 PDCH scenario the delay increases dramatically in the analytical model, while the TCP source is adapting the offered traffic to the available channel capacity leading to a lower offered traffic with shorter queue lengths and lower delays than in the analytical model.

Evaluating the effect of the mean C/I ratio to the mean IP packet delay, which is visible in the left part of Fig. 7, we see that the simulation model is less sensitive to a degradation of channel quality. Again, this is obviously caused by TCP flow control, the argument for this is the same as before. An increasing BLEP leads to performance degradation, because additional time needed for retransmission adds to the IP delay. As a reaction TCP reduces the offered traffic, which partly compensates the performance degradation. But in this case, another aspect becomes effective. The simulation model incorporates a fast retransmission procedure, which especially under bad conditions on the radio channel significantly reduces the IP delay. When an RCL block is transmitted for the first time, it is marked as *Pending ACK* at the transmitting entity. The next time radio resources are allocated to this MS, the radio block is transmitted again, provided that no new data has arrived meanwhile. Thus, the receiving entity does not have to discover the loss of radio blocks and request the radio block for retransmission in case the first transmission attempt was not successful.

In Fig. 7 we illustrate the effect of scaling the system load. For a scenario with aggregated WWW users, 4 PDCHs, MSC 4, CS-2 and an error-free radio channel we increase the number of users from 1 MS to 9 MS. Again we see the influence of TCP flow control. When system load increases, the induced traffic load of a single user is gradually decreased (see Fig. 10), which results in a slower increase of the total system load, compared to linear superposition of multiple users. Besides that, the use of TCP also allows a higher number of aggregated users before the system becomes instable.

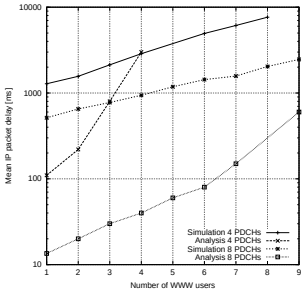


Fig. 9. Mean IP packet delay vs. no of WWW users, MSC 4, CS-2, 30 dB C/I. Analysis see [2]

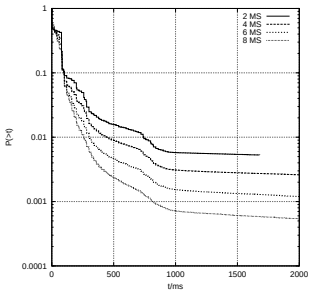


Fig. 10. IP packet inter-arrival time CCDF obtained by simulation, CS-2, 8 PDCHs, 30 dB C/I

As we have already seen in our examination of the C/I ratio's impact, the simulation and the analysis benefit from additional system capacity in different ways. We understand this effect as an indication that the efficiency of the GPRS implementation decreases with increasing system capacity. The understanding of this effect is an important issue for future research.

VII. CONCLUSIONS

Our examination has shown that even extremely complicated analytical modelling approaches are not able to depict the GPRS performance quantitatively, when the regarded traffic sources show an elastic behavior. The TCP slow start and congestion avoidance procedures have been identified to be partly responsible for the deviations between analytical and simulation results. Nevertheless we have also shown that there have to be some other influences to the performance of the simulation system. Our simulation results indicate the existence of some kind of inefficiency in the way the combination of TCP and GPRS utilizes increasing system capacity. Finding the reason for this behavior is subject to future research, the reader may kindly refer to our future publications concerning this subject. Additionally we propose that future research in the area of analytical modelling should primarily focus on embedding the elastic property of TCP-based Internet traffic into the source model under retainment of the ability to closely approximate heavy-tailed on and off phase duration distributions.

REFERENCES

[1] U. Vornefeld, "Analytical performance evaluation of mobile internet access via GPRS networks," in *Proc. of the European Wireless 2002*, (Florence, Italy), 2002.

[2] U. Vornefeld, "Analytical concepts for gprs network dimensioning," in *Proc. of the International Communications Conference '02*, (New York, US), April 2002.

[3] L. Kleinrock, *Queueing Systems: Theory*, vol. 1. New York, London, Sydney, Toronto: John Wiley & Sons, 1 ed., 1975.

[4] H. K. Choi and J. O. Limb, "A behavioral model of web traffic," in *Proc. of the 7th International Conference on Network Protocols (ICNP 99)*, (Ontario, Canada), October 1999.

[5] P. Stuckmann, "Simulation Environment GPRSim: Tool for Performance Analysis, Capacity Planning and QoS Enhancement in GPRS/EDGE Networks," Technical Report, available via WWW: `http://www.comnets.rwth-aachen.de/~pst`.

[6] P. Stuckmann and O. Paul, "Dimensioning GSM/GPRS networks for circuit- and packet-switched services," in *Proceedings of the 10th Symposium on Wireless Personal Multimedia Communications, ISBN 87-988568-0-4*, (Aalborg, Denmark), pp. 597–602, September 2001.

[7] P. Tran-Gia, D. Staehle, and K. Leibnitz, "Source traffic modeling of wireless applications," *Int. J. Electron. Commun. (AE)*, vol. 55, no. 1, pp. 27–36, 2001.

[8] R. Stevens, *TCP/IP Illustrated*, vol. 1. Massachusetts: Addison-Wesley, October 1996.

[9] A. Furuskär, S. Mazur, F. Müller, and H. Olofsson, "EDGE: Enhanced Data Rates for GSM and TDMA/136 Evolution," *IEEE Personal Communications*, pp. 56–65, June 1999.

[10] J. Wigard, T. T. Nielesen, P. H. Michaelsen, and P. Mogensen, "BER and FER Prediction of Control and Traffic Channels for a GSM Type of Air-Interface," in *Vehicular Technology Conference (VTC)*, pp. 1588–1592, 1998.

[11] ETSI, "Digital cellular telecommunications system (Phase 2+) (GSM); Radio transmission and reception (GSM 05.05)," Technical Specification 5.2.0, European Telecommunications Standards Institute, Sophia Antipolis, France, Jan. 1996.

[12] Q. He, "Queues with marked customers," *Adv. Appl. Prob., 28*, vol. 28, pp. 567–587, 1996.

[13] M. F. Neuts, "A versatile markovian point process," *Journal of Applied Probability*, vol. 16, pp. 764–779, 1979.

[14] S. Asmussen, O. Nerman, and M. Olsson, "Fitting phase type distributions via the EM algorithm," *Scand. J. Statist*, vol. 23, pp. 419–441, 1996.

[15] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tailed distributions to analyze network performance models," *Performance Evaluation*, vol. 31, pp. 245–279, 1998.