Analytical Concepts for GPRS Network Dimensioning

Ulrich Vornefeld
Aachen University of Technology
Chair of Communication Networks
Kopernikusstr. 16, 52064 Aachen, Germany

***Abstract*— This paper deals with analytical concepts for the dimensioning of packet switching mobile radio networks. A realistic characterization of source traffic with means of marked Markovian arrival processes (MMAP) serves as an input process for a queuing model. The analysis of this model permits to derive the quantities of interest and allows to determine the achievable quality of service, which is quantified in terms of utilization and packet delays. The applicability of the approach is demonstrated by embedding our analysis into the framework of GSM/GPRS.**

I. Introduction

The advent of the 3rd generation mobile radio communication systems of the IMT 2000 family such as UMTS or CDMA2000 is accompanied by a change of the inherent switching technology. Migrating from 2nd generation channel switching systems via hybrid systems such as General Packet Radio Service (GPRS), next generation systems rely completely on packet switching. Besides the radio aspects like propagation and interference, further attention has to be paid to traffic issues. While wired systems are meanwhile able to provide bandwidth in abundance, the bandwidth for radio communication is limited by the available frequency spectrum, which is a scarce and sometimes tremendously expensive resource. Thus, in contrast to fixed packet switching networks, enhanced traffic engineering in mobile radio networks for an efficient and effective use of capacity is worth the effort of a closer look. Besides voice transmission, the access to Internet services is the driving application in communication networks. For our analytical investigations, we consider the most popular Internet service – web browsing. We will show the applicability of our analytical concepts embedded into the framework of GPRS. We derive the system's utilization, packet delays and discuss an the effects of the correlation structure of the arrival process in an environment with unsteady radio transmission quality.

II. Traffic Modeling

To allow for analytical modeling, the incoming traffic streams have to be characterized by stochastic processes at the point of interest within the system. Dealing with packet traffic, simple Poisson process modeling fails [1], since Poisson processes cannot represent inherent correlation structures or effects like long range dependence. Another requirement for the stochastic process is the ability to integrate batch arrivals. On its way through the protocol stack, an application packet undergoes several segmentation procedures, which result in batch arrivals of Service Data Units (SDUs) seen by the next layer. A stochastic process that is able to represent correlation structures and a large group of interarrival time distributions together with analytical tractability is the Markovian Arrival Process (MAP). The potential power of a MAP is demonstrated by a result due to Asmussen and Koole [2]. They proved that any stationary point process can be approximated arbitrarily close by a MAP. Performance analysis in cellular networks makes it necessary to analyze each traffic flow individually, since the current environment in terms of radio coverage determines the service process, which might be different for each user. The extension of the MAP by He [3] makes it possible to assign an individual service process to each single arrival event. With the *Markovian arrival process with marked arrivals (MMAP)*, introduced by He, we are able to model the required individual arrival *and* service process of each user.

A. The Marked Markovian Arrival Process (MMAP)

The MMAP is controlled by an M-state finite Markov chain with an irreducible generator matrix. The state space consists of k transient phases and comprises an absorbing state. The duration of the transient phases i, $i = 1, \ldots, k$, is negative exponentially distributed with parameter λ_i. The MMAP is completely described by a set of matrices $\mathbf{D}_0$ and $\mathbf{D}_n$, $1 \le n \le K$, where $\mathbf{D}_n$, $n \ge 0$ are $M \times M$ matrices. $\mathbf{D}_0$ contains negative diagonal elements, nonnegative off-diagonal elements and is assumed to be non-singular. The latter requirement ensures that infinitely many transition epochs are marked and the arrival process never ceases. The elements of $\mathbf{D}_n$ with $1 \le n \le K$ are nonnegative. In addition, we have

$$\mathbf{D} = \mathbf{D}_0 + \sum_{n=1}^{K} \mathbf{D}_n, \tag{1}$$

where $\mathbf{D}$ is the irreducible infinitesimal generator of the Markov process. When the arrival is marked by n, it is called an arrival of a class n packet. The MMAP permits the analysis of queuing models with multiple non-Poissonian arrival streams with individual service time distributions for each stream. The service times of class n packets ($n = 1, \ldots, K$) that arrive with the transition from state i to state j can be i.i.d. according to a distribution function $H_{n,i,j}(t)$ with mean $h_{n,i,j}$. In our model, it is not necessary to assign an individual service time distribution to each state transition. It is sufficient, when each arrival class n can have its own service time distribution. Thus, the analysis of our model is the analysis of an $MMAP_i/G_i/1$ queue. Following the analysis in [4], we define $\mathbf{D}_n(t)$ as an $M \times M$ matrix whose (i,j)th element $D_{n,i,j}(t)$ is given by

$$D_{n,i,j}(t) = D_{n,i,j} H_n(t) \quad i,j = 1,\ldots,M \tag{2}$$

The stationary probability vector $\boldsymbol{\pi}$ of the underlying Markov chain satisfies

$$\boldsymbol{\pi} \sum_{n=0}^{K} \mathbf{D}_n = \mathbf{0} \quad \text{and} \quad \boldsymbol{\pi e} = \mathbf{1}, \tag{3}$$

where $\mathbf{e}$ denotes an $M \times 1$ column vector whose elements are all equal to one. The mean arrival rate λ_n and the load or utilization ρ_n of the queue caused by class n packets is given by

$$\lambda_n = \boldsymbol{\pi} \mathbf{D}_n \mathbf{e} \quad \text{and} \quad \rho_n = \boldsymbol{\pi} \int_{0-}^{\infty} t \, d\mathbf{D}_n(t) \mathbf{e}. \tag{4}$$

The overall arrival rate and the overall utilization are given by

$$\lambda = \sum_{n=1}^{K} \lambda_n \quad \text{and} \quad \rho = \sum_{n=1}^{K} \rho_n. \tag{5}$$

1) The Virtual Waiting Time: In the following, we consider a stable queue ($\rho < 1$), i.e. the queue serves all requests in finite time. The generator of the underlying Markov chain after excising the busy periods is given by the $M \times M$ matrix $\boldsymbol{Q}$

$$\boldsymbol{Q} = \boldsymbol{D}_0 + \int_{0-}^{\infty} d\boldsymbol{D}(t) \, e^{\boldsymbol{Q}t} \tag{6}$$

and the stationary state vector of the chain $\boldsymbol{\kappa}$, given the system is idle, obeys

$$\boldsymbol{\kappa Q} = \mathbf{0}, \qquad \boldsymbol{\kappa e} = \mathbf{1}. \tag{7}$$

The the LST vector of the workload in state j is given by

$$\boldsymbol{v}^*(s) = (1-\rho) s \boldsymbol{\kappa} [s\boldsymbol{I} + \boldsymbol{D}_0 + \boldsymbol{D}^*(s)]^{-1}, \quad \mathrm{Re}\{s\} > 0. \tag{8}$$

and (9) provides the LST of the virtual waiting time of each class k

$$\boldsymbol{w}_k^*(s) = \frac{\boldsymbol{v}^*(s)\boldsymbol{D}_k}{\lambda_k}, \quad \mathrm{Re}\{s\} > 0. \tag{9}$$

The Complementary Cumulative Distribution Function (CCDF) of the virtual waiting time is obtained by inverting the following LST:

$$G_{W_k}^*(s) = \frac{1 - \boldsymbol{w}_k^*(s)\boldsymbol{e}}{s}, \quad \mathrm{Re}\{s\} > 0. \tag{10}$$

2) The Sojourn Time: The sojourn time of a packet inside the system consists of the waiting time or queuing delay and the service time. Both values are described by random variables, i.e. the sojourn time is the convolution of both for each class k

$$\boldsymbol{r}_k^*(s) = \frac{\boldsymbol{v}^*(s)\boldsymbol{D}_k^*(s)}{\lambda_k}, \quad \mathrm{Re}\{s\} > 0. \tag{11}$$

The CCDF of the sojourn time is obtained by inverting the LST

$$G_{R_k}^*(s) = \frac{1 - \boldsymbol{r}_k^*(s)\boldsymbol{e}}{s}, \quad \mathrm{Re}\{s\} > 0. \tag{12}$$

3) MMAP Representation of PH-Distributions: The MAP comprises as a special case phase-type (PH) renewal processes, introduced by Neuts [5], which is another useful property for modeling arrival processes and estimating arrival time distributions. PH processes contain Erlang, E_k and hyper-exponential, H_k processes, as well as finite mixtures of those. A PH renewal process with representation $(\alpha, \mathbf{T})$ is a MAP with $\mathbf{D_0} = \mathbf{T}$ and $\mathbf{D_1} = -\mathbf{T}\mathbf{e}\alpha$.

B. The WWW-Model

To derive an analytical tractable model for web-traffic, we follow the recent work of Choi [6]. The traffic model is an on/off-model with alternating phases of packet generation and silence. An on-phase starts after the arrival and acceptance of a web request. During this phase, the model generates the packets corresponding to 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 viewing times, respectively.

During the on-phase the page's objects are loaded. We distinguish two types of objects: the main object containing the document's HTML code and inline objects, such as linked objects, images or JAVA applets. To fetch all those inline objects modern browsers open several TCP-connections in parallel after the successful retrieval of the main object. Choi uses the following random variables to describe the object's sizes, the number of inline objects and the length of the viewing time:

TABLE I
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

Since we aim at a representation of the web-model, which completely consists of exponential phases in order to be analytical tractable, we fit a PH-distribution to the simulated on-phase durations, using the EM algorithm of [7]. The duration of the off-phase of Choi's model is described by a heavy tailed Weibull distribution. Since this heavy tailed property is essential in performance evaluation, it has to be kept in the analytical model. Based on the work of Feldmann and Whitt [8], we have extended their algorithm to represent heavy tailed distributions by hyper-exponential distributions. Deviating from [8], we fit the exponential phases at the points determined by the algorithm to the original distribution function and to its derivative, which is available in closed form for the Weibull-distribution. For further optimization, we piece-wisely linearize the CCDF and use linear programming and the simplex

TABLE II
PARAMETERS OF THE WWW-MODEL

Maximum Segment Size (MSS)	536 byte
TCP-header	20 byte
IP-header	20 byte
On-phase arrival rate λ_{On}	6.667 1/s
Overall arrival rate λ_{WWW}	1.94226 1/s
Mean Data Rate	8.95 kbit/s

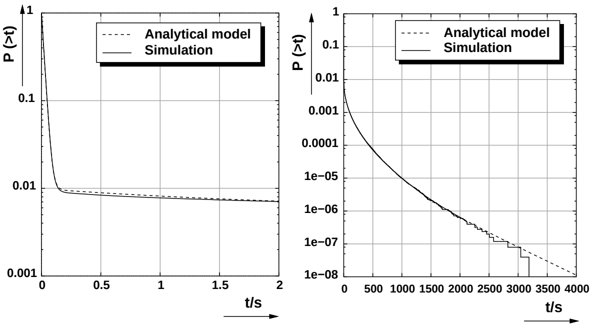

Fig. 1. Comparison of the analytical derived interarrival time CCDF of Choi's model with the CCDF obtained by simulation

algorithm. Fig. 1 shows a good agreement between the arrival time's CCDFs obtained by simulation and the analytical model.

To obtain the packet interarrival time distribution of the superposition of two or more WWW-users, the single processes can be aggregated by means of Kronecker sums. Since our model comprises 10 states, it becomes intractable with the aggregation of 3 or more processes. Therefore we approximate two aggregated distributions using our approximation algorithm with optimization, before another process is superimposed. The resulting PH-distributions of the superposition of up to 10 WWW-processes are depicted in Fig. 2. The right part shows the CCDF's tails and it can be clearly seen that the heaviness of the tails declines from a nearly parallel run to the x-axis for one source to a more exponential behavior for the aggregation of 10 sources. We use the MMAP representation of the PH distribution for describing the arrival process, see Sec. II-A.3.

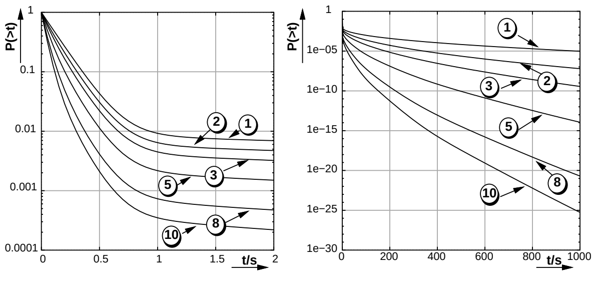

Fig. 2. CCDF of packet interarrival times for simultaneous operation of 1 to 10 WWW-users

III. THE UNRELIABLE RADIO LINK

Radio links suffer from unreliable transmission causing packet error rates several orders of magnitude higher than in fixed networks. Error recovery schemes based on retransmissions of flawed packets create additional load on the radio link that cannot be neglected due to the comparable low bandwidth of the radio link. Several parameters, such as signal strength, co- and inter-channel interference and fading processes, influence the transmission quality. These effects impair the transmission quality, lead to bit errors that, if not possible to correct, accumulate to packet errors. Automatic Repeat Request (ARQ) schemes try to recover packet or block errors by retransmitting flawed packets. Unfortunately also retransmitted packets are in danger to be corrupted once again. Hence, when describing the retransmission process, not only the first retransmissions have to be taken into account, but also the succeeding transmission attempts have to be considered.

In our analytical model, we incorporate packet transmission, transmission of acknowledgments and packet retransmissions into the service process and describe these effects with means of individual service time distributions. Link level simulations provide some insight into the wireless channel's behavior and serve as an input to derive the corresponding service time distributions. In [9], we use n-point distributions to model the required number of transmission slots service, while in the following we apply PH-distributions fitted to stochastic link level simulation results using the EM-algorithm [7].

IV. GPRS

The *General Packet Radio Service* (GPRS) in GSM provides packet-switching logical channels (Packet Data CHannel, PDCH) for data applications. At the air-interface with its time division multiple access (TDMA) scheme the radio resources are assigned to the mobile station only temporarily on a per-packet basis. The PDCH's basic transmission unit is a *radio block* that requires four time slots in four consecutive TDMA frames. The length of a TDMA-frame is 4.615 ms and every 13th burst is not used for transmission. Thus, the mean transmission time of a radio block of 456 bits sums up to 20 ms, provided only 1 PDCH per frame is available for GPRS. Four different coding schemes (CS) are defined providing data rates from 9.05 kbit/s to 21.4 kbit/s, see Tab. III. 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 the single mobile station its Multi Slot Capability (MSC) defines how many consecutive slots within the TDMA-frame may be used. Currently defined are MSC1, MSC2 and MSC4 [10] but we also consider MSC8.

A. Modeling the Service Process

We use a non-frequency selective wireless channel model that comprises slow and fast fading, Doppler shifts caused by the mobile station's movement and appropriate power spectral densities. Our model is based on the work of Loo [11]. The channel model provides a stochastic description of the fading

TABLE III
GPRS CODING SCHEMES (CS)

	CS 1	CS 2	CS 3	CS 4
Code rate	1/2	≈ 2/3	≈ 3/4	1
Data rate [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)
Number of MAC/RLC blocks per IP-packet	31	21	17	12

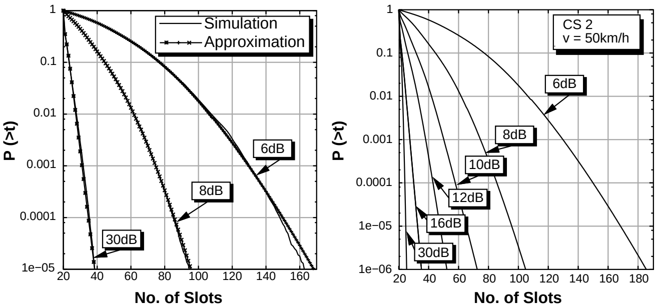

Fig. 3. No. of required slots for the transmission of one IP-packet, CS 2 and v=50 km/h. Left: Approximation vs. Simulation. Right: Approximation

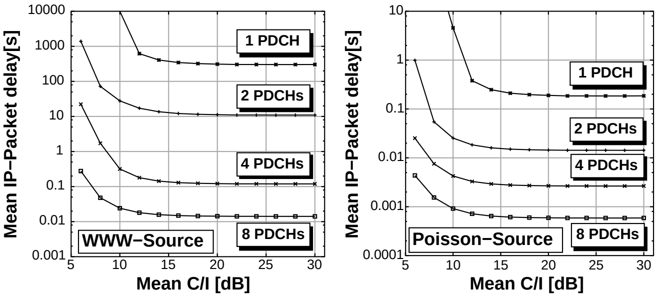

Fig. 4. Mean IP-packet delay for one WWW-user vs. C/I for 1,2,4 and 8 PDCHs, MSC1, MSC2, MSC4 and MSC8, CS 2 and v= 50 km/h

process that serves as an input for a GMSK-demodulator and a channel decoder. The decoder's parameters are set according to the four CS, see [12]. With the parameters given in Tab. II the arrival of an IP-packet is equivalent to the arrival of 576 bytes of data at the Subnet Dependent Convergence Protocol (SNDCP) level. Further segmentation, header expansion and the generation of a Frame Check Sequence (FCS) of 3 bytes takes place at Logical Link Control (LLC), Medium Access Control and Radio Link Control (RLC) level. Thus, the arrival of an IP-packet corresponds to a batch arrival of Protocol Data Units (PDU) at RLC-level. The CS determines the size of those batches, depending on the different number of payload bits each RLC-PDU can carry. The last row of Tab. III shows how many RLC-PDUs are needed to carry one IP-packet. E.g., with CS 1, 31 RLC-PDUs, transmitted in 124 burst, are needed for an errorfree transmission of one IP-packet. Starting with an segmented IP-packet, we determine the n-point distribution of the overall number of slots necessary to transmit the corresponding number of RLC-blocks. With degrading channel quality, the packet error probability rises and as a result the number of required slots attempts increases. In the next step we approximate the obtained n-point distribution by a continuous PH-distribution. Although an n-point distributed service process is analytical tractable, see [9], convolution and mixture operations are more convenient dealing with PH-distributions. Fig. 3 shows the CCDFs of the required numaber of slots to transmit one IP-packet with CS 2 and at a mobile speed of v=50 km/h. The left part of the figure compares the simulated distribution with the fitted distribution. The PH-distributions comprise 8 to 14 states and although Fig. 3 shows a good agreement, the first 3 moments and the area between the two curves are considered as an additional goodness of fit measure. For starting a retransmission the sender needs to receive an negative acknowledgment (NACK). Generating and transmitting the NACK also takes time. In our calculations, we have not considered this additional delay, but, once quantified in terms of a probability density, it would fit into our model by convoluting its density with the derived PH service time density.

B. Analytical Dimensioning

To gain some insight into the system, we consider the downlink and model one GPRS radio cell as a queuing system with a single server. For each user, individual service time distributions are applied, reflecting different coding schemes and radio channel conditions. Due to space restrictions, we limit our discussion to CS 2, which is currently the most frequently used CS. For a discussion of the coding scheme's impact, see [9]. We are interested in calculating the system's utilization caused by the IP-traffic originating from web-browsing and in experienced delays. We calculate these quantities for CS 2 depending on the mean C/I ratio and the number of available PDCHs. The system's utilization is calculated using (4) and (5). The waiting time and sojourn time CCDF are given as their LST in (10) and (12). We obtain the moments by differentiation and use numerical inversion algorithms, such as in [13] to reveal the run of the CCDF in the time domain.

C. Discussion of Results

In the following, we select the delay experienced by an IP-packet as the quantity of interest for dimensioning the system. The delay comprises the queuing delay and the transmission delay, i.e. the delay is given as the sojourn time of the packet in the queue.

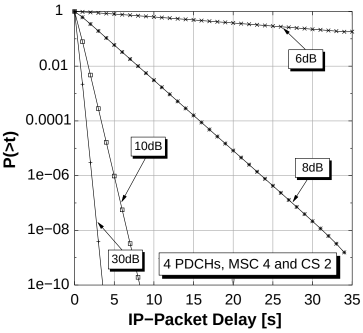

Fig. 5. CCDF of IP-packet delay for one WWW-user at 6,8,10 and 30 dB with CS 2, 4 PDCHs, MSC 4 and v=50 km/h

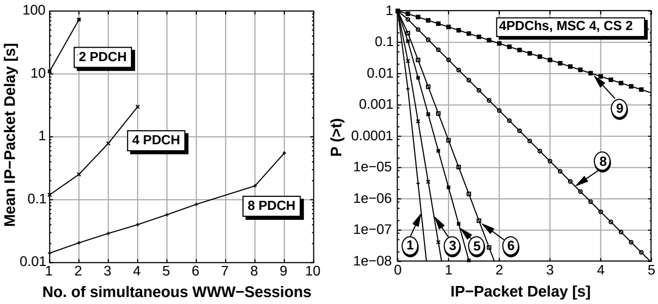

Fig. 6. IP-packet delay v. no of WWW-users for 4 PDCHs, MSC 4, CS 2, 30 dB mean C/I and v= 50 km/h

The left part of Fig. 4 shows the mean delay of a single IP-packet vs. mean C/I for 1,2,4 and 8 PDCHs. With decreasing C/I the system's load increases, since more and more packets have to be retransmitted. Depending on the number of PDCHs, the delay increases asymptotically (note the logarithmic scaling) with decreasing C/I, since the system is flooded by an avalanche of retransmission until the queue becomes unstable ($\rho > 1$). The mean delay for a single IP-packet is remarkable high. Even for nearly errorfree transmission at C/I values around 30 dB, it takes more than 10 s on average to transmit a single IP-packet even with 2 PDCHs. Since the queue is only moderately loaded, this cannot cause the high delay. A comparison of the arrival process to a Poisson arrival process that creates the same load (right part of Fig. 4), reveals that the correlation structure of the WWW arrival process is responsible for the queuing delay. This effect should be quantified and studied in more detail.

In Fig. 5, we examine the CCDF of the delay for a single user, 4 PDCHs, CS 2 and different C/I values. Although the arrival process exhibits some heavy tailed properties and despite the PH service time distribution, the CCDF of the delay shows an exponentially decaying behavior. The service process determines the run of the curve for probabilities near one. Originating from Fig. 3, a small decrease in the C/I value leads to a large increase in delays, especially for low C/I values.

Fig. 6 exhibits, how the packet delay is affected by scaling the scenario. Starting with 1 user the load is increased by increasing the number of WWW users. With 8 PDCHs, MSC 8, CS 2 and errorfree transmission, the system is able to support up to 9 WWW-users. Increasing the load near the system's capacity limit has to be paid by an unproportional high increase in delay. The left part of Fig. 6 shows the scalability for 2, 4 and 8 PDCHs (note the logarithmic scaling).

V. Conclusions

We presented some analytical concepts for dimensioning packet switching mobile radio networks and showed how to describe the system's behavior aiming at analytical tractable models for source traffic and radio link characterization. Based on the concepts outlined in [9], we replaced the somewhat theoretical service time description by phase type distributions obtained from link level simulations and made another step closer to reality. The properties of PH service time distributions permit the extension of the model e.g. by mixtures or convolutions of service time distributions without loosing its analytical tractability. The results reveal the importance of the correlation structure of the arrival process and its impact on the queuing delay. It can be expected that correlated service time distributions also affect the performance. Therefore, link level investigations should also address the correlation of packet errors. Since MMAP-modeling permits the assignment of an individual service time to each arriving packet, correlated service time processes can be incorporated into the modeling approach.

References

[1] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[2] S. Asmussen and G. Koole, "Marked point processes as limits of markovian arrival streams," *J. Appl. Probab.*, , no. 30, pp. 365–372, 1993.

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

[4] T. Takine, "Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions," 1999, Submitted to Queueing System. Available at: citeseer.nj.nec.com/75314.html.

[5] M.F. Neuts, *Matrix-Geometric-Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, 1981.

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

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

[8] 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.

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

[10] 3GPP TS 25.322, "Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2," Technical Specification V4.0.0, Release 4, 3rd Generation Partnerschip Project, Technical Specification Group GERAN, Jan. 2001.

[11] C. Loo and J.S. Buttworth, "Land mobile satellite channel measurements and modeling," *Proc. of the IEEE*, vol. 86, no. 7, pp. 1442–1463, Juli 1998.

[12] 3GPP TS 45.003, "Channel Coding," Technical Specification V4.0.0, Release 4, 3rd Generation Partnerschip Project, Technical Specification Group GERAN, Jan. 2001.

[13] G. Choudhury and W. Whitt, "Computing distributions and moments in polling models by numerical transform inversion," *Performance Evaluation*, vol. 25, pp. 267–292, 1996.