Call Level Performance Analysis of 3^{rd} Generation Mobile Core Networks

Sándor Rácz, Miklós Telek
Technical University of Budapest, Pázmány Péter S. 1, H-1111, Budapest, Hungary,
Tel: +36 1 4632084, Fax: +36 1 4633266
raczs@ttt-atm.ttt.bme.hu, telek@hit.bme.hu

Gábor Fodor
Ericsson Research, Torshamnsg. 23, Kista, SE-164 80 Stockholm, Sweden
Tel: +46 8 4043084, Fax: +46 8 4047020
Gabor.Fodor@era-t.ericsson.se

Abstract—**In this paper we develop a call level model of UMTS** *core networks* **where calls belonging to one of the four UMTS service classes arrive randomly. Arriving calls are granted service depending on the call's service class, the required maximum- and minimum bandwidth, and the available network resources at the arrival instance. We use a Markov model of transmission links to derive GoS (blocking probability) and QoS (throughput) measures under two reasonable and technologically feasible bandwidth sharing policies. We conclude that one of these policies is able to provide GoS/QoS guarantees for a wide range of traffic mixes. We argue that the results are applicable to the** *all-IP/MPLS* **based new UMTS architecture.**
Key words: 3^{rd} generation mobile networks, UMTS networks, bandwidth sharing objectives, multi-rate loss models, blocking probabilities, Markov models.

I. Introduction

One of the key architectural aspects of 3^{rd} generation mobile networks is the separation of the *access* part from the *core network*. This separation supports the evolution of various access technologies and the continuous development of new services provided by the core network [3], [4]. As such, the core of UMTS/IMT-2000 networks is a multi-service network providing GoS (blocking probability) and QoS (throughput) guarantees to four types of service classes [14]:

- The *conversational class* provides high quality access to a range of different services including high bit rate services. This class is suitable for the demanding user who wishes to receive bandwidth guarantees similar to that of the CBR class in ATM.
- The *streaming class* is designed to carry high bandwidth, variable bit rate services, such as a medium- or high quality video- or teleconferencing service. One common feature of the calls belonging to this class is that their holding time is independent of the actual throughput received during the residency time in the system.
- The *interactive class* supports less demanding services typically supported by today's best effort IP networks, including file transfer, web browsing or telnet applications. UMTS networks are expected to provide some form of throughput guarantee even to these types of services. The holding time of the interactive class calls typically depends on the throughput (the transfer of a file, for instance, would take half of the time with doubled throughput).
- The *background class* is of the best effort type, meaning that background calls receive the whatever bandwidth is "left over" by the calls of the higher priority service classes. Examples of this class include e-mail or low quality file transfers. With respect to their holding time, this class is similar to the interactive class.

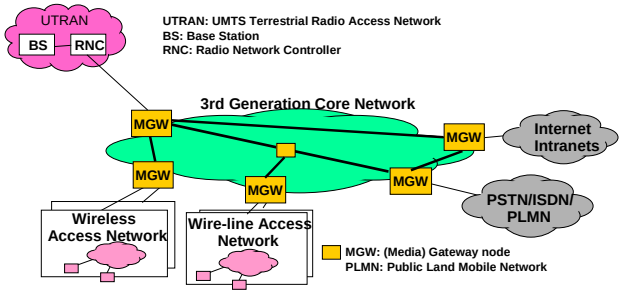

Fig. 1. Different access networks with a common UMTS core network

Figure 1 highlights some aspects of the UMTS core network, notably the separation of the different access networks from the core network through *gateway* nodes that exercise call admission control at the edges. Regarding the candidate technologies, the core is expected to be based on fast packet switching techniques with some connection oriented resource reservation mechanism, like ATM or IP/MPLS [12], [15]. In any case, the dimensioning and performance analysis of UMTS networks requires that call level models that take account of the *blocking probability* and *throughput* trade-off in a dynamic environment (where calls arrive and depart randomly in time) are available. As it is detailed in Section II, the widely available call level

models for circuit switched [10] or ATM [5] networks are not directly applicable in the UMTS environment, because they do not take into account the UMTS-specific service class definitions. Whereas a number of papers have studied various access networks [11] - [13], in this paper we focus on the core network, where the main objective is to study bandwidth sharing alternatives, such that the GoS/QoS requirements of the UMTS services classes are met and the bandwidth utilization is kept high.

The contribution of this paper is two-fold. Firstly, we extend the classical multi-rate models to include the UMTS service classes. Secondly, by focusing on the performance of the background class, while guaranteeing QoS for the first three classes, we propose a simple bandwidth sharing policy that performs better than complete sharing in terms of throughput and blocking probabilities.

We organize the paper as follows. Section II describes the network model. Section III defines the performance measures and associated throughput constrains. Section IV discusses simple possible bandwidth sharing policies in the UMTS core network. Section V discusses numerical results assuming different traffic mixes. We conclude in Section VI.

II. Modeling the UMTS Core Network

Since the core network exercises admission control on a link-by-link basis, we formulate the Markov model of a single transmission link. At this abstraction level our model is suitable for an ATM link as well as for a link of an MPLS label switched path. From the description of the UMTS service classes it is clear that *we only need to model the service classes with throughput guarantees* (i.e., conversational, streaming and interactive) *and calculate the (mean and the distribution of the) bandwidth left over for the background class.*

The system under consideration consists of a transmission link of capacity C, which is supposed to be an integer number in some suitable bandwidth unit, say $Mbps$. Calls arriving at the link belong to one of the following three traffic classes:

- *Conversational* service class calls are characterized by their peak bandwidth requirement b_1, flow arrival rate λ_1 and departure rate μ_1;
- *Streaming* class calls are characterized by their peak bandwidth requirement b_2, minimum bandwidth requirement b_2^{min}, flow arrival rate λ_2 and departure rate μ_2. Although the bandwidth occupied by streaming calls may fluctuate as a function of the link load, their actual holding time is not influenced by the received throughput throughout their residency in the system. This is the case for instance with a streaming video codec, which, in the case of throughput degradation decreases the quality of the video images and thereby occupies less bandwidth.
- *Interactive* class flows are characterized by their peak bandwidth requirement b_3, minimum bandwidth requirement b_3^{min}, flow arrival rate λ_3, and their *ideal* departure rate μ_3. The ideal departure rate is experienced when the peak bandwidth is available. The real instantaneous departure rate is proportional to the bandwidth of the flows.

We denote the actual bandwidth allocated (reserved) to a flow of class-2 (streaming) and class-3 (interactive) in a given system state with b_2^r and b_3^r, both of which vary in time as flows arrive and depart. We will also use the quantity $r_{min} := b_{min}/b$ associated with the streaming and the interactive classes. One may think of an interactive class call as one that upon arrival has an associated amount of data to transmit (W) sampled from an exponentially distributed service requirement, with distribution $G(x) = 1 - e^{-\frac{b_3}{\mu_3}x}$, which in the case when the peak bandwidth b_3 is available during the entire duration of the call gives rise to an exponentially distributed service time with mean $1/\mu_3$. Since the free capacity of the link fluctuates in time according to the instantaneous number of calls in service, the bandwidth given to the interactive calls may drop below the peak bandwidth requirement, in which case the actual holding time of the call increases.

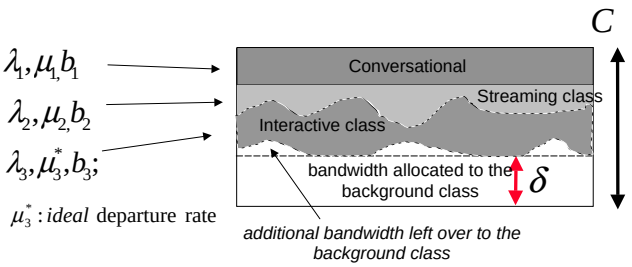

Fig. 2. Modeling a single transmission link with the four types of UMTS service classes

As illustrated in Figure 2, the first three types of flows with throughput guarantees arrive according to independent Poisson processes, and the holding time for the conversational and streaming class calls are exponentially distributed and the (phase type distributed) holding time of the interactive class calls are determined using the Markov model of the transmission link.

In what follows we are interested in the performance of two simple bandwidth sharing strategies that provide GoS/QoS bounds for the first three service classes and also throughput guarantee for the lower priority background class.

III. Performance Measures and Constrains

In contrast to classical multi-rate models [10], our link model allows the fluctuation of the occupied bandwidth, and therefore the relevant performance measures include the blocking probabilities and the class-wise throughput values. In order to evaluate the performance of bandwidth

sharing, we need the rigorous definitions of the class-wise throughput values and associated constrains.

A. Throughput Definitions

The throughput of the conversational calls is simply the allocated bandwidth which is a constant value (b_1). It is intuitively clear that the residency time of the interactive class flows depends not only on the amount of data they want to transmit (which is a random variable), but also on the bandwidth they receive during their holding times. Similarly, the amount of data transmitted through a streaming class flow depends on the received bandwidth. In order to specify this relationship we define the following quantities:

- $\theta_2(t)$ and $\theta_3(t)$ defines the instantaneous *throughput* of streaming and interactive flows at time *t*, respectively, (e.g., if there are n_1, n_2, n_3 conversational, streaming, and interactive flows in the system at time *t*, respectively, the instantaneous throughput are $\min(b_2, (C - n_1 b_1 - n_3 r_3 b_3)/n_2)$ and $\min(b_3, (C - n_1 b_1 - n_2 r_2 b_2)/n_3)$) for streaming and interactive flows, respectively. Note that $\theta_2(t)$, and $\theta_3(t)$ are discrete r.v. for any $t \geq 0$.
- $\tilde{\theta}_t = \frac{1}{t}\int_0^t \theta_2(\tau)d\tau$ defines the *throughput* of the streaming flow whose holding time is t.
- $\tilde{\theta} = \int_0^\infty \tilde{\theta}_\tau \; dF(\tau) = \mu_2 \int_0^\infty \tilde{\theta}_\tau \; e^{-\mu_2 \tau} \; d\tau$ (r.v.) defines the *throughput* of the streaming flow, where $F(t)$ denotes the cdf. of the exponentially distributed flow holding time.
- $T_x = \inf\{t \mid \int_0^t \theta_3(\tau)d\tau \geq x\}$ (r.v.) gives the time it takes to transmit x amount of data of an interactive flow,
- $\hat{\theta}_x = x/T_x$ defines the *throughput* of the interactive flow during the transmission of x data unit. Note that θ_x is a continuous r.v.
- $\hat{\theta} = \int_0^\infty \hat{\theta}_x \; dG(x) = \mu_3/b_3 \int_0^\infty \hat{\theta}_x \; e^{-x \; \mu_3/b_3} \; dx$ (r.v.) defines the *throughput* of the interactive flow, where the amount of transmitted data is exponentially distributed with parameter μ_3/b_3.

In the following we assume that the applied CAC is such that the maximum accepted blocking probability associated with the conversational, streaming and interactive traffic classes (B_1^{max}, B_2^{max} and B_3^{max} respectively) and the minimum accepted throughput for th streaming and interactive calls ($\tilde{\theta}^{min}$, $\hat{\theta}^{min}$) are ensured [9]. Since there is no minimum rate guarantee for the background class, for this class we will consider the following two throughput constrains.

B. Throughput Constrains for the Background Class

When there are n_1, n_2, n_3 conversational, streaming, and interactive flows in the system at time *t*, respectively, the instantaneous bandwidth provided for (the lowest priority) background traffic class is

$$\theta_4(t) = \max(0, C - n_1 b_1 - n_2 b_2 - n_3 b_3)$$

The steady state bandwidth of the background traffic class is denoted by θ_4. (θ_4 is a discrete r.v.)

In oder to characterize the bandwidth "left over" for the background class, we define the following two throughput constrains: Mean throughput constraint: $E(\theta_4) > \theta_4^{min}$, Throughput threshold constraint: $\Pr(\theta_4 > \theta_4^{min}) > \epsilon$.

IV. Bandwidth Sharing Policies

Bandwidth sharing in telephony and ATM networks have long been studied, but most models ignore the elastic nature of traffic classes such as the streaming or interactive classes. Papers that do model elastic traffic on the call level include the ones by Altman *et al.* [1], Andersen *et al.* [2] and Nunez *et al.* [6], but these papers restrict themselves to two traffic classes only. Also, they restrict attention to the moments of the elastic traffic rather than considering its distribution.

With the above throughput and constraint definitions we are now in the position to establish the following bandwidth sharing rules. In the presentation we let $C - \delta$ denote the portion of the link capacity that is dedicated to the first three service classes and denote δ the portion of the link capacity which is dedicated to the background class. In the numerical section we will let $\delta = 0$ (Policy I: complete sharing) and $\delta > 0$ (Policy II: complete partitioning). Our bandwidth sharing policies are motivated by technological feasibility and - as we will indeed see in the numerical section - by the intuition that since high priority elastic classes tend to occupy all available link capacity, the background class may need a dedicated link capacity to perform acceptably under various offered traffic mixes.

- If there is enough bandwidth for all flows to get their respective peak bandwidth demands, then class-2 and class-3 flows occupy b_2 and b_3 bandwidth units respectively.
- If there is a need for bandwidth compression, i.e. $n_1 \cdot b_1 + n_2 \cdot b_2 + n_3 \cdot b_3 > C - \delta$, then the bandwidth compression of the flows is such that $r_2 = r_3$, where $r_2 = b_2^r/b_2$ and $r_3 = b_3^r/b_3$, as long as the minimum rate constraint is met for both classes (i.e. $b_2^{min}/b_2 \leq r_2 \leq 1$ and $b_3^{min}/b_3 \leq r_3 \leq 1$).
- If there is still need for further bandwidth compression, but either one of the two classes does not tolerate further bandwidth decrease, (i.e. r_i is already b_i^{min}/b_i for either $i = 2$ or $i = 3$) at the time of the arrival of a new flow, then the service class which tolerates further compression decreases equally the bandwidth occupied by its flows, as long as the minimum bandwidth constraint is kept for this traffic class.

Three underlying assumptions of the above rules (and the model) are noteworthy. Firstly, we assume that both the streaming and interactive flows are greedy, in the sense that they always occupy the maximum possible bandwidth on the link, which is the smaller of their peak bandwidth requirement (b_2 and b_3 respectively) and the equal share (in the above sense) of the bandwidth left for them by the rigid flows (which will depend on the link allocation policy). Secondly, we assume that all streaming and interac-

tive flows in progress share proportionally equally (i.e. the r_i's are equal) the available bandwidth among themselves, i.e. the newly arrived flow and the in-progress flows will be squeezed to the same r_i value. If a newly arriving flow decreased the flow bandwidth below b_2^{min} and b_3^{min} (i.e. both the streaming and the interactive classes were compressed to their respective minima), that flow is not admitted into the system, but it is blocked and lost. Note that all arriving flows are allowed to "compress" the in-service streaming and interactive flows, as long as the minimum bandwidth constrains are kept. Thirdly, the model assumes that the rate control of the streaming and interactive flows in progress is ideal, in the sense that an infinitesimal amount of time after any system state change (i.e. flow arrival and departure) these sources readjust their current bandwidth on the link. While this is clearly an idealizing assumption, we assume that the buffers at the IP packet layer are large enough to absorb the IP packets until e.g. TCP or any other upper layer protocol throttles the senders.

V. Numerical Results

A. Generating and Solving the Markov Model

The QoS/GoS parameters of different communication streams using a common resource have already been analyzed through Markov models [6], [1], [9]. There are two main difficulties applying this approach, the automatic generation of the (commonly large) Markov model and its solution. We have used the Web accessible Markovian analysis tool named MRMSolve that performs both of these two steps automatically based on a high level description of the model [8].

B. Discussion

We define the "offered" traffic load of the conversational, streaming and interactive traffic classes and the total offered load as

$$\rho_i = \frac{\lambda_i}{\mu_i}\, b_i, \quad i = 1,2,3; \quad \rho = \sum_{i=1}^{3} \rho_i,$$

respectively. The traffic load realized on the transmission link differs from the offered load due to call blocking and bandwidth reduction. When there is not enough available capacity even with the maximum possible compression of the on-going streaming and interactive calls, new call arrivals are blocked. The load lost due to call blocking is bounded by the maximum accepted blocking probability values (B_1^{max}, B_2^{max} and B_3^{max}).

The bandwidth of the streaming and the interactive flows fluctuates according to the traffic load of the link. The fluctuation of the interactive flows does not result in any load reduction because the amount of data transmitted through an interactive class connection is independent of the available throughput (lower throughput results in longer service time). However, the fluctuation of the streaming flows results in a further load reduction, since the amount of data transmitted through a streaming class connection is proportional to the available bandwidth during the connection.

Therefore, we define two traffic loss measures to evaluate the different ways of losses:

$$B_{call} = B_1\frac{\rho_1}{\rho} + B_2\frac{\rho_2}{\rho} + B_3\frac{\rho_3}{\rho};$$

$$B_{bw} = B_1\frac{\rho_1}{\rho} + \left(B_2 + (1-B_2)\frac{b_2 - E(\theta_2)}{b_2}\right)\frac{\rho_2}{\rho} + B_3\frac{\rho_3}{\rho}.$$

B_{call} accounts only for losses by call blocking, while B_{bw} is a combined measure that accounts for both call blocking and bandwidth reduction.

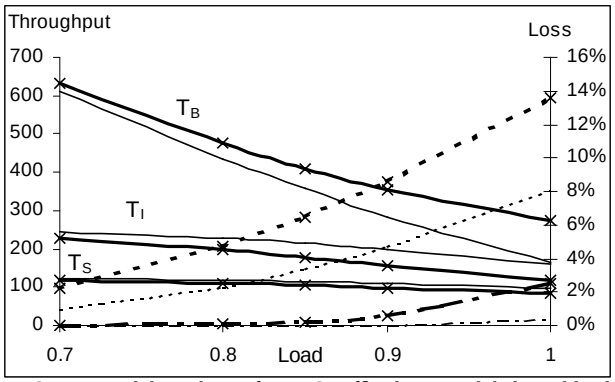

Fig. 3. Loss and throughput of UMTS traffic classes with balanced load

Figure 3 shows the throughput and loss measures as a function of the total offered load. The considered traffic scenario is as follows: $C = 2000, b_1 = 12, b_2 = 128, b_2^{min} = 64, b_3 = 256, b_3^{min} = 51.2, \mu_1 = 1/180, \mu_2 = 1/300, \mu_3 = 1/30$, the high priority classes provide the same load, $\rho_1/\rho = \rho_2/\rho = \rho_3/\rho = 1/3$. (The call arrival intensities are set accordingly.) Two policies are compared. In the first case the total link capacity is used by all service classes (complete sharing), $\delta = 0$, while in the second case a given bandwidth is reserved for the background traffic $\delta = 200$. Results associated with the first (second) case are drawn by thin (thick) lines. The throughput, B_{call}, and B_{bw} curves are drawn with solid, dash and dot, and dot lines, respectively.

It can be seen that for reasonable load, $\rho < 0.9$, we have some bandwidth reduction for the variable bandwidth classes, and a low call blocking probability $B_{call} < 2\%$; but the combined loss measure, B_{bw}, shows a much higher value. For high load, $\rho > 0.9$, the trends are continued. Interestingly, we cannot see significant advantage of the bandwidth reservation, $\delta = 200$, on the mean throughput of the background traffic ($E(\theta_4)$). But the mean throughput is not a "fine enough" measure of the throughput available for the background traffic. The real advantage of the bandwidth reservation becomes visible only on Figure 4, that provides

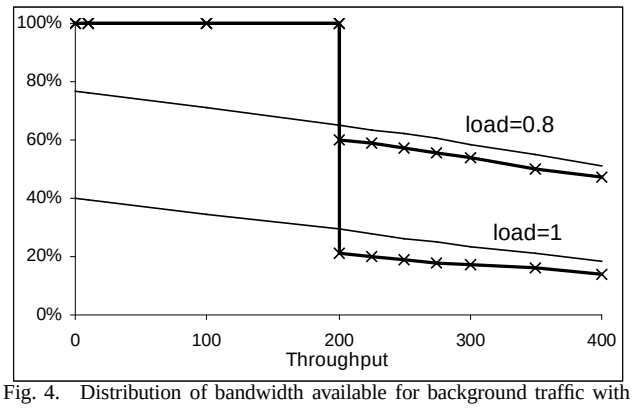

Fig. 4. Distribution of bandwidth available for background traffic with balanced load

the distribution of θ_4 (i.e., $\Pr(\theta_4 \geq x)$). Even the mean throughput of the background traffic does not differ significantly when $\delta = 200$; we have a completely different distribution of θ_4. The most important difference, that has a significant consequence on the background traffic services, is the probability of having no bandwidth available for background traffic. In case of $\delta = 0$ there is a high probability that the higher priority services completely utilize the link capacity and the background services are completely starving. However, when $\delta = 200$ a minimal bandwidth is always provided for the background class.

The high probability of complete link utilization of the high priority services is due to the adaptive nature of the streaming and the interactive class. This services set their bandwidth such that the link capacity is most utilized. Figure 5 and 6 shows how the probability of complete link utilization increases when the traffic mixture changes and the portion of 'adaptive' traffic classes increase. On these figures we depict the same result when the traffic mixture is $\rho_1/\rho = 0.1, \rho_2/\rho = 0.1, \rho_3/\rho = 0.8$. In Figure 6 it can be seen that the probability of no bandwidth left for the background traffic is much higher with this traffic mixture.

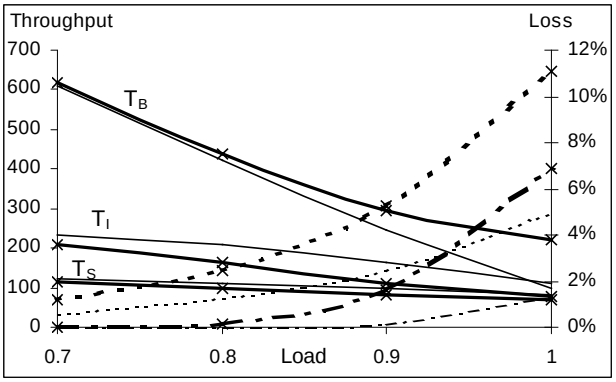

Fig. 5. Loss and throughput of UMTS traffic classes with dominant interactive

VI. Conclusions

In this paper we extended the widely used multi-rate loss model in terms of the modeled traffic classes. Specifically, we included the two main types of the elastic classes (streaming and interactive) standardized for 3^{rd} generation mobile (such as the UMTS) core networks. The differentiation between these two classes is important, (but often overlooked) because they differ in terms of how their holding times (and their actual carried traffic) depends on the system load. We then used the model to derive performance measures for the blocking probabilities and throughput values under two simple bandwidth sharing policies. We have found that dedicating a portion of the link capacity to the lower priority background class is useful for ensuring some throughput for this class without causing significant blocking or throughput degradation for the higher priority classes. We have noted that this somewhat counter-intuitive result comes from the elastic nature of the streaming and interactive classes: as opposed to circuit switched systems, these classes tend to greedily consume all available link capacity. We believe that the model and the numerical results can be used to dimension the core network of future multiservice systems.

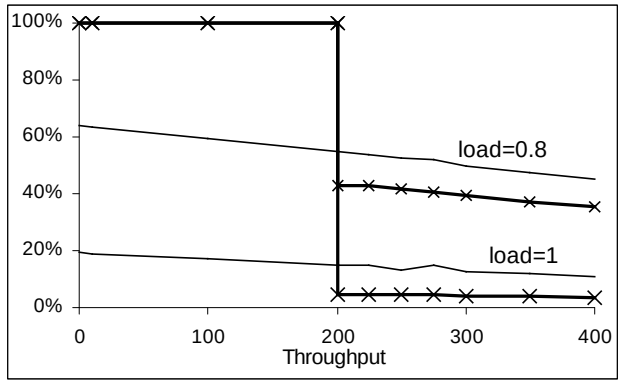

Fig. 6. Distribution of bandwidth available for background traffic with dominant interactive load

References

[1] Eitan Altman, Damien Artiges and Karim Traore, "On the Integration of Best-Effort and Guaranteed Performance Services", *INRIA Research Report No. 3222*, July, 1997.

[2] Allan T. Andersen, Sören Blaabjerg, Gábor Fodor and Miklós Telek, "A Partially Blocking-Queueing System with CBR/VBR and ABR/UBR Arrival Streams", 5^{th} *IFIP International Conference on Telecommunications System*, Nashville, TN, USA, March 1997.

[3] Emano Berruto, Giovanni Colombo, Pantelis Monogioudis, Antonella Napolitano and Kyriacos Sabatakis, "Architectural Aspects for the Evolution of Mobile Communications Toward UMTS", *IEEE Journal of Selected Areas in Communications*, Vol. 15, No. 8, pp. 1477-1487, October 1997.

[4] G. Fodor, G. Malicsko, S. Malomsoky, "A Joint Radio- and IP Resource Reservation Scheme in All-IP 3^{rd} Generation Networks", accepted to the *IEEE Wireless Communications and Networking Conference*, Chicago, IL, USA, September 2000.

[5] D. Mitra, J. A. Morrison and K. G. Ramakrishnan, "ATM Network Design and Optimization: A Multirate Loss Network Framework", *IEEE/ACM Transactions on Networking*, Vol. 4, No. 4, pp. 531-543, August, 1996.

[6] R. Nunez Queija, J. L. van den Berg, M. R. H. Mandjes, "Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic", *International Teletraffic Congress*, UK, 1999.

[7] T. Ojanpera and R. Prasad, *eds.*, "Wideband CDMA for Third Generation Mobile Communications, *Artech House*, 1998.
[8] S. Rácz, B. P. Tóth, and M. Telek, "MRMSolve: Numerical analysis of large markov reward models," in *Tools 2000*, pp. 337–340, Springer, LNCS 1786, 2000.
[9] S. Rácz, M. Telek, and G. Fodor, "Link capacity sharing between guaranteed- and best effort services on an atm transmission link under GoS constraints," in *System Performance Evaluation: Methodologies and Applications* (E. Gelenbe, ed.), pp. 69–79, CRC Press, 2000. Ch. 5.
[10] K. W. Ross, "Multi-service Loss Models for Broadband Telecommunication Networks", *Springer Verlag London Limited*, ISBN 3-540-19918-7, 1995.
[11] "Third Generation Mobile Systems in Europe", *IEEE Personal Communications Magazine*, April 1999.
[12] "The Evolution of TDMA to 3G", *IEEE Personal Communications Magazine*, June 1999.
[13] "Multiple Access for Broadband Wireless Networks", *IEEE Communications Magazine*, July 2000.
[14] 3GPP TS 23.107, "Services and System Aspects: QoS Concept and Architecture", ver. 3.2.0, 2000.
[15] 3GPP TR 23.922, "Architecture for an All-IP Network", ver. 1.0.0, 1999.