# Link Capacity Sharing Between Guaranteed- and Best Effort Services on an ATM Transmission Link under GoS Constraints

Sándor Rácz

*Department of Telecommunications and Telematics, Technical University of Budapest, Hungary,*
E-mail: raczs@ttt-atm.ttt.bme.hu

Miklós Telek

*Department of Telecommunications, Technical University of Budapest, Hungary,*
E-mail: telek@hit.bme.hu

Gábor Fodor

*Mobile Networks- and Systems Research, Ericsson Radio Systems, Sweden,*
E-mail: Gabor.Fodor@era-t.ericsson.se

While link allocation policies in multi-rate circuit switched loss models have drawn much attention in recent years, it is still an open question how to share the link capacity between service classes in a fair manner. In particular, when an ATM link is offered calls from service classes with/without strict QoS guarantees one is interested in link capacity sharing policies that maximize throughput and keep the per-class blocking probabilities under some GoS constraints. In this paper we propose a model and associated computational technique for an ATM transmission link to which CBR/VBR and ABR classes offer calls. We also propose a simple link allocation rule which takes into account blocking probability constraints for the CBR/VBR calls and a throughput constraint for the ABR calls and attempts to minimize the blocking probability of ABR calls. Numerical examples demonstrate the effectiveness of the policy and of the applied computational technique.
**Keywords:** multi-rate loss models, link capacity sharing, blocking probabilities, ATM service categories, Markov reward models.

## 1. Introduction

In recent years the various aspects of the coexistence of different service classes in ATM gained much attention and significant advances in the management of ATM traffic have been achieved. Most of the ATM traffic management efforts both within the major standardization bodies and the industry have been focusing on the *cell level* aspects of ATM, such as devising efficient congestion control- and policing mechanisms, and also call admission control (CAC), buffer allocation- and cell scheduling rules. Although *call level* issues in the multi-rate environment, like the computation of the blocking probabilities and establishing link capacity sharing policies have also been addressed by many papers [6,8,10,12,14,15,22], very few papers deal with the problem of blocking probability calculation and link allocation policy when service classes with/without congestion control and with/without cell level QoS guarantees are present in a system simultaneously. The investigation of the call level aspects is important, since the blocking

probability constraint is the primary input to the network dimensioning process. The hardship of this type of problem lies in the fact that the classical method of the *equivalent bandwidth* connecting the cell- and call level aspects is not directly applicable to an ATM link supporting CBR/VBR and ABR service classes simultaneously. This is because while it has been possible to associate a bandwidth-like quantity even with the VBR class, it is difficult to do the same for the ABR service class, because

- ABR does not provide the same level of QoS as the CBR/VBR classes
- there is very limited or no resource allocation prior to the information transfer phase
- the bandwidth available for the ABR calls fluctuates in time in accordance with the load on the link [1,5,7,21,25].

Since we have to dispose of the direct application of the equivalent bandwidth based approach when devising and analyzing link capacity sharing policies, we seek alternative methods to do this. This problem has been raised by for instance [18] and [19] as well, without providing an analytical approach or a modeling framework. While many interesting contributions have proposed link allocation- and associated performance analysis methods for complete sharing, complete partitioning, partial overlap, trunk reservation [17], class limitation [20,22] and Markov decision [9,11,13,16,23], very few papers propose efficient computational technique for ATM with CBR[1] and ABR classes, especially when the state space becomes large, say in the order of $10^4 - 10^6$. Thus, we in this paper have it as an objective to (1) extend the widely used multi-rate model such that they allow the ABR bandwidth to fluctuate between the minimal and the peak bandwidth during the call's holding time, (2) propose a simple and yet efficient method for link capacity sharing between calls coming from different ATM service classes and (3) devise efficient computational technique for the calculation of the throughput and blocking probabilities, applicable for large systems.

The rest of this paper is organized as follows. In section 2, the extension of the multi-rate Erlang blocking model for the best effort environment is described. Here we also give the definitions of the performance measures of interest. Also in this Section, we propose a link capacity allocation rule, which can handle CBR and ABR traffic classes. The proposed method guarantees the required call blocking probability for CBR calls, and provides the required throughput for ABR calls. In Section 3 we develop a computational technique to determine the call blocking probabilities of CBR and ABR calls. In Section 4 different throughput measures of ABR calls is calculated. We formulate the problem as a Markov reward model [4], which is a very useful modeling tool for any stochastic accumulation process when the actual rate of accumulation (e.g. the accumulation of the bytes of a file during an ftp file transfer session) depends only on the state of a Continuous Time Markov Chain (CTMC). As we will see in Section 5, the efficiency of this method allows us to apply it for systems with large state space (cardinality of the state space: $\sim 10^4 - 10^6$). Finally, in Section 6, we illustrate the results with numerical examples.

[1] Because we model the system on the call level, in the rest of this paper we use the CBR service class as one which represents strict QoS guarantees, with the understanding that by adopting the notion of equivalent bandwidth, this class could as well be the VBR class.

## 2. The Multi-service Model of an ATM Link

In this Section we formulate the Markovian model of a single ATM transmission link receiving CBR and ABR traffic. In the presentation we restrict ourselves to two CBR classes and a single ABR class, but the model is extensible to more general cases. More traffic classes increase both the complexity and the size of the state space, and the numerical results become more difficult to interpret, and therefore we believe that it is reasonable to start with these restrictions. It should be pointed out, however, that both the basic idea of the model extension to include ABR traffic and the results in Section 3 and 4 are applicable to more general cases as well.

### *2.1. Assumptions and Notation*

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

- Narrow-band CBR calls are characterized by their peak bandwidth requirement $b_1$, call arrival rate $\lambda_1$ and departure rate $\mu_1$;
- Wide-band CBR calls are characterized by their peak bandwidth requirement $b_2$, call arrival rate $\lambda_2$ and departure rate $\mu_2$;
- ABR calls are characterized by their peak bandwidth requirement $b_3$, call arrival rate $\lambda_3$, minimal bandwidth requirement $b_3^{min}$, and *ideal* departure rate $\mu_3$. By ideal we mean that the peak bandwidth is available during the entire duration of the call.

One may think of an ABR 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 $G(x) = 1 - e^{-\frac{\mu_3}{b_3}x}$, 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 in accordance with the instantaneous number of CBR and ABR calls in service, the bandwidth given to the ABR calls may drop below the peak bandwidth requirement, in which case the actual holding time of the call increases.

All three types of calls arrive according to independent Poisson processes, and the holding time for CBR calls are exponentially distributed. As we will see, the moments of the holding time of the ABR calls can be determined using the theory of Markov reward processes.

Three underlying assumptions of the above model are noteworthy. First of all, we assume that the ABR calls 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 $b_3$ and the equal share of the bandwidth left for ABR calls by the CBR calls (which will depend on the link allocation policy). The greediness assumption entails that the ABR call's holding time is lower bounded by our model, since if the source does not always send with the maximum possible value, the actual call holding time will be longer. In fact, the greediness assumption may be replaced by an ABR source model that takes into account an upper layer protocol, such as the TCP/IP stack over the ABR service. In such cases, the link utilization will necessarily be lower than in the greedy case, since the non-greedy sources do not always make use of

the free link capacity. It is left for future work to compare our results with those investigating e.g. the TCP protocol over ABR [3]. Second of all, we assume that all ABR calls in progress share equally the available bandwidth among themselves, i.e. the newly arrived ABR call and the in-progress ABR calls will be squeezed to the same bandwidth unless each of them gets their peak bandwidth. Note that if a newly arriving call decreased the ABR bandwidth below $b_3^{min}$, that call is not admitted into the system, but it is blocked and lost. Also note, that arriving CBR as well as ABR calls are allowed to "compress" the in-service ABR calls, as long as the minimal bandwidth constraint is kept. Third of all, the model assumes that the rate control of the ABR calls in progress is ideal, in the sense that an infinitesimal amount of time after any system state change (i.e. call arrival and departure) the ABR sources readjust their current bandwidth on the link. Clearly, this call level model does not take into consideration some of the cell level parameters including the initial cell rate and the tagged cell rate [2].

It is intuitively clear that the residency time of the ABR calls in this system not only depends on the amount of data they want to transmit, but also on the bandwidth they receive during their service. In order to specify this relationship we define the following quantities:

- $\theta(t)$ defines the instantaneous *throughput* of the ABR calls at time $t$ (e.g., if there are $n_1, n_2, n_3$ narrow-band CBR, wide-band CBR, and ABR calls in the system at time $t$, respectively, the instantaneous throughput is $min\,[b_3, (C - n_1 b_1 - n_2 b_2)/n_3]$ ). Note that $\theta(t)$ is a discrete random variable (r.v.) for any $t \geq 0$.
- $T_x = \inf\{t \mid \int_0^t \theta(\tau) d\tau \geq x\}$ (r.v.) gives the time it takes for the system to transmit $x$ amount of data through an ABR call.
- $\theta_x = x/T_x$ defines the *throughput* of the ABR call during the transmission of $x$ data unit. Note that $\theta_x$ is a continuous r.v.
- $\theta = \int_0^\infty \theta_x \; dG(x)$ (r.v.) defines the *throughput* of the ABR call.

In addition, we associate the maximal accepted blocking probabilities with both CBR classes, i.e., $B_1^{max}$ and $B_2^{max}$ respectively and the minimal accepted throughput $\theta^{min}$ with the ABR class. We refer to the set of arrival rates $(\lambda_1, \lambda_2, \lambda_3)$, departure rates $(\mu_1, \mu_2, \mu_3)$[2], bandwidths $(b_1, b_2, b_3)$, the minimal ABR bandwidth $(b_3^{min})$, the blocking probabilities $(B_1^{max}, B_2^{max})$ and the ABR throughput constraint $(\theta^{min})$ as the *input parameters* of the system.

### *2.2. System Description*

The system under investigation (with the above assumptions on the arrival process and holding time/transmission requirement) is a Continuous Time Markov Chain (CTMC) whose state is uniquely characterized by the triple $i = (n_1, n_2, n_3)$, where $n_1$ and $n_2$ are the number of narrow-band and wide-band CBR calls in the system, respectively, and $n_3$ is the number of ABR calls in the system.

It is clear that in order to obtain the performance measure of this system we need to determine the CTMC's generator matrix $\mathbf{Q}$ and solve $\mathbf{Q}^T \cdot \underline{P} = \underline{0}$ and $\underline{P}^T \cdot \underline{h} = 1$, where $\underline{P} = [p_i]$ is the steady

[2] $\mu_3$ is the maximum departure rate of the ABR class assuming that the bandwidth of the ABR connection equals to $b_3$.

state distribution and $\underline{h}$ is the column vector of 1's. However, the structure of $\mathbf{Q}$ reflects the applied link allocation policy and therefore we first need to define it.

We would like to define the link allocation policy such that it is able to minimize the call blocking probability for the ABR calls while it is able to take into account the GoS (blocking probability) constraints for the CBR calls and the minimal throughput constraint for the ABR calls. Because of its flexibility (in that it is able to take into account the above constraints) and simplicity (in that the performance measures of interest can be determined even for large systems) we in this paper adopt the *partial overlap, POL* link allocation policy from the multi-rate circuit switched modeling paradigm [27].

According to the POL policy the link capacity $C$ is divided into two parts, the $C_{COM}$ common part and the $C_{ABR}$ part, which is reserved for the ABR calls only, such that $C = C_{COM} + C_{ABR}$. Under the considered POL policy the number of calls in progress on the link is subject to the following constraints:

$$n_1 \cdot b_1 + n_2 \cdot b_2 \leq C_{COM} \tag{2.1}$$

$$N_{ABR} \cdot b_3^{min} \leq C_{ABR} \tag{2.2}$$

$$n_3 \leq N_{ABR} \tag{2.3}$$

where $N_{ABR}$ stands for the maximum number of ABR calls in the system (sometimes referred to as the *cut-off* parameter [24]) and will be determined later. Note that this policy has two free parameters, ($C_{COM}$ and $N_{ABR}$) which allows for the easy dimensioning of a system with blocking and throughput constraints. Furthermore, we find it relatively easy to analyze systems with large state space as well.

The set of such triples which satisfy these constraints constitutes the set of *feasible states* of the system which we denote by $\mathcal{S}$. Cardinality of the state space can be determined with (2.4).

$$\#\mathcal{S} = (N_{ABR} + 1) \cdot \sum_{l=0}^{\lfloor C_{COM}/b_1 \rfloor} \left\lfloor \frac{C_{COM} - l \cdot b_1}{b_2} + 1 \right\rfloor \tag{2.4}$$

In (2.1) the ABR calls are protected from CBR calls. In (2.2,2.3) the maximum number of ABR calls is limited by two constraints. The constraint in (2.2) protects the CBR calls from ABR calls while (2.3) protects the ABR calls from the new ABR calls, because if too many ABR calls were admitted into the system then $\theta$ could decrease below $\theta^{min}$. Clearly, $\theta$ can be modified by changing the value of $N_{ABR}$.

It is easy to realize that the generator matrix, $\mathbf{Q}$, possesses a nice structure, because only transitions between "neighboring states" are allowed in the following sense. Let $q_{i,j}$ denote the transition rate from state $i$ to state $j$. Then, taking into account the above constraints associated with the proposed POL policy, the non-zero transition rates between the states are:

$$q_{i,i_{k+}} = \lambda_k, \qquad k = 1, 2, 3 \tag{2.5}$$

$$q_{i,i_{k-}} = n_k \cdot \mu_k, \qquad k = 1, 2 \tag{2.6}$$

$$q_{i,i_{3-}} = r_i \cdot \mu_3, \tag{2.7}$$

where $i_{1+} = (n_1 + 1, n_2, n_3)$ when $i = (n_1, n_2, n_3)$; $i_{k+}$ and $i_{k-}$ $(k = 1, 2, 3)$ are defined similarly; and

$$r_i = \min\left(n_3 \ , \ \frac{C - (b_1 \cdot n_1 + b_2 \cdot n_2)}{b_3}\right). \tag{2.8}$$

Equation (2.5) represents the state transitions due to call arrivals, while (2.6) and (2.7) represent the transitions due to call departures. Based on (2.8) the $r_i b_3$ quantity is the total bandwidth of the ABR calls when the system is in state $i$. The generator matrix of the CTMC, $\mathbf{Q}$, is constructed based on the transition rates defined in (2.5), (2.6) and (2.7). Note that the POL policy as described above is fully determined by specifying its two parameters: the $C_{COM}$ common part, and the $N_{ABR}$ maximal number of ABR calls. We refer to the $C_{COM}$ and the $N_{ABR}$ parameters of the POL policy as the *output parameters* of the system. In the next subsection, we introduce an example on the structure of the state space of the CTMC described above, and at the end of this section we describe the considered constraints which determine the output parameters. The actual determination of the output parameters from the input parameters is left for the subsequent sections.

### *2.3. Example on the State Space Under the POL Policy*

In this subsection we consider a small system for illustration purposes.

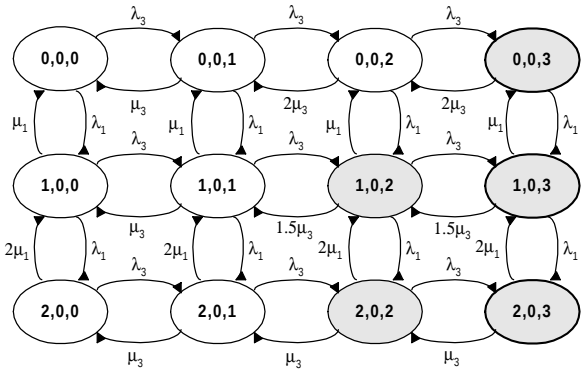


Figure 1. The state space of the small example when $n_2 = 0$

Figure 1 depicts the state space of a system with capacity $C = 4$ and with a CBR and an ABR class (i.e., for ease of presentation $n_2 = 0$ is kept fixed). We let $C_{COM} = 2$, $b_1 = 1$ and $b_3 = 2$. The ABR class is further characterized by its *minimal* accepted bandwidth, which we here let $b_3^{min} = 2/3$. This setting gives rise to 12 feasible states, out which there are 5 (gray) states where the ABR bandwidth is compressed below the peak bandwidth specified by $b_3$. In for instance state $(1, 0, 3)$ each of the 3 ABR calls receive 1/2 bandwidth, which gives rise to an aggregated ABR death rate $1.5\mu_3$.

The CTMC in Figure 1 is not reversible due to the possible compression of ABR bandwidth. Based on this small example in Figure 1 it can be seen that, in general, the CTMC of the considered communication link is not reversible due to the possible compression of ABR bandwidth. Hence, the steady state distribution does not obey a product form solution. However, the generator matrix, as we will see next possesses a nice birth-death structure allowing for efficient numerical solution approaches.

### *2.4. Constraints for Determining the Parameters of the POL Policy*

The POL policy is easy to dimension. It has two free parameters, with which the performance of the system can be tuned. It guarantees call level GoS for CBR calls and throughput for ABR services.

The GoS of CBR calls is guaranteed by the proper setting of $C_{COM}$. In case of a change in the ABR load (i.e. the call arrival intensity ($\lambda_3$) or the parameter of required data to transmit ($\mu_3/b_3$)), the $N_{ABR}$ parameter has to be adjusted to keep the required throughput. We divide the problem of determining the output parameters of the POL policy into two steps. In the first step we determine the minimum required capacity for CBR calls, that guarantees the required blocking probabilities:

$$\mathbf{min}\Big\{C_{COM} \ : \ B_1 \leq B_1^{max} \ , \ B_2 \leq B_2^{max}\Big\} \tag{2.9}$$

where $B_1$ ($B_2$) is the blocking probability of the narrow-band (wide-band) CBR class. In the second step we determine the maximum number of ABR calls simultaneously present in the system. In fact, we minimize the blocking probability of the ABR calls (by determining the maximum number of admissible ABR calls) applying constraints on the throughput of the ABR connections. The current ATM standards ([1,2]) define the notion of the *minimum cell rate* (MCR) to describe the minimum instantaneous bandwidth accepted by the user at the user-network interface (UNI). In order for our call level model to take into account the impact of the MCR on the call blocking probabilities we need a call level abstraction of this quantity.

The MCR in ABR corresponds to the instantaneous throughput $\theta(t)$ as defined in Subsection 2.1. Seen from the call level, we will first approximate this quantity with the overall throughput, (also defined in subsection 2.1), which expresses the average (expected) value of the instantaneous throughput over the lifetime of a given call. This approximation is motivated by the fact that many applications' performance - as perceived at the call level - ultimately depends on the average throughput, like in the case of an ftp file transfer, where the transfer time is determined by the average throughput.

In many cases, however, we need a finer granularity call-level abstraction of the MCR value. In TCP/IP over ABR applications the TCP performance depends on the probability that the connection's throughput temporarily falls below a certain value [29]. Also, although the ABR service is primarily intended for non-real-time applications that tolerate delay and delay variation, depending on the operator's pricing strategy some users may want to use this service for other type of applications. For instance, in the case of real time video transfers we also need a finer granularity approximation of the MCR value. This is because such applications' user-perceived performance depends on the probability that the instantaneous throughput drops below a certain threshold. Ultimately, the user may want to set this probability to zero in order to avoid temporary quality degradation in the particular application in question.

Therefore, we also consider the throughput threshold constraint as an alternative call-level dimensioning constraint. As we will see from its definition, this constraint allows us to model the impact of the instantaneous throughput (and thereby the MCR) on the call level by setting its associated $\epsilon$ value to zero.

The considered two constraints are defined as:

- constraint on the average throughput:

$$\mathbf{max}\Big\{N_{ABR} \ : \ E(\theta) \geq \theta^{min}, \ N_{ABR} \leq \frac{C - C_{COM}}{b_3^{min}}\Big\} \tag{2.10}$$

  i.e., the average throughput of ABR connections can not be less than $\theta^{min}$.

To make a plausible interpretation of this constraint let us assume that the distribution of $\theta$ is fairly symmetric around $E(\theta)$, i.e. the median of $\theta$ is close to $E(\theta)$. In this case the probability that an ABR call obtain less bandwidth than $\theta^{min}$ is around 0.5. Users (even with ABR traffic) often prefers more informative throughput constraints like the next one.

- constraint on throughput threshold:

$$\mathbf{max}\Big\{N_{ABR} \;:\; Pr(\theta_x \leq \theta^{min}) \leq \varepsilon,\ \forall x,\ N_{ABR} \leq \frac{C - C_{COM}}{b_3^{min}}\Big\} \tag{2.11}$$

This throughput threshold constraint requires that the throughput of ABR connections is greater than $\theta^{min}$ with a predefined probability $(1-\varepsilon)$ independent of the associated service requirements $(x)$. Hence, if the (input) parameter $\theta^{min}$ is much less than $E(\theta)$ then this second constraint is much more informative for the user about the expectable minimal level of the ABR throughput.

Note that in case of applying throughput threshold constraint $\varepsilon$ is also an input parameter of the model.

## 3. Analysis of Call Blocking Probabilities

The call blocking probabilities of the CBR and ABR calls are obtained from the steady state distribution ($\underline{P} = [p_i]$) of the CTMC specified by its generator matrix $\mathbf{Q}$. Different numerical methods can be used to obtain the steady state distribution [26]. It is noteworthy that the nice structure of the generator matrix (as described by the equations (5)-(7)) is retained when there are more than 3 bandwidth classes in the system. However, the cardinality of the state space rapidly increases with the number of considered bandwidth classes, which entails that as the number of bandwidth classes increases the computation of the blocking probabilities becomes numerically hard. This computational problem in stochastic knapsack based systems has been recognized by e.g. [15], where efficient techniques are studied to overcome these problems. Also, a comprehensive list of different numerical methods to obtain the steady state distribution is given by [26].

*Direct methods* compute the solution in a fixed number of operations, e.g. the Gaussian elimination computes the result employing $(n^3/3 + n^2/2 - 5 \cdot n/6)$ multiplications and additions and $(n^2/2 + n/2)$ divisions, where $n$ is the cardinality of $\mathcal{S}$. In contrast, *iterative methods* begin from some initial guess and produce a sequence of intermediate results, which converge to the solution. Most often it is not known how many iterations are required to achieve a certain accuracy, hence the number of necessary numerical operations depends on the model and is not known explicitly in advance. With large models, where the cardinality of state space exceeds $\sim 10^4$, the required computation time of direct methods is usually unacceptable. Therefore, we have implemented an iterative method to calculate the steady state of the system.

The computation time of the iterative methods depends mostly on three factors; namely the speed of convergence, the complexity of an iteration step and the initial guess. The convergence speed depends on (the dominant eigenvalue of) the generator matrix $\mathbf{Q}$, which is hard to evaluate for large models. Usually, the unbalancedness (or stiffness) of $\mathbf{Q}$ results in a slower convergence. In the case when the arrival and service rates $(\lambda_i, \mu_i; i = 1, 2, 3)$ of the considered example are in the same order of magnitude

**Q** is fairly balanced. The complexity of an iteration step is mainly determined by the sparseness of **Q**. The generator of our model contains no more than 6 in- and outgoing state transitions per state, hence **Q** is very sparse and the complexity of an iteration step is low.

The computation time of the iterative methods also depends on the initial guess. For this reason a heuristic direct method is applied to calculate the initial guess of the iterative method. Due to the occasional reduction of the ABR bandwidth (and corresponding departure rate) the CTMC of the studied system does not exhibit nice properties like reversibility or product form solution, but we can utilize the fact that ABR calls do not affect the service of the CBR calls. Hence, the Markov chain that describes the number of CBR calls in the system (a Markovian finite capacity multi-rate model) is reversible, and $\hat{p}(n_1, n_2) = \sum_{n_3=0}^{N_{ABR}} p_{(n_1,n_2,n_3)}$ is obtained from :

$$\hat{p}^*_{(0,0)} = 1 \ , \tag{3.1}$$

$$\hat{p}^*_{(n_1,n_2)} = \hat{p}^*_{(n_1-1,n_2)} \cdot \frac{\lambda_1}{n_1 \cdot \mu_1} = \hat{p}^*_{(n_1,n_2-1)} \cdot \frac{\lambda_2}{n_2 \cdot \mu_2} \ , \tag{3.2}$$

$$\hat{p}_{(n_1,n_2)} = \frac{\hat{p}^*_{(n_1,n_2)}}{\sum\limits_{(n'_1,n'_2)\in\mathcal{S}'} p^*_{(n'_1,n'_2)}} \ , \tag{3.3}$$

where the $\hat{p}^*_{(n_1,n_2)}$ unnormalized steady state probabilities are auxiliary variables of the iterative method, and are defined by equations (3.1)-(3.3) and $\mathcal{S}' = \{(n_1, n_2) \ : \ \exists (n_1, n_2, n_3) \in \mathcal{S}\}$.

From the steady state distribution of the CBR class ($\hat{p}_{(n_1,n_2)}$) the overall steady state behavior ($p_{(n_1,n_2,n_3)}$) is obtained by fixing the number of narrow-band CBR calls ($n_1 = m$) and assuming that the obtained Markov chain is reversible, even it is not the case. This assumption allows to evaluate an initial guess for the iterative method as follows. For all possible fixed value of $n_1$ ($n_1 = m$):

$$p^*_{(m,0,0)} = 1 \tag{3.4}$$

$$p^*_{(m,n_2,n_3)} = \frac{p^*_{(m,n_2-1,n_3)} \cdot \lambda_2 + p^*_{(m,n_2,n_3-1)} \cdot \lambda_3}{2(n_2 \cdot \mu_2 + \mu_3 \cdot r_{(m,n_2,n_3)})} \tag{3.5}$$

$$p_{(m,n_2,n_3)} = p^*_{(m,n_2,n_3)} \frac{\sum\limits_{n'_2 \ : \ (m,n'_2)\in\mathcal{S}'} p_{(m,n'_2)}}{\sum\limits_{n''_2,n''_3 \ : \ (m,n''_2,n''_3)\in\mathcal{S}} p^*_{(m,n''_2,n''_3)}} \ . \tag{3.6}$$

The obtained distribution can be used as a rude approximation of the steady state distribution, and it is used as the initial guess for the iterative method, which refines the results for the appropriate accuracy.

Based on the steady state distribution of the CTMC, the call blocking probabilities are calculated as:

$$B_k = \sum_{i\in\mathcal{S}, i_{k+}\notin\mathcal{S}} p_i \ , \quad k = 1, 2, 3 \tag{3.7}$$

## 4. Analysis of ABR Throughput Measures

Once the steady state distribution of the CTMC has been found, we can determine the required throughput measures the *average throughput* and the *throughput threshold* defined by equations (2.10) and (2.11), respectively.

### *4.1. Average Throughput Constraint*

The calculation of the average throughput of the ABR calls is straightforward, since

$$E(\theta) = \frac{\sum_{(n_1,n_2,n_3)\in\mathcal{S}} b_3 \; p_{(n_1,n_2,n_3)} \; r_{(n_1,n_2,n_3)}}{\sum_{(n_1',n_2',n_3')\in\mathcal{S}} n_3' \; p_{(n_1',n_2',n_3')}} \; . \tag{4.1}$$

### *4.2. Throughput Threshold Constraint*

Unfortunately, it is much harder to check the throughput threshold constraint in (2.11), since neither the distribution nor the higher moments of $\theta_x$ can be analyzed based on the steady state distribution of the above studied Markov chain. Hence, in this section, a different approach is applied to analyze the system with the throughput threshold constraint.
The constraint in (2.11) can be analyzed based on the distribution of $T_x$ applying:

$$Pr\left(\theta_x \le \theta^{min}\right) = Pr\left(\frac{x}{T_x} \le \theta^{min}\right) = Pr\left(T_x \ge \frac{x}{\theta^{min}}\right) \; . \tag{4.2}$$

Since it is hard to evaluate the distribution of $T_x$ directly, but there are effective numerical methods to obtain its moments, we check the throughput threshold constraint in (2.11) applying the Markov inequality[3], which gives the following relations:

- if applied for $T_x^n \ge \frac{x^n}{b_3^n}$:

$$Pr\left(T_x \ge \frac{x}{\theta^{min}}\right) \le \frac{\frac{E(T_x^n)}{x^n} - \frac{1}{b_3^n}}{\frac{1}{{\theta^{min}}^n} - \frac{1}{b_3^n}} \tag{4.3}$$

- if applied for $(T_x - E(T_x))^{2n} \ge 0$:

$$Pr\left(T_x \ge \frac{x}{\theta^{min}}\right) \le \frac{M^{(2n)}(T_x)}{\left(\frac{x}{\theta^{min}} - E\left(T_x\right)\right)^{2n}} \tag{4.4}$$

where $n \in \mathbb{N}$ and $M^{(2n)}(T_x) = E\left([T_x - E(T_x)]^{2n}\right)$ denotes the $2n$-th central moment of $T_x$.
The inequalities (4.3) and (4.4) provide approximations on the distribution of $T_x$ based on its $n$-th and $2n$-th central moments, respectively, i.e., a bound of the distribution of $T_x$ is provided based on its $n$-th and $2n$-th central moments. Using different moments (different $n$) the inequalities provide different upper

[3] For any r.v. $X$ with support on $(c,\infty)$ $\quad Pr(X \ge a) \le \frac{E(X)-c}{a-c}$

bounds for $Pr\left(T_x \geq \frac{x}{\theta^{min}}\right)$. Some of these upper bounds can be greater than 1, i.e., meaningless. Of course, we are interested in finding the tightest (lowest) upper bound. If at least one of the upper bounds of $Pr\left(T_x \geq \frac{x}{\theta^{min}}\right)$, obtained by (4.3) or (4.4) for any considered $n$, is less than $\varepsilon$ then the throughput threshold constraint is fulfilled. To check if a given $N_{ABR}$ fits to the throughput threshold constraint for a given $x$ value we apply a numerical method to evaluate the moments of $T_x$ and based on the known moments of $T_x$ the tightest (lowest) upper bound provided by (4.3) or (4.4) for the known moments is chosen, as:

$$\min_n \min \left[ \frac{\frac{E(T_x^n)}{x^n} - \frac{1}{b_3^n}}{\frac{1}{\theta^{min^n}} - \frac{1}{b_3^n}}, \frac{M^{(2n)}(T_x)}{\left(\frac{x}{\theta^{min}} - E(T_x)\right)^{2n}} \right] \tag{4.5}$$

### *4.3. Customer Tagging and System Behavior During ABR Service*

The method we follow to evaluate the moments of $T_x$ is based on tagging an ABR call arriving to the system, which can be in one of the states $i$ such that $i \in \mathcal{S}, i_{3+} \in \mathcal{S}$, and carefully examining the possible transitions from the moment this tagged call enters the system until it acquires the required service and leaves the system.

To illustrate the system behaviour during the service of the tagged ABR call we use the same example as in Section 2. Figure 2 shows the state transition diagram of the same system as in Figure 1 an infinitesimal amount of time after the tagged ABR call entered the system until it leaves the system, which is represented by the state transitions to the trapping state. The initial state probabilities of the CTMC of Figure 2 are obtained by considering the steady state probabilities of the system just before the tagged customer arrived. The dashed arrows indicate the system state before arrival.

The time the tagged ABR call spends in the system ($T$) is a Phase-type distributed r.v. [4], hence its distribution is easy to evaluate. Unfortunately, the distribution of $T$ does not provide any useful information about the distribution of $\theta$, hence no practical throughput constraint can be evaluated based on it.

In order to evaluate $T_x$, instead of $T$, we consider the modified Markov chain of the system assuming that the tagged ABR call never leaves it (e.g., the tagged call has infinite data to transmit) and we evaluate the time needed to transmit $x$ unit of data through the tagged ABR call. (Figure 3 shows the state transition diagram of the system in Figure 2 assuming the tagged customer never leaves it.)

The system introduced in Section 2, is specified by a CTMC over the state space $\mathcal{S}$ with generator matrix $\mathbf{Q}$. The *modified system* used to evaluate $T_x$ has the following properties:

- Since we assume that at least the tagged ABR call is now in the system we exclude states where $n_3 = 0$.
- With each state of the state space there is an associated entrance probability, which is the probability of the event that the modified CTMC starts from that state. When the tagged arriving ABR call finds the system in state $(n_1, n_2, n_3)$ it will bring the system into state $(n_1, n_2, n_3 + 1)$ unless $(n_1, n_2, n_3)$ happens to be an ABR blocking state.

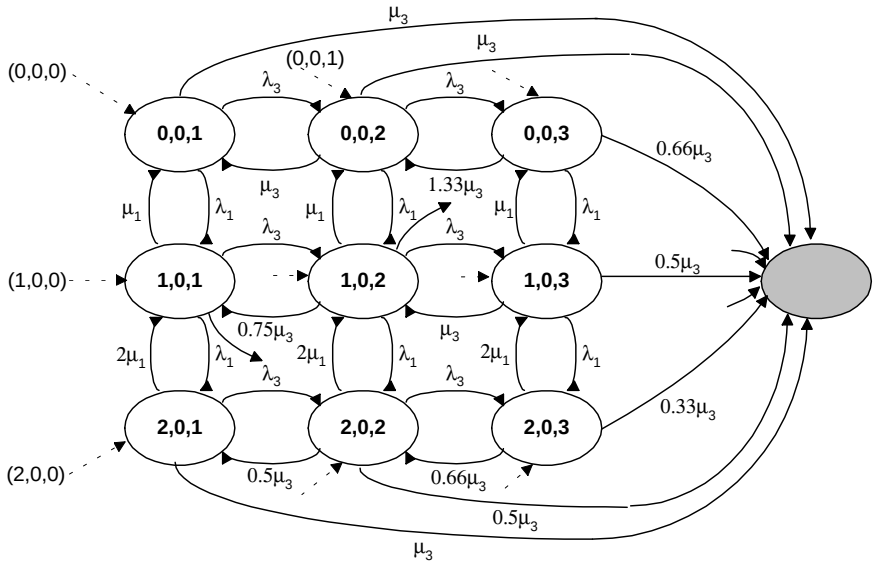


Figure 2. The modified CTMC describes the system during the service of the tagged ABR call

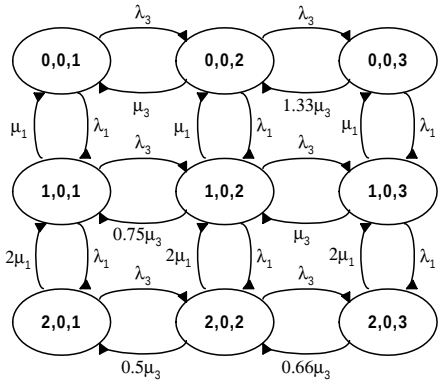


Figure 3. The modified CTMC of Fig. 1 to evaluate $T_x$

Let $\{\mathcal{Z}(t),\ t \geq 0\}$ be the modified CTMC assuming the tagged ABR call never leaves the system over the finite state space $\mathcal{F}$ with generator $\mathbf{B}$. The state space $\mathcal{F}$ can be defined as:

$$n_1 \cdot b_1 + n_2 \cdot b_2 \leq C_{COM} \tag{4.6}$$

$$N_{ABR} \cdot b_3^{min} \leq C_{ABR} \tag{4.7}$$

$$n_3 \leq N_{ABR} \tag{4.8}$$

$$n_3 \geq 1\ . \tag{4.9}$$

Indeed, $\mathcal{F} = \mathcal{S} \backslash \mathcal{S}_0$ where $\mathcal{S}_0$ is the states in $\mathcal{S}$ where $n_3 = 0$. The state transition rates in $\mathbf{B}$ are closely related to the appropriate rates in $\mathbf{Q}$:

$$b_{i,i_{k+}} = \lambda_k, \qquad k = 1,2,3 \tag{4.10}$$

$$b_{i,i_{k-}} = n_k \cdot \mu_k, \qquad k = 1,2 \tag{4.11}$$

$$b_{i,i_{3-}} = \frac{n_3 - 1}{n_3} \cdot r_i \cdot \mu_3 \ . \tag{4.12}$$

The probability $\check{p}_{(n_1,n_2,n_3)}$ that the system is in state $i = (n_1, n_2, n_3)$ immediately after the arrival of the tagged ABR call can be evaluated based on the steady state distribution of the original CTMC with generator $\mathbf{Q}$:

$$\check{p}_{(n_1,n_2,n_3)} = \frac{p_{(n_1,n_2,n_3-1)}}{\sum\limits_{(n_1',n_2',n_3')\in\mathcal{F}} p_{(n_1',n_2',n_3')}}$$

*4.4. The Moments of $T_x$*

To obtain the moments of $T_x$ a Markov reward model over $\{\mathcal{Z}(t),\ t \geq 0\}$ is defined. Let $\mathbf{S}$ be the diagonal matrix whose diagonal elements are So $s_i = r_i b_3/n_3$. $s_i$ is the bandwidth of the tagged ABR connection in state $i$. $T_x$ is a random variable which depends on the (random) arrival and departure of the CBR and ABR calls described by $\mathbf{B}$. The rate matrix $\mathbf{B}$ and the "reward rate" matrix $\mathbf{S}$ determine a Markov reward model over $\mathcal{F}$, and $T_x$ is the time to "accumulate $x$ amount of reward" by this model [4].

Let $\underline{T}_x = \{T_{ix}\}$ be the vector composed by the elements $T_x$ assuming that the system is in state $i$ immediately after the arrival of the tagged ABR call, i.e., $T_{ix} = T_x \mid Z(0^+) = i$.

With the generator matrix $\mathbf{B}$ and the reward rate matrix $\mathbf{S}$ in hand, the moments of $\underline{T}_x$ are obtained using the following effective iterative numerical method [28]. The $n$-th moment of $\underline{T}_x$ is computed as

$$E(\underline{T}_x^n) = n!\beta^n \sum_{i=0}^{K} \underline{N}^{(n)}(i) \cdot \frac{(\alpha x)^i}{i!} e^{-\alpha x} + \underline{\varepsilon}(K, n, \alpha x, \beta) \tag{4.13}$$

where $\alpha = \max\limits_{i,j\in\mathcal{F}}(|b_{ij}|)$; $\beta = \dfrac{1}{\alpha} \cdot \max\limits_{i,j\in\mathcal{F}}(1/s_{ij})$;

$$\mathbf{V} = \frac{\mathbf{S}^{-1} \cdot \mathbf{B}}{\alpha} + \mathbf{I}; \tag{4.14}$$

$$0 \leq \varepsilon_i(K, n, \alpha x, \beta) \leq n! \cdot \beta^n \cdot \alpha x \cdot \left(1 - \sum_{j=0}^{K} \frac{(\alpha x)^j}{j!} e^{-\alpha x}\right); \quad \forall i \in \mathcal{F} \tag{4.15}$$

and $\underline{N}^{(n)}(i)$ is defined as

$$\underline{N}^{(n)}(i) = \begin{cases} \underline{h} & \text{if } i \geq 0, n = 0 \ , \\ \underline{0} & \text{if } i = 0, n \geq 1 \ , \\ \mathbf{V} \cdot \underline{N}^{(n)}(i-1) + & \\ \dfrac{1}{\alpha \cdot \beta} \cdot \mathbf{S}^{-1} \cdot \underline{N}^{(n-1)}(i-1) & \text{if } i \geq 1, n \geq 1 \ ; \end{cases} \tag{4.16}$$

Finally the unconditional moments of $T_x$ can be computed as:

$$E(T_x^n) = \underline{\check{p}}^T \cdot E(\underline{T}_x^n) \ . \tag{4.17}$$

The two terms in the *rhs* of (4.13) are a finite approximation of $E(\underline{T}_x^n)$ and an error term. Equation 4.15 provides an upper bound of the error term. An iteration step defined in (4.16) is composed by matrix vector multiplications, where the matrix is as sparse as $\mathbf{B}$ is (no more than 6 outgoing transitions per state). The numerical method given by (4.13)-(4.16) to determine $E(T_x^n)$ has two notable advantages. Its numerical complexity is very low, hence large models ($\sim 10^4 - 10^6$ states) can be evaluated with it, and the required number of iteration steps ($K$) to achieve a given absolute accuracy is easy to compute in advance (4.15). The proof and a detailed explanation of this numerical method are provided in [28].

In (2.11) it is required that the inequality $Pr(\theta_x \leq \theta^{min}) \leq \varepsilon$ holds for $\forall x \geq 0$, but in the numerical approach we followed some representative $x$ values were considered only. The extreme cases result:

$$\lim_{x\to 0} E\left(\frac{T_x^n}{x^n}\right) = \underline{\check{p}} \cdot \mathbf{S}^{-n} \cdot \underline{h}^T \tag{4.18}$$

$$\lim_{x\to \infty} E\left(\frac{T_x^n}{x^n}\right) = (\underline{\tilde{p}} \cdot \mathbf{S} \cdot \underline{h}^T)^{-n} \tag{4.19}$$

where $\underline{\tilde{p}}$ denotes the steady state distribution of $\{\mathcal{Z}(t),\ t \geq 0\}$. Since $T_x/x$ tends to the constant $\underline{\tilde{p}} \cdot \mathbf{S} \cdot \underline{h}^T$ as $x$ tends to $\infty$, i.e., its variance decreases to 0, the representative $x$ values are "close" to 0 with respect to the transient time of $\{\mathcal{Z}(t),\ t \geq 0\}$.

### *4.5. The Complete Link Allocation Procedure*

Finally, the link allocation procedure is summarized in Figure 4.

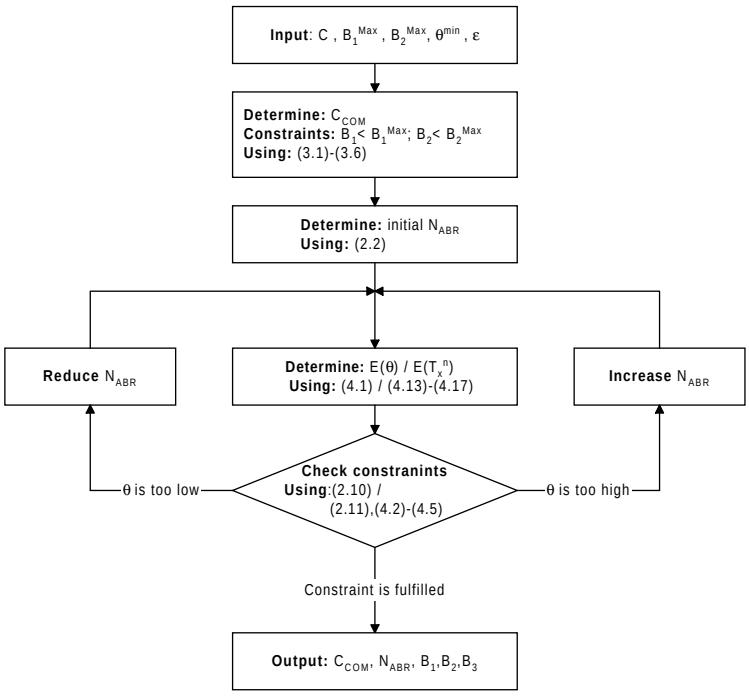


Figure 4. The block diagram of the link allocation procedure

| | 1. (CBR) | 2. (CBR) | 3. (ABR) |
|---|---|---|---|
| Peak bandwidth $b_i$ (Mbps) | 3 | 6 | 10 |
| Arrival intensity $\lambda_i$ (1/s) | 6 | 3 | 12 |
| Mean holding time $1/\mu_i$ (s) | 1 | 1 | 1 |

Table 1
The input parameters of the system under investigation

| $N_{ABR}$ | 10 | 20 | 40 | 60 | 80 | 100 | 150 |
|---|---|---|---|---|---|---|---|
| $B_3$ | 0.310 | 0.0811 | 0.0320 | 0.0212 | 0.00141 | 0.0112 | 0.000461 |
| E($\theta$) | 9.99 | 7.9 | 4.83 | 3.45 | 2.69 | 2.2 | 1.52 |

Table 2
Impact of the $N_{ABR}$ on the ABR throughput and blocking probability

## 5. Numerical Examples on the Application of the POL Link Allocation Policy

In this Section we consider and discuss numerical results, which reveal the blocking probability and the throughput features of the POL link allocation scheme.

### *5.1. Average Throughput Constraint*

Consider an ATM transmission link of capacity 155 Mbps, which is offered calls according to Poisson processes belonging three different service classes, two CBR service classes and one ABR service class, as described in Section 2. The input parameters of this system are given in Table 1.
The value of $b_3^{min}$ is 0.1Mbps. Furthermore, we require that the blocking probabilities of the narrow-band and wide-band CBR calls are less than $B_1^{max} = 2\%$ and $B_2^{max} = 4\%$, respectively. From these parameters and from (3.7), it follows that the minimal bandwidth necessary to provide these blocking probabilities is $C_{COM} = 60Mbps$, which leaves $C_{ABR} = 95Mbps$ for the ABR traffic. The size of the state space is 18271 when $N_{ABR} = 150$.

#### *5.1.1. Examining the Trade-off Between Throughput and Blocking Probability*

Table 2 contains the throughput (E($\theta$)) and blocking probability ($B_3$) constraints for the ABR service class. From Table 2 we clearly note the trade-off between the ABR throughput and blocking probability. In the POL policy this trade-off is conveniently controlled by means of the $N_{ABR}$ cut-off parameter. For instance, when the required average throughput of the ABR class is set to $\theta^{min}$=2.2Mbps, the maximum number of active ABR connections ($N_{ABR}$) is limited to 100.

#### *5.1.2. Comparing the POL and the CP Policies*

Table 3 compares the throughput (E($\theta$)) and the blocking probabilities ($B_1, B_2, B_3$) for the service classes under the POL and the Complete Partitioning (CP) policies. Under the CP policy the link capacity is partitioned among the service classes such that each partition is dedicated to one (or a subset) of the

| | $\lambda_3$ | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|
| POL | $B_1$ | 0.017 | 0.017 | 0.017 | 0.017 | 0.017 | 0.017 |
| | $B_2$ | 0.0411 | 0.0411 | 0.0411 | 0.0411 | 0.0411 | 0.0411 |
| | $B_3$ | $4.11 \cdot 10^{-12}$ | $2.74 \cdot 10^{-8}$ | $5.00 \cdot 10^{-5}$ | 0.0115 | 0.0766 | 0.142 |
| | E($\theta$) | 9.02 | 7.93 | 5.64 | 2.20 | 1.37 | 1.28 |
| CP | $B_1$ | 0.017 | 0.017 | 0.017 | 0.017 | 0.017 | 0.017 |
| | $B_2$ | 0.0411 | 0.0411 | 0.0411 | 0.0411 | 0.0411 | 0.0411 |
| | $B_3$ | $3.26 \cdot 10^{-4}$ | 0.0503 | 0.136 | 0.208 | 0.269 | 0.321 |
| | E($\theta$) | 3.86 | 1.16 | 1.01 | 0.98 | 0.97 | 0.97 |

Table 3
Comparing CP and POL policies with different load of ABR traffic

| $\lambda_1$ | 7.2 | 6 | 4.8 | 3.6 | 2.4 | 1.2 |
|---|---|---|---|---|---|---|
| $\lambda_2$ | 3.6 | 3 | 2.4 | 1.8 | 1.2 | 0.6 |
| $B_1$ | 0.0417 | 0.0170 | 0.00436 | $5.12 \cdot 10^{-4}$ | $1.46 \cdot 10^{-5}$ | $1.46 \cdot 10^{-8}$ |
| $B_2$ | 0.0940 | 0.0411 | 0.0114 | 0.00150 | $5.05 \cdot 10^{-5}$ | $6.80 \cdot 10^{-8}$ |
| $B_3$ | 0.0448 | 0.0111 | $6.29 \cdot 10^{-4}$ | $1.16 \cdot 10^{-5}$ | $1.51 \cdot 10^{-7}$ | $1.85 \cdot 10^{-9}$ |
| E($\theta$) | 1.45 | 2.20 | 4.36 | 6.66 | 8.02 | 8.83 |

Table 4
Impact of the CBR load on the ABR throughput and blocking probabilities

service classes. The capacity associated with the CBR classes is $C_{CBR} = 60Mbps$ while the capacity associated with the ABR class is $C_{ABR} = 95Mbps$. We observe that the POL policy is superior to the CP policy under all loads both in terms of blocking probabilities and ABR throughput. It is intuitively also clear that this is because the POL policy utilizes the link capacity better than the CP policy, since it allows the ABR class to make use of any unused CBR bandwidth.

*5.1.3. Examining the Impact of Varying the Traffic Mix*

In the example presented in Table 4 the ABR class arrival intensity is kept fixed at $\lambda_3 = 12$. The total CBR load (i.e. $\sum \lambda_i/\mu_i * b_i, i = 1, 2$) varies from 43.2 Erlang*Mbps down to 7.2 Erlang*Mbps, while the ABR load is kept fixed at 120 Erlang*Mbps. As expected, as the load in the system decreases, so increases the ABR throughput and decreases the ABR blocking probability.

In the example presented in Table 5 we keep the total CBR load fixed at 36 Erlang*Mbps and we are interested of the ABR class blocking probability and throughput when the CBR traffic mix varies. We note that when the CBR load comes from the "wide-band" CBR class, the ABR blocking probability is somewhat lower and its throughput is higher than when the CBR load is mainly due to the "narrow-band" CBR class. We explain this result by noting that the narrow-band CBR class arrival intensity is higher than the wide-band CBR class arrival intensity (when the offered loads are the same), and it is well-known that the blocking probability of the more intensive but narrower band service classes is

| $\lambda_1$ | 0 | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|---|
| $\lambda_2$ | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| $B_1$ | 0.043 | 0.0204 | 0.0188 | 0.0170 | 0.0149 | 0.0119 | 0.00979 |
| $B_2$ | 0.0431 | 0.0450 | 0.0454 | 0.0411 | 0.0369 | 0.0320 | 0.0261 |
| $B_3$ | 0.00979 | 0.00969 | 0.0105 | 0.0111 | 0.0124 | 0.0134 | 0.0144 |
| E($\theta$) | 2.32 | 2.32 | 2.26 | 2.20 | 2.14 | 2.09 | 2.05 |

Table 5
Impact of the CRBs load on the ABR throughput and blocking probabilities

| $C_{COM}$ | 69 | 66 | 63 | 60 | 57 | 54 |
|---|---|---|---|---|---|---|
| $B_1$ | 0.00498 | 0.00770 | 0.0116 | 0.0171 | 0.0244 | 0.0342 |
| $B_2$ | 0.0126 | 0.0192 | 0.0284 | 0.0411 | 0.0578 | 0.0794 |
| $B_3$ | 0.0149 | 0.0141 | 0.0129 | 0.0115 | 0.00973 | 0.00773 |
| E($\theta$) | 2.04 | 2.08 | 2.13 | 2.20 | 2.31 | 2.46 |

Table 6
Impact of the $C_{COM}$ on the ABR throughput and blocking probabilities

lower than that of the wide-band classes. Thus, in the case of Table 5, the CBR classes consume more bandwidth when $\lambda_1 = 12, \lambda_2 = 0$ than when $\lambda_1 = 0, \lambda_2 = 6$.

*5.1.4. Examining the Impact of the $C_{COM}$ Parameter*

The $C_{COM}$ parameter, as the other output parameter of the POL policy (i.e. the one being $N_{ABR}$), offers a method to control the trade-off between the CBR blocking probabilities and the ABR blocking probability and throughput (Table 6). Unlike in the case of tuning the $N_{ABR}$ parameter as in Table 2, in this case both the ABR throughput and blocking probability are improved at the expense of some CBR blocking probability degradation.

*5.2. Throughput Threshold Constraint*

In this subsection we consider a system with a single CBR and a single ABR service class, where the system input parameters are as follows: $C = 100Mbps$, $C_{COM} = 80Mbps$, $b_1 = 2Mbps$, $\lambda_1 = 20\ (1/s)$, $\mu_1 = 1\ (1/s)$, $b_3 = 5Mbps$, $\lambda_3 = 10\ (1/s)$, $\mu_3 = 1\ (1/s)$, $b_3^{min} = 0.166Mbps$. In the throughput threshold constraint case one is interested in the probability that the $\theta_x$ ( the ABR throughput during the transmission of $x$ amount of data) falls below the $\theta^{min}$ throughput threshold constraint. In the numerical analysis below we evaluated the throughput threshold constraint with three different value of $x$, $0.1, 1, 10$ Mb. The worst upper bound was obtained with $x = 0.1$ Mb, in all cases. Under the POL policy, this probability is controlled by the $N_{ABR}$ cut-off parameter, as it is shown in Table 7. For instance, we can see that if we allow 4% probability for the event that the ABR throughput drops below

| $\theta^{min}$ | | 4 | 3 | 2.5 | 2 | 1.5 | 1 |
|---|---|---|---|---|---|---|---|
| $N_{ABR}$ | 10 | 2.3% | 0.22% | 0.12% | 0.053% | 0 | 0 |
| | 15 | 27.9% | 5.6% | 1.9% | 0.2% | 0.014% | 0 |
| | 20 | 59.3% | 21.2% | 8.5% | 1.16% | 0.065% | 0 |
| | 25 | 82.4% | 30.9% | 18.2% | 4.7% | 0.3% | 0.011% |
| | 30 | * | 36.5% | 24.3% | 10.5% | 1.4% | 0.041% |
| | 35 | * | 40.1% | 26.7% | 14.3% | 2.6% | 0.1% |
| | 40 | * | 42.2% | 28.1% | 16.4% | 4% | 0.2% |
| | 80 | * | 44.9% | 29.9% | 19.1% | 7.6% | 0.86% |
| | 120 | * | 45% | 30% | 19.1% | 7.6% | 1% |

Table 7
Upper bounds for $\max_{\forall x \geq 0} P\left(\theta_x \leq \theta^{min}\right)$;

1.5 Mbps for each $x$ value, then it is sufficient to set the $N_{ABR}$ parameter to 40. In Table 7, '$*$' denotes the meaningless ($> 1$) upper bounds and the 0s are determined by the structure of the state space $\mathcal{F}$.

To compare the behaviour of the average throughput constraint and the throughput threshold constraint, let $\theta^{min} = 4$ Mbps. In this case the average throughput constraint allows $N_{ABR} = 30$, while if $\varepsilon$ is 30% (3%) then the throughput threshold constraint allows $N_{ABR} = 15$ ($N_{ABR} = 10$). It is visible that the throughput threshold constraint is a stronger constraint than the one on the average throughput if $\varepsilon$ is small, hence it allows less ABR calls to enter the system, i.e., $N_{ABR}$ is less.

Next, we let $N_{ABR} = 40$ be fixed and we compare the Type 1 and Type 2 bounds defined by the equations (4.3) and (4.4), respectively. The actual upper bound for each case is given in the last row of Table 8. ('$*$' denotes the meaningless ($> 1$) upper bounds.) From Table 8 it turns out that none of the two bounds is superior to the other in each case. For instance, when the upper bound is set to 16.4%, the Type 1 bound is better, whereas at 4% the Type 2 gives the better result. This table also shows that the best bound can be obtained by different moments. Not always the higher moments provide the better upper bounds.

## 6. Conclusion

In this paper we have proposed an ATM call level model, which is an extension of the classical multi-rate loss model in that it allows one to model service class whose bandwidth fluctuates in time in accordance with the instantaneous load on the link. This is achieved by allowing such service classes to specify their minimal accepted bandwidth in addition to their peak bandwidth requirement. Furthermore, this type of calls specify their ideal mean call holding time, which corresponds to the total amount of required service, rather than specifying the mean call holding time.

We have used this model to investigate the performance of the adoption of the Partial Overlap link allocation policy for an ATM transmission link which is offered CBR and ABR calls. By employing efficient numerical methods to find the steady state of the system and the reward measures of a modified system, we have found that the POL policy is relatively easy to dimension and is able to take into account GoS and throughput constraints and to minimize the ABR class blocking probability.

| $\theta^{min}$ | | 3 | 2.5 | 2 | 1.5 | 1 |
|---|---|---|---|---|---|---|
| Type 1 defined in (4.3) | $n=1$ | 42.2% | 28.1% | 18.7% | 12% | 7.0% |
| | $n=2$ | 50.5% | 29.9% | 17.1% | 8.8% | 3.7% |
| | $n=3$ | 66.2% | 34.3% | 16.4% | 6.6% | 1.9% |
| | $n=4$ | 97.5% | 43.6% | 17.2% | 5.3% | 1.0% |
| | $n=5$ | * | 62.4% | 20% | 4.7% | 0.62% |
| | $n=6$ | * | * | 26.1% | 4.6% | 0.4% |
| Type 2 defined in (4.4) | $2n=2$ | * | 49.4% | 17.2% | 6.0% | 1.8% |
| | $2n=4$ | * | * | 32.2% | 4.0% | 0.37% |
| | $2n=6$ | * | * | * | 7.2% | 0.2% |
| Upper bound | | 42.2% | 28.1% | 16.4% | 4% | 0.2% |

Table 8
Upper bounds for $\max_{\forall x \geq 0} P\left(\theta_x \leq \theta^{min}\right)$; $(N_{ABR} = 40)$

## Acknowledgements

The authors wish to thank Gergely Mátéfi for the implementation of the numerical method in Section 5. M. Telek was partially supported by OTKA F-23971. G. Fodor gratefully acknowledges the support from the Technical University of Denmark in Lyngby, Denmark.

## References

[1] ATM Forum, "Traffic Management Specification Version 4.0", April, 1996.

[2] ITU-T Recommendation I.371, "Traffic Control and Congestion Control in B-ISDN"

[3] M. Ajmone Marsan, K. Begain, R. Lo. Cigno and M. Munafo, "Performance of TCP File Transfers over the Explicit Rate ABR ATM Service Category", *5th International Conference on Telecommunication Systems - Modeling and Analysis*, pp. 248-258, Nashville, TN, USA, March 1997.

[4] A. Bobbio and K.S. Trivedi, "Computation of the Distribution of the Completion Time when the Workload Requirement is a PH Random Variable", *Stochastic Models*, 6:133-149, 1990.

[5] F. Bonomi and K. W. Fendick, "The Rate-Based Flow Control Framework for the Available Bit Rate ATM Service", *IEEE Network*, pp. 25-39. March/April, 1995.

[6] S. C. Borst and D. Mitra, "Virtual Partitioning for Robust Resource Sharing: Computational Techniques for Heterogeneous Traffic", *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 5, pp. 668-678, June, 1998.

[7] T. M. Chen, S. L. Liu and V. K. Samalam, "The Available Bit Rate Service for Data in ATM Networks", *IEEE Communications Magazine*, pp. 56-71, May 1996.

[8] G. Choudhury, K. K. Leung and W. Whitt, "Efficiently Providing Multiple Grade of Service with Protection Against Overloads in Shared Resources", *AT&T Technical Journal*, July/August, pp. 50-63, 1995.

[9] Z. Dziong and L. G. Mason, "Call Admission and Routing in Multi-Service Loss Networks", *IEEE Transactions on Communications*, Vol. 42, No. 2, 1994.

[10] A. Farago, S. Blaabjerg, L. Ast, G. Gordos and T. Henk, "A New Degree of Freedom in ATM Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, September, 1995.

[11] G. Fodor, E. Nordström and S. Blaabjerg, "Revenue Optimization and Fairness Control of Priced Guaranteed and Best Efort Services on an ATM Transmission Link", in the Proc. of the *IEEE International Conference on Communications, ICC '98*, Vol. 3, pp. 1696-1705, Atlanta, GA, USA, June, 1998.

[12] J. Kaufman, "Blocking in a Shared Resource Environment", *IEEE Trans. on Comm.*, pp. 1474-1481, 1981.

[13] D. Mitra, M. I. Reiman and J. Wang, "Robust Dynamic Admission Control for Unified Cell and Call QoS in Statistical Multiplexers", *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 5, pp. 692-707, June, 1998.

[14] 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.

[15] A. A. Nilson and M. J. Perry, "Multirate Blocking Probabilities: Numerically Stable Computations", *International Teletraffic Congress, ITC-15*, pp. 1359-1368.

[16] E. Nordström, "Near-Optimal Link Allocation of Blockable Narrow-Band and Queueable Wide-Band Call Traffic in ATM Networks", in the *Proceedings of the* $15^{th}$ *International Congress on Telecommunications, ITC '15*, Washington D.C., USA, June, 1997.

[17] T. Oda and Y.Watanabe, "Optimal Trunk reservation for a group with multislot traffic streams" *IEEE Transactions on Communications*, Vol. 38, No. 7, pp 1078-1084, 1990.

[18] J. W. Roberts, "Realizing Quality of Service Guarantees in Multi-service Networks", in *Performance Management of Complex Communication Networks*, eds.: T. Hasegawa, H. Takagi, Y. Takahashi, IFIP., Chapman & Hall, pp. 277-293, 1998.

[19] J. W. Roberts, "Quality of Service Guarantees and Charging in Multiservice Networks", *IEICE Transactions on Communications, Spec. Issue on ATM Traffic Control and Performance Evaluation*, Vol. E81-B, No. 5, pp. 824-831, May, 1988.

[20] J. W. Roberts (ed), "Methods for the Performance Evaluation and Design of Broadband Multiservice Networks", *Published by the Commission of the European Communities, Information Technologies and Sciences, COST 242 Final Report*, 1996.

[21] S. Rosenberg, M. Aissaoui, K. Galway and N. Giroux, "Functionality at the Edge: Designing Scalable Multi-service ATM Networks", *IEEE Communications Magazine*, pp. 88-99, May, 1998.

[22] K. W. Ross, "Multi-service Loss Models for Broadband Telecommunication Networks", *Springer Verlag London Limited*, ISBN 3-540-19918-7, 1995.

[23] K. W. Ross and D. K. H. Tsang, "Optimal Circuit Access Policies in an ISDN Environment: A Markov Decision Approach", *IEEE Transactions on Communications*, Vol. 37, No. 9, September, 1989, pp. 934-939.

[24] Y. D. Serres and L. G. Mason, "A Multi-server Queue with Narrow- and Wide-band Customers and Wide-band Restricted Access", *IEEE Transactions on Communications*, Vol. 36, pp. 675-684, 1988.

[25] A. Smith, J. Adams and G. Tagg, "Available Bit Rate - A New Service for ATM", *Computer Networks and ISDN Systems*, 28, pp. 635-640, 1996.

[26] W. J. Stewart, "Introduction to the Numerical Solution of Markov Chains", *Princeton University Press*, Princeton, New Jersey, ISBN 0-691-03699-3, 1994.

[27] E. D. Sykas, K. M. Vlakos, I. S. Venieris and E. N. Protonotarios, "Simulative Analysis of Optimal Resource Allocation and Routing in IBCN's", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 3, 1991.

[28] M. Telek and S. Rácz, "Numerical Analysis of Large Markov Reward Models", *Performance Evaluation*, Vol. 36&37, pp. 95-114, 1999;

[29] Wu-chang Feng, Dilip D. Kandlur, Debanjan Saha and Kang G. Shin, "Understanding and Improving TCP Performance Over Networks with minimum Rate Guarantees", *IEEE/ACM Transactions on Networking*, pp. 173-187, Vol. 7, No. 2, April 1999.