A Partially Blocking-Queueing System with CBR/VBR and ABR/UBR Arrival Streams

Søren Blaabjerg
Institute of Telecommunications, Building 343, Technical University of Denmark,
DK-2800, Lyngby, Denmark, sbb@tele.dtu.dk, Fax: +45 45930355

Gábor Fodor
Department of Telecommunications and Telematics, Technical University of Budapest,
Sztoczek 2, H-1111 Budapest, Hungary, fodor@ttt-atm.ttt.bme.hu, Fax: +36 14633107

Allan T. Andersen
Department of Mathematical Modelling, Building 321, Technical University of Denmark,
DK-2800, Lyngby, Denmark, ata@imm.dtu.dk, Fax: +45 45881397

Miklós Telek
Department of Teleommunications, Technical University of Budapest,
Sztoczek 2, H-1111 Budapest, Hungary, telek@hit.bme.hu, Fax: +36 14633266

Abstract

In this paper we consider a single *Asynchronous Transfer Mode (ATM)* transmission link, to which *Constant Bit Rate (CBR)* or *Variable Bit Rate (VBR)* and *Available Bit Rate (ABR)* or *Unspecified Bit Rate (UBR)* calls arrive according to independent Poisson processes. CBR/VBR calls (characterized by their *equivalent bandwidth*) are blocked and leave the system if the available link capacity is less than required at the time of arrival. ABR/UBR calls, however, accept *partial blocking* [3], meaning that they may enter service even if the available capacity is *less* than the specified *required peak bandwidth*, but greater than the so called *minimal accepted bandwidth.* Partially blocked ABR/UBR calls instead experience longer service time, since smaller given bandwidth entails longer time spent in the system, as detailed in [3]. Throughout the life time of an ABR/UBR connection, its bandwidth consumption fluctuates in accordance with the current load on the link but always at the highest possible value up to their peak bandwidth (greedy sources). Additionally, if this minimal accepted bandwidth is unavailable at the time of arrival ABR/UBR calls are allowed to wait in a *finite queue.* This system is modelled by a *Continuous Time Markov Chain* and the blocking probabilities, the moments and the distribution of the ABR/UBR waiting and service time are derived.

1 Introduction

One of the main concerns regarding the *Asynchronous Transfer Mode (ATM)* is the integration of services having strict *Quality of Service (QoS)* gurantees [41] (*Constant Bit Rate (CBR)* and *Variable Bit Rate (VBR)*), with services of limited (*Available Bit Rate (ABR)*) or without (*Unspecified Bit Rate (UBR)*) such gurantees. Since ATM networks are connection oriented and by adopting the concept of *Equivalent Bandwidth*, it is possible to model ATM networks carrying CBR and VBR traffic as multirate circuit switched networks on the call level [33, 34, 10, 36]. Thus, the multirate Erlang Blocking Model [36] has been successfully used to analyze such networks. With the introduction of the "best effort" type service classes (ABR and UBR) these models need to be extended, because (1) the traditional equivalent bandwidth based approach for bandwidth estimation is not directly applicable (since there are less or no QoS parameters at all), (2) there is either very limited or no resource allocation made prior to the information transfer phase and (3) the traditional models disregard the rate based closed loop flow control mechanism which is an essential feature of the ABR service category. [41, 38].

A generalization of the multirate circuit switched loss model has been presented in [3], where it was argued that with the introduction of *partial blocking* into the Multirate Erlang Model it is possible to model ABR/UBR services (best effort services) on the call scale. The key feature of such a system is that calls accepting partial blocking specify (in addition to their peak bandwidth requirement, B_r) a so called *minimal accepted service rate*, r_{min} (in ABR terminology the *minimum cell rate* MCR). During the call negotiation process an ABR/UBR (best effort) connection is accepted if, and only if, the available bandwidth (B_a) at the time of arrival satisfies: $r_{min} * B_r \leq B_a$. During the life time of such a connection the instantenous service rate $r(t)$, defined as $B_a(t)/B_r$, fluctuates according to the current load and the available capacity on the link, capturing the behaviour of an ideally working rate based ABR control algorithm. An underlying assumption here is that the best effort source is greedy in the sense that as long as the connection is established the source will always transmit with maximum possible rate which is the smallest of its peak rate B_r and its equal share of the bandwidth left for the ABR/UBR service category.

Since we in this paper want to develop a new *call level* model to include these new service classes, we will use the terms "best effort type service classes" and "service classes with QoS guarantees" quite loosely and interchangebly with the terms "ABR/UBR" and "CBR/VBR" service classes.

Since the given bandwidth-residency time is kept constant, and since the given (available) bandwidth, B_a, may fluctuate, this model can be seen as a generalization of the "Erlang Blocking Model with Retrials" analyzed by Kaufman in [21]. There, a type i call can specify "retry parameters" $(B_{ir}, 1/\mu_{ir})$, where $B_{ir} < B_i$ (B_i is the original bandwidth requirement of a type i call). If such retry parameters are specified a blocked type i call will immediately re-attempt, but now requesting reduced bandwidth B_{ir} with a mean residency time $1/\mu_{ir}$. Therefore the non real-time message types (e.g. file transfers) may, upon being blocked obtain service but with smaller bandwidth (B_{ir}) and larger residency time ($1/\mu_{ir}$), as long as the bandwidth-residency time product is the same as originally requested ($B_i * 1/\mu_i$), [21].

It is expected from second generation ATM switches that they will allow the fluctuation of the actual bandwidth to ABR and UBR calls in accordance with the available capacity. For the ABR service category this is made possible by the ATM forum standardization of the rate based flow control framework [41, 5]. For both ABR and UBR it is further supported by the introduction of fair queueing cell scheduling schemes like virtual spacing scheduling [35], which are implementable approximations of the generalized processor sharing scheduling discipline and its packetized version (PGPS) [30, 31].

It has been observed in many papers [2, 28, 47, 15, 25, 40], that in a multirate network, where services with large difference between the bandwidth requirements are present wide band calls suffer much higher blocking probabilities than narrow band calls. By applying either *trunk reservation* or *class limitation* it is possible to level out the blocking probabilities. However, in most cases, the disadvantage put on the narrow band traffic is much bigger than the advantage obtained for the wide band traffic. Employing these fairness procedures therefore does not solve the problem of how to achieve good network performance for all traffic types and high utilisation at the same time.

If it is possible to allow calls requiring a large amount of bandwidth to wait in a queue (i.e. to allow for *call queueing*) until resources become available, there is a hope to significantly reduce the blocking probability for these calls on the expense that sometimes a wide band call will have to wait until a connection can be established for the call. Allowing some traffic classes to wait in a queue implies that we have a *mixed system with both loss and delay.*

Therefore in this paper we consider a system where both *partial blocking, (PB)* and *call queueing* are allowed for best effort calls. Specifically, we investigate an ATM link of capacity C, to which *Constant Bit Rate (CBR)* or *Variable Bit Rate (VBR)* and *Available Bit Rate (ABR)* or *Unspecified Bit Rate (UBR)* calls arrive according to independent Poisson processes with intensities λ_{CV} and λ_{AU}. Calls belonging to the CBR/VBR class are blocked and leave the system if the available link capacity is less than the required, B_{CV}, at the time of arrival. Calls belonging to the best effort class, however, accept partial blocking [3] down to a minimum bandwidth requirement of $r_{min} * B_{AU}$ on the expense of a longer time spent in the system. Additionally, best effort calls are allowed to wait in a *finite queue* of length Q, if even this minimal accepted bandwidth is unavailable at the time of arrival. If the current queue length, q, is already Q, then the best effort call is blocked and lost. Queued best effort calls enter service as soon as the available bandwidth reaches $r_{min} * B_{AU}$. Since all in-service best effort calls always receive the same *instantenous service rate*, $r(t)$, a new call (be it of guaranteed service or best effort) is allowed into service even if the link is "full", provided that the system is able to "compress" the in-service best effort calls such, that the new service rate for the best effort calls still remains greater than r_{min}.

Section 2 presents the Markovian model where both partial blocking and queueing are allowed. Calls with guaranteed service compete with best effort calls for the bandwidth on a link and the underlying Quasi-Birth-Death structure (QBD) of the transition matrix is described. From the usual steady state analysis blocking probabilities for the two traffic types are derived and by an application of Little's result also the mean time a best effort call spends in the system is derived. In section 3 a tagged best effort call is modelled from it enters the system (in either the queue or directly into service) and until it leaves the system by a transient Markov Chain with an absorbing state. Thereby the distribution of the time a best effort call spends in the system is derived. In section 4 the simplified system without queueing is analysed and it is shown how the distribution of the time a best effort call spends in the system can be derived by applying techniques from Markov driven workload processes. Finally, section 5 discusses a number of numerical results enlightning how the relevant performance measures varies as a function of e.g. *minimal accepted service rate*, r_{min}, and the finite queue's length, Q, for best effort calls.

2 The Partially Blocking-Queueing System

2.1 Model and Assumptions

In this section we formulate the Markov model in which a single link is offered calls from two classes of traffic.

- Calls with guaranteed service characterized by their arrival rate λ_{CV} their departure rate μ_{CV} and their *equivalent* bandwidths B_{CV} and
- Best effort calls characterized by their their arrival rate λ_{AU} their departure rate μ_{AU} their *peak* rate B_{AU} and *minimum required* rate $r_{min}B_{AU}$

Both types of calls arrive according to Poisson processes and the holding time for CBR/VBR (guaranteed service) calls are exponentially distributed with departure rate μ_{CV}. Each arriving ABR/UBR (best effort) call brings with it an exponentially distributed service requirement which in case the peak bandwidth is available throughout the entire duration of the connection gives rise to a departure intensity of μ_{AU}. In case the peak bandwidth is not available all best effort connections in progress on the link share the available bandwidth equally among them. The rate at which the best effort calls are receiving service then fluctuates in accordance with the bandwidth that is available on the link, the response time assumed to be zero corresponding to an ideally working closed loop ABR rate based flow control without propagation delay. The best effort calls in progress on the link are not allowed to receive service at a rate smaller than $r_{min}B_{AU}$. Instead incoming call attempt are blocked or queued.

2.2 System Description

The system under investigation is characterised by $(n_{CV}(t), n_{AU}(t))$ where $n_{CV}(t)$ is the number of guaranteed service calls on the link at time t and $n_{AU}(t)$ is the number of best effort calls in the system (on the link and in the queue) at time t. The vector $(n_{CV}(t), n_{AU}(t))$ uniquely specifies how many best effort calls are waiting in the queue (q), and what service rate r the in-service best effort calls receive. Under the assumption of Poisson arrivals and exponential holding times $(n_{CV}(t), n_{AU}(t))$ constitutes a two dimensional Markov Chain and to obtain the performance measures we need to find the generator matrix G and to solve $\pi G = 0$ and $\pi e = 1$ where $e = (1, ..., 1)$ and π is the steady state probability distribution to be found.

The Markov Chain is not time reversible and does not obey a product form solution. However, as will be shown next, it does have a nice quasi-birth-death (QBD) structure which allows for efficient methods for deriving the steady state distribution π.

Let $p = \lfloor \frac{C}{B_{AU} * r_{min}} \rfloor$. The two dimensional state space can be partitioned into $Q + p + 1$ "macro-states" $S_0, .., S_Q, S_{Q+1}, .., S_{Q+p}$ where $S_j = \{(i,j) | i = 0, .., \lfloor \frac{C}{B_{CV}} \rfloor\}$ for $0 \leq j \leq Q$ and $S_j = \{(i,j) | i = 0, .., \lfloor \frac{C - (j-Q) * r_{min} * B_{AU}}{B_{CV}} \rfloor\}$ for $Q < j \leq Q + p$. Adopting this organisation of the state space and utilising the fact that in a continuous time Markov Chain multiple events have probability zero, the generator matrix will have the following quasi-birth-death structure:

$$G = \begin{bmatrix} A_0 & C_0 & 0 & 0 & .. & 0 & 0 & 0 \\ B_1 & A_1 & C_1 & 0 & .. & 0 & 0 & 0 \\ 0 & B_2 & A_2 & C_2 & .. & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & .. & A_{Q+p-2} & C_{Q+p-2} & 0 \\ 0 & 0 & 0 & 0 & .. & B_{Q+p-1} & A_{Q+p-1} & C_{Q+p-1} \\ 0 & 0 & 0 & 0 & .. & 0 & B_{Q+p} & A_{Q+p} \end{bmatrix} \tag{1}$$

where the square matrix A_j represents all possible transitions when the number of best effort calls in the system is kept at j, the B_j (in general not square) matrix represents the possible transitions when a best effort call is removed from the system, and the C_j (also not square in general) matrix represents the transitions when a best effort call enters the system.

The maximum number of best effort calls the system can handle simultaneously is $N_{AU}^{max} = \lfloor \frac{C}{B_{AU} * r_{min}} \rfloor + Q$. Let $S_{TOT}(k)$ denote the number of *states* (i.e. the maximum number of guaranteed service calls + 1) when the number of best effort calls in the system (i.e. in service *and* in the queue) is fixed at just k, $k = 0, ..., N_{AU}^{max}$. Then

$$S_{TOT}(k) = \lfloor \frac{C - (max\{0, k - Q\} * r_{min} * B_{AU})}{B_{CV}} \rfloor + 1 \tag{2}$$

Thus, the total number of states, S, is simply:

$$S = \sum_{l=0}^{N_{AU}^{max}} S_{TOT}(l) \tag{3}$$

Next, we will number (assign scalar indexes to) the states in the two dimensional state space from $0...S-1$, such that when the system is in state (i, j), it will have the index $s = \sum_{l=0}^{j-1} S_{TOT}(l) + i$. That is, when the system is in the state with index s, there are i guaranteed service and j best effort calls in the system:

$$j = f_{AU}(s) := inf\{I; \sum_{l=0}^{I} S_{TOT}(l) > s\} \qquad (4)$$

$$i = f_{CV}(s) := s - \sum_{l=0}^{f_{AU}(s)-1} S_{TOT}(l) \qquad (5)$$

Thus $f_{AU}(s)$ gives from the index s the unique number of best effort calls in state s while $f_{CV}(s)$ gives the unique number of guaranteed service calls in state s. In any state (either given by its index, s, or by the tuple (i,j)), we need to find the current queue length (i.e. the number of best effort calls in the queue) and the current service rate, r, associated with that state. Note that when $j = 0$, or when $j = q$, i.e. all best effort calls are queued (if any in service), we naturally have $r = 0$. It is important to note that arriving guaranteed service calls are not allowed to "squeeze out" in-service best effort calls, even if there were space for them in the queue. That is, an arriving CBR/VBR call cannot increase the current queue length, q.

$$q = f_q(i,j) := sup\{q; C - i * B_{CV} \geq r_{min} * (j-q) * B_{AU}\} \qquad (6)$$

$$r = f_r(i,j) := \begin{cases} \frac{C - i * B_{CV}}{(j-q) * B_{AU}} & \text{if } j \neq q \text{ and } j \neq 0 \\ 0 & otherwise \end{cases} \qquad (7)$$

With these equations at hand it becomes easy to specify element (s_1, s_2) of the generator matrix G, since it represents a transition from a state of index s_1 to a state with index s_2. Denote $j_1 = f_{AU}(s_1), j_2 = f_{AU}(s_2)$ the number of best effort calls, and $i_1 = f_{CV}(s_1), i_2 = f_{CV}(s_2)$ the number of guaranteed service calls in the system when it is in states of indexes s_1, s_2 respectively. Further, let q_1, q_2, r_1, r_2 denote the queue length and service rate in states s_1 and s_2 respectively. Note that because of the above restriction of no squeeze out of best effort calls by guaranteed service calls, a transition from a state with i_1 guaranteed service calls to another state with i_2 guaranteed service calls is only allowed if that transition does not imply a queue length increase, i.e. it is only allowed if $q_1 = q_2$ in these two states.

Thus the generator matrix G has the following form:

$$G(s_1, s_2) = \begin{cases} \lambda_{CV} & \text{if } i_2 = i_1 + 1, j_2 = j_1 \\ & \text{ and } q_2 = q_1; \\ \lambda_{AU} & \text{if } i_2 = i_1, j_2 = j_2 + 1; \\ i_1 * \mu_{CV} & \text{if } i_2 = i_1 - 1, j_2 = j_1; \\ r_{s_1} * j_1 * \mu_{AU} & \text{if } i_2 = i_1, j_2 = j_1 - 1; \\ D_{s_1} & \text{if } i_2 = i_1, j_2 = j_1; \\ 0 & \text{otherwise} \end{cases}$$

The matrix diagonal D is determined such that the sum of the elements in each row is 0.

2.3 State Space Example

Consider the partially blocking-queueing system in Figure 1a.

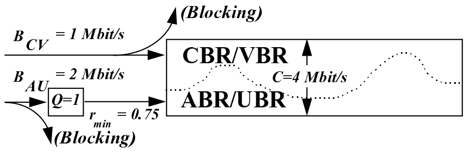

Fig.1a: The ATM link with two types of arrival streams

For illustration purposes we consider an ATM transmission link of capacity $C = 4$ Mbit/s, with two types of calls as described above, with bandwidth demands $B_{CV} = B_1 = 1$ Mbit/s and $B_{AU} = B_2 = 2$ Mbit/s. The second type calls accept partial blocking with minimal service rate $r_{min} = 0.75$. Additionally, these calls are allowed to wait in a finite queue of capacity $Q = 1$. Let the arrival rates be $\lambda_{CV} = \lambda_1 = 1$ and $\lambda_{AU} = \lambda_2 = 1$ $1/s$ and let us assume equal mean holding times $1s$ for both classes.

The state transition diagram of this system is depicted by Figure 1b. According to Figure 1b the system can be in one of 15 different $(feasible)$ states, where each state is uniquely characterized by the tuple (n_1, n_2), where n_1 and n_2 are the number of the guaranteed service and best effort calls respectively. As described above, from this state descriptor tuple the associated r service rate and q queue length may easily (and uniquely) be derived. Indeed, in e.g. state $(n_1 = 1, n_2 = 2)$ both best effort calls receive service with service rate $r = r_{min} = 0.75$ occupying $n_2 * r * B_2 = 3$ Mbit/s link capacity leaving 1 Mbit/s capacity for the guaranteed service call. Note that in e.q. state $(n_1 = 1, n_2 = 3)$ there is one best effort call waiting in the queue and two receiving service (with rate $r = 0.75$).

Regarding the transitions between "neighbouring" states, i.e. states where only one of the state descriptors differ with only 1 (the other being equal), the lack of two kinds of transitions are noteworthy. First, guaranteed service call arrivals are never allowed to increase the best effort calls queue length, q, since no preemption of the best effort calls are allowed. (Such a policy is analyzed and compared to the partially blocking scheme in [49].) Indeed, there is no transition from state $(n_1 = 2, n_2 = 1)$ to state $(n_1 = 3, n_2 = 1)$.

Secondly, (and obviously): from a state, where all best effort calls in the system are waiting in the queue not receiving service, there cannot be a state transition "downwards", i.e. to a state with one less best effort call. (Consider e.g. state $(n_1 = 3, n_2 = 1)$ with $q = n_2 = 1$.)

Finally, note that if the system is in a state where $q = Q$, further arriving best effort calls will be blocked, and thus we refer to these states as best effort *blocking states*. In the figure both the guaranteed service and the best effort blocking states are shown.

2.4 Obtaining Blocking Probabilities and the Mean Time in System for Best Effort Calls

Once the steady state distribution $\pi(s)$ has been found, we can obtain the guaranteed service class and the best effort class blocking probabilities (P_{CV} and P_{AU}) by identfying the indexes of the guaranteed service and best effort blocking states.

When the system is in state of index s, there are $j = f_{AU}(s)$ best effort and $i = f_{CV}(s)$ guaranteed service calls in the system. A state s is clearly a guaranteed service blocking state if $i = S_{TOT}(j) - 1, j = 0, ..., N_{AU}^{max}$. Additionally, because guaranteed service calls may not squeeze out in-service best effort calls, a state (i, j) is also a guaranteed service blocking state if $q_{i+1,j} = q_{i,j} + 1$. A state (i, j) is a best effort blocking state if the following inequality holds:

$$C - i * B_{CV} \leq r_{min} * (j + 1 - q) * B_{AU} \quad (8)$$

Then let $S_{CV,Bl}$ and $S_{AU,Bl}$ be the sets of guaranteed service and best effort blocking states respectively. The blocking probabilities are then given by $P_{CV} = \sum_{s \in S_{CV,Bl}} \pi(s)$ and $P_{AU} = \sum_{s \in S_{AU,Bl}} \pi(s)$.

The *distribution* of the time spent in the system for the best effort class can be obtained as shown in the next section. However, the determination of the *mean* time spent in the system is due to Little's famous result much easier.

Let S_q denote the set of states where the queue size is q. Then from the steady state distribution π we can easily calculate the mean queue length. It is

$$q_{MEAN} = \sum_{q=0}^{N_{AU}^{max}} (q * \sum_{s \in S_q} \pi(s)) \quad (9)$$

From Little's equation the mean waiting time of the best effort calls is therefore $q_{MEAN}/(\lambda_{AU} * (1 - P_{AU}))$. Similarly, let T_w denote the set of states where the number of best effort calls in the system is w. Then the mean number of best effort calls in the system is

$$W_{MEAN} = \sum_{w=0}^{N_{AU}^{max}} (w * \sum_{s \in T_w} \pi(s)) \quad (10)$$

And from Little's equation the mean time a best effort call spends in the system is $W_{MEAN}/(\lambda_{AU} * (1 - P_{AU}))$.

3 Customer Tagging and State Space Expansion

The method we follow here is based on (1) *tagging* a best effort call arriving to the system, which can, as we have seen, be in one of the feasible states; and (2) carefully examining the possible transitions from the moment this tagged call enters the system until it acquires the required service and leaves the system. Finally, unconditioning on all possible entrance state probabilities, and applying results from [17], the moments of the best effort service time can be determined.

Consider the *expanded state space* of the system depicted of Subsection 2.3. Figure 1c shows the state transition diagram *from the tagged call's point of view* of the same system an *infinitesimal amount of time after* the tagged call entered the system. Since we assume that at least the tagged tagged best effort call is now in the system we exclude states where $n_2 = 0$. Figure 1c also shows the entrance probabilities, with which the tagged call will find the system in *that* state. Thus, in Figure 1c, the tagged arriving best effort call will find the system in state (n_1, n_2) with probability $p^{(n_1,n_2)}$, and will bring the system into state (n_1, n_2+1) unless (n_1, n_2) happened to be a best effort blocking state. Note that $p^{(n_1,n_2)}$ gives the non-zero elements of the initial probability vector $P_R(0)$ of Section 4.

Since we are now focusing on our tagged call, Figure 1c also shows that we have to introduce a third state descriptor in order to uniquely characterize the state of the system. The third state descriptor specifies the position of the tagged call in the queue (with the understanding that queue position 0 means that the call is in service.) We also define a *trapping* (*absorbing*) *state* [17], which corresponds to the state where the tagged call has acquired the requested amount of service (Figure 1c) and leaves the system. In this expanded state space the time until absorbtion [17] corresponds to the time the tagged call spends in the system. Indexing the new state space in a similar manner as we did with the original state space in Section 2 the new generator matrix, G_E, will have the following structure:

$$G_E = \begin{bmatrix} B & T \\ 0 & w \end{bmatrix} \quad (11)$$

where the B matrix represents the transitions between the non-trapping states, the T vector contains the transitions *to* the trapping state, the 0 vector indicates that no transitions are allowed *from* the trapping state, and $w = 0$. Once the structure of the expanded state space and the associated transition rates together with the initial probability vector, $P_R(0)$, are determined, we can apply the result of [6, 20, 27] for the determination of the rth moment of T_x:

$$\{T_x\}^{(r)} = r! * P_R^T(0) * (-B)^{-r} * e \quad (12)$$

and specifically for the mean:

$$E\{T_x\} = P_R^T(0) * (-B)^{-1} * e \quad (13)$$

In order to determine the generator matrix G_E and the corresponding B matrix of the new state space, we first need to determine the number of states in the expanded state space. From Equation (3) by omitting the states from the original state space where $n_2 = 0$, we have:

$$S_E = S - S_{TOT}(0) + \sum_{k=S_{TOT}(0)}^{S-1} CH(k) \qquad (14)$$

where $CH(k)$ stands for the number of *children* states of state k in the original state space. The children states of a given state in the original state space are those that have the same n_1 and n_2 state variables. Of course, all states in the original state space, except for the ones where no best effort calls are present, i.e. $n_2 = 0$ will have at least 1 child state in the expanded space. However, the states where $n_2 \geq 2$ and the corresponding queue length $q \geq 1$ will have the following number of children states (all with different n_3 state variable value):

$$CH(k) = \begin{cases} q & \text{if } n_2 = q \\ q+1 & \text{otherwise} \end{cases} \qquad (15)$$

This equation expresses that the number of children states must be equal to the number of possible queue positions of the tagged call ($n_3 = 1..q$) plus 1 for the state when the tagged call is in service ($n_3 = 0$), unless in the original state all best effort calls are in the queue ($n_2 = q$) and none in service, in which case n_3 cannot be 0.

It is important to note that because of the FIFO policy, neither a guaranteed service nor a best effort call arrival can influence the position of the tagged call in the queue. However, it is possible that either a guaranteed service or a best effort call *departure* affects the tagged queue position, if the tagged call can either enter service or advance in the queue due to link capacity increase after the departure. Let $n_3(s_i)$ and $q(s_i)$ denote the tagged queue position and the queue length in state s_i respectively. Then n_3 after a departure will either decrease with 1 or remain the same:

$$n_3(s_2) = \begin{cases} n_3(s_1) - 1 & \text{if } q(s_2) < q(s_1) \\ & \quad \text{and } n_3(s_1) > 0 \\ n_3(s_1) & \text{otherwise} \end{cases} \qquad (16)$$

Similarly, an arriving best effort call may join the queue only at the last position. Indeed, in Figure 1c if there are several states with identical n_1 and n_2, but different n_3 values (i.e. if there are several children states of the same original state (n_1, n_2)) then only the one with highest n_3 queue position can be entered by an arriving best effort call. This latter observation is essential in determining the $P_R(0)$ initial state probability vector. Because of the Poisson arrival process assumption, the non-zero elements of this vector are clearly the steady state probabilities of the best effort non-blocking states in the original state space.

In Figure 1c, for instance, if an arriving best effort call cannot enter service, but it finds the system at the time of arrival with at least one free queue place ($q < Q$ like in state $(n_1 = 2, n_2 = 1)$, then it will join the queue setting $n_3 = q{+}1$ and incrementing n_2, otherwise (when $q = Q$ like e.g. in state $(n_1 = 2, n_2 = 2)$) it will be blocked. However, *after* an arrival there is no way for the system to be in state $(n_1 = 2, n_2 = 2, n_3 = 0)$ (see Figure 1c) and therefore the corresponding value in the initial probability vector $P_R(0)$ has to be 0. In our example $P_R(0)$ can be easily derived from Figure 1c by normalizing with the best effort class blocking probability:

$$P_R(0) = \{p^{(0,0)}; p^{(1,0)}; p^{(2,0)}; p^{(3,0)}; p^{(4,0)}; \\ p^{(0,1)}; p^{(1,1)}; 0; p^{(2,1)}; 0; p^{(0,2)}; 0; p^{(1,2)}\}/(1 - P_{AU})$$

As an illustrative example let us now follow an arriving best effort call from the moment it enters the system until it is served and leaves it. Suppose that this tagged call finds the system in state $(n_1 = 1, n_2 = 2, q = 0)$; this will happen with probability $p^{(1,2)}$ on Figure 1c. In this state no further service rate r decrease is allowed, but there is room in the queue, so the tagged call will join it bringing the system into state $(n_1 = 1, n_2 = 3, n_3 = 1)$.

Suppose now that a guaranteed service call gets served and the system moves into state $(n_1 = 0, n_2 = 3, n_3 = 1)$, indicating that the tagged call still has to wait in the queue. Since this state is a best effort blocking state, only a guaranteed service call arrival, or a best effort departure can happen. Assuming this latter event, the system moves to state $(n_1 = 0, n_2 = 2, n_3 = 0)$, that is the tagged call has now entered service. The tagged and untagged best effort calls are now in fact receiving service with equal μ_{AU} rate, and therefore we need to distinguish between the tagged departure (into the trapping state) and the untagged departure (into state $(n_1 = 0, n_2 = 1, n_3 = 0)$).

Suppose this latter happens. Let us also assume that the tagged call gets served before any kind of arrival, then finally from state $(n_1 = 0, n_2 = 1, n_3 = 0)$ the system enters the trapping state. Note that no transitions are possible to the trapping state from e.g. the states $(n_1 = 3, n_2 = 1, n_3 = q = 1)$ and $(n_1 = 4, n_2 = 1, n_3 = q = 1)$, since the tagged call does not receive service.

With the above considerations it is possible to derive the G_E and B matrices as well as the $P_R(0)$ vector from the original state space and the G generator matrix, even though it is quite complicated to specify it symbolically. Using the state indexes of Figure 1c the structure of G_E matrix in the above example is shown below:

$$\begin{bmatrix}
D & a & & & & A & & & & & & & & U \\
u & D & a & & & & A & & & & & & & U \\
 & 2u & D & & & & & A & & & & & & U \\
 & & 3u & D & a & & & & & & & & & \\
 & & & 4u & D & & & & & & & & & \\
U & & & & & D & a & & & A & & & & U \\
 & rU & & & & u & D & & & & & A & & rU \\
 & & & & & & 2u & D & & & & & & U \\
 & & U & & & & 2u & & D & & & & & \\
 & & & & & U & & & & D & & a & & U \\
 & & & & & 2U & & & & & D & & a & \\
 & & & & & & rU & & & u & & D & & rU \\
 & & & & & & 2rU & & & & u & & D & \\
 & & & & & & & & & & & & & D
\end{bmatrix}$$

where $a = \lambda_{CV}$, $A = \lambda_{AU}$, $u = \mu_{CV}$, $U = \mu_{AU}$ and D is detemined such that the rows sum up to 0. Note that the last column lists the transition rates to the trapping state from the different states. For instance, from state 6 (i.e. $(n_1 = 1, n_2 = 1)$) (Figure 1c) the tagged call will enter the trapping state with rate rU, since it is a partially blocking state, as indicated in Figure 1c. Note that in this particular example the service rate in all partially blocking states are equal ($r = 0.75$).

We have developed a C program to generate numerically these matrices and this vector. With respect to the mean, the next Section provides an efficient method to check the correctness of the above reasoning by applying Little's theorem. Additionally, the numerical results presented herein have been confirmed by simulation, too.

4 The Partially Blocking Loss System

When best effort calls are not allowed to wait in a queue if insufficient bandwidth are available at time of arrival, the Markov model simplifies in two ways. First, the size of the state space becomes smaller since $Q = 0$. Second and more important, the computation of the time spent in the system simplifies because with zero queue the need for children states in the extended state diagram which models the behaviour of a tagged best effort call disappears.

4.1 The Time in System Conditioning on Service Requirement x

By removing the absorbing state and all transitions to it we obtain an irreducible Markov Chain with a generator matrix which we denote M. Assuming that a best effort call has just arrived and conditioning that its service requirement is x, the Laplace transform of the time this best effort call will spend in the system can be found by applying the technique of Markov driven workload processes. The computation is detailed in the Appendix and the Laplace transform of the time a best effort call spends in the system conditioning that its service requirement is x is:

$$P_R(0)s^*(x,s)e = P_R(0)\exp[R^{-1}(M - sI)x][I - M/s]^{-1}e \tag{17}$$

where M is the generator of reduced irreducible Markov process given above, R is a diagonal matrix where enty k, k gives the service rate available for a best effort call in state k, and $P_R(0)$ is the probability vector given the probabilities by which a best effort call enters the system. Finally, $s^*(x,s)$ is the matrix of Laplace Transforms of the time spent where in entry (i,j) entrance to the system is in state i and departure from the system is in state j.

4.2 The Unconditional Time to Completion when Service Requirement is Exponential

Conditioning that the required inital workload is x we have from (30) when summing over all final states

$$s^*(x,s)e = \exp[R^{-1}(M - sI)x]e \tag{18}$$

Assuming that the initial service requirement is exponentially distributed with parameter μ then unconditioning x yields

$$\begin{aligned}
s^*(s)e &= \int_0^\infty s(x,s)e\mu e^{-\mu x}dx \\
&= \int_0^\infty \exp[R^{-1}(M - sI)x]\mu e^{-\mu x}dxe \\
&= \int_0^\infty \exp[-R^{-1}(sI - (M - \mu R))x]dx\mu e
\end{aligned}$$

The integration yields

$$s^*(s)e = [sI - (M - \mu R)]^{-1}R\mu e \tag{19}$$

Let $P_R(0)$ denote the initial probabilities in which the Markov chain is started. Then the density function T_{exp} for the time until completion of an exponentially distributed workload has transform

$$P_R^T(0)s^*(s)e = P_R^T(0)[sI - (M - \mu R)]^{-1}R\mu e \tag{20}$$

which is seen to correspond to a *phase type distribution* with initial probability vector $P_R(0)$, transient matrix $M - \mu R$ and vector $R\mu$ of rates to the absorbing state. This result is in accordance with theorem 3 in [4].

5 Numerical Results

5.1 Numerical Solution Approach

We have employed a sparse implementation of a direct matrix method called the GTH algorithm [14] (named after the authors Grassman, Taksar and Heyman).

The GTH algorithm is a specific variant of the Gaussian elimination for the calculation of the steady-state probability vector of a Markov chain. This algorithm makes the calculation of the steady-state probability vector of a Markov chain numerically stable. The complexity of the algorithm is of the same order as the standard Gaussian elimination and the GTH algorithm consists of only minor modifications of the standard Gaussian elimination procedure without pivoting and it makes the elimination procedure cancellation free i.e. no subtractions are performed. In [14] and [16] numerical evidence was given that the cancellation free scheme did in fact facilitate the accurate computation of steady-state probability vectors of large Markov chains (10^6 states).

Recently in [29] it has been shown formally that the algorithm is stable and that the algorithm computes each component in the steady-state vector with low relative error. The latter is clearly of extreme importance when considering large Markov chains and when the individual elements in the steady-state probability vector differ by many orders of magnitude. The analysis in [29] indicates that a careful sparse implementation of the GTH algorithm for the calculation of the steady-state probability vector of a finite state QBD matrix could be done accurately even for as much as $10^5 - 10^6$ states.

5.2 Results

In this section we consider a single link of capacity C=60 Mbit/s, which is offered calls according to a Poisson process and with exponential holding time belonging to two different service classes. To obtain interesting numerical results and emphasizing the role of call queueing and partial blocking we assume that (guaranteed service) CBR/VBR calls are narrow band, while best effort calls are wide band. Specifically, any guaranteed service class calls have a bandwidth demand of $B_{CV} = 1$ Mbit/s and mean holding time $1/\mu_N = 1$ s. Guaranteed service class calls do not accept partial blocking, i.e. they are either given the required bandwidth B_{CV} or blocked and lost. Wide band calls are of the best effort (ABR/UBR) type characterized by the (maximal) bandwidth demand $B_{AU} = 10$ Mbit/s and mean holding time $1/\mu_W = 1$ s, and, as discussed above, by the minimal accepted service rate $r_{min} < 1$. Best effort calls do accept partial blocking, i.e. they are admitted into the system if, at the time of arrival the available bandwidth is at least $r_{min} * B_{AU}$. In the examples below we assume that all in-service best effort calls receive the same instantenous service rate $r(t) = max[\frac{C - n_{N(t)} * B_{CV}}{n_{W(t)} * B_{AU}}, 1] > r_{min}$, where $n_N(t)$ and $n_W(t)$ denote the number of (narrow band) guaranteed service and (wide band) best effort calls in the system at time t.

Figure 2 shows the performance measures of this partially blocking (PB) system (where we let $\lambda_{CV} = 10 * \lambda_{AU} = 30$ 1/s). As r_{min} decreases form 1.0 to 0.4, ABR (wide band, WB) class blocking also decreases from 40% to 13% ! Additionally, CBR/VBR (narrow band, NB) class blocking decreases, too, even though this decrease is not so significant. This performance increase in blocking probabilities is, of course, at the expense of the best effort class calls increased time spent in the system. This time increase is less than 20% at $r_{min} = 0.6$, but reaches almost 60% at $r_{min} = 0.4$. To assess the performance of the system we define the overall performance measure in the spirit of [25], as follows:

$$Perf = \frac{1 - P_N - P_W}{1 + \Delta T} \qquad (21)$$

where ΔT stands for the mean additional time spent in the system as compared to the mean time spent in the system if no partial blocking or queueing were allowed (i.e. the "original" mean holding time of the best effort calls, $1/\mu_{AU}$). This performance measure takes into account the tradeoff between blocking probabilities and best effort call time spent in the system. It is maximal around $r_{min} = 0.7$, indicating that choosing a smaller value for r_{min} results in a relatively great increase in service time for the best effort calls, and it "doesn't pay off" in terms of blocking probability decrease. Another popular extension of the Erlang Loss Model has been the so called mixed delay and loss (MDL) systems [12, 2, 28]. Even though it is fundamentally different from the PB model, its performance measures are comparable to ours. This is because mixed delay and loss systems also attempt to decrease blocking probability at the expense of increased time spent in the system, i.e. in the queue and in service.

This motivates the comparison of the performance measures of a PB and an MDL system. Figure 3 shows the perfomance measures of a system with the same system parameters as of Figure 2. Here, instead of partially blocking best effort calls, they are placed in a finite queue (the size of which varies form 0 to 6) in case of insufficient bandwidth at the time of arrival. As the queue length increases, the wide band class blocking decreases, as expected, from 40% to 9% - roughly the same decrease in the blocking probability as in the PB system. Note, however, that the blocking of the narrow band class here increases to 18% ! This explains why the combined performance measure (Perf) of the PB system is strictly superior to that of the MDL system, which is an important advantage of PB systems.

To combine the advantages of MDL and PB systems, we now consider a system where both queueing and partial blocking are allowed for the wide band best effort calls. Here we consider a link of capacity $C = 30$ Mbit/s. Narrow band calls require $B_{CV} = 1$ Mbit/s bandwidth, wide band calls require $B_{AU} = 12$ Mbit/s (case I) or 6 Mbit/s peak bandwidth (case II). In case I the total offered load $B_{CV} * \lambda_N * (1/\mu_N) + B_{AU} * \lambda_W * (1/\mu_W)$ is 16 Mbit/s*Erlang, in case II it is 30 Mbit/s*Erlang. Figures 4 and 5 show the different performance measures when r_{min} decreases from 1.0 to 0.5. The behaviour of the system is investigated in six subcases as the maximal queue length (buffer size), Q for wide band calls changes from 0 up to 5.

Figures 4a and 5a show the wide band class blocking probabilities in these two cases (I and II). The wide band class blocking probability drastically decreases as r_{min} decreases when there is call queueing, or when the queue size is small, $Q = 1$ or $Q = 2$. Providing for a single queue place and accepting 50% partial blocking in case II., for instance, decreases blocking from 42% under 10%. Further increase of the queue capacity, or, when the buffer space is kept at 2 or more, further decrease of r_{min} doesn't have significant impact on wide band blocking.

Figures 4b and 5b depicts the narrow band class blocking probabilities. Naturally, wide band call queueing causes an increase in narrow band blocking, but this increase can be compensated somewhat by permitting narrow band calls to "squeeze" the in-service wide band calls. In case II., for instance, providing a single queue place for the wide band calls, increases narrow band blocking from 6 to 14 %, but as r_{min} decreases to 0.5, blocking decreases to 11%.

It is interesting how the time a wide band call spends in the system depends on r_{min} and Q, as seen in Figures 4c and 5c. Clearly, the longer the queue, the longer the mean queueing time (and smaller their blocking probability) becomes. Choosing r_{min} is a clear trade off between (1) how long a call has to wait in the queue (the smaller r_{min} becomes, the faster wide band calls get into service, becuase they accept smaller bandwidth, and (2) how "fast" service they get (the greater r_{min} is, the smaller the in-service time becomes).

The overall performance measure as defined by (21) is shown in Fig. 4.d and 5.d. According to this performance measure we conclude that very short queues ($Q = 0, 1, 2$) combined with moderate minimal service rates $r_{min} = 0.7, 0.8$ give satisfactory performance.

6 Conclusions

We have investigated a mixed queueing and loss system where calls with guaranteed service paramters (CBR/VBR) and best effort (ABR/UBR) calls require service. Assuming that wide band calls subscribe for best effort service, we model these calls as ones which tolerate *partial blocking* of their required peak bandwidth. Further, these calls also accept non-zero connection setup time modelled as the waiting time in a finite capacity queue. Narrow band calls have been assumed to be of CBR/VBR sources. These calls themselves do not accept partial blocking, but they are allowed to decrease the given bandwidth for best effort service users. With a Markov analysis we have found that in terms of the most important performance measures this system performs better than systems without call queueing or partial blocking. Furtheremore, it has been shown that in our model the time spent in the system by the best effort (i.e. ABR/UBR) calls is a phase type distributed random variable. Future works include the investigation of optimal call admission procedures in the mixed best effort - QoS guaranteed environment [49] on the link level and the investigation of optimal routing strategies on the network level [48].

Appendix: Time Spent in System: Approach Based on Markov Driven Workload Processes

Let M denote the infinitesimal generator of the Continuous Time Markov Chain (CTMC) X_t and let the steady state distribution, π fulfil: $\pi e = 1$ and $\pi M = 0$, where e denotes the vector with all unit elements: $e = \{1..1\}$. Furthermore let R be a diagonal matrix in which diagonal element r_k denotes the rate at which fluid is emitted (in our application the rate at which service is accomplished) when the process is in state k.

If W_t denotes the total amount of accomplished service at time t, then $T_x = inf\{t|W_t > x\}$ will be the time it takes for the Markov process to accomphish a total service requirement of x. Then the events $\{W_t \leq x\}$ and $\{T_x \leq t\}$ are mutually exclusive and their union gives the event of certainty which implies:

$$Pr\{W_t \leq x\} + Pr\{T_x \leq t\} = 1$$

and

$$Pr\{W_t \leq x, X_t = j\} + Pr\{T_x \leq t, X_t = j\} = \pi_j$$

that is

$$P_{ij}(x,t) + S_{ij}(x,t) = \pi_{ij}(t) \tag{22}$$

where we let

$$\begin{aligned} P_{ij}(x,t) &= Pr\{W_t \leq x, X_t = j | X_0 = i\} \\ S_{ij}(x,t) &= Pr\{T_x \leq t, X_t = j | X_0 = i\} \end{aligned} \tag{23}$$

and

$$\pi_{ij}(t) = Pr\{X_t = j | X_0 = i\}$$

Now, using the well known connection between M and $\Pi(t) = [\pi_{ij}(t)]$:(see e.g. [9]) $\Pi(t) = exp[Mt]$, and using the matrix notation $P(x,t) = [P_{ij}(t)]$ and $S(x,t) = [S_{ij}(t)]$ we obtain:

$$P(x,t) + S(x,t) = exp[M(t)] \tag{24}$$

Next considering (23), and making use of the exponential state sojourn times in a CTMC, and applying arguments from [1], we'll get a differential equation, which describes the system dynamics.

Derivation of Transform of Distributions

From an argument analogue to the argument pp. 1875 in [1] we get

$$\frac{\partial P}{\partial t}(x,t) + \frac{\partial P}{\partial x}(x,t)R = P(x,t)M \tag{25}$$

where $R = diag(r_1, ..., r_n)$

Multiplication with $\exp(-zx)\exp(-st)$ and integration over t and x on the positive line gives after a few algebraic manipulations

$$P^{**}(z,s) = \frac{1}{z}[sI + zR - M]^{-1} \tag{26}$$

where $F^{**}(z,s) = \int_0^\infty \int_0^\infty \exp(-zx)\exp(-st)F(x,t)dtdx$ is the double Laplace transform.

Inversion in the s-parameter immediately yields $P^*(z,t) = \frac{1}{z}\exp[(M - zR)t]$ and the Laplace Transform of the density function for the workload at time t is

$$p^*(z,t) = \exp[(M - zR)t] \qquad (27)$$

From this equation we can easily get the mean acquired service (W_t) generated at time t. It is

$$m(t) = -\pi \left.\frac{\partial p^*(z,t)}{\partial z}\right|_{z=o} e$$

where, just as before, $e = \{1..1\}$ is the n-dimensional vector of 1's.

Since $\exp[(M - zR)t] = \sum_{n=0}^{\infty} \frac{t^n}{n!}(M - zR)^n$ we get

$$\frac{\partial \exp[(M - zR)t]}{\partial z} = \sum_{n=1}^{\infty} \frac{t^n}{n!} \sum_{j=0}^{n-1} (M-zR)^j R (M-zR)^{n-1-j}$$

Evaluation of the derivative in $z = 0$ and the fact that $\pi M = 0$ leads to

$$m(t) = t\pi Re = t\sum_{i=0}^{n} \pi_i r_i \qquad (28)$$

just as we would expect! Combining (22) and (26) gives

$$S^{**}(z,s) = [zI + R^{-1}(sI - M)]^{-1}[sI - M]^{-1} \qquad (29)$$

An inversion in z gives $S^*(x,s) = \exp[R^{-1}(M - sI)x][sI - M]^{-1}$ yielding the following Transform for the density function of T_x

$$s^*(x,s) = \exp[R^{-1}(M - sI)x][I - M/s]^{-1} \qquad (30)$$

Since $[I - M/s]^{-1}e = e$ because $e = (I - M/s)e$ we get $s^*(x,0) = \sum_{n=0}^{\infty} \frac{x^n}{n!}(R^{-1}M)^n$ implying $s^*(x,0)e = e$ showing that $ps^*(x,0)e = 1$ for any probability vector p since $Me = 0$.

Furthermore, and more interestingly

$$\begin{aligned} -\frac{\partial s^*}{\partial s}(x,s) &= \exp[R^{-1}(M - sI)x][I - M/s]^{-2}M/s^2 \\ &+ \sum_{n=1}^{\infty} \frac{x^n}{n!} \sum_{j=0}^{n-1} [R^{-1}(M - sI)]^j R^{-1} \\ &\quad [R^{-1}(M - sI)]^{n-1-j}[I - M/s]^{-1} \end{aligned}$$

Again, since $Me = 0$ we get

$$-\frac{\partial s^*}{\partial s}(x,s)e = \sum_{n=1}^{\infty} \frac{x^n}{n!}[R^{-1}M]^{n-1}R^{-1}e \qquad (31)$$

At an arbitrary point in time the distribution of the underlying Markov process is π and an arrival (infinitesimal amount of fluid arrival) will see a probability a_i for being in state i where $a_i = \frac{\pi_i r_i}{\sum_1^n \pi_j r_j}$. Written in vector notation we get $a = \frac{\pi R}{\pi Re}$.

From these considerations we finally get that the mean time until worklevel x is reached seen from an arbitrary arrival is

$$\begin{aligned} E\{T_x\} &= a(-\frac{\partial s^*}{\partial s}(x,0))e = \frac{\pi R}{\pi Re} \sum_{n=1}^{\infty} \frac{x^n}{n!}[R^{-1}M]^{n-1}R^{-1}e \\ &= x\frac{\pi e}{\pi Re} = \frac{x}{\pi Re} \end{aligned} \qquad (32)$$

just as could be expected !

For the variance the computation is a bit more complicated and the final formula unfortunately also.

Put $F_x(s) = \exp[R^{-1}(M - sI)x]$ and $G(s) = [I - M/s]^{-1}$. Then $s^*(x,s) = F_x(s)G(s)$ and

$$\frac{\partial^2 s^*}{\partial s^2}(x,s) = F_x''(s)G(s) + 2F_x'(s)G'(s) + F_x(s)G''(s)$$

Therefore $\frac{\partial^2 s^*}{\partial s^2}(x,0)e = F_x''(s)G(s)e = F_x''(s)e$ since $G(0)e = e$ and since $G'(0)e = G''(0)e = 0$ because $Me = 0$. The second derivative of the $\exp[R^{-1}(M - sI)x]$ gives

$$\begin{aligned} F_x''(s) &= \sum_{n=2}^{\infty} \frac{x^n}{n!} \sum_{j=1}^{n-1} \sum_{i=0}^{j-1} [R^{-1}(M - sI)]^i (R^{-1}) \\ &\quad [R^{-1}(M - sI)]^{j-1-i}(R^{-1})[R^{-1}(M - sI)]^{n-1-j} \\ &+ \sum_{n=2}^{\infty} \frac{x^n}{n!} \sum_{j=0}^{n-1} \sum_{i=0}^{n-2-j} [R^{-1}(M - sI)]^j (R^{-1}) \\ &\quad [R^{-1}(M - sI)]^i (R^{-1})[R^{-1}(M - sI)]^{n-2-j-i} \end{aligned}$$

From this we get after some algebraic manipulations

$$\begin{aligned} F_x''(s)e &= 2\sum_{n=2}^{\infty} \frac{x^n}{n!} \sum_{i=0}^{n-2} [R^{-1}(M - sI)]^i (R^{-1}) \\ &\quad [R^{-1}(M - sI)]^{n-2-i}R^{-1}e \end{aligned}$$

Recalling that $a = \frac{\pi R}{\pi Re}$ we get

$$\begin{aligned} E\{T_x^2\} &= aF_x''(0)e \\ &= \frac{2}{\pi Re} \sum_{n=2}^{\infty} \frac{x^n}{n!} \pi [R^{-1}M]^{n-2}R^{-1}e \end{aligned}$$

Applying that $R^{-1}M[R^{-1}M - ea]^2 = [R^{-1}M]^3$ and the non-singularity of $[R^{-1}M - ea]$ see e.g. p. 238 in [26] gives

$$\begin{aligned} E\{T_x^2\} &= \frac{2\pi}{\pi Re}(\frac{x^2}{2}I + \\ &\quad \sum_{n=3}^{\infty} \frac{x^n}{n!}[R^{-1}M]^n[R^{-1}M - ea]^{-2})R^{-1}e \end{aligned}$$

$$
\begin{aligned}
= \ & \frac{2\pi}{\pi Re}(\frac{x^2}{2}I + exp[R^{-1}Mx][R^{-1}M - ea]^{-2} \\
& -[R^{-1}M - ea]^{-2} - xR^{-1}M[R^{-1}M - ea]^{-2} \\
& -\frac{x^2}{2}[R^{-1}M]^2[R^{-1}M - ea]^{-2})R^{-1}e
\end{aligned}
$$

Noting that $ea[R^{-1}M - ea] = -ea$ it is not difficult to show that $R^{-1}M[R^{-1}M - ea]^{-1} = I - ea$. This gives the following simplifications

$$
\begin{aligned}
E\{T_x^2\} = \ & \frac{2\pi}{\pi Re}(\frac{x^2}{2}I + exp[R^{-1}Mx][R^{-1}M - ea]^{-2} \\
& -[R^{-1}M - ea]^{-2} - x(I - ea)[R^{-1}M - ea]^{-1} \\
& -\frac{x^2}{2}(I - ea))R^{-1}e \\
= \ & \frac{2\pi}{\pi Re}(exp[R^{-1}Mx] - I)[R^{-1}M - ea]^{-2}R^{-1}e \\
& +(\frac{x}{\pi Re})^2 - x\frac{2\pi}{\pi Re}[R^{-1}M - ea]^{-1}R^{-1}e \\
& -x\frac{2}{(\pi Re)^2}
\end{aligned}
$$

From this it is clear that

$$
\begin{aligned}
Var\{T_x\} = \ & \frac{2\pi}{\pi Re}(exp[R^{-1}Mx] - I)[R^{-1}M - ea]^{-2}R^{-1}e \\
& -x\frac{2}{\pi Re} \cdot \\
& (\pi[R^{-1}M - ea]^{-1}R^{-1}e + \frac{1}{\pi Re})
\end{aligned} \tag{33}
$$

Acknowledgements

Gábor Fodor is grateful to Miklós Boda at Ericsson Radio Systems, Sweden, and to Tamás Henk at the Technical University of Budapest, Hungary, for the support which made this study possible. Søren Blaabjerg would like to acknowledge the financial support from TeleDanmark. Allan T. Andersen acknowledges the financial support of the Technical Research Council of Denmark under grant no. (9500947). Miklós Telek acknowledges the financial support of OTKA under grant no. (T-16637).

References

[1] D. Anick, D. Mitra and M.M. Sondhi, "Stochastic Theory of a Data-handling System with Multiple Sources", *The Bell System Technical Journal,* Vol. 61, 1871-1894, 1982.

[2] H. Akimaru, H. Kuribayashi and T. Inoue, "Approximate Evaluation for Mixed Delay and Loss Systems with Renewal and Poisson Inputs", *IEEE Transactions on Communications*, Vol. 36, No. 7, pp. 850-854, 1988.

[3] S. Blaabjerg and G. Fodor, "A Generalization of the Multirate Circuit Switched Loss Model to Model ABR Services in ATM Networks", in the *Proc. of the IEEE International Conference on Communication Systems, ICCS '96, Singpore*, Singapore, November, 1996.

[4] A. Bobbio, 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, K.W. Fendick, "The Rate Based Flow Control Framework for the Available Bit Rate Service", IEEE Networks, March 1995.

[6] J. A. Buzacott, "Markov Approach to Finding Failure Times of Repairable Systems", *IEEE Trans. on Reliability*, Vol. R19, Nov. 1970, pp. 152-156.

[7] 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.

[8] C-P. Chung, K.W. Ross, "Reduced Load Approximations for Multi-Rate Loss Networks", *ACM/IEEE Trans. on Networking*, 1993, pp. 1222- 1231.

[9] E. Cinlar, "Introduction to Stochastic Processes", *Prentice Hall, Englewood Cliffs*, 1975.

[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] R. Guérin, "Queueing Blocking System with Two Arrival Streams and Guard Channels", *IEEE Trans. on Comm.* 1988, pp. 153-163.

[12] R. Guérin, "Queueing-Blocking System with Two Arrival Streams and Guard Channels", *IEEE Transactions on Communications*, Vol. 36, No. 2, February 1988, pp. 153-163.

[13] L.A. Gimpelson, "Analysis of mixtures of wide- and narrow-band trafic", *IEEE Trans. on Commun. Technol.*, Sept 1965, pp. 258-266.

[14] Winfried K. Grassmann, Michael I. Taksar and Daniel P. Heyman, "Regenerative Analysis and Steady State Distributions for Markov Chains", *Operations Research*, 1985, vol. 33, No. 5 pp. 1107-1116.

[15] R. Guérin, Hamid Ahmadi and Mahmoud Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", *IEEE J-SAC*, Vol. 9, No. 7, pp. 968-981, Sept 1991.

[16] Daniel P. Heyman "Further Comparisons of Direct Methods for Computing Stationary Distributions of Markov Chains", *SIAM J. Alg. Disc. Math.* ,1987, vol. 8, No. 2 pp. 226-232.

[17] R.A. Howard, "Dynamic Probabilistic Systems and Markov Decision Processes" The MIT press 1960,

[18] J. Y. Hui, "Resource Allocation for Broadband Networks", *IEEE J-SAC*, vol. SAC-6, pp. 1598-1608, Dec. 1988.

[19] A. Hung and G. Kesidis, "Buffer Design for Wide-Area ATM Networks Using Virtual Finishing Times", *IEEE International Conference on Communications*, 1995.

[20] L. Jereb, "Reliability Analysis of Complex Markovian Systems", *Ph.D. Thesis, in Hungarian, Technical University of Budapest*, 1987.

[21] J.S. Kaufman, "Blocking in a Completely Shared Resource Environment with State Dependent Resource and Residency Requirements", *IEEE Infocom*, 1992.

[22] F.P. Kelly, "Routing in Circuit Switched Networks: Optimization, Shadow Prices and Decentralization", *Adv. Appl Prob.*, 1988.

[23] D.K. Kin, C.K. Un, "Performance analysis of bandwidth allocation strategy with state-dependent Bernoulli accesss and preemptive priority in wideband integrated networks", *Telecommunications Systems* 4, (1995), pp. 97-111.

[24] Leonard Kleinrock, "Queueing Systems, Vol. I:Theory", *Wiley*, ISBN 963 10 2725 2, 1975.

[25] Do Kyy Kim and Chong Kwan Un, "Performance Analysis of Bandwidth Allocation Strategy with State-Dependent Bernoulli Access and Preemptive Priority in Wideband Integrated Networks", *Telcommunications System Journal*, 4(1995)97-111, 1995.

[26] M. F. Neuts, “Structured Stochastic Matrices of M/G/1 Type and Their Applications”, volume 5 of *Probability: Pure and Applied Marcel Dekker, Inc*, 1989.

[27] B. F. Nielsen, "Modelling of multiple access systems with Phase Type Distributions", Ph.D. Thesis, Technical University of Denmark, IMSOR no. 49 1988.

[28] Z. Niu and H. Akimaru, "Studies of Mixed Delay and Nondelay Systems in ATM Networks", *International Teletraffic Congress, ITC-13*, Elsevier Science Publishers B. V., pp515-520, 1991.

[29] Colm Art O'Cinneide, "Entrywise perturbation theory and error analysis for Markov chains", *Numer. Math.*, 1993, vol. 65, pp. 109-120.

[30] A. K. Parekh, R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case", *IEEE/ACM Transactions on Networking*, 1, 344-357, 1993.

[31] A.K. Parekh, R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case", *IEEE/ACM Transactions on Networking*, 1, 137-150, 1994.

[32] Jing-Fei Ren, Jon. W. Mark and Johnny W. Wong, "A Dynamic Priority Queueing Approach to Traffic Regulation and Scheduling in B-ISDN", *IEEE GLOBECOM*, 1994.

[33] J. W. Roberts (ed), "Performance Evaluation and Design of Multiservice Networks", *Published by the Commission of the European Communities, Information Technologies and Sciences, COST 224 Final Report*, ISBN 92-826-3728-X, October, 1991.

[34] 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.

[35] J. W. Roberts, "Resource Sharing in B-ISDN Using Virtual Spacing Scheduling", *NETWORKS '94*, pp. 241-246, 1994.

[36] Keith W. Ross, "Multiservice Loss Models for Broadband Telecommunication Networks", *Springer Verlag London Limited*, ISBN 3-540-19918-7, 1995.

[37] Y.D. Serres, L.G. Mason, "A Multiserver Queue with Narrow- and Wide-Band Customers and Wide-Band Restricted Access", *IEEE Trans. on Comm.* Vol 36, 1988, pp. 675-684.

[38] Avril Smith, John Adams and Geoff Tagg, "Available Bit Rate - A New Service for ATM", *Computer Networks and ISDN Systems*, 28, 635-640, 1995.

[39] Ching-fong Su and Gustavo de Veciana, "On the Capacity of Multi-Service Networks", *IEEE International Conference on Communications*, 1995.

[40] E. D. Sykas, K. M. Vlakos, I. S. Venieris, E. N. Protonotarios, "Simulative Analysis of Optimal Resource Allocation and Routing in IBCN's", *IEEE J-SAC*, Vol. 9, No. 3, 1991.

[41] *ATM User Network Interface Specification Version 3.1*, September, 1994.

[42] ITU-T Recommendation I.356, *B-ISDN ATM Layer Cell Transfer Performance*, Draft, February 1996.

[43] ITU-T Recommendation I.371, *Traffic Control and Congestion Control in B-ISDN*, Draft, December 1995.

[44] ATM Forum, *ATM User Network Interface Specification Version 3.1*, September, 1994.

[45] ATM Forum, *Traffic Management Specification Version 4.0*, April, 1996.

[46] ATM Forum, *B-ISDN Inter Carrier Interface (B-ICI) Specification Version 1.0*, August, 1993.

[47] S. Blaabjerg, G. Fodor and A. T. Andersen, "Reducing Wide Band Blocking by Allowing Wide Band Calls to Queue", *COST Technical Report, available on request*, COST-TD(14)1996

[48] S. Blaabjerg, G. Fodor, A. Racz and K. Szarkowicz, "Simulative Analysis of Optimal Routing Strategies in ATM Networks Supporting ABR and UBR Services", accepted to the *IEE/IEEE* 5th *International Conference on Telecommunications, ICT '97*, Melbourne, Australia, April, 1997.

[49] E. Nordström, S. Blaabjerg and G. Fodor, "Call Admission Control of CBR/VBR and ABR/UBR Arrival Streams: A Markov Decision Approach", *submitted to the Northern Operational Research Conference, NORC '97*, August, 1997.