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

Allan T. Andersen
Nokia Telecommunications, Frederikskaj
DK-1790, Copenhagen V., Denmark, allan.andersen@ntc.nokia.com

Søren Blaabjerg
TeleDanmark, Telegade 2, DK-2630 Taastrup, Denmark
sbb@tele.dtu.dk, Fax: +45 43341672

Gábor Fodor
Mobile Network and Systems Research, Ericsson Radio Systems,
Torshamnsgatan 23, Kista, SE-164 80 Stockholm, Gabor.Fodor@era-t.ericsson.se
Fax: +46 8 4047020

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

Abstract

In this paper we consider an ATM transmission link, to which CBR or VBR and ABR or 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*, 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 proportionally longer time spent in the system, as first suggested in [3] and analyzed in details herein. 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 modeled by a Continuous Time Markov Chain (CTMC) and the CBR/VBR and ABR/UBR blocking probabilities and the mean ABR/UBR waiting- and service times 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) guarantees (such as the Constant Bit Rate (CBR) and the Variable Bit Rate (VBR) service categories), with services of limited or without such guarantees (such as the Available Bit Rate (ABR) and the Unspecified Bit Rate (UBR) service categories). *On the call level* ATM networks have traditionally been modeled as multi-rate circuit switched networks. This is possible by adopting the concept of *equivalent bandwidth*. Thus, the multi-rate Erlang Blocking Model has been successfully used to analyze such networks [7, 15, 20, 24, 25, 26]. 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 [28].

A generalization of the multi-rate circuit switched loss model (without call queueing) to include best effort traffic like ABR has been presented in [3], where it was argued that with the introduction of *partial blocking* into the Multi-rate Erlang Model it is possible to model ABR/UBR services (best effort services) on the call level. 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 ratio*, r_{min}. In ABR terminology this would correspond to the fraction of the minimum cell rate and the peak cell rate i.e. $\frac{MCR}{PCR}$. 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 ratio $r(t)$, defined as $min[1, B_a(t)/B_r]$, 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 ABR source is greedy in the sense that as long as the connection is established the source will always transmit with the 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. Another assumption here is that the given bandwidth-residency time product is kept constant as in [17], but the "given bandwidth" fluctuates, so the calculation of the residency time becomes complicated.

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 [17]. 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 ($B_i * 1/\mu_i$), [17].

It has been observed in many papers [5, 8, 12, 18, 24, 25, 27], that in a multi-rate 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 in terms of blocking probability increase incurred on the narrow-band traffic is much bigger than the advantage in terms of blocking probability decrease obtained for the wide-band traffic (see e.g. Figure 2). 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.

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 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 4 discusses a number of numerical results enlightning how the relevant performance measures vary as the function of the *minimal accepted service ratio*, 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. The link capacity is denoted by C, which is assumed to be an integer number in some suitable bandwidth unit.

- The first type is CBR traffic and it is supposed to represent a service with QoS guarantees. By adopting the notion of equivalent bandwidth this type could as well be VBR traffic. CBR calls are characterized by their arrival rate λ_1 their departure rate μ_1 and their *equivalent* bandwidths B_1 and

- The second type is ABR traffic and it is supposed to represent best effort traffic. ABR calls are characterized by their their arrival rate λ_2 their departure rate μ_2 their *peak rate* B_2 and *minimum required rate* $r_{min}B_2$

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_1. Each arriving ABR/UBR (best effort) call brings with itself an exponentially distributed service requirement, which in the case when the peak bandwidth is available throughout the entire duration of the connection gives rise to a departure intensity of μ_2. In the case when the peak bandwidth is not available, all ABR connections in progress on the link share the available bandwidth (which is the link capacity minus the total bandwidth occupied by the CBR calls) equally.

The rate at which the ABR calls are receiving service thus 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 ABR calls in progress on the link are not allowed to receive service at a rate smaller than $r_{min}B_2$. In such link states the incoming call attempts are blocked or queued. Note that arriving CBR/VBR calls are also allowed to "compress" the in-service ABR calls as long as the r_{min} constraint is not violated.

2.2 System Description

The system under investigation is characterised by $(n_1(t), n_2(t))$ where $n_1(t)$ is the number of CBR calls on the link at time t and $n_2(t)$ is the number of ABR calls in the system (on the link and in the queue) at time t. The vector $(n_1(t), n_2(t))$ uniquely specifies how many ABR calls are waiting in the queue (q), and what service ratio r the in-service ABR calls receive.

Under the assumption of Poisson arrivals and exponential holding times $(n_1(t), n_2(t))$ 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 $e = (1, ..., 1)^T$ is a column vector 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 quasi-birth-death (QBD) structure which allows for efficient methods for deriving the steady state distribution π.

Let P denote the maximum number of ABR calls in the system, i.e.

$$P = \lfloor \frac{C}{B_2 * r_{min}} \rfloor + Q. \tag{1}$$

Let p denote the actual number of ABR calls in the system, i.e. p takes integer values between $0 \dots P$. Furthermore, let $I(p)$ denote the maximal number of CBR calls in macro state p:

$$I(p) = \lfloor \frac{C - max(0, p - Q) * r_{min} * B_2}{B_1} \rfloor, \quad p = 0 \dots P \tag{2}$$

When there are p ABR calls in the system and there are i CBR calls in the system, we say that the system is in *macro state* p and *micro state* i. Thus, in a given macro state the number of ABR calls is fixed, and there are only CBR arrivals and departures. The macro states will be characterized by the matrices A_p.

When p runs through the macro states, i.e. $p = 0 \dots P$, the G generator matrix takes the form:

$$G = \begin{bmatrix} A_0 & C_0 & 0 & 0 & .. & 0 & 0 & 0 & 0 \\ B_1 & A_1 & C_1 & 0 & .. & 0 & 0 & 0 & 0 \\ 0 & B_2 & A_2 & C_2 & .. & 0 & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & .. & B_{P-2} & A_{P-2} & C_{P-2} & 0 \\ 0 & 0 & 0 & 0 & .. & 0 & B_{P-1} & A_{P-1} & C_{P-1} \\ 0 & 0 & 0 & 0 & .. & 0 & 0 & B_P & A_P \end{bmatrix} \tag{3}$$

Let $q(p,i)$ and $r(p,i)$ denote the actual queue length and the service ratio when the system is in macro state p and micro state i. Then:

$$q(p,i) = min[max(0, p - \lfloor \frac{C - i * B_1}{r_{min} * B_2} \rfloor), Q]$$

$$r(p,i) = \begin{cases} min(1, \frac{C - i*B_1}{(p-q(p,i))*B_2}) & \text{if } p \neq q(p,i) \text{ and } p \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

for all $p = 0 \ldots P$ and $i = 0 \ldots I(p)$. The A_p matrices represent the CBR arrivals and departures in state p, $p = 0 \ldots P$. All A_p matrices are square matrices. We define the following auxiliary multiplication factor, which is zero if an arriving CBR call is blocked, because it would increase the ABR queue length, which is forbidden.

$$a(p,i) = \begin{cases} 1 & \text{if } q(p,i) = q(p,i+1) \text{ and } i+1 \leq I(p) \\ 0 & \text{otherwise} \end{cases}$$

where $i = 0 \ldots I(p)$.

The size of the square A_p matrix is $(I(p)+1) \times (I(p)+1)$, i.e. the index range is : $0 \ldots I(p) \times 0 \ldots I(p)$, $p = 0 \ldots P$.

$$A_p[i,i+1] = a(p,i) * \lambda_1, \quad i = 0 \ldots I(p) - 1; \quad A_p[i,i-1] = i * \mu_1, \quad i = 1 \ldots I(p) \tag{4}$$

Note that the A_p matrices are tri-diagonal, which implies:

$$A_p[i,j] = 0 \quad \text{for any } i,j \text{ pair for which } |i-j| \geq 2. \tag{5}$$

The size of the (in general not square) B_p matrix: $(I(p)+1) \times (I(p-1)+1)$, i.e. the index range is : $0 \ldots I(p) \times 0 \ldots I(p-1)$, $p = 1 \ldots P$. Thus we have:

$$B_p[i,i] = (p - q(p,i)) * r(p,i) * \mu_2, \quad i = 0 \ldots I(p) \tag{6}$$

and

$$B_p[i,j] = 0 \quad \text{for any } i,j \text{ pair for which } |i-j| \geq 1. \tag{7}$$

The size of the (in general not square) C_p matrix: $(I(p)+1) \times (I(p+1)+1)$, i.e. the index range is : $0 \ldots I(p) \times 0 \ldots I(p+1)$, $p = 0 \ldots P-1$.

$$C_p[i,i] = \lambda_2, \quad i = 0 \ldots I(p+1) \tag{8}$$

and

$$C_p[i,j] = 0 \quad \text{for any } i,j \text{ pair for which } |i-j| \geq 1. \tag{9}$$

These equations along with the understanding that each row in the generator matrix must sum up to zero, uniquely determine each matrix element in G. Note that the total number of states S in the system is given by:

$$S = \sum_{p=0}^{P} (I(p)+1) \tag{10}$$

2.3 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 CBR class and the ABR class blocking probabilities (P_1 and P_2) by identifying the indexes of the CBR and ABR blocking states.

In order to identify the blocking states, we will number (assign scalar indexes to) the states in the two dimensional state space from $0 \ldots S-1$, such that when the system is in state (i,p), it will have the index $s = \sum_{l=0}^{p-1}(I(l)+1)+i$. That is, when the system is in the state with index s, there are i CBR and p ABR calls in the system:

$$p = f_2(s) := inf\{J; \sum_{l=0}^{J}(I(l)+1) > s\} \tag{11}$$

$$i = f_1(s) := s - \sum_{l=0}^{f_2(s)-1} (I(l)+1) \tag{12}$$

Thus $f_2(s)$ gives from the index s the unique number of ABR calls in state s while $f_1(s)$ gives the unique number of CBR calls in state s.

When the system is in state of index s, there are $p = f_2(s)$ ABR and $i = f_1(s)$ CBR calls in the system. A state s is clearly a CBR blocking state if $i = I(p)$, $p = 0, ..., P$. Additionally, because CBR calls may not squeeze out in-service ABR calls, a state (i,p) is also a CBR blocking state if $q(i+1,p) \geq q(i,p)+1$. A state (i,p) is an ABR blocking state if the following inequality holds for $p = 0, ..., P-1$:

$$C - i * B_1 \leq r_{min} * (p + 1 - q(p+1, i)) * B_2 \tag{13}$$

If $p = P$ we obviously always have an ABR blocking state. Then let $S_{1,Bl}$ and $S_{2,Bl}$ be the sets of CBR service and ABR blocking states respectively. The blocking probabilities are then given by $P_1 = \sum_{s \in S_{1,Bl}} \pi(s)$ and $P_2 = \sum_{s \in S_{2,Bl}} \pi(s)$.

The *distribution* of the time spent in the system for the ABR 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

$$q_{MEAN} = \sum_{q=0}^{P} (q * \sum_{s \in S_q} \pi(s)) \tag{14}$$

From Little's equation the mean waiting time of the ABR calls is therefore $q_{MEAN}/(\lambda_2 * (1 - P_2))$.

Similarly, let T_w denote the set of states where the number of ABR calls in the system is w. Then the mean number of ABR calls in the system is

$$W_{MEAN} = \sum_{w=0}^{P} (w * \sum_{s \in T_w} \pi(s)) \tag{15}$$

And from Little's equation the mean time an ABR call spends in the system is $W_{MEAN}/(\lambda_2 * (1 - P_2))$.

3 The Partially Blocking Loss System

When the ABR calls are not allowed to wait in a queue if sufficient bandwidth is not available at time of arrival, the Markov model simplifies in two ways. Firstly, the size of the state space becomes smaller since $Q = 0$. Secondly and more important, the computation of the time spent in the system simplifies because with zero queue the ABR calls always receive service with positive rate, and therefore it is possible to apply the theory of Markov driven workload processes. As it is shown in this section, in this case it is possible to determine the distribution of the ABR service time.

3.1 The Time in System Conditioning on Service Requirement x

Assuming that an ABR call has just arrived and conditioning upon that its service requirement is x, the Laplace transform of the time this ABR 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 an ABR call spends in the system conditioning that its service requirement is x is:

$$P_R(0)s^*(x,s)e = P_R(0)\exp[R^{-1}(M - sI)x][I - M/s]^{-1}e \tag{16}$$

where M is the generator of reduced irreducible Markov process given above when $Q = 0$, R is a diagonal matrix where entry (k,k) gives the service ratio available for an ABR call in state k, and $P_R(0)$ is the probability vector giving the probabilities by which an ABR call enters the system [9]. Finally, $s^*(x,s)$ is the matrix of Laplace Transforms of the time spent where entry (i,j) denotes the entrance to the system in state i and departure from the system in state j.

3.2 The Unconditional Time to Completion when Service Requirement is Exponential

Conditioning upon that the required initial workload is x we have from (29) when summing over all final states

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

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

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

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 the completion of an exponentially distributed workload has transform

$$P_R^T(0)s^*(s)e = P_R^T(0)[sI-(M-\mu R)]^{-1}R\mu e \tag{19}$$

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].

4 Numerical Results

4.1 Numerical Solution Approach

The generator matrix describing the QBD has a band structure. This sparse property of the generator can be exploited when using a direct solution approach for obtaining the steady state probability vector. We have employed a sparse implementation of a direct matrix method called the GTH algorithm (named after the authors Grassman, Taksar and Heyman) [13]. The GTH algorithm is a simple modification of standard Gaussian elimination for the calculation of the steady-state probability vector of a Markov chain. The modification makes the procedure numerically stable unlike the standard Gaussian elimination without pivoting. The complexity of the algorithm is of the same order as standard Gaussian elimination. The key to the numerical stability of the GTH approach is that all operations performed are cancellation free i.e. the approach ensures that all additions are done with elements that have the same sign. In [13] and [16] numerical evidence was given that the cancellation free scheme did in fact allow for the accurate computation of steady-state probability vectors of large Markov chains (non sparse matrices with around 1000 states). Recently in [23] 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. In the subsection below we examine models with generators of up to around 250000 states and find loss probabilities etc. for these models. We have run all our examples in a standard HP-UNIX computing environment consisting of 5 powerful servers and a number of smaller workstations. The system has around 100 users. We ran our examples on HP 9000/800/K460 servers. The largest examples - 250000 states - used around 5 CPU minutes, and around 250 Mb of RAM. On a powerful server the memory consumption seems to be the limiting factor with regards to the size of the problems we can solve rather than the execution time. We did not reconfigure the system in any way to be able to run our examples - obviously we used a large part of the servers 2 Gigabytes of RAM when executing the program.

4.2 Results

This section consists of two subsections where we examine examples of systems defining generators with moderate and large state spaces respectively.

4.2.1 Results for systems with moderate state space

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 CBR calls are narrow-band, while ABR calls are wide-band. Note that the usefulness of the model does not depend on

CBR and ABR being respectively narrow-band and wide-band calls. The model is also valid for systems where CBR is broadband and ABR is narrow-band. Specifically for our example, any CBR calls have a bandwidth demand of $B_1 = 1$ Mbit/s and mean holding time $1/\mu_N = 1$ s. CBR class calls do not accept partial blocking, i.e. they are either given the required bandwidth B_1 or blocked and lost. Wide band calls are of the ABR type characterized by the (maximal) bandwidth demand $B_2 = 10$ Mbit/s and mean holding time $1/\mu_W = 1$ s, and, as discussed above, by the minimal accepted service ratio $r_{min} < 1$. 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 $r_{min} * B_2$. In the examples below we assume that all in-service ABR calls receive the same instantaneous service ratio $r(t) = max[\frac{C - n_{N(t)} * B_1}{n_{W(t)} * B_2}, 1] > r_{min}$, where $n_N(t)$ and $n_W(t)$ denote the number of (narrow-band) CBR and (wide-band) ABR calls in the system at time t.

Figure 1 shows the performance measures of this partially blocking (PB) system (where we let $\lambda_1 = 10 * \lambda_2 = 30$ 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 ABR 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 [19], as follows:

$$Perf = \frac{1 - P_N - P_W}{1 + \Delta T} \qquad (20)$$

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 ABR calls, $1/\mu_2$). This performance measure takes into account the tradeoff between blocking probabilities and ABR 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 of the service time for the the ABR 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 [2, 8, 14, 22]. Even though it is fundamentally different from the PB model, its performance measures are comparable to those of 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 2 shows the performance measures of a system with the same system parameters as of Figure 1. Here, instead of partially blocking ABR calls, they are placed in a finite queue (the size of which varies form 0 to 6) in the case when 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.

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 ABR calls. Here we consider a link of capacity $C = 30$ Mbit/s. Narrow band calls require $B_1 = 1$ Mbit/s bandwidth, wide-band calls require $B_2 = 12$ Mbit/s (case I) or 6 Mbit/s peak bandwidth (case II). In case I the total offered load $B_1 * \lambda_N * (1/\mu_N) + B_2 * \lambda_W * (1/\mu_W)$ is 16 Mbit/s*Erlang, in case II it is 30 Mbit/s*Erlang. Figures 3 and 4 show the different performance measures when r_{min} decreases from 1.0 to 0.5. The behavior of the system is investigated in six sub-cases as the maximal queue length (buffer size), Q for wide-band calls changes from 0 up to 5.

Figures 3.a and 4.a 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 3.b and 4.b depict 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

3.c and 4.c. 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, because they accept smaller bandwidth, and (2) how "fast" service they get (the greater r_{min} is, the smaller the in-service time becomes).

4.2.2 Results for large systems

In this section we consider two large systems. Both have a link capacity of 350 Mbit/s. For the CBR calls the mean holding times are $1/\mu_N = 1$ s and the bandwidth requirements are $B_1 = 2$ Mbit/s (in both cases). For the ABR calls the mean holding times are $1/\mu_B = 1$ s and the (maximal) bandwidth demands are $B_2 =$ 2.5 Mbit/s (in both cases). The first system has CBR arrival rate $\lambda_N = 87.5 calls/s$ and ABR arrival rate $\lambda_B = 70 calls/s$. This reflects a system that is offered traffic equal to its capacity. The second system has CBR arrival rate $\lambda_N = 150 calls/s$ and ABR arrival rate $\lambda_B = 120 calls/s$. This reflects a system that is offered traffic equal to 171 % of its capacity, i.e. an heavily overloaded system. We examine both systems for $r_{min} = 0.05, 0.10, 0.20, 0.30, 0.50, 0.75, 1$ and for Q $= 0, 1, 2, 5, 10, 20$. In both cases the maximum number of simultaneous CBR calls is readily seen to be 175. When $r_{min} = 0.05$ the maximum number of simultaneous ABR calls is 2800 in both cases. When $r_{min} = 0.05$ the number of states in the QBD is app. 250000. It is easy to see that doubling r_{min} approximately reduces the number of states with a factor 2. Notice that in both cases the offered traffic is evenly divided between CBR and ABR traffic.

In Figure 5.a we have the ABR blocking probabilities of the first system. Not surprisingly increasing the number of queueing places and/or decreasing r_{min} lowers the blocking probabilities. It is interesting to note that very low blocking probabilities can be obtained when using appropriate combinations of number of queueing places and r_{min}. Given 5, 10 or 15 queueing places there is a huge benefit in going $r_{min} = 1$ to $r_{min} = 0.75$. Figure 5.b shows the CBR blocking probabilities. Here, not surprisingly, increasing the number of queueing places also increases the CBR blocking probabilities. Implicitly increasing the number of queueing places protects ABR traffic against losses and hence inflicts losses on the CBR traffic. It is interesting to note that lowering r_{min} also lowers the blocking probability of the CBR traffic. Clearly both the ABR and CBR blocking probabilities decrease towards 0 as $r_m in$ decreases. In Figure 5.c we have the ratio of the additional mean time an ABR call spends in the system to the mean service time if the requested bandwidth had been available (2.5 Mbit/s). It is apparent that the number of queueing places do not play a significant role here. For r_{min} lower than 0.3 more than twice the mean service time is spent in the system on average. Our model permits the quantification of the obvious QOS trade off between low blocking probability and low additional time spent in the system.

In Figure 6.a we have the ABR blocking probabilities for the second system. Again very low blocking probabilities can be obtained when selecting small values of r_{min} and having a modest number of queueing places. Given that the system is loaded with 171 % of its capacity the implicit services protection of ABR calls is of practical interest. Figure 6.b illustrates the CBR blocking in the second system. Unlike with the previous system the CBR blocking probability increases when r_{min} is lowered. Looking at the numerical results it is evident that when $r_{min} \leq 0.75$ the carried traffic is close to the capacity of the system (first 5 digits agree) which is a major benefit of allowing even a modest squeezing of ABR calls i.e. no bandwidth is wasted. This also explains why the CBR blocking grows - lowering r_{min} moves blocking from ABR to CBR in an overloaded system. In Figure 6.c we have the ratio of the additional time an ABR call spends in the system to its mean service time when no squeezing and queueing is available. It is evident that on average the ABR calls are squeezed close to the limit i.e. as much as $r_m in$ allows for. Again, the additional time spent in the system is not noticeably influenced by the number of queueing places.

For the purposes of the two systems examined here the previously used performance measure appears to be less informative - it seems to have limited value when describing the value of protecting ABR traffic in critically loaded systems.

5 Conclusions

We have investigated a mixed queueing and loss system where calls with guaranteed service parameters (CBR/VBR) and best effort (ABR/UBR) calls require service. Assuming that wide-band calls subscribe for the ABR 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 modeled as the waiting time in a finite capacity queue. Narrow band calls have been assumed to be of CBR 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. We have demonstrated that it is possible to analyze large systems - we have looked at system defining generators with approximately 250000 states. It is interesting to note that a high degree of service protection can be given to the ABR traffic even for overloaded systems by making appropriate choices of number of queueing places and the degree of squeezing allowed. Also the utilization in critically loaded systems is very close maximum when even a modest amount of squeezing is allowed - hence better network performance than without squeezing. The implicit service protection achieved for ABR calls clearly has practical implications for system design. It has been shown that in our model the time spent in the system by the ABR calls is a phase type distributed random variable. Future works include the investigation of optimal call admission procedures (bandwidth sharing strategies) in the mixed best effort - QoS guaranteed environment [11] on the link level and the investigation of optimal routing strategies on the network level [10].

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 fulfill: $\pi e = 1$ 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 accomplish 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

$$P_{ij}(x,t) + S_{ij}(x,t) = \pi_{ij}(t) \tag{21}$$

where we let

$$\begin{aligned} P_{ij}(x,t) &= Pr\{W_t \leq x, X_t = j|X_0 = i\} \\ S_{ij}(x,t) &= Pr\{T_x \leq t, X_t = j|X_0 = i\} \end{aligned} \tag{22}$$

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. [6]) $\Pi(t) = exp[Mt]$, and using the matrix notation $P(x,t) = [P_{ij}(t)]$ and $S(x,t) = [S_{ij}(t)]$ we obtain:

$$P(x,t) + S(x,t) = exp[M(t)] \tag{23}$$

Next considering (22), 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

$$\frac{\partial P}{\partial t}(x,t) + \frac{\partial P}{\partial x}(x,t)R = P(x,t)M \tag{24}$$

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

$$P^{**}(z,s) = \frac{1}{z}[sI + zR - M]^{-1} \tag{25}$$

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

$$p^*(z,t) = \exp[(M-zR)t] \tag{26}$$

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

$$m(t) = t\pi R e = t\sum_{i=0}^{n}\pi_i r_i \tag{27}$$

just as we would expect! Combining (21) and (25) gives

$$S^{**}(z,s) = [zI + R^{-1}(sI-M)]^{-1}[sI-M]^{-1} \tag{28}$$

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

$$s^*(x,s) = \exp[R^{-1}(M-sI)x][I-M/s]^{-1} \tag{29}$$

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

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

Again, since $Me = 0$ we get

$$-\frac{\partial s^*}{\partial s}(x,s)e = \sum_{n=1}^\infty \frac{x^n}{n!}[R^{-1}M]^{n-1}R^{-1}e \tag{30}$$

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 $a = \frac{\pi R}{\pi R e}$.
From these considerations we finally get that the mean time until work-level x is reached seen from an arbitrary arrival is

$$\begin{aligned} E\{T_x\} &= a(-\frac{\partial s^*}{\partial s}(x,0))e = \frac{\pi R}{\pi R e}\sum_{n=1}^\infty \frac{x^n}{n!}[R^{-1}M]^{n-1}R^{-1}e \\ &= x\frac{\pi e}{\pi R e} = \frac{x}{\pi R e} \end{aligned} \tag{31}$$

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 [21] gives

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

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

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

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

From this it is clear that

$$
\begin{aligned}
Var\{T_x\} &= \frac{2\pi}{\pi Re}(exp[R^{-1}Mx] - I)[R^{-1}M - ea]^{-2}R^{-1}e \\
&- x\frac{2}{\pi Re} \cdot \\
&(\pi[R^{-1}M - ea]^{-1}R^{-1}e + \frac{1}{\pi Re})
\end{aligned} \tag{32}
$$

Acknowledgments

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). Most of the work was done while Allan T.Andersen was with the Department of Mathematical Modelling, The Technical University of Denmark. 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, pp. 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 Multi-rate Circuit Switched Loss Model to Model ABR Services in ATM Networks", in the *Proc. of the IEEE International Conference on Communication Systems, ICCS '96, Singapore*, Vol. 2, pp. 17.4.1-17.4.5, 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] 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.

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

[7] 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.

[8] G. Fodor, S. Blaabjerg and A. T. Andersen, "Modeling and Simulation of Mixed Queueing- and Loss Systems", *Wireless Personal Communications*, Kluwer Academic Publishers, Issue 8:3, pp. 233-256, August, 1998.

[9] G. Fodor, S. Blaabjerg, A. T. Andersen and M. Telek, "A Partially Blocking-Queueing System with CBR/VBR and ABR/UBR Arrival Streams", *IFIP WG 7.3* 5^{th} *International Conference on Telecommunication Systems Modeling and Performance Analysis*, pp. 411-424, Nashville, TN, USA, March, 1997.

[10] G. Fodor, A. Racz and S. Blaabjerg, "Simulative Analysis of Routing and Link allocation Strategies in ATM Networks Supporting ABR and UBR Services", *IEICE Transactions on Communications, Special Issue on ATM Traffic Control and Perf. Eval.*, pp. 985-995, Vol. E81-B, No. 5, May, 1998.

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

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

[13] 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.

[14] 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.

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

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

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

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

[21] 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.

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

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

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

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

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

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

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