Call Admission Control in the presence of Point-to-Multipoint Best Effort Connections

Gülgün Alpan, Eitan Altman, Hedi Magroun
INRIA B.P. 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France
Daniel Kofman
Network Department, E.N.S.T. , 46 rue Barrault, 75634 Paris Cedex, France

Abstract— **In this paper we consider two types of sessions: Guaranteed Performance (GP) and point-to-multi point Best Effort (BE). The existence of different types of users sharing the same resource raises the need for a Call Admission Control (CAC) which will enable an efficient use of the shared network. As is commonly done in todays telecommunications network, CAC is assumed to be imposed only on the GP session. We analyse the effect of bandwidth reservation for BE sessions on the performance of the BE sessions (in terms of the expected delay) and of the GP sessions (in terms of their rejection probabilities). Our analysis is based on a matrix-geometric approach [17].**

Keywords— **Quality of Service,guaranteed performance, best effort calls, matrix-geometric analysis.**

I. Introduction

High speed telecommunication networks facilitate the accommodation of different types of traffic within a single communication network. The existence of different types of users sharing the same resource raises the need for an admission control which will enable an efficient use of the shared network. In this paper, we will consider two types of users: Guaranteed Performance (GP) customers and Best Effort (BE) customers.

The GP calls are typically real time applications which require a certain level of Quality of Service (QoS) such as a reserved fixed bandwidth throughout the session. BE calls, on the other hand, are typically non-real time applications such as file transfers. They are more adapted to the variations in the available bandwidth. In that respect, BE calls are analogous to ABR (Available Bit Rate) service category in ATM using the bandwidth leftover from the GP traffic such as CBR (Constant Bit Rate) and rt-VBR (real time Variable Bit Rate) service categories [1].

The purpose of this paper is to analyse the impact that Connection Admission Control (CAC) performed on GP traffic has on point-to-multipoint BE traffic, in order to properly allocate resources in a dynamical manner as well as guarantee QoS. In particular, we are interested in analysing the effect of reservation of some fixed amount of bandwidth to be shared among the of BE sessions only.

In ATM networks, a large part of the architecture has been standardised [1]. Implementation of the CAC is however not standardized and is left to the network manager. This has motivated much research and a large literature has emerged. Several approaches have been used in designing CACs. The first approach is based on characterising the input traffic of a source by some simple parameters, such as the effective bandwidth; basically it estimates the bandwidth that should be given to a source to guarantee that the cell loss rate does not exceed (in some asymptotic sense) some small number. Examples of this approach are [9], [7], [11], [12], [16], [13], [22].

A recent approach to admission control estimates (online) the consequence of possible admission, using real-time measures of the network activity [8], [19], [20]. An optimal-control approach for call admission, using a Markov Decision Processes approach has been used in a number of papers [5], [10], [18]. All the cited papers focus on the impact of admission control on GP traffic.

A number of recent papers consider the design of the CAC for the GP sessions in terms of the performance of both the BE and the GP sessions: [2], [3], [4]. Our paper is the first to analyse the performance of GP and point-to-multipont BE sessions when CAC is performed on GP sessions. The CAC is not intended to be performed on BE traffic, since BE traffic is sufficiently flexible so as not to deteriorate QoS that are required by GP connections. In particular, the ATM forum has decided [1] that ABR sessions will not be subject to CAC (unless they require from the network a guarantee on minimum cell rate).

Our analysis is based on a matrix-geometric approach [17] which was successfully used in the case of a single queue [2] and point-to-point sessions, see also [14], [15]. In the case of point-to-multipoint BE traffic that we study here, the transmission rate of the BE traffic sources is limited to the *minimum* bandwidth available along the different transmission paths. We shall show in this paper that this constraint makes the performance of the BE traffic even more sensitive to the CAC mechanisms used for the GP sessions. It will be shown that by slightly limiting the acceptance of GP sessions we can gain considerably in the performance of BE sessions; moreover, the improvement in the BE performance can be significantly higher than in the case of point-to-point BE sessions [2].

The structure of the paper is as follows. In the next section we present the model. In Section III we present the blocking probabilities of GP sessions and in Section IV we present the expected delay for the BE sessions. A numerical analysis is presented in Section V. The paper concludes with a conclusion section.

II. Model Description

We consider two service stations (ATM switches), S_1 and S_2, each with a bandwidth capacity of C_1 and C_2, respectively. GP calls arrive to the stations S_1 and S_2 according to a Poisson process with rate λ_{G1} and λ_{G2} (this assumption is a good description of reality, since we are working at the session level, where this property has often been observed). [1] If the number of GP calls exceed the station capacity, the incoming calls will be rejected and lost. For the accepted GP calls a point-to-point connection is established. For simplicity, we assume that each GP session utilizes the same amount of bandwidth, $L = 1$. The service times (which correspond to the amount of information to be transmitted within a session) are exponential with parameter μ_{GP} in both stations (this assumption is realistic since we are at a call level, rather than a packet level). We use the notation $X_{G1}(t)$ and $X_{G2}(t)$ to denote the number of GP sessions in stations 1 and 2, respectively.

The BE calls arrive to the system according to a Poisson process with rate λ_{BE}. All incoming BE calls are accepted into the system, and are served according to the processor sharing discipline (this assumption is compatible with the maxmin fairness assumption on ABR traffic, see [1]). We assume that a point-to-multipoint connection is established for the BE calls. The GP sessions is assumed to have priority over the BE sessions. Hence, the BE sessions can transmit at a rate that is bounded by the *lowest* bandwidth leftover from the GP sessions in the two switches. In other words, the bandwidth available for the BE sessions at time t is $min(C_1 - X_{G1}(t), C_2 - X_{G2}(t))$. We will use the notation $X_{BE}(t)$ to define the number of BE sessions in the system.

If there is a large number of GP sessions in at least one of the stations, the BE sessions suffer from large waiting times. To avoid this drawback, we can put a threshold, N_1 (resp. N_2), on the station S_1 (resp. S_2), for which the incoming GP calls are accepted only if $X_{G1}(t) < N_1$ (resp. $X_{G2}(t) < N_2$). All other GP calls beyond these thresholds are rejected and lost. Naturally, if N_1 and N_2 is set to a low value, we risk to reject a lot of GP calls. Therefore, the trade-off between the two performance measures are of interest: 1. The blocking probabilities for the GP sessions for given N_1 and N_2. 2. The expected waiting time for BE sessions. In the following sections we will present the numerical methods that are employed for the calculations of these performance measures.

III. Distribution of the number of GP sessions

Due to the priority assumption of GP calls over BE calls, the distributions of $X_{G1}(t)$ and $X_{G2}(t)$ are independent of the arrivals and departures of the BE calls. Therefore, BE calls can be ignored for the calculations that concern GP calls. Here, we will derive only the distribution of $X_{G1}(t)$. The calculations for $X_{G2}(t)$ is the same. At each service station we observe a $M/M/c/c$ queue associated with the GP calls. From the balance equations of the underlying Markov Chain, the stationary probabilities $\pi_i = P[X_{G1}(t) = i]$ can be written recursively as:

$$\pi_{i+1} = \frac{\rho}{i+1}\pi_i \tag{1}$$

where $\rho = \frac{\lambda_{G1}}{\mu_{GP}}$. Using equations (1) and the relation $\sum_{i=0}^{N_1} \pi_i = 1$, each π_i can be computed as:

$$\pi_i = \frac{\frac{\rho^i}{i}}{\sum_{i=0}^{N_1} \frac{\rho^i}{i!}}$$

Hence, π_{N_1} gives the blocking probability for GP traffic arriving to station 1.

IV. Expected time of the BE sessions

As we mentioned before, the BE sessions use the remaining capacity from the GP sessions. Thus, the waiting time distribution of the BE sessions depends on the arrivals and departures of both processes $X_{G1}(t)$ and $X_{G2}(t)$.

We note that the expected waiting time is unchanged if we assume that the BE sessions used FIFO discipline (but the distribution of the waiting time may change). We thus use the FIFO discipline for waiting time computations.

In order to obtain the performance measures on the BE sessions we need to study the Markov process $X(t) = (X_{G1}(t), X_{G2}(t), X_{BE}(t))$ where $X_{G1}(t)$, $X_{G2}(t)$ and $X_{BE}(t)$ are as defined before. As a first step, we represent the state of the system - which is given by the triplet (X_{G1}, X_{G2}, X_{BE})- by a single parameter, X, as follows:

$$X = X_{G1} + (N_1+1)X_{G2} + (N_1+1)(N_2+1)X_{BE}$$

where, $0 \le X_{G1} \le N_1$, $0 \le X_{G2} \le N_2$ and $0 \le X_{BE}$. We note that, this transformation assigns a unique value for each X as well as a unique inverse, i.e. there are unique values $X_{BE}[i], X_{G1}[i], X_{G2}[i]$ for each given value i of X.

The process $X(t)$ is a quasi birth and death process. Given the above transformation, its infinitesimal generator Q can be written in the following block matrix form:

$$Q = \begin{pmatrix} B & A_0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \cdots \\ . & . & . & . & \cdots \\ . & . & . & . & \cdots \end{pmatrix}, \tag{2}$$

The matrices, B, A_0, A_1, and A_2 are square matrices of dimension $(N_1+1)(N_2+1)$. Henceforth, we will use N to denote $(N_1+1)(N_2+1)$.

$A_0 = diag(\lambda_{BE})$ corresponds to the arrivals of BE calls; it is a diagonal matrix with all elements on its diagonal equal to λ_{BE}.

Similarly, A_2 corresponds to the departures of the BE sessions and is a diagonal matrix whose ith $(0 \le i \le N-1)$ element on the diagonal equals;

$$A_2[i][i] = \min(C_1 - X_{G1}[i], C_2 - X_{G2}[i])\mu_{BE}$$

The matrix A_1 corresponds to the processes $X_{G1}(t)$ and $X_{G2}(t)$. A_1 can be written as $A_1 = T - D$ where matrix T is a matrix with diagonal zero. The off-diagonal elements of T with offset 1 is identical to the infinitesimal generator of the process $X_{G1}(t)$ and with offset N_1+1 is identical to that of process $X_{G2}(t)$. The matrix T is defined as

$$\begin{aligned} T[i][i+1] &= \lambda_{G1}\mathbb{I}(X_{G1}[i] \ne N_1) \\ T[i][i-1] &= X_{G1}[i]\mu_{GP}\mathbb{I}(X_{G1}[i] \ne 0) \\ T[i][i+(N_1+1)] &= \lambda_{G2}\mathbb{I}(X_{G2}[i] \ne N_2) \\ T[i][i-(N_1+1)] &= X_{G2}[i]\mu_{GP}\mathbb{I}(X_{G2}[i] \ne 0) \end{aligned}$$

[1] See. e.g., Liu (INRIA, Sophia Antipolis, France), Measurements Over the Web, private communication, to be submitted.

where $0 \leq i \leq N-1$, and all other entries are 0.

Matrix D is a diagonal matrix with entries such that the rows of the infinitesimal generator Q will add up to zero:

$$\begin{aligned} D[i][i] &= \lambda_{BE} + (min(C_1 - X_{G1}[i], C_2 - X_{G2}[i])\mu_{GP} \\ &+ \lambda_{G1}\mathbb{I}(X_{G1}[i] \neq N_1) + \lambda_{G2}\mathbb{I}(X_{G2}[i] \neq N_2) \\ &+ X_{G1}[i]\mu_{GP}\mathbb{I}(X_{G1}[i] \neq 0) \\ &+ X_{G2}[i]\mu_{GP}\mathbb{I}(X_{G2}[i] \neq 0) \end{aligned}$$

Finally, the matrix $B = T - D_B$ is similar to the matrix A_1 where T is defined as above and D_B is a diagonal matrix similar to D and guarantees that the rows of the generator Q add up to zero:

$$\begin{aligned} D_B[i][i] &= \lambda_{BE} + \lambda_{G1}\mathbb{I}(X_{G1}[i] \neq N_1) \\ &+ \lambda_{G2}\mathbb{I}(X_{G2}[i] \neq N_2) \\ &+ X_{G1}[i]\mu_{GP}\mathbb{I}(X_{G1}[i] \neq 0) \\ &+ X_{G2}[i]\mu_{GP}\mathbb{I}(X_{G2}[i] \neq 0) \end{aligned}$$

Let the stationary distribution $\pi = [\pi(0), \cdots, \pi(k), \cdots]$, where each $\pi(k)$ is a vector of dimension N and given by

$$\begin{aligned} \pi(k) &= [\pi(0,0,k), \pi(1,0,k), \cdots, \pi(N_1,0,k), \pi(0,1,k), \\ &\cdots, \pi(N_1,N_2,k)] \end{aligned}$$

From the description for π and the block matrix of (2), the balance equations, $\pi Q = 0$ can be written as:

$$\pi(0)B + \pi(1)A_2 = 0 \quad (3)$$

$$\pi(j)A_0 + \pi(j+1)A_1 + \pi(j+2)A_2 = 0 \quad (4)$$

In order to solve the equations (3) and (4) for π, we will refer to [17]. In [17] a general solution procedure is provided for stochastic processes with infinitesimal generators having special structures. In the case of quasi-birth-death processes, as is the case in this study, the infinitesimal generator Q has the block diagonal form as shown in equation (2). The first step of the general solution is to write the matrix characteristic function of the process as follows:

$$A_0 + RA_1 + R^2A_2 = 0 \quad (5)$$

In equation (5), R is a finite non-negative characteristic root matrix which needs to be obtained. Once R is found, the steady state distribution π is obtained by:

$$\pi(j) = \pi(0)R^j$$

Efficient numerical methods are readily available to compute matrix R. After substituting $(T - D)$ instead of A_1 and rearranging the terms, equation (5) can be rewritten as

$$R = (A_0 + RT + R^2A_2)D^{-1} \quad (6)$$

Given equation (6), R can be calculated by a successive substitution algorithm similar to [6]. The iterative procedure is governed by the following equations:

$$\begin{aligned} R_0 &= 0 \\ R_{n+1} &= (A_0 + R_nT + R_n^2A_2)D^{-1} \quad (7) \end{aligned}$$

According to this procedure, an estimate is made for matrix R. This estimate is then used in the right hand side of equation (7) to obtain the next estimate which will be used in the following iteration. This substitution procedure continues until convergence. Once the matrix R is found, we calculate the N eigenvalues and eigenvectors of R which we will denote by r_i , $i = 1,..,N$. The final step of the matrix geometric approach involves calculation of a vector $a = [a_0, a_1, \cdots, a_N]$ which derives from the boundary condition (3). Vector a is a solution to the linear system of equations:

$$aB + aRA_2$$

Then the probability of having j BE sessions in the system can be calculated as:

$$P[X_{BE}(t) = j] = \sum_{i=0}^{N} a_i(r_i)^j$$

and the expected number of BE sessions is calculated as:

$$\mathbb{E}[X_{BE}] = \sum_{i=0}^{N} a_i \frac{r_i}{(1-r_i)^2} \quad (8)$$

for the BE sessions, T_{BE} can be calculated as

$$T_{BE} = \mathbb{E}[X_{BE}]/\lambda_{BE}.$$

V. Numerical Results

In order to obtain the performance measures mentioned in sections III and IV we modified an already existing C code which has been implemented for a single service station handling both BE and GP sessions [2]. The numerical procedures which are mentioned above are implemented through the meschach library for linear algebra [21]. We collected and analysed performance data under light, heavy and saturated traffic. For each traffic intensity, we assumed that the capacity of each station is equal and is set to $C_1 = C_2 = 10$. We obtained mean BE call duration and GP blocking probabilities with respect to the varying acceptance threshold values, N_1 and N_2, from 0 to C_1 and C_2.

Throughout the numerical experiments we took the load factors $\frac{\lambda_{G1}}{\mu_{GP}}$ and $\frac{\lambda_{G2}}{\mu_{GP}}$ to be constant which leads to a constant mean occupation level of servers by GP sessions. By varying the mean duration, μ_{GP}^{-1}, of GP sessions we were also able to observe the reaction of mean BE connection duration to the mean GP service time, as well. In the following examples the service rate for BE traffic is taken as $\mu_{BE} = 0.2$.

A. Light traffic

Light traffic is the case when the total traffic intensity of both GP and BE sessions is small compared to the station capacities, C_1 and C_2. As an example of light traffic situation, we chose the traffic intensities to be $\frac{\lambda_{G1}}{\mu_{GP}} = \frac{\lambda_{G2}}{\mu_{GP}} = 2$ and $\frac{\lambda_{BE}}{\mu_{BE}} = 3$. Figure 1 illustrates the changes in the blocking probabilities of GP sessions arriving to station 1 with respect to different threshold values. Since there is an excess capacity in the system, the blocking probability of the GP sessions remains small even if we decrease the acceptance threshold. (However, it decreases quite rapidly when increasing the acceptance threshold.) Furthermore, as we observe from figure 2 that, the gain in terms of expected call duration for BE sessions that we obtain by reducing the acceptance threshold of GP sessions is quite small.

In summary, because of the excess service capacity, we observe good performance for both call types without necessitating a reservation policy for the BE sessions.

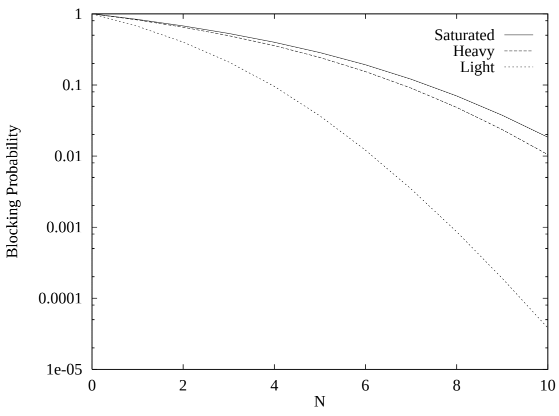

Fig. 1. Blocking probability of GP sessions.

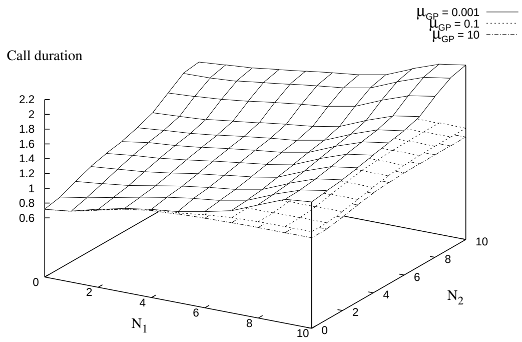

Fig. 2. Mean BE call duration under light traffic.

B. Heavy traffic

Heavy traffic occurs when the total traffic intensity is less than but is close to the station capacity. This regime corresponds to an efficient utilisation of the system's bandwidth, although it may yield large delays and rejection probabilities. It is therefore of interest to study how bandwidth reservation for BE performs in this regime.

We assumed the following traffic intensities: $\frac{\lambda_{G1}}{\mu_{GP}} = \frac{\lambda_{G2}}{\mu_{GP}} = \frac{\lambda_{BE}}{\mu_{BE}} = 4.5$ for both GP and BE calls.
The results of our numerical example show that the BE sessions may suffer from extremely long service times when the system load is high (see figure 3). In this case,it is possible to find appropriate N_1 and N_2 values which will decrease the connection durations of the BE sessions significantly while not much degrading the blocking probabilities for the GP sessions.

For instance ,in figure 4, by setting $N_1 = N_2 = 7$, it is possible to reduce the waiting time of BE sessions from order of 10^{16} to the order of 10 for $\mu_{GP} = 0.1$ (resp. to the order of 10^2 for $\mu_{GP} = 0.001$). On the other hand, limiting the GP call acceptance to 7 at both service stations

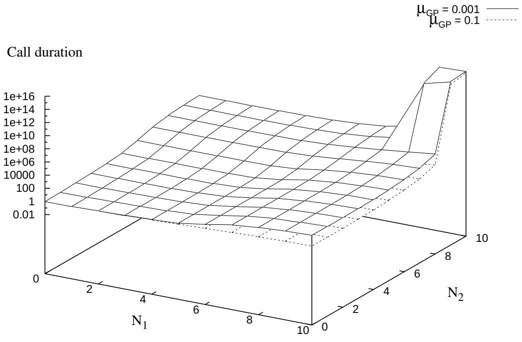

Fig. 3. Mean BE call duration under heavy traffic.

will result in a rejection rate of GP calls of 9 percent at both service stations (as opposed to 1 percent without any threshold)

One other possibility is to cluster the GP calls. For instance, let's assume that the GP calls (class G2) arriving to station S_2 has higher priority with respect to the GP calls (class G1) arriving to S_1 and their rejection rate should be kept lower than that of the calls in station S_1. Then, the connection time of BE sessions may be reduced by setting a low acceptance threshold only for class G1. This way, the rejection rate of class G2 is also kept low. For instance, by setting $N_1 = 6$, $N_2 = 9$ the waiting time of BE sessions is reduced to the order of 10 for $\mu_{GP} = 0.1$ (resp. to the order of 10^2 for $\mu_{GP} = 0.001$) with a blocking probability 0.024 for class G2 and 0.09 for class G1.

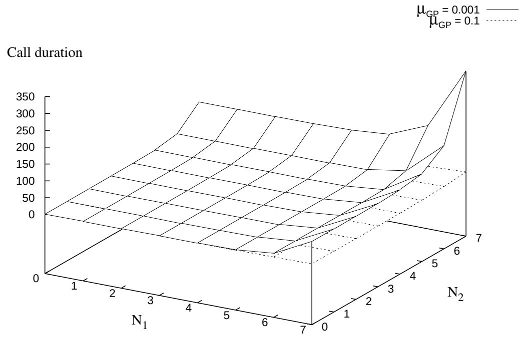

Fig. 4. Mean BE call duration under heavy traffic for $N_1 = 7$ and $N_2 = 7$.

For heavy traffic conditions we observe, furthermore, that the mean connection duration of BE calls are much more affected as the mean service time of GP sessions increase. In this case, the threshold policy can be adjusted based on the mean GP call duration as well.

C. Saturated traffic

The traffic is saturated when the total load of both GP and BE sessions together exceeds the service capacity. Under saturated traffic, when $\mu_{BE}(min(C_1 - \mathbb{E}[X_{G1}], C_2 - \mathbb{E}[X_{G2}]) < \lambda_{BE}$ the connection time for BE will grow without bounds resulting in an unstable system. In this case, it becomes even more important to put an acceptance threshold on GP sessions if by doing so, this can stabilise the system. To simulate the saturated traffic, the traffic intensities of GP and BE calls are chosen to be $\frac{\lambda_{G1}}{\mu_{GP}} = \frac{\lambda_{G2}}{\mu_{GP}} = 5$

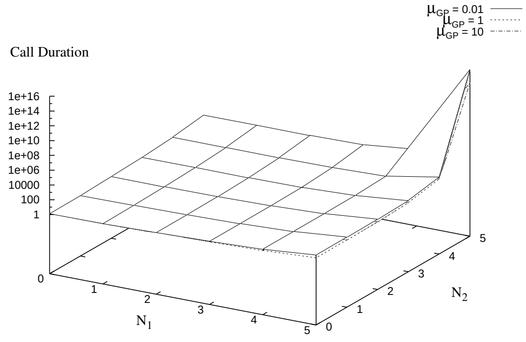

Fig. 5. Mean BE call duration under saturated traffic.

and $\frac{\lambda_{BE}}{\mu_{BE}} = 6$, respectively. The behaviour of the system under these traffic intensities is illustrated in the figures 5 and 1. From these figures we observe that both the performance of BE and GP sessions are sensitive to the value of the acceptance threshold for GP sessions.

For this example, the mean connection time for BE calls is infinite if N_1 and N_2 exceeds 5. In the case of $N_1 = N_2 = 5$ the BE sessions are still subject to very high waiting times (in the order of 10^{15}). As illustrated in figure 6, we can reduce the mean connection time of a BE call from an order of 10^{15} to the order of 10 by pulling the acceptance threshold to a lower value. It is also possible to make a tradeoff for only a priority GP class as we discussed in the previous section.

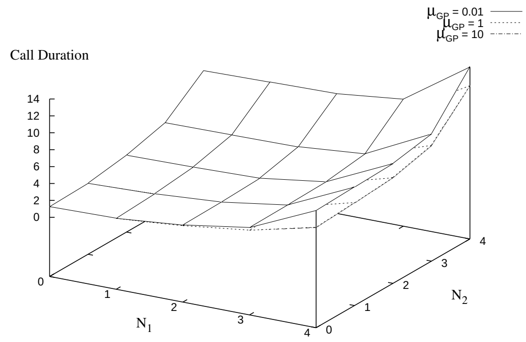

Fig. 6. Mean BE call duration under saturated traffic when $N_2 = 4$ and $N_1 = 4$.

VI. Conclusion

In this paper we considered as part of the call admission control for GP sessions a reservation of some bandwidth to be shared among BE sessions only; this means rejection of GP sessions even if there is still available bandwidth in the system. We justified this by showing that there are cases of interest - the heavy traffic regime, in which a bandwidth reservation for BE can considerably improve the performance of BE sessions without degrading much the rejection probabilities of GP sessions. In light traffic regime, although the reservation of bandwidth did not improve much the delays of BE sessions (which were anyhow low), they did not harm the GP sessions which had anyhow low rejection probabilities.

We compared the delay of BE sessions versus blocking probabilities for the GP sessions in the point-to-multipoint case to the point-to-point case analysed in [2] and observed that the bandwidth reservation for BE results in a considerably larger gain in the performance of BE sessions in the point-to-multipoint case, where as the rejection rate of the GP sessions is of the same order.

References

[1] The ATM Forum Technical Committee, *Traffic Management Specification*, Version 4.0, April 1996.

[2] Altman, E., D. Artiges and K. Traore, "On the integration of best effort and guaranteed performance services," INRIA research report No. RR-3222, 1997.

[3] E. Altman and H. J. Kushner, "Call admission control and routing in integrated service networks using reinforcement learning", *IEEE Conference on Decision and Control*, Dec. '98, Tampa, Florida, USA.

[4] E. Altman, A. Orda and N. Shimkin, "Bandwidth allocation for Guaranteed versus Best Effort service categories" *IEEE Infocom*, San-Fransisco, California, USA, March 29 - April 2, 1998.

[5] E. A. Feinberg and M. I. Reiman, "Optimality of randomized trunk reservation", *Probability in the Engineering and Informational Sciences*, **8**, pp. 463-489, 1994.

[6] Feldman, M. R. and C. Valdez-Flores, *Applied Probability and Stochastic Processes*, PWS Publishing Company, 1995.

[7] R. J. Gibbens and P. J. Hunt, "Effective Bandwidth for the Multi-Type UAS Channel", *Queueing Systems*, **9**, pp. 17-28, 1991.

[8] M. Grossglauser and D. Tse, "A Framework for Robust Measurement-Based Admission Control", Proc. *ACM SIGCOMM 97*, Cannes, France, September 1997.

[9] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and its application to bandwidth allocation in High-speed networks",*IEEE J. Selected Areas Commun.* **9**, pp. 968-981,1991.

[10] A. Hordijk and F. Spieksma (1989), "Constrained admission control to a queuing system", *Advances of Applied Probability* **21**, pp. 409-431.

[11] F. P. Kelly, "Effective Bandwidth at multi-class queues", *Queueing Sytems*, **9**, pp. 5-16, 1991.

[12] G. Kesidis, J. Walrand and C.-S. Chang, "Effective Bandwidth for multiclass Markov fluids and other ATM sources", *IEE/ACM Trans. on Networking*, **1**, pp. 424-428, 1993.

[13] Z. Liu, P. Nain and D. Towsley, " Exponential bounds with application to call admission." *Journal of the ACM*, Vol. 44, No. 3, May 19.

[14] R. Núñez Queija and O.J. Boxma, "Analysis of a multi-server queueing model of ABR", *J. Appl. Math. Stoch. Anal.*, Vol. 11, No.3., 1998.

[15] O.J. Boxma, "A queueing model with varying service rate for ABR", *Lecture Notes in Computer Science: Proceedings of TOOLS'98*, Springer Verlag.

[16] R. Mazumdar, "On call admission control", *2nd. IFIP Workshop on Traffic Management and Synthesis of ATM Networks*, Montreal, September 24-26, 1997.

[17] Neuts, M. F., *Matrix-geometric solutions in stochastic models : an algorithmic approach.* Johns Hopkins series in the mathematical sciences. The Johns Hopkins University press, 1981.

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

[19] D. Tse and M. Grossglauser, "Measurement-Based Call Admission Control: Analysis and Simulation", *Proc. IEEE Infocom '97*, Kobe, Japan, April 1997.

[20] J. Walrand, "Measurement and control of ATM networks", *2nd. IFIP Workshop on Traffic Management and Synthesis of ATM Networks*, Montreal, September 24-26, 1997.

[21] Stewart, D. E., *Meschach: Matrix Computations in C.* Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, 1994. CMA Proceedings #32.

[22] W. Whitt, "Tail probabilities with statistical multiplexing and Effective Bandwidth in multi class queues" *Telecommunication Systems*, **2**, pp. 71-107, 1993.