Robust Admission Control for Heterogeneous ATM Systems with both Cell and Call QoS Requirements

Debasis Mitra[a] and Martin I. Reiman[a] and Jie Wang[b]

[a] Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974
{mitra,marty}@research.bell-labs.com

[b] AT&T Labs, Holmdel, NJ 07733
jie@buckaroo.att.com

The principal goal of call admission control in an ATM system is to simultaneously maintain the Quality of Service (QoS) for several traffic streams with different characteristics. A well-known concept used in designing ATM call admission control schemes is *effective bandwidth*. However, almost all of the notions of effective bandwidths in the literature are based on cell level QoS alone and the point of view of this paper is that this is not adequate. We propose a new method for ATM call admission control, which is based on the notion of *virtual partitioning* (VP), where a measure of bandwidth requirement for connections is used that takes into account both cell and call level QoS requirements. This measure can be obtained by analyzing a single class system. The effective bandwidth thus calculated is less conservative than traditionally calculated values, which ignore call dynamics. We describe an asymptotic analysis of a single class system, which combines two qualitatively different scalings associated, respectively, with the cell and call QoS measures. The call admission control policy based on VP reduces to *complete sharing* when the traffic of all classes are light, and otherwise it gives priority to *underloaded* classes, i.e., those that are using less than their engineered bandwidth allocations. As a consequence, our call admission control is robust in the face of deviations from engineered load.

1. INTRODUCTION

Connection admission control is a key to ensuring Quality of Service (QoS) in ATM networks. There is a substantial literature in this area (see [1], [8], [10], [17], [18] and references therein). With very few exceptions, the published work has focused on cell level QoS, such as cell loss ratio, cell delay and cell delay variation. The well known concept of effective bandwidth, which is the basis of many ATM call admission control schemes, has been studied extensively (see [3], [5], [9], [11]). The formulation of effective bandwidth in literature is also mainly based on cell level QoS. The basic call admission principle is tacitly *complete sharing* (CS), i.e., a new call is admitted if by doing so the previously agreed cell level QoS of this and other calls already in progress will not be violated. However, in order to take advantage of statistical multiplexing, it is necessary to carry traffic with different

statistical characteristics on the same network and CS can cause a significant difference in call blocking probabilities between sources with different bandwidth requirements. Hence, in order to obtain efficient resource sharing, it is necessary to consider not only the cell level QoS but also the call level QoS, in the form of call blocking probability. This leads naturally to call admission control schemes that are not CS. In this paper we propose two ways to take call level QoS into consideration when studying call admission control. First, we introduce a new measure of bandwidth requirement for connections that takes into account both cell and call level QoS requirements. Secondly, based on the new bandwidth requirement, which we call (yet another) *effective bandwidth*, we propose a new method for ATM call admission control, which guarantees not only cell level QoS, but also regulates traffic at the call level to meet the call level QoS requirements in a robust manner. Importantly, the effective bandwidth as calculated here is less conservative than traditionally calculated values that ignore call dynamics. This is because in the latter the worst case call configuration dominates, while here the call distribution is taken into account. The call level QoS requirements of all sources are met when the load on the system is in the engineered range, and the underloaded traffic is protected when the total load is above the engineered range. Which class is underloaded is determined by comparing the offered loads with the engineered loads. The method is based on the notion of *virtual partitioning* (VP) ([4], [15]). Instead of each class of traffic having a fixed priority, as in traditional trunk reservation ([2], [12]), in virtual partitioning the priorities depend on the state of the system. The call admission control policy based on VP reduces to CS when the traffic of all classes are light, and otherwise it gives priority to underloaded traffic.

The work presented here has an experimental aspect, with emphasis given to issues such as robustness and simplicity of implementation. There is considerable reliance on numerical investigations for illustrations and validations. The asymptotic analysis of a single class system which is described here is new for its combination of two quite different scalings that are associated with cell and call QoS requirements.

The rest of this paper is organized as follows. In Section 2 we first study a single class system to obtain some insights. The asymptotic analysis of the single class system is presented in Section 3. In Section 4 we discuss the notion of feasible region in the space of call arrival rates of various classes, and the issue of call admission for feasible call arrival rates. The new call admission policy based on VP is introduced in Section 5. Some numerical results are presented in Section 6. In Section 7 we make some concluding remarks.

2. A SINGLE CLASS SYSTEM

Both cell and call level QoS are considered in the following study of a single class system. The insights gained from this study provide the basis for the techniques for multiclass systems.

2.1. The Model

The system and traffic model can be summarized as follows:

- A bufferless system composed of a single link with transmission rate R.

- Poisson call arrivals with arrival rate λ.
- After a call is admitted into the system, its dynamical behavior can be described by a Markov process. It spends a random amount of time, which is independent and exponentially distributed with mean $1/\alpha$, in the "on" state; then with probability q it leaves the system, and with probability $(1-q)$ it enters the "off" state; after spending a random amount of time, which is independent and exponentially distributed with mean $1/\beta$ in the "off" state, the call enters the "on" state again; it repeats the on-off cycle until it leaves the system from the "on" state.
- When a call is "on", it generates cells as a fluid at rate ν.

There are two QoS criteria:

- The average cell loss ratio does not exceed p^{cell} (cell level).
- The call blocking probability does not exceed p^{call} (call level).

Let k be the number of calls in progress and n the number of calls in the "on" state. Given a stationary call admission control policy (depending on both k and n), the dynamics of the system can be described by a Markov process with state (k, n).

With time scale decomposition between the call and burst levels and using the notion of Nearly Completely Decomposable (NCD) Markov chains, we can reduce the dimensionality of the Markov chain from two to one, with the latter having state (k). As described in [16], the NCD limit corresponds to having $q \to 0$, $\alpha \to \infty$, and $\beta \to \infty$ in such a way that α/β and $q\alpha$ stay constant. In this limit the call holding time distribution becomes exponential, so the number of calls in progress becomes a Markov process. As shown in [16], $\mu = q\alpha\beta/(\alpha+\beta)$, and $p = \alpha/(\alpha+\beta)$. Along with ν, μ and p are the key parameters for a call. Intuitively, this limiting regime is a good approximation when the process describing the number of calls in the "on" state reaches equilibrium between any change in the number of calls in progress due to call arrival/departure. In this situation we can use the one-dimensional Markov chain resulting from the NCD limit, instead of the two-dimensional Markov chain. Moreover, the NCD limit allows the call admission control problem formulated as a Semi-Markov Decision Process to be treated as a one-dimensional problem.

It seems intuitively clear that the reduction in dimensionality that arises in the NCD limit does not require the exponential distribution for on and off times. The underlying mathematical results apply for any phase-type distribution, and any distribution can be approximated to any desired degree of accuracy by a phase-type distribution.

In the NCD limit, the holding time distribution of a call will be exponential for any on and off time distribution, as long as the memoryless process contained in our model for terminating a call is used. Although it would be possible to incorporate non-exponential call holding times, this would not be entirely straightforward. A Semi-Markov Decision Process formulation would require that we keep track of the elapsed holding time for each call in progress. (The current phase of each call would be sufficient for a phase-type holding time distribution.) The optimal policy would typically depend in a non-monotone manner on this elapsed holding time information, making it difficult to implement. An alternative would be to implement a policy that only depends on the number of calls in progress, and ignores the elapsed holding time information. Such a policy might do well, but it might not.

Our goal is feasibility and robustness, so we do not provide any cost function to be optimized. Although the *optimal* admission control policy with a reward for each accepted call normally has a randomized threshold due to the cell and call level QoS constraints, nonrandomized threshold policies, which are simpler, are sufficiently accurate for our purpose here.

2.2. Some Basic Relations

Let $\phi(\kappa)$ be the admission control policy based on the threshold κ, i.e., a new call is admitted if there are fewer than κ calls in progress, otherwise it is rejected. Based on the above model, we can calculate the following performance measures under policy $\phi(\kappa)$. The mean cell arrival rate with k calls permanently in progress is $s(k) = k\nu p$. The mean cell loss rate with k calls permanently in progress is $b(k) = \sum_{n=0}^{k} \binom{k}{n} p^n (1-p)^{k-n} [n\nu - R]^+$. Let $g(\kappa) = \sum_{k=0}^{\kappa} (\lambda/\mu)^k/k!$. The stationary distribution for the number of calls in progress is $P(k \text{ calls}) = [(\lambda/\mu)^k/k!]/g(\kappa), \quad k = 0, 1, \ldots, \kappa$. The average cell arrival rate is $s_\kappa = \left[\sum_{k=1}^{\kappa} (\lambda/\mu)^k s(k)/k!\right]/g(\kappa)$. The average cell loss rate is $b_\kappa = \left[\sum_{k=1}^{\kappa} (\lambda/\mu)^k b(k)/k!\right]/g(\kappa)$. The average cell loss ratio is

$$L(\lambda, \kappa, R) = \frac{b_\kappa}{s_\kappa} = \frac{\sum_{k=1}^{\kappa} \frac{(\lambda/\mu)^k}{k!} b(k)}{\sum_{k=1}^{\kappa} \frac{(\lambda/\mu)^k}{k!} s(k)}. \tag{1}$$

The call blocking probability is given by the Erlang loss formula $B(\lambda, \kappa) = [(\lambda/\mu)^\kappa/\kappa!]/g(\kappa)$.

2.3. Feasibility of Call Arrival Rate

We define a call arrival rate λ to be feasible if there exists an admission control policy from the class of threshold policies that meets both QoS requirements. That is, λ is feasible if there exists κ such that $\phi(\kappa)$ ensures both cell and call QoS. Recall that our cell and call QoS requirements are, respectively: $L(\lambda, \kappa, R) \le p^{cell}$, and $B(\lambda, \kappa) \le p^{call}$. For a given λ, let $\kappa_1(\lambda) = \max\{\kappa : L(\lambda, \kappa) \le p^{cell}\}$ and $\kappa_2(\lambda) = \min\{\kappa : B(\lambda, \kappa) \le p^{call}\}$. If $\kappa_2(\lambda) \le \kappa_1(\lambda)$, then there exists a threshold type policy $\phi(\kappa)$ with $\kappa_2(\lambda) \le \kappa \le \kappa_1(\lambda)$, such that under $\phi(\kappa)$ both cell and call QoS requirements are met, and we say λ is feasible.

Given p^{cell}, p^{call} and the traffic parameters described in Section 2.1, there is a maximum feasible call arrival rate, λ_{max}, such that for any $\lambda > \lambda_{max}$, $\kappa_2(\lambda) > \kappa_1(\lambda)$. It is not surprising that for most cases of practical interest, $\kappa_2(\lambda_{max}) \le \kappa_1(\lambda_{max})$, which is assumed throughout this paper.

2.4. An Effective Bandwidth

Let $e = R/\kappa_1(\lambda_{max})$. Then e is a measure of the resources a call requires to satisfy *both* cell level and call level QoS requirements. Note that e is independent of the call arrival rate λ.

Note that a traditional effective bandwidth definition, see [5] for instance, assumes that calls last in perpetuity and addresses only cell level QoS: $e_{static} = R/\kappa_{static}$, where $\kappa_{static} = \max\{k : \ b(k)/s(k) \le p^{cell}\}$. Note that $e \le e_{static}$. Since e_{static} ignores call level dynamics it is more conservative.

Example 1 Consider a system with $R = 45.0$, and homogeneous sources with parameters $\mu = 0.1$, $\nu = 6.0$, $p = 0.025$, $p^{cell} = 10^{-9}$ and $p^{call} = 0.01$. In Figure 1 we illustrate the

relation between $\kappa_1(\lambda)$, $\kappa_2(\lambda)$ and λ_{max}. $\kappa_1(\lambda)$ and $\kappa_2(\lambda)$ intersect at 20 when $\lambda = 1.125$. Hence, $\lambda_{max} = 1.125$, and $e = \frac{45}{20} = 2.25$. Note that $\kappa_{static} = 14$, hence $e_{static} = 3.21$.

Figure 1. Calculating Effective Bandwidth for Example 1

Example 2 Now consider a system with the same R, p^{cell}, p^{call} as in Example 1, $\mu = 1.0$, $\nu = 1.5$ and $p = 0.1$. $\kappa_1(\lambda)$ and $\kappa_2(\lambda)$ intersect at 111 when $\lambda = 94.0$. Hence, $\lambda_{max} = 94.0$, and $e = \frac{45}{111} = 0.4054$. Now $\kappa_{static} = 100$, and $e_{static} = 0.45$.

3. ASYMPTOTIC ANALYSIS

Our goal here is to provide fundamental insights into the joint behavior of the cell loss ratio and call blocking probability for purposes of sizing and operations. The investigation is in the asymptotic framework of large systems, i.e., as $(\lambda, \kappa, R) \to \infty$ in a manner consistent with practical QoS requirements. Specifically, p^{cell} is expected to be in the range $10^{-6} - 10^{-9}$, while p^{call} is expected to be in the neighborhood of 10^{-2}. These numbers suggest the following important dichotomy: cell loss ratios decay exponentially in the large parameter, say κ, while call blocking probabilities decay polynomially, more specifically as $1/\sqrt{\kappa}$. While both elements are separately recognized in the literature, see for instance [19], we do not know of any prior analysis in which both elements are simultaneously present. The loadings at the cell and call levels are required to be such that the cell and call performances are, respectively, exponential and polynomial in κ. We obtain such loading guidelines.

First, we write the expression for the cell loss ratio in (1) as $L(\lambda, \kappa, C)$, where $C = R/\nu$ by definition. Also, we assume for convenience that C is an integer, and select the unit of time to be such that $\mu = 1$. Then, $L(\lambda, \kappa, C) =$ Num/Den, where Num $= 0$ if $\kappa \leq C$,

$$\text{Num} = \sum_{k=C+1}^{\kappa} \frac{\lambda^k}{k!} \sum_{n=C+1}^{k} (n - C) \binom{k}{n} p^n (1-p)^{k-n} \qquad \text{if } \kappa > C \tag{2}$$

and

$$\text{Den} = p \sum_{k=0}^{\kappa} k\lambda^k/k! \ . \tag{3}$$

Note that we may assume that $\kappa > C$, as otherwise $L = 0$.

As mentioned above, we let $(\lambda, \kappa, C) \to \infty$, while p and ν are held fixed. For loading at the cell level we assume that the system is underloaded, and that critical loading holds at the call level, i.e.,

$$\rho \triangleq \frac{\lambda p}{C} < 1 \,, \quad \text{and} \quad \gamma \triangleq \left(1 - \frac{\kappa}{\lambda}\right)\sqrt{\lambda} = O(1) \,, \tag{4}$$

i.e., γ is bounded. At the loss of some small generality we will make the convenient assumption that γ is a fixed constant, which can be either positive or negative. Hence $\kappa = \lambda - \gamma\sqrt{\lambda}$, so that, to leading order, $\kappa/\lambda \sim 1$.

We let $\alpha \triangleq (1 - p/\rho)/(1 - p)$, which arises since $(\kappa - c)/\{\lambda(1 - p)\} \sim \alpha$. Note that $0 < \alpha < 1$. Our main result, which is proved in [14], is that

$$L(\lambda, \kappa, C) \sim \frac{Ae^{-\delta\kappa}}{\kappa^2} \ , \tag{5}$$

where

$$\begin{aligned} A &= A(\rho, p, \gamma) = \frac{\rho\alpha}{\pi p(1-\alpha)(1-\rho\alpha)^2} \ \frac{1}{\sqrt{(1-p/\rho)p/\rho}} \cdot \frac{e^{-\gamma^2/2}}{\text{erfc } (\gamma/\sqrt{2})} \ , \\ \delta &= \delta(\rho, p) = \log\left[\left(\frac{1-p/\rho}{1-p}\right)^{1-p/\rho} \ \left(\frac{1}{\rho}\right)^{p/\rho}\right] \ , \end{aligned} \tag{6}$$

and erfc is the complementary error function.

It is easy to verify that for all (ρ, p) such that $0 < \rho < 1$ and $0 < p < 1$, $\delta(\rho, p) > 0$. Clearly $A(\rho, p, \gamma) = O(1)$. These facts prove the important result that for the asymptotic scaling in (4), the cell loss ratio $L(\lambda, \kappa, C)$ is exponentially small in the large parameter κ, with δ the constant in the exponential. The asymptotic call blocking probability for our scaling is well known to exhibit $1/\sqrt{\lambda}$ (and therefore $1/\sqrt{\kappa}$) behavior.

4. AN ADMISSION CONTROL PROBLEM

Now consider an ATM system with two classes of on-off sources. For $i = 1, 2$, let λ_i, μ_i, p_i, ν_i, p_i^{cell} and p_i^{call} denote the traffic and QoS parameters corresponding to those in Section 2.1 for class i calls.

The two class system with parameters specified in Examples 1 and 2, respectively, has been studied in [16]. Figure 2, which is from [16], shows, for a given link capacity R, the feasible region in (λ_1, λ_2) space, i.e., the set of (λ_1, λ_2) such that the cell and call QoS requirements for both classes can be met by at least one stationary call admission policy. Observe that the feasible region has an almost linear boundary. The feasible region is calculated as follows: for each λ_1, we increase the value of λ_2 until we cannot

find a feasible policy that meets both cell and call level QoS requirements. Since the Semi Markov Decision Process formulation of the connection admission control problem is equivalently a Linear Programming Problem, the existence of a feasible policy is equivalent to feasibility in the latter (see [6], [7], [16] for more detail).

Figure 2. Feasible Region of Call Arrival Rates

We know that when (λ_1, λ_2) is on or very close to the boundary, the call admission policies that meet all the QoS requirements necessarily have complicated structures. For just this reason, the engineered operating points are designed not to be too close to the boundary. Under the assumption that (λ_1, λ_2) are not very close to the boundary, the question of interest in this paper is whether we can find a *simple* connection admission policy that meets *both cell and call QoS requirements*. We want our policy to be robust in the sense that if the operating point drifts away from the engineered loads due to the unexpected high arrival rate of one class, the admission policy should be able to protect the other class. We should point out that the construction of policies was not a subject of study in [16].

The most widely studied call admission policy type is CS. Although it is easy to implement, it always favors calls requiring less bandwidth capacity, thus it can drive the blocking probability of calls with a larger bandwidth requirement up when the arrival rates of the calls with smaller bandwidth requirement are high. A very undesirable situation is when the class 1 blocking probability exceeds the QoS constraint while the blocking probability for class 2 calls is much lower than its constraint. With CS, regardless which class exceeds its engineered load, class 1 will suffer in terms of call blocking probability.

Trunk Reservation (TR) policies have been known to be able to provide protection to wideband calls. With properly chosen trunk reservation parameters, class 1 calls will

not have unacceptably high call blocking probabilities due to high arrival rates of class 2 calls. However, the traditional trunk reservation gives priority to a fixed class, hence cannot protect the other class when the favored class has high arrival rate. Of particular interest is the following trunk reservation policy which we study numerically in Section 6: for single link systems with two classes of calls, narrowband and wideband, if we admit a new call (regardless of its class) only when the spare capacity in the system is at least the bandwidth of wideband calls, then this special trunk reservation policy balances the call blocking probabilities of the two classes, which can be easily seen with PASTA [13]. Furthermore, this policy is optimal among all the trunk reservation policies that balance the call blocking probabilities in the sense that the call blocking probabilities are the smallest. However, because this policy balances the call blocking probabilities, if one class has high call arrival rate, both classes will have high call blocking probabilities. In other words, it is not robust, which is far from desirable in many situations.

It is intuitively apparent that if (λ_1, λ_2) are small enough (close to the origin), a complete sharing policy will be sufficient. The interesting case is when (λ_1, λ_2) is not very close to the boundary of the feasible region and also not small. In this case, we believe a simple policy based on virtual partitioning with properly chosen parameters gives satisfactory performance.

5. VIRTUAL PARTITIONING

We now describe a call admission policy based on the notion of VP. Let e_1 and e_2 be the bandwidth requirements of class 1 and 2 calls, and K_1 and K_2 be the partitioning parameters which are two positive integers such that $K_1 e_1 + K_2 e_2 \geq R$. A call admission policy based on VP is summarized as follows: *When a call of class* 1 *arrives and finds* (k_1, k_2) *calls in progress, it is accepted if*

$$k_1 < K_1 \quad \textit{and} \quad e_1 k_1 + e_2 k_2 \leq R - e_1,$$

$$\textit{or,} \quad k_1 \geq K_1 \quad \textit{and} \quad e_1 k_1 + e_2 k_2 \leq R - t_2 e_2 - e_1,$$

where $t_2 e_2$ *is the bandwidth reserved for (underloaded) class* 2 *calls. Similarly, a class* 2 *call is accepted if*

$$k_2 < K_2 \quad \textit{and} \quad e_1 k_1 + e_2 k_2 \leq R - e_2 \quad \textit{or}$$

$$k_2 \geq K_2 \quad \textit{and} \quad e_1 k_1 + e_2 k_2 \leq R - t_1 e_1 - e_2,$$

where $t_1 e_1$ *is the bandwidth reserved for (underloaded) class* 1 *calls.*

Note that the call admission is performed on a set in (k_1, k_2) space defined by $k_1 e_1 + k_2 e_2 \leq R$. The motivation for selecting this set is derived from the linearity implicit in the notion of effective bandwidths. Admittedly there is no sound theoretical basis for the linearity at this time. In the absence of such a theory we take the precautionary step of verifying in our numerical investigations that cell level QoS requirements are satisfied by our admission control policies.

By choosing parameters K_1 and K_2, we partition the bandwidth between the two classes. The trunk reservation parameters t_1 and t_2 allow us to block calls from the overloaded

class so as to reserve bandwidth for the underloaded class. The nature of the policy is to give the underloaded class higher priority, and the consequence is that the underloaded class is protected.

We expect t_1 and t_2 to be small nonnegative integers. When $t_1 = t_2 = 0$, the policy becomes a complete sharing policy over the whole admissible region (regardless of K_1 and K_2). If in this case the call blocking probabilities for *both* classes are still greater than the allowed limits (say, 1%), then the arrival rates are too high for there to exist any feasible policy. When t_1 and t_2 are not zero, the policy is complete sharing on the set $\{(k_1, k_2) : k_1 \leq K_1,\ k_2 \leq K_2\}$, and dynamically prioritized outside the set. With complete sharing, the less bursty traffic enjoys lower call blocking, and consequently when the set in which complete sharing applies is bigger, fewer calls of the less bursty traffic are blocked.

6. NUMERICAL RESULTS

In our numerical study, we attempt to show that we can use VP to design simple call admission policies such that when the system is subject to load at or below the engineered call arrival rates $(\lambda_1^e, \lambda_2^e)$ the admission policy will meet all the cell and call QoS requirements, and when the system is subject to load above the engineered loads due to a high arrival rate from one class, the other class is protected. More specifically, we observe the call blocking probabilities of the two classes, denoted by (B_1, B_2), under CS, TR and VP to illustrate the robustness of VP. The following load scenarios are studied: both classes are below the engineered load; both classes are at the engineered load; one of the two classes is at the engineered load while the other is at least 20% higher; both classes are 10% higher than the engineered loads.

The system and traffic sources are the ones specified in Section 4. From the analysis of the two single class problems in Section 2.4, we have $e_1 = 2.25, \quad e_2 = 0.4054$, which are used for all the cases in this section.

6.1. Case 1

The first set of numerical results are for the engineered load: $(\lambda_1^e, \lambda_2^e) = (0.6, 35.0)$. The bandwidth partitioning parameters used are $(K_1, K_2) = (9, 42)$ and the trunk reservation parameters used are $(t_1, t_2) = (2, 0)$. The call blocking probabilities for several different loads are listed in Table 1.

Table 1
Numerical Results for Case 1

Call arrival rates	Call blocking Probabilities (B_1, B_2)		
(λ_1, λ_2)	CS	TR	VP
(0.60, 35.0)	(0.0064, 0.0008)	(0.0031, 0.0031)	(0.0049, 0.0033)
(0.72, 35.0)	(0.0184, 0.0026)	(0.0096, 0.0096)	(0.0153, 0.0076)
(0.60, 45.0)	(0.0232, 0.0032)	(0.0106, 0.0106)	(0.0083, 0.0202)
(0.66, 38.5)	(0.0174, 0.0024)	(0.0085, 0.0085)	(0.0110, 0.0110)
(0.55, 33.0)	(0.0026, 0.0003)	(0.0013, 0.0013)	(0.0021, 0.0012)

When the load is below or at the engineered load, all three policies give satisfactory call blocking probabilities. However, when one of the two classes has load higher than the engineered load, class 1 always suffers under CS and the blocking probabilities are driven above the allowed limit 1%. TR in general gives better performance than CS. However, when $(\lambda_1, \lambda_2) = (0.6, 45.0)$, TR gives balanced call blocking probabilities of 0.0106. Hence class 1 suffers because class 2 has an arrival rate higher than the engineered load. VP is able to protect the underloaded class. Because the engineered load is near the middle of the boundary of the feasible region, we need to protect each class when the other class exceeds its engineered load. Since on the complete sharing set class 2 is favored, by choosing the set large enough for VP we can protect class 2. Protection for class 1 in VP is achieved by choosing the complete sharing set not too large and selecting proper trunk reservation parameter t_1, which is set to 2 in this case.

6.2. Case 2

The second set of numerical results are for the engineered load $(\lambda_1^e, \lambda_2^e) = (0.2, 70.0)$. The bandwidth partitioning parameters used are $(K_1, K_2) = (7, 60)$ and the trunk reservation parameters used are $(t_1, t_2) = (1, 0)$. Table 2 summarizes the call blocking probabilities corresponding to the five scenarios.

Table 2
Numerical Results for Case 2

Call arrival rates	Call blocking Probabilities (B_1, B_2)		
(λ_1, λ_2)	CS	TR	VP
(0.20, 70.0)	(0.0150, 0.0018)	(0.0048, 0.0048)	(0.0004, 0.0059)
(0.24, 70.0)	(0.0241, 0.0031)	(0.0079, 0.0079)	(0.0010, 0.0096)
(0.20, 84.0)	(0.1049, 0.0153)	(0.0327, 0.0327)	(0.0018, 0.0386)
(0.22, 77.0)	(0.0536, 0.0073)	(0.0171, 0.0171)	(0.0013, 0.0205)
(0.18, 63.0)	(0.0026, 0.0003)	(0.0009, 0.0009)	(0.0001, 0.0011)

Now the engineered load is near the upper left corner of the feasible region. Again at $(\lambda_1, \lambda_2) = (0.20, 84.0)$, class 1 suffers under CS and TR when class 2 exceeds the engineered load, while VP is able to protect it. When the class 1 arrival rate exceeds the engineered load by 20%, the impact on class 2 is very small. On the other hand, increasing the arrival rate of class 2 has more dramatic impact on class 1 traffic, hence it seems now the focus should be on protecting class 1 against class 2.

6.3. Case 3

The third set of results are for the engineered load $(\lambda_1^e, \lambda_2^e) = (0.9, 10.0)$. The bandwidth partitioning parameters used are $(K_1, K_2) = (10, 25)$ and the trunk reservation parameters used are $(t_1, t_2) = (3, 0)$. Table 3 summarizes the call blocking probabilities for the five scenarios.

Table 3
Numerical Results for Case 3

Call arrival rates (λ_1, λ_2)	Call blocking Probabilities (B_1, B_2)		
	CS	TR	VP
(0.9, 10.0)	(0.0038, 0.0005)	(0.0030, 0.0030)	(0.0038, 0.0005)
(1.08, 10.0)	(0.0156, 0.0022)	(0.0128, 0.0128)	(0.0156, 0.0022)
(0.90, 15.0)	(0.0073, 0.0010)	(0.0052, 0.0052)	(0.0072, 0.0017)
(0.99, 11.0)	(0.0093, 0.0013)	(0.0073, 0.0073)	(0.0093, 0.0014)
(0.81, 9.0)	(0.0013, 0.0002)	(0.0010, 0.0010)	(0.0013, 0.0002)

This is a case where traffic is light so that VP is almost identical to CS. At $(\lambda_1, \lambda_2) = (1.08, 10.0)$, class 2 suffers under TR while both CS and VP are able protect class 2.

7. CONCLUSIONS

We propose a new method for ATM call admission control that is based on the notion of virtual partitioning and dynamic priorities. Our numerical results indicate that with properly selected bandwidth partitioning and trunk reservation parameters, this new admission control method is able to protect the underloaded class.

Although our results have been presented for a two-class system, we believe that they would apply for more than two classes. The key to our approach is the (almost) linearity of the feasible region depicted in Figure 2. If the feasible region is linear for the multiclass system (we have not checked this), then the $L(\geq 2)$ class system can be analyzed using L single class systems.

ACKNOWLEDGEMENT

We are grateful to our colleague John Morrison for his invaluable help with the analysis in Section 3.

REFERENCES

1. *Special Issue on* Advances in the Fundamentals of Networking - Part I*, IEEE Journal of Selected Area in Communications*, 13, 1995.
2. J.M. Akinpelu. The overload performance of engineered networks with nonhierarchical and hierarchical routing. *AT&T Bell Labs Technical Journal*, 63:1261–1281, 1984.
3. N.G. Bean. Effective bandwidths with different quality of service requirements. In *Integrated Broadband Communication Networks and Services*. V.B. Iverson (Ed.), IFIP, 1993.
4. S. Borst and D. Mitra. Virtual partitioning for resource sharing by state-dependent priorities: analysis, approximations, and performance for heterogeneous traffic. In these proceedings.

5. A.I. Elwalid and D. Mitra. Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking*, 1:329–343, 1993.
6. E.A. Feinberg. Constrained Semi-Markov decision processes with average rewards. *ZOR-Mathematical Methods of Operations Research*, 39:257–288, 1994.
7. E.A. Feinberg and M.I. Reiman. Optimality of Randomized Trunk Reservation. *Probability in Engineering and Informational Sciences*, 8:463–489, 1994.
8. R.J. Gibbens, F.P. Kelly, and P.B. Key. A decision-theoretic approach to call admission control in ATM networks. *IEEE J. Sel. Areas Commun.*, 13:1101–1114, 1995.
9. J.Y. Hui. Resource allocation for broadband networks. *IEEE J. Selected Areas in Communications*, 6, 1988.
10. J.M. Hyman, A.A. Lazar, and G. Pacifici. A separation principle between scheduling and admission control for broadband switching. *IEEE Journal on Selected Areas in Communications*, 11:605–616, 1993.
11. F.P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–16, 1991.
12. P. Key. Optimal control and trunk reservation in loss networks. *Probability in the Engineering and Informational Sciences*, 4:203–242, 1990.
13. K. Lindberger. Dimensioning and designing methods for integrated ATM networks. In *Proceedings of ITC-14*, pages 897–906, 1994.
14. D. Mitra, M.I. Reiman, and J. Wang. Robust admission control for heterogeneous ATM systems with both cell and call QoS requirements. Technical report, 1996.
15. D. Mitra and I. Ziedins. Virtual partitioning by dynamic priorities: Fair and efficient resource-sharing by several services. In B. Plattner, editor, *Broadband Communications:* Proceedings of 1996 International Zurich Seminar on Digital Communications, pages 173–185. Springer, 1996.
16. M.I. Reiman, J. Wang, and D. Mitra. Dynamic call admission control of an ATM multiplexer with on/off sources. In *Proceedings of the 34th IEEE Conference on Decision and Control*, pages 1382–1388, 1995.
17. J.W. Roberts (Ed.). *COST 224: Performance evaluation and design of multiservice networks.* ECSC-EEC-EAEC, Brussels, 1992.
18. H. Saito. Call admission control in an ATM network using upper bound of cell loss probability. *IEEE Trans. on Communications*, 40:1512–1521, 1992.
19. A. Shwartz and A. Weiss. *Large Deviation for Performance Analysis.* Chapman & Hall, London, 1995.