Admission Control for Combined Guaranteed Performance and Best Effort Communications Systems Under Heavy Traffic

Eitan Altman
INRIA
2004 Route des Lucioles
B.P. 93, 06902 Sophia-Antipolis Cedex
FRANCE

Harold J. Kushner*
Division of Applied Mathematics
Brown University
Providence RI 02912
USA

August 11, 1999

*The work was supported by contract DAAH04-96-1-0075 from the Army Research Office and NSF grant ECS9703895.

Abstract

Communications systems often have many types of users. Since the users share the same resource, there is a conflict in their needs. This conflict leads to the imposition of controls on admission or elsewhere. In this paper, there are two types of customers, GP (Guaranteed Performance) and BE (Best Effort). We consider an admission control of GP customer which has two roles. First, to guarantee the performance of the existing GP customers, and second, to regulate the congestion for the BE users. The optimal control problem for the actual physical system is difficult. A heavy traffic approximation is used, with optimal or nearly optimal controls. It is shown that the optimal values for the physical system converge to that for the limit system and that good controls for the limit system are also good for the physical system. This is done for both the discounted and average cost per unit time cost criteria. Additionally, asymptotically, the pathwise average (not mean) costs for the physical system are nearly minimal when good nearly optimal controls for the limit system are used. Numerical data show that the heavy traffic optimal control approach can lead to substantial reductions in waiting time for BE with only quite moderate rejections of GP, under heavy traffic. It also shows that the controls are often linear in the state variables. The approach has many advantages. It is robust, simplifies the analysis (both analytical and numerical) and allows a more convenient study of the parametric dependencies. Even if optimal control is not wanted, the approach is very convenient for a systematic exploration of the possible tradeoffs among the various cost components. This is done by numerically solving a series of problems with different weights on the costs. We can then get the best tradeoffs, and the control policies which give them.

Key words: admission control, control of communications systems, control of queueing networks, heavy traffic limits, ergodic control, singular control, weak convergence
AMS numbers: 90B22, 60K25, 60K30, 60F17, 93E20, 93E25

1 Introduction

Broadband-ISDN (integrated services data network) high speed networks allows the possibility of integrating different services in one single telecommunications network. In particular, they handle applications that require guaranteed QoS (quality of service), such as bounded delays bounded cell loss rates, and a guarantee on the throughput. Such a service is called Guaranteed Performance (GP). On the other hand, they support also more flexible applications, such as data transfer, that are less sensitive to instantaneous variations in available bandwidth, in delays, jitter etc., and which do not require guarantees on throughputs or delays. We call the latter "Best Effort" (BE) traffic. In the context of ATM (asynchronous transfer mode), this corresponds to the Available Bit Rate (ABR) service category [1], which can adapt to the bandwidth unused by the GP service classes. In context of the Internet, the BE traffic are the TCP/IP connections, which use a congestion avoidance mechanism [17] so as to adapt to the available bandwidth, in contrast with real-time applications that use UDP (UDP is a protocol which does not adapt its transmission rate to the current congestion state of the system) such as phone over Internet [5]. In particular, we suppose that the BE users share the remaining bandwidth, as for example, in the standard Internet. (Note that, unlike ATM, the Internet does not provide performance guarantees for real time-applications. However, in practice, the TCP/IP connections do adapt their transmission rate to the available bandwidth left over after the UDP sessions, see [3]).

Since the GP and BE share the same resources of the network, there is a conflict in their needs, and the question of admission control arises. The purpose of this paper is to analyze the impact that Connection Admission Control (CAC) performed on GP traffic has on the BE traffic, in order to properly allocate resources in a dynamical manner as well as guarantee QoS.

In ATM networks, a large part of the architecture has been standardized [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 characterizing 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 [10, 15, 13, 18, 19, 31, 32, 37]. A diffusion based statistical CAC was proposed in [12].

A recent approach to admission control estimates (on-line) the consequence of possible admission, using real-time measures of the network activity [14, 35, 36]. An optimal-control approach for call admission, using a Markov Decision Processes) approach has been used in a number of papers [11, 16, 34]. All the cited papers focused on the impact of admission control on GP traffic. To the best of our knowledge, our paper is the first to consider the optimal design of the CAC for the GP sessions in terms of the performance of both the BE and the GP sessions.

The CAC was 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). We shall therefore assume that BE sessions are always admitted into the network, unless some very large limit is reached. This limit might represent the number of sessions that can be handled simultaneously by the switches. This limit will not depend on the available bandwidth.

Quite often, CACs are designed taking only into account the performance requirements of GP sessions. However, as already noted in [2, 4], under appropriate operating conditions in the network (in particular, the heavy-traffic conditions, which correspond to an efficient utilization of the resources), it is possible to improve substantially the performance (delays and throughputs) of the BE sessions by modifying *slightly* the CAC for the GP sessions.

References [2, 4] focused on analyzing the performance of GP and BE sessions for given CACs (performed on GP sessions) that take into account also the performance of BE sessions. In this paper, we go one further step and provide a design based on an optimization approach. We consider the *optimal* admission control arising when the number of sessions and the available bandwidth are taken to be large, and scaled in an appropriate way so as to perform in the desirable heavy traffic region.

The approach taken is that of optimal control in the heavy traffic regime. In this regime, the system has little spare bandwidth over what is needed to handle the "average load." It is shown that the physical system can be well approximated by a controlled (reflected) diffusion process, and that good controls for this diffusion process are good for the physical system under heavy traffic. The resulting controls are easily implementable (and are often simply linear in the state variables) and can give a significant improvement of the performance of the BE sessions by only a small rejection of GP sessions. The work is a continuation of the past work on the control of communications and queueing systems under heavy traffic conditions [29, 27, 26, 22, 30, 25, 28, 23].

The extensive numerical results in [25] illustrated the power of such an approach. It obtained excellent controls and system performance and obtained very useful information which would not otherwise be available. The paper [30], concerned with trunk line problems, also showed the power of the heavy traffic approach for modeling, simplifying and controlling very complex systems.

The control problem for the actual physical model is quite hard, if not virtually impossible to solve. The system is not always Markovian. Even in the Markovian case, the number of states can be extremely large. Additionally, even for an uncontrolled Markovian system, the computation of the steady state overflows (losses) and delays is difficult. The heavy traffic approach provides much analytical and computational simplification. The optimally controlled costs for the physical problem converge to the optimally controlled cost for the limit problem, and a nearly optimal control for the limit problem is also nearly optimal for the physical problem if the traffic is heavy.

In applications, the traffic is not always "heavy." But, this regime is one of

the crucial ones for design, analogous to the case of the trunk line problem for long distance teletraffic. The analysis illustrates efficient use of resources. In any large system, there are many alternative uses of the resources, and continuous tradeoffs among them. Here we can see, for example, the effect of marginal reallocations of bandwidth within a control context. Of course, serious control problems do exist even outside of the heavy traffic regime; for example, even if the system is heavily overbuilt from the heavy traffic perspective, the structure of the burstiness in arrivals can lead to important control problems. In this regard, it is worth noting that, in the numerical study in [25] of controlled multiplexers in the heavy traffic regime, the buffer is empty or nearly empty most of the time, and it is the burstiness of the input which yields the losses and the control problem.

There are many other advantages to the combined heavy traffic and optimal control approach. Under broad conditions, it yields the appropriate dimensioning of the system, and shows that good performance can often be achieved with only modestly extra bandwidth (over average requirements). The "limit" variables can be interpreted as "aggregated" states. The analysis takes advantage of the laws of large numbers and central limit theorems that come into play as the size of the system grows. The "limit" or aggregate equations can be used to compute nearly optimal controls for the physical system, and to get good estimates of various measures of performance. The relative simplicity of the form of the heavy traffic limit process facilitates understanding the parametric dependencies. As in [25], we can obtain both qualitative and quantitative information which is often very hard to get otherwise . To do so otherwise, one would need to study many particular cases of systems with different parameters and sizes whose relationships would remain obscure. The reference [33] contains examples of other types of applications.

The method allows convenient numerical approximations (see [24, 25]) whose complexity is much less than that of even the physical Markov systems. In fact, the basic Markov chain approximation method of [24] has been coded for problems of this type and is publicly available. [1] The format is robust: bursty data, priorities, time dependence, dependence of arrivals on the system state, and other useful extensions can be readily handled.

An important question concerns the tradeoff between gains in QoS for BE versus losses for GP. The heavy traffic formulation allows a systematic exploration of this. One solves the optimization problem with various weights for the associated losses and costs, and computes the values of the individual components of the cost under the optimal controls. This yields the set of possible tradeoffs, under good operating policies. One can then choose the control which get the best performance for one component, with a constraint on the others. Usually, optimization for its own sake is not the main interest. But the use of optimizing or optimal control methods for exploration of the possibilities is of greater interest and provides a powerful tool for design. The value of this

[1] See the home page of the Lefschetz Center for Dynamical Systems, Brown University Dept. of Applied Mathematics home page http://www.dam.brown.edu/lcds.html. Select the software link to get the documentation and codes.

approach was amply demonstrated in [25], and the large gains due to the use of feedback control were demonstrated. This is equally valid for the present problem. Indeed, numerical data shows that substantial reductions (say, 30% or even much more) in BE delays can be obtained with only very modest levels of rejection of GP customers. The percent rejected depends on the system parameters and the arrival rates, and it goes to zero as the system grows, for any fixed percentage gain in delay.

The "state space" collapse phenomenon illustrated in Section 5, when there are many GP and BE subclasses, allows one to explore the effects on performance and control of complex customer requirements. One can do the numerics for aggregations of several classes as well as for the original problem to get insights into robustness of the model and performance under varying conditions. This would be very hard under any current alternative approach. Indeed, the limit allocations to the different classes are consistent with the (ad-hoc) weighted fair-shares for different types of BE customers which were defined by the ATM forum for sharing bandwidth (see the Section I.3 in Appendix I "Implementation Examples on ABR Service" [1]).

The model In Section 4 can be used to study the effects of finite bandwidth constraints on individual users. The multicast problem of Section 6 shows how seemingly complex forms of the problem can be put into the context of a general theory.

The structure of the paper is as follows. The basic model is presented in Section 2, and the input-output equation is written in a way that is convenient for the heavy traffic analysis. In the basic model, the BE customers share the available bandwidth equally, whatever it is. The general methods used for the analysis of this case also apply with little modification to the subsequent cases. Section 3 deals with the weak convergence of the heavy traffic approximations and the discounted cost function as the traffic intensity approaches unity. Many of the ideas have been used in the references in various ways, despite the difference in the problems considered. Because of this and to save space, we provide detailed outlines with references where possible. There is a new problem with "tightness" of the set of the (singular) controls, but we show that they can be approximated such that they are tight.

Section 4 considers the case where there is an upper limit to the bandwidth that any one of the BE customers can use. This changes the dynamics and the scaling, but the previous analysis can be carried over. In order to illustrate the power and versatility of the approach, in Section 5 we consider extensions where there are several classes of BE customers, which might be allocated bandwidth in class dependent ways. There is a surprising degeneracy, which can be exploited to simplify the analysis and subsequent numerical work.

Section 6 extends the basic model to a multicast case. Here there are two channels (any number could be used), each with its own class of GP customers, and own control. But the BE customers must be transmitted simultaneously on each of the channels. Many variations of the format are possible. The ergodic cost problem is dealt with in Section 7. In Section 7 we impose a constraint on the rate at which GP arrivals can be rejected, so that previous results on

the ergodic cost problem can be exploited. The form of the heavy traffic limit equations implies that this constraint is not very restrictive and this is borne out by numerical data. But, in Section 8 we show that it is not a restriction.

The optimal mean cost per unit time for the physical system converges to the optimal "ergodic" value for the limit system as the size of the system and time go to infinity in any way at all. Furthermore, the limit of the pathwise (not mean) average costs per unit time cannot be better than the optimal ergodic value for the limit system. The pathwise average costs for the physical system can be made to be arbitrarily close to this ideal limit by using a nice nearly optimal (for the limit system) control on the physical system, under heavy traffic. The pathwise results are important, since in any single application, we have just one sample path, and the mean values are less important than the pathwise values. Furthermore, a "nice" nearly optimal control for the limit system provides nearly optimal values for the physical system, under heavy traffic, in both a mean and pathwise sense.

Some of the development is similar to that in recent works on control under heavy traffic (e.g., [29, 22, 25, 23]), although the exact form of the adaptation is not entirely obvious for the present problem. This is particularly true of the derivation of the basic setup in the first part of Section 3. Owing to this, we have referred to the literature whenever possible. There remain many new methodological features, apart from the novelty of an important application in a broad context: The approximation of the singular controls, both for the discounted and the ergodic problem, the degeneration (or state space collapse) for the multi BE/GP class problem, the adaptation of the formulation to multicast, and other extensions, with state dependent dynamics.

2 Problem 1: The Basic Setup

In this section, we set up the notation and evolution equations for the *basic problem*. The development for the more complex problem classes is similar. There are two classes of users, which we refer to as class 0 and class 1. Class 1 being the GP traffic and class 0 the BE traffic. As usual in the analysis of systems under conditions of heavy traffic, the mean service capacity is slightly greater than the mean demand. The system is parameterized by a parameter N. Both the system capacity, excess capacity (over what is required for the average demands) and demand grow as $N \to \infty$, with the *relative* excess capacity) going to zero. This will be formalized below. A consequence of the heavy traffic analysis is that these relationships represent good design in that very good performance can be obtained. A well designed heavy traffic operating regime implies that the system is not overbuilt and can well handle the demands placed on it.

In this and in the next section, the bandwidth is normalized so that each member of class 1 requires one unit of bandwidth. The members of class 0 share whatever bandwidth is left. An applications paradigm is that each arrival is the work for a "session," whose work arrives essentially at once, and is buffered on

arrival. This models well the more general situation in which the input rate of packets within a session is not less than the transmission rate, so that when we allocate a given bandwidth to a session then this bandwidth is indeed used by it. In Section 4 we shall relax this assumption. The channel is time shared, with a guaranteed time going to the class 1 customers, and the remaining to the class 0 customers. Thus, there is no limitation on the rate at which work can be done on the set of class 0 customers except for the the available capacity.

Let $\alpha_l^{i,N}, l = 1, \ldots,$ denote the interarrival times for the members of class i, and set $E\alpha_l^{i,N} = \bar{\alpha}^{i,N}$ To conform with standard usage, we also define the normalized mean "rates" $\lambda^{i,N} = 1/N\bar{\alpha}^{i,N}$. The service times for the members of class 1 are exponentially distributed with constant rate $\mu^{1,N}$. Class 1 can be controlled in that any requested admission can be denied by the controller. The system is a "loss" system in that any customer denied admission disappears from the system. For analytic convenience, we suppose further that there is a constant $B_+^0 > 0$ such that no class 0 customers are admitted to the system if the current number of class 0 customers in the system is greater than $\sqrt{N}B_+^0$. It follows from the heavy traffic limit theorems (and from the numerical data) that this is inconsequential for large enough B_+^0. In fact, the condition is useful only to simplify some of the analytic details for the ergodic cost criterion, and B_+^0 can be set equal to infinity for the discounted cost function. But we include the finite upper bound here to unify the development. Let $F^{0,N}(t)$ denote $1/\sqrt{N}$ times the number of class 0 customers not admitted by time t.

The time requirements for the members of class 0 are also exponentially distributed. But the rate at which a customer of class 0 departs from the system depends on the (random) resources available to it while it is in the system. The "conditional mean instantaneous departure rate" for a class 0 customer at time t is defined to be $\mu^{0,N}$ times the bandwidth available to that customer at that time. I.e., if at time t, $B(t)$ is the total bandwidth unused by members of class 1, and there are $A(t)$ members of class 0 in the system, then the probability (conditioned on the data up to that time) of a single departure of a member of class 0 in the time interval $[t, t+\delta)$ is $\mu^{0,N}B(t)\delta/A(t) + o(\delta)$, and the probability of more than one departure in that interval is $o(\delta)$. The set of interarrival times, and the service times for class 1 are assumed to be mutually independent. This independence of the interarrival times is a good description of reality, since we are working at the session level, where this property has often been observed. [2] Long range dependence in the arrival process is not relevant as it might be at the packet level.

We assume that there is b (parameterizing the "excess capacity," which might be negative) such that the channel capacity is

$$C_N = N\left[\frac{\lambda^{0,N}}{\mu^{0,N}} + \frac{\lambda^{1,N}}{\mu^{1,N}}\right] + b\sqrt{N}. \tag{2.1}$$

Suppose that $\lambda^{i,N} \to \lambda_i$ and $\mu^{i,N} \to \mu_i$, all positive. Thus, the mean arrival

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

rates and the channel capacity are all $O(N)$, and the channel capacity is $O(\sqrt{N})$ greater than the mean requirements. For appropriate b, this will be seen to be sufficient for good behavior. We also make the innocuous assumptions that

$$\left\{ \left| \alpha_l^{i,N} / \bar{\alpha}^{i,N} \right|^2 ; l, i, N \right\} \text{ is uniformly integrable,} \tag{2.2}$$

and that there are σ_i^2 such that

$$E\left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}\right]^2 \to \sigma_i^2. \tag{2.3}$$

In common cases, the arrival processes are assumed to be Poisson. Then all moments of the terms in (2.2) are uniformly (in i, l, N) bounded. If the interarrival intervals for class i are constant, then $\sigma_i^2 = 0$. If the arrival stream for class i is Poisson, then $\sigma_i^2 = 1$. Define the scaled mean arrival process

$$S^{i,N}(t) = \frac{1}{N} [\# \text{ of class } i \text{ arrivals by time } t]$$

and the normalized and centered sum

$$W^{i,N}(t) = \frac{1}{\sqrt{N}} \sum_{l=1}^{[Nt]} \left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}\right],$$

where $[Nt]$ denotes the integer part.

As usual in heavy traffic scaling, all of the basic system variables are scaled by $1/\sqrt{N}$. Define $X^{0,N}(t)$ to be $1/\sqrt{N}$ times the number of members of class 0 in the system at time t, and set

$$X^{1,N}(t) = \frac{1}{\sqrt{N}} \left[\# \text{ of class 1 in system at time } t - N\frac{\lambda^{1,N}}{\mu^{1,N}}\right].$$

Thus $X^{1,N}(t)$ is the scaled number of class 1 customers centered about the mean, if there were no rejections, and an infinite channel capacity. Define $A^{i,N}(t)$ (resp., $D^{i,N}(t)$) to be $1/\sqrt{N}$ times the number of arrivals (resp., departures) of class i by time t, and $F^{1,N}(t)$ denotes $1/\sqrt{N}$ times the number of arrivals of class 1 up to time t which the controller did not admit. Thus, $A^{i,N}(t) = \sqrt{N} S^{i,N}(t)$. The (non decreasing) control process $F^{1,N}(\cdot)$ is assumed to be *admissible* in that it is (ω, t) measurable and its value at time t depends only on the data which is available up to time t.

The system balance equations are

$$X^{0,N}(t) = X^{0,N}(0) + A^{0,N}(t) - D^{0,N}(t) - F^{0,N}(t), \tag{2.4a}$$

$$X^{1,N}(t) = X^{1,N}(0) + A^{1,N}(t) - D^{1,N}(t) - F^{1,N}(t) - U^{1,N}(t), \tag{2.4b}$$

where $U^{1,N}(t)$ is $1/\sqrt{N}$ times the number of class 1 customers that could not be admitted due the *entire* channel being occupied by class 1 customers. This last term will disappear in the limit. We will suppose that the set

$$\{X^{1,N}(0), N\} \text{ is tight (bounded in probability).} \tag{2.5}$$

If this set is not tight, then there will be a long delay before the heavy traffic regime is entered. The set $\{X^{0,N}(0)\}$ is always tight, since $B_+^0 < \infty$. It is also supposed that the initial condition is independent of the subsequent arrival times, and service times for class 1.

In order to simplify the convergence proofs, one needs to put the input and output processes into a more convenient form.

Representation of the input processes. Following a common practice in weak convergence analysis, we decompose the arrival process as

$$\begin{aligned} A^{i,N}(t) &= \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} 1 \\ &= \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} \left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}\right] + \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}. \end{aligned} \tag{2.6a}$$

Note that

$$\sum_{l=1}^{NS^{i,N}(t)} \alpha_l^{i,N}$$

equals t minus the time since the last arrival before or at t. Thus, by (2.6a),

$$A^{i,N}(t) = W^{i,N}(S^{i,N}(t)) + t\lambda^{i,N}\sqrt{N} - \frac{\rho^{i,N}(t)}{\sqrt{N}}, \tag{2.6b}$$

where $\rho^{i,N}(t)$ is the time since the last arrival before or at t, divided by the mean interarrival interval. The sequence of processes $\rho^{i,N}(\cdot)/\sqrt{N}$ converges weakly to the "zero" process. It does not affect any of the subsequent calculations and for the sake of notational simplicity, it will be *omitted* in all of the subsequent system equations after (2.11).

Representation of the output processes. Owing to the exponential distribution of the service time for the class 1 customers, $D^{1,N}(\cdot)$ can be decomposed into the sum of the integral of the *instantaneous conditional mean rate* at which $D^{1,N}(\cdot)$ increases, and a martingale process $\tilde{D}^{i,N}(\cdot)$. To do this, first note that the conditional mean instantaneous rate of increase of $D^{1,N}(\cdot)$ at t is $\mu^{1,N}/\sqrt{N}$ times the number of class 1 customers in the system at t.

We have the decomposition

$$D^{1,N}(t) = \mu^{1,N} \int_0^t \left[X^{1,N}(s) + \sqrt{N}\frac{\lambda^{1,N}}{\mu^{1,N}}\right] ds + \tilde{D}^{1,N}(t). \tag{2.7}$$

The Doob-Meyer increasing process associated with the martingale is just $1/\sqrt{N}$ times the integral in (2.7); namely,

$$\langle \tilde{D}^{1,N} \rangle(t) = \frac{\mu^{1,N}}{\sqrt{N}} \int_0^t \left[X^{1,N}(s) + \sqrt{N} \frac{\lambda^{1,N}}{\mu^{1,N}} \right] ds. \tag{2.8}$$

The factor $1/\sqrt{N}$ appears due to the definition of $D^{i,N}(t)$ as $1/\sqrt{N}$ times the number of departures by time t. Similar decompositions were used in [22, 27].

Let $I^{0,N}(t)$ denote the indicator function of the event that there are class 0 customers in the system at t. The departure process for class 0 is similarly decomposed into the integral of the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases, and a martingale $\tilde{D}^{0,N}(\cdot)$.

In preparation for this, first note that the available bandwidth per class 0 customer at time t is

$$\frac{C_N - \left[\sqrt{N} X^{1,N}(t) + N\lambda^{1,N}/\mu^{1,N} \right]}{\sqrt{N} X^{0,N}(t)} I^{0,N}(t),$$

which equals

$$\frac{N\lambda^{0,N}/\mu^{0,N} + b\sqrt{N} - \sqrt{N} X^{1,N}(t)}{\sqrt{N} X^{0,N}(t)} I^{0,N}(t).$$

Thus the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is

$$\left[\sqrt{N} \frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(t) \right] I^{0,N}(t).$$

Hence,

$$D^{0,N}(t) = \mu^{0,N} \int_0^t \left[\sqrt{N} \frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s) \right] I^{0,N}(t) ds + \tilde{D}^{0,N}(t), \tag{2.9a}$$

which we rewrite as

$$D^{0,N}(t) = \mu^{0,N} \int_0^t \left[\sqrt{N} \frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s) \right] ds + \tilde{D}^{0,N}(t) - Y^{0,N}(t). \tag{2.9b}$$

The term $Y^{0,N}(\cdot)$ is a reflection term; it corrects for the difference in the integrals in (2.9a) and (2.9b), and it can increase only when $X^{0,N}(t) = 0$. The Doob-Meyer increasing process associated with $\tilde{D}^{0,N}(\cdot)$ is

$$< \tilde{D}^{0,N} > (t) = \frac{\mu^{0,N}}{\sqrt{N}} \int_0^t \left[\sqrt{N} \frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s) \right] I^{0,N}(s) ds,$$

which is written more conveniently as

$$\mu^{0,N} \int_0^t \left[\frac{\lambda^{0,N}}{\mu^{0,N}} + \frac{b}{\sqrt{N}} - \frac{X^{1,N}(s)}{\sqrt{N}} \right] I^{0,N}(s) ds. \tag{2.10}$$

Now, putting all of the above representations together and canceling the $\pm\sqrt{N}\lambda^{i,N}$ terms yields the forms:

$$X^{0,N}(t) = X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) - \mu^{0,N}\int_0^t g_0(X^N(s))ds \\ -\tilde{D}^{0,N}(t) - F^{0,N}(t) + Y^{0,N}(t) - \frac{\rho^{0,N}(t)}{\sqrt{N}}, \tag{2.11a}$$

$$X^{1,N}(t) = X^{1,N}(0) + W^{1,N}(S^{1,N}(t)) - \mu^{1,N}\int_0^t X^{1,N}(s)ds \\ -\tilde{D}^{1,N}(t) - F^{1,N}(t) - U^{1,N}(t) - \frac{\rho^{1,N}(t)}{\sqrt{N}}, \tag{2.11b}$$

where we define

$$g_0(x) = b - x^1, \tag{2.12}$$

and the $Y^{0,N}(t)$ term compensates for the fact that there are no departures of class 0 customers at time t if $X^{0,N}(t) = 0$. It assures that $X^{0,N}(t)$ will stay non negative.

Comments on weak convergence. The path space for all of the random processes is $D^k[0,\infty)$, the space of functions which are right continuous, have left hand limits and take values in Euclidean $k-$space for appropriate integers k. The Skorohod topology is used [6, 9]. This is the most common and convenient choice in heavy traffic analysis. The following is a convenient criterion for tightness in this space. It will be used implicitly. Let $\{Y^n(\cdot)\}$ be a sequence of processes with paths in $D[0,\infty)$, with probability one. Let $\mathcal{T}^n(t)$ denote the stopping times with respect to the filtration engendered by $Y^n(\cdot)$ and which are no larger than t. If

$$\lim_{\delta\to 0}\sup_n \sup_{\tau\in\mathcal{T}^n(t)} E\left(1 \wedge |Y^n(\tau+\delta) - Y^n(\tau)|\right) = 0, \tag{2.14}$$

for each t, and

$$\{Y^n(t) : n, t\} \text{ is tight}, \tag{2.15}$$

then $\{Y^n(\cdot)\}$ is tight [9].

3 Discounted Cost Function and Weak Convergence

The convergence theorem implies that only a bandwidth excess of order $O(\sqrt{N})$ is needed for good performance.

Although the set of applications is new, the setup so far is similar to that of many other problems of control of queues under heavy traffic. See, for example, [29, 22, 25, 23]. The main new questions concern the control functions $F^{1,N}(\cdot)$. We will be concerned with two types of cost functions; the discounted and the

average cost per unit time (to be called the ergodic cost function). Until Section 7, we concentrate on the discounted problem. Much of the analysis carries over to the ergodic case with little change. Also, analogous procedures are used for the convergence proofs for all of the problem formulations in the following sections.

Let $\beta > 0, c_i > 0$, where β can be as small as we wish, and let $k(\cdot)$ be a non negative continuous function with $k(0) = 0$. The discounted cost function is defined by

$$C_\beta^N(x, F^{1,N}) = E\int_0^\infty e^{-\beta s}k(X^{0,N}(s))ds + E\int_0^\infty e^{-\beta s}\sum_i c_i dF^{i,N}(s). \quad (3.1)$$

The second term penalizes the rejections. We do not penalize the loss $U^{1,N}(\cdot)$, since it is zero in the limit as $N \to \infty$ no matter what the controls are. The first term can be quite general. If $k(\cdot)$ is linear, then it simply penalizes the waiting time for class 0 customers. More generally, it can be non linear. For example, it might be zero for small values of the argument (if the delays at small values are considered to be unimportant), or it might increase superlinearly to discourage long delays. The allowed generality of the cost function is one of the advantages of the approach. Define $V_\beta^N(x) = \inf_{F^{1,N}} C_\beta^N(x, F^{1.N})$, where the inf is over the admissible controls.

The basic structure of the convergence proofs is similar to those in the cited references to controlled queues in heavy traffic, except for the questions of tightness and approximation of the control functions. After stating the convergence theorems, the proofs will be outlined and references given for many of the details.

Define $B^{i,N}(\cdot) = W^{i,N}(S^{i,N}(\cdot)) - \tilde{D}^{i,N}(\cdot)$.

Theorem 3.1. *Let $\epsilon > 0$ be small but arbitrary, and let $F^{1,N}(\cdot)$ be $\epsilon-$optimal controls. Suppose that $\{F^{1,N}(\cdot), N\}$ is tight*[3] *. Then the set*

$$\left\{X^{i,N}(\cdot), B^{i,N}(\cdot), F^{i,N}(\cdot), i = 1, 2; Y^{0,N}(\cdot), N\right\}$$

is tight. The weak sense limit of any weakly convergent subsequence satisfies

$$X^0(t) = X^0(0) - \mu_0\int_0^t g_0(X(s))ds + B^0(t) + Y^0(t) - F^0(t), \quad (3.2a)$$

$$X^1(t) = X^1(0) - \mu_1\int_0^t X^1(s)ds + B^1(t) - F^1(t), \quad (3.2b)$$

where the $B^i(\cdot)$ are mutually independent Wiener processes with variance parameters $\lambda_i(1 + \sigma_i^2)$. $0 \le X^0(t) \le B_+^0$, and $Y^0(\cdot)$ is the reflection term at zero. $F^0(\cdot)$ is the reflection term at the upper bound. The other processes are non anticipative[4] *with respect to the Wiener processes.*

[3]By Theorem 3.3, this tightness assumption entails no loss of generality.

[4]The associated filtration here and in subsequent uses of "non anticipative" is that generated by all of the processes $(X^i(\cdot), B^i(\cdot), F^i(\cdot), Y^i(\cdot), i = 0, 1)$.

Comments on the proof. Details for similar results are in [22, 23, 25, 29, 27], and we will only outline the sequence of ideas. First, by the renewal theorem $S^{i,N}(\cdot)$ converges weakly to the deterministic (limit scaled mean arrival rate) process with values $\lambda_i t$. By Donsker's Theorem [6, 9], $W^{i,N}(\cdot)$ converge weakly to mutually independent Wiener processes with variance parameters σ_i^2. Hence the $W^{i,N}(S^{i,N}(\cdot))$ converge to mutually independent Wiener processes $\tilde{W}^i(\cdot)$ with variance parameters $\lambda_i \sigma_i^2$.

The set of martingales $\{\tilde{D}^{i,N}(\cdot); i, N\}$ can readily be proved to be tight via the criterion (2.14), (2.15). This is done by a direct computation using the fact that their associated Doob-Meyer increasing processes (2.8) and (2.10) are bounded by a constant times t. The fact that the scaled discontinuities go to zero as $N \to \infty$ implies that their weak sense limits have continuous paths with probability one.

With the above tightness results available on the "driving terms" $W^{i,N}(S^{i,N}(\cdot)), \tilde{D}^{i,N}(\cdot)$, and the tightness assumption on the control terms, the tightness of $\{X^{1,N}(\cdot), U^{1,N}(\cdot), N\}$ can be readily proved if the $X^{1,N}$ in (2.11b) were replace by some bounded functions. For the general case, the tightness follows by a standard truncation argument [21]. This tightness implies that the sequence $U^{1,N}(\cdot)$ has "zero" weak sense limits, since the "upper boundary" for $X^{1,N}$ gets pushed to infinity as $N \to \infty$. The tightness of $\{X^{1,N}(\cdot)\}$ also implies the tightness of $\{X^{0,N}(\cdot), Y^{0,N}(\cdot), F^{0,N}(\cdot), N\}$. In fact, the tightness of the $\{F^{0,N}(\cdot)\}$ can be shown even without tightness of the controls, since the controls only decrease $X^{1,N}(\cdot)$ (hence $X^{0,N}(\cdot)$) and $F^{0,N}(\cdot)$ is a reflection process at an upper boundary. Indeed, one can show that $F^{0,N}(\cdot)$ is asymptotically continuous with probability one. If it were not asymptotically continuous then the asymptotic discontinuity and the properties of the other driving terms for $X^{0,N}(\cdot)$ would imply that asymptotically there is a jump to the interior of $[0, B_+^0]$ from the upper boundary, which is impossible, since the individual steps go to zero as $N \to \infty$ and $F^{0,N}(\cdot)$ can increase only on the boundary. The tightness of $\{Y^{0,N}(\cdot), N\}$ implies that $I^{0,N}(\cdot)$ can be non zero only on set whose Lebesgue measure goes to zero a $N \to \infty$. Hence, $I^{0,N}(\cdot)$ has no asymptotic influence on the Doob-Meyer increasing process associated with the martingale $\tilde{D}^{0,N}(\cdot)$.

Now, given the tightness, extract a weakly convergent subsequence, with the weak sense limits being denoted by $X^i(\cdot)$, etc. It can be shown that the weak sense limits $\tilde{D}^i(\cdot)$ are mutually independent Wiener processes which are independent of the $\tilde{W}^i(\cdot)$, and have the variance parameters λ_i. The Wiener property of the limits of the $\tilde{D}^{i,N}(\cdot)$ is proved as in [22, Theorem 3.1]. The mutual independence is a consequence of the independence conditions and the definition of "conditional mean instantaneous rate" via a conditional expectation, and an analogous computation is in [22, Theorem 3.1]. The fact that (3.2) holds follows from the weak convergence. The non anticipativenss properties are proved by standard "martingale" means as in [29, Theorem 5.1], [22, Theorem 3.1] or [23, Theorem 2.1] ■

Definition. The discounted cost function for (3.2) is

$$C_\beta(x, F^1) = E\int_0^\infty e^{-\beta s}k(X^0(s))ds + E\int_0^\infty e^{-\beta s}\sum_i c_i dF^i(s). \tag{3.3}$$

Define an *admissible* control $F^1(\cdot)$ for (3.2) to be a non decreasing process which is non anticipative with respect to the Wiener processes. If $F^1(\cdot)$ is absolutely continuous, with derivative $u(\cdot)$, then we say that $u(\cdot)$ is admissible if it is non negative, measurable and non anticipative. Define $V_\beta(x) = \inf_{F^1} C_\beta^N(x, F^1)$, where the inf is over the admissible controls. We say that $F^{1,N}(\cdot)$ has a *derivative* which is bounded by R if for each t and $s > 0$, $F^{1,N}(t+s) - F^{1,N}(t) \leq Rs + 1/\sqrt{N}$. The $1/\sqrt{N}$ term is needed since $F^{1,N}(\cdot)$ is piecewise constant with jumps $1/\sqrt{N}$.

Theorem 3.2. *Let $X^N(0) \Rightarrow X(0)$ (weak convergence) Then*

$$V_\beta^N(X^N(0)) \to V_\beta(X(0)). \tag{3.4}$$

Discussion of the Proof. By Theorem 3.3, it can be assumed without loss of generality that for each fixed $\epsilon > 0$, there is some set of $\epsilon-$optimal controls for (2.11), (3.2), with uniformly bounded derivatives; hence the set is tight. Owing to the discounting, as noted in the proof of Theorem 3.3, we can suppose that there is $T_\epsilon < \infty$ such that the $F^{1,N}(\cdot)$ do not change after time T_ϵ. It is sufficient to work with this set below.

Given $\epsilon > 0$, let $F^{1,N}(\cdot)$ be a sequence of $\epsilon-$optimal controls. Let $X^N(0) \to X(0)$. By Theorem 3.1, $\{X^{0,N}(\cdot), X^{1,N}(\cdot), F^{1,N}(\cdot)\}$ is tight. Let N_k index a weakly convergent subsequence with weak sense limits $(X^0(\cdot), X^1(\cdot), F^1(\cdot))$. Then, by the weak convergence and Fatou's Lemma

$$\liminf_k V_\beta^{N_k}(X^{N_k}(0)) \geq V_\beta(X(0)).$$

This and the $\epsilon-$optimality of $F^{1,N}(\cdot)$ for each N implies that

$$\begin{aligned}\epsilon + \liminf_N V_\beta^N(X^N(0)) &\geq \liminf_N C_\beta^N(X^N(0), F^{1,N}) \\ &\geq C_\beta(X(0), F^1) \geq V_\beta(X(0)).\end{aligned} \tag{3.5}$$

Since ϵ is arbitrary, (3.5) implies that

$$\liminf_N V_\beta^N(X^N(0)) \geq V_\beta(X(0)). \tag{3.6}$$

Now we prove the reverse inequality to (3.5), namely,

$$\limsup_N V_\beta^N(X^N(0)) \leq V_\beta(X(0)). \tag{3.7}$$

To do this we apply the following widely useful approach. Note that, for each initial condition, the distribution of the limit process $X(\cdot)$ depends on the pair

$(F^1(\cdot), B(\cdot))$. For each fixed $\epsilon > 0$, we first find an $\epsilon-$optimal pair $(F^{\epsilon,1}(\cdot), B^\epsilon(\cdot))$ for the limit process with the property that there is a sequence of admissible controls $F^{\epsilon,1,N}(\cdot)$ for (2.11) such that (define $B^{\epsilon,N}(\cdot)$ as $B^N(\cdot)$ was defined but under control $F^{\epsilon,1,N}(\cdot)$) $\{X^N(0), F^{\epsilon,1,N}(\cdot), B^{\epsilon,N}(\cdot))\}$ converges weakly to $(X(0), F^{\epsilon,1}(\cdot), B(\cdot))$ and also that

$$C_\beta^N(X^N(0), F^{\epsilon,1,N}) \to C_\beta(X(0), F^{\epsilon,1}) \leq V_\beta(X(0)) + \epsilon. \tag{3.8}$$

For any $\epsilon > 0$, there is an $F^{\epsilon,1}(\cdot)$ and a sequence of admissible controls $\{F^{\epsilon,1,N}(\cdot)\}$ satisfying the requirement (3.8). See, e.g., [28, Section 5]. The sequence of controls given by the reference is admissible and the $F^{\epsilon,1,N}(T_\epsilon)$ are uniformly bounded. It does not necessarily have a bounded derivative, and might not even be tight in the Skorohod topology. [A time transformation methods was used in the reference to circumvent the tightness problem. But the method used here is simpler in the current case, since Theorem 3.3 shows how to alter the control sequence (without loss of generality) so that we have bounded derivatives (hence tightness) with (3.8) holding (with perhaps ϵ replaced by 2ϵ).]

Now, (3.8) and the fact that (due to the non optimality of $F^{\epsilon,1,N}(\cdot)$)

$$V_\beta^N(X^N(0)) \leq C_\beta^N(X^N(0), F^{\epsilon,1,N})$$

yields (3.4). ■

Comment on the optimal controls. The comments on the form of the optimal control in Section 7 also hold for the discounted cost problem. In particular, numerical data show that there is a piecewise linear or nearly linear switching curve such that the optimal control is to reject above and accept below, with any decision allowed when on the curve. For large N, this control will be nearly optimal for the physical system. These comments on the shape of the switching curves are based on (very consistent) numerical computations, but not on proofs.

Comment on tightness. An $\epsilon-$optimal sequence $\{F^{1,N}(\cdot), N\}$ need not be tight in general. Let $X^{1,N}(t_0) > 0$. Consider the example of a control which rejects until $X^{1,N}(t)$ reaches the value zero, and then stops. Since the required time for $X^{1,N}(\cdot)$ to reach zero is of the order of $X^{1,N}(t_0)/\sqrt{N}$, the control clearly converges to a step function in an obvious way. But, owing to the fact that it increases in small steps (of size $1/\sqrt{N}$), the sequence is not tight in the Skorohod topology. There are many ways of dealing with this problem. We can adapt the time transformation method of [26, 24]. But, owing to the relative simplicity of the dynamics and cost function there is a simpler way, which avoids the extra notation and concepts.

Theorem 3.3. *It can be supposed that the controls in Theorems* 3.1 *and* 3.2 *are tight. In fact, it can be supposed that they have uniformly bounded derivatives for each* $\epsilon > 0$.

Proof. Fix $\epsilon > 0$. Since $B^0_+ < \infty$, $\{k(X^{0,N}(t)), k(X^0(t)); N, t\}$ is uniformly bounded. [5] Thus, due to the discounting, there is $T_\epsilon < \infty$ and a sequence of $\epsilon/2-$optimal controls which do not change after time T_ϵ. Note that the control which is identically zero has uniformly (in N) bounded costs.

The existence of $T_\epsilon < \infty$ can be seen from the following argument. From any time T on, and with no control after that time, the limit cost is

$$E\int_T^\infty e^{-\beta t}k(X^0(t)dt + E\int_T^\infty e^{-\beta t}c_0 dF^0(t).$$

The first term obviously goes to zero as $T \to \infty$, uniformly in the past values of the control. The same thing can be said of the second term, owing to the properties of the solution to (3.2a). A similar argument applies to the physical system and controls $F^{1,N}(\cdot)$. We need only work with controls for which the sequence of costs is uniformly bounded.

For large enough $K < \infty$, the sequence defined by $F^{1,N}_K(\cdot) = F^{1,N}(\cdot) \wedge K$ will be $3\epsilon/4-$optimal, and similarly for $F^1_K(\cdot) = F^1(\cdot) \wedge K$. To see this, note that whatever the controls in the previous paragraph are, the boundedness of the costs imply that

$$\sup_N EF^{1,N}(T_\epsilon) < \infty, \quad EF^1(T_\epsilon) < \infty. \tag{3.9}$$

Furthermore,

$$\begin{array}{c} \lim_K \sup_N P\left\{\left[F^{1,N}(T_\epsilon) - F^{1,N}(T_\epsilon) \wedge K\right] \neq 0\right\} = 0, \\ \lim_K P\left\{\left[F^1(T_\epsilon) - F^1(T_\epsilon) \wedge K\right] \neq 0\right\} = 0. \end{array} \tag{3.10}$$

(3.9) and (3.10) and a straightforward analysis using Fatou's Lemma implies that the costs and systems associated with the use of $F^{i,N}(\cdot) \wedge K$ (resp, $F^i(\cdot) \wedge K$ for the limit system) are asymptotically (as $K \to \infty$) no worse (uniformly in N) than the costs for the original untruncated controls.

Now that we know that there are $3\epsilon/4-$optimal controls $F^{1,N}(\cdot), F^1(\cdot)$ which do not increase after T_ϵ and that are uniformly bounded, we can show that we can bound the derivative as well: More precisely, we can show that there are $R < \infty$ and $\epsilon-$optimal controls which satisfy:

$$F^1_R(t+s) - F^1_R(t) \leq Rs, \tag{3.11}$$

$$F^{1,N}_R(t+s) - F^{1,N}_R(t) \leq Rs + 1/\sqrt{N}, \tag{3.12}$$

for all $t, s > 0$. In particular, let $F^1_R(\cdot)$ denote the largest control which satisfies (3.11), but is no greater than $F^1(\cdot)$, and let $F^{1,N}_R(\cdot)$ be the largest control which satisfies (3.12) and which is no greater than $F^{1,N}(\cdot)$.

[5] Even if $B^0_+ = \infty$, by imposing a growth rate $k(x^0) = O(|x^0|^{1+\delta})$ for large x^0, $0 \leq \delta < 1$, and assuming that $\sup_N E|X^{0,N}(0)|^2 < \infty$, $Ek(X^{0,N}(t))$ is at most $O(t^2)$ for large t, uniformly in the control (and analogously for the limit system).

Since $K < \infty$ and R is as large as desired, any jump in $F^{1,N}(\cdot)$ can be reached by $F_R^{1,N}(\cdot)$ in an arbitrarily short (uniformly in N and in the realization) time afterwards Thus, excluding a set of measure which goes to zero (uniformly in N) as $R \to \infty$, $F^{1,N}(t) - F_R^{1,N}(t)$ goes to zero (uniformly in N) as $R \to \infty$. This and the boundedness of the F functions imply that $X^{1,N}(t) - X_R^{1,N}(t)$ is bounded and (excluding sets of arbitrarily small measure) converges to zero uniformly in N. Similar remarks hold for $F^{1,N}(\cdot), F_R^{1,N}(\cdot)$. These results imply that the costs converge as well.

The $1/\sqrt{N}$ term appears in (3.12) since $F^{1,N}(\cdot)$ is piecewise constant with an increment of $1/\sqrt{N}$ at each rejection. ■

4 Upper Limit to the Bandwidth for the (BE) Sharing Customers

In the model of the previous two sections, the class 0 customers shared the available bandwidth, whatever it was, and used *all* of it. In general, it might not be possible for all of the available bandwidth to be used. For example, there might be local restrictions on the rate at which data can enter the channel buffer (e.g., bounded modem speed, etc.). This possibility changes the problem a little, and we will indicate the few required adjustments. Such examples are further illustrations of the versatility of the approach.

Suppose that the *maximum* bandwidth that any single class 0 customer can use is C_0. The main difference in the development concerns the departure process for class 0 customers and the structure of the appropriate cost function. We now *redefine*

$$X^{0,N}(t) = \frac{\#\text{ of class 0 customers at time } t - N\lambda^{0,N}/(C_0\mu^{0,N})}{\sqrt{N}}.$$

Note that now $X^{0,N}(t)$ is centered around a *mean value,* assuming that each class 0 customer uses exactly C_0 units of bandwidth. In Sections 2 and 3, the number of class 0 customers in the system was $O(\sqrt{N})$, and $X^{0,N}(t)$ measured that actual number, scaled by $1/\sqrt{N}$. Now, the number in the system will be $O(N)$, and $X^{0,N}(t)$ measures the deviation from the mean number, scaled by $1/\sqrt{N}$. We suppose that class 0 customers are rejected if $X^{0,N}(t) \geq B_+^0$, where $B_+^0 < \infty$. Theoretical and numerical results show that this will have negligable effect if B_+^0 is large.

The martingale decomposition (2.9) remains valid, but the instantaneous conditional mean departure rate is different, being determined by whether or not the available capacity per class 0 customer is greater than C_0. The conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at time t is

$$\frac{\mu^{0,N}}{\sqrt{N}}[\#\text{ of class 0 in system at } t]\times\min\left[\frac{\text{available BW at } t}{\#\text{ of class 0 in system at } t},\ C_0\right] I^{0,N}(t),$$

which equals

$$
\begin{aligned}
&I^{0,N}(t)\frac{\mu^{0,N}}{\sqrt{N}} \times \min\left[\text{available BW at } t,\ C_0(\#\text{ of class 0 in system at } t)\right] \\
&= I^{0,N}(t)\frac{\mu^{0,N}}{\sqrt{N}} \times \min\left[\frac{N\lambda^{0,N}}{\mu^{0,N}} + b\sqrt{N} - \sqrt{N}X^{1,N}(t),\ C_0\left(\frac{\lambda^{0,N}N}{C_0\mu^{0,N}} + \sqrt{N}X^{0,N}(t)\right)\right] \\
&\quad = I^{0,N}(t)\lambda^{0,N}\sqrt{N} + I^{0,N}(t)\mu^{0,N}g_1(X^N(t),
\end{aligned}
\tag{4.1}
$$

where we define

$$
g_1(x) = \min\left[b - x^1, C_0x^0\right]. \tag{4.2}
$$

Now, analogously to what was done in Section 2, we can write the decomposition as

$$
D^{0,N}(t) = \lambda^{0,N}\sqrt{N}t + \int_0^t \mu^{0,N}g_1(X^N(s))ds + \tilde{D}^{0,N}(t) - Y^{0,N}(t),
$$

where the Doob-Meyer increasing process associated with the martingale $\tilde{D}^{0,N}(\cdot)$ is

$$
<\tilde{D}^{0,N}>(t) = \int_0^t \left[\lambda^{0,N} + \frac{1}{\sqrt{N}}\mu^{0,N}g_1(X^N(s))\right] I^{0,N}(s)ds.
$$

The dynamical equation is (2.11) with (2.11a) replaced by

$$
\begin{aligned}
X^{0,N}(t) &= X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) \\
&\quad -\mu^{0,N}\int_0^t g_1(X^N(s))ds - \tilde{D}^{0,N}(t) - F^{0,N}(t) + Y^{0,N}(t),
\end{aligned}
\tag{4.3}
$$

and the limit equation is (3.2) with (3.2a) replaced by

$$
X^0(t) = X^0(0) - \mu_0\int_0^t g_1(X(s))ds + B^0(t) - F^0(t). \tag{4.4}
$$

In the cost function (3.1), the function $k(\cdot)$ was assumed to be non negative. This made sense since $X^{0,N}(t)$ was non negative. Now, since $X^{0,N}(t)$ can take any sign, we suppose that $k(x^0)$ takes the sign of x^0 and is zero if $x^0 = 0$. Note that for large negative x^0, the departure rate is essentially limited by the C_0 limitation, and the control has little effect. Also, suppose that

$$
|k(x^0)| = O(|x^0|), \quad \sup_N E\left|X^{0,N}(0)\right|^2 < \infty. \tag{4.5}
$$

Under the given conditions, Theorems 3.1 to 3.3 hold.

The savings in waiting time for the class 0 customers (for the controlled problem) are of the order of that of the model in Sections 2 and 3. But, since there are many more customers in the system at any time, the savings per customer are less. We are essentially concerned with "marginal" savings, at a "marginal" cost.

Comments on the proofs. The proofs outlined in Section 3 work here as well, with essentially the same details. The only differences are due to the fact that in the present case the $X^{0,N}(t)$ are not bounded below. But in the proofs, the second moment bound

$$\sup_{N,t,F^{1,N}} E\left|X^{0,N}(t)\right|^2 < \infty \tag{4.6}$$

is used in lieu of the zero lower bound. (4.6) is proved by use of a dominating system. The second moments are bounded by a constant plus the second moments of the following system, which is defined on the interval $(-\infty, 0]$, and the reflection term $F^{0,N}(\cdot)$ now acts at the origin, and keeps the state non positive:

$$X^{0,N}(t) = X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) - \mu^{0,N}\int_0^t X^{0,N}(s)ds - \tilde{D}^{0,N}(t) - F^{0,N}(t). \tag{4.7}$$

The proof of (4.6) for the model (4.7) is done by a Liapunov function technique, and can be found in [22, page 771]

5 Extension. Several BE and GP Subclasses

The developments in the previous sections can be extended to the case where there are multiple subclasses of any of the classes (and similarly for the models in the subsequent sections). We will illustrate only one of the many possibilities, working with the setup in Sections 2 and 3. Suppose that there are now two types of BE (class 0) customers, called class 01 and class 02, with parameters $\lambda^{0i,N}, \mu^{0i,N}, i = 1, 2$. We suppose that the natural analogs of the conditions in Sections 2 and 3 hold. There is a surprisng and very useful degeneracy, which simplifies both the analytical problem and the numerical analysis.

Analogously to the formula (2.1), we let the channel capacity be

$$C_N = N\left[\sum_i \frac{\lambda^{0i,N}}{\mu^{0i,N}} + \frac{\lambda^{1,N}}{\mu^{1,N}}\right] + \sqrt{N}b. \tag{5.1}$$

The bandwidth available at time t for both subclasses 01 and 02 is

$$\sum_i N\frac{\lambda^{0i,N}}{\mu^{0i,N}} - \sqrt{N}X^{1,N}(t) + \sqrt{N}b, \quad b > 0.$$

The processes are defined analogously to what was done in Section 2. E.g., $D^{0i,N}(\cdot)$ is the number of departures of subclass $0i$ by time t, divided by $\sqrt{N}$.

Until otherwise noted, let us assume that the available bandwidth is shared equally among all class 0 customers, irrespective of the subclass. Then the total conditional mean instantaneous rate at which $D^{0i,N}(\cdot)$ increases at t is (following the idea used in Section 2 and (assuming that there are class 0 customers in the system)

$$\frac{\mu^{0i,N}}{\sqrt{N}}[\#\text{ of subclass } 0i \text{ in system at } t]\frac{\text{avail BW at } t}{\text{total } \#\text{ of class 0 in system at } t},$$

which equals

$$\mu^{0i,N}X^{0i,N}(t)\frac{\sum_j N\frac{\lambda^{0j,N}}{\mu^{0j,N}} - \sqrt{N}X^{1,N}(t) + \sqrt{N}b}{\sqrt{N}\sum_j X^{0j,N}(t)}. \tag{5.2}$$

Define

$$\bar{a}^N = \frac{\lambda^{01,N}/\mu^{01,N}}{\sum_j \lambda^{0j,N}/\mu^{0j,N}}, \quad \bar{a} = \lim_N \bar{a}^N. \tag{5.3}$$

The system is degenerate in that if the costs are bounded in N, then the ratios $X^{01,N}(t)/(X^{01,N}(t) + X^{02,N}(t))$ converge to $\bar{a}$ as $N \to \infty$. Thus we need only analyze the system with class 1 and one of the subclasses $0i$. Only an informal argument will be given. We note in passsing that this convergence relation is an example of what is called state space collapse in the heavy traffic analysis of queueing systems. It is not the usual type, which is concerned with multiclass queues under the workload formulation.

Let us examine the mean rates of increase of $A^{0i,N}(\cdot)$ and $D^{0i,N}(\cdot)$. The "mean rate" at which $A^{0i,N}(\cdot)$ increases is $\sqrt{N}\lambda^{0i,N}$. Define $B^N = \sum_i[\lambda^{0i,N}/\mu^{0i,N}]$. Set

$$a^N(t) = \frac{X^{01,N}(t)}{X^{01,N}(t) + X^{02,N}(t)}.$$

Using the fact that the available bandwidth is partitioned equally among all the class 0 customers, the conditional mean instantaneous rates at which $D^{0i,N}(\cdot), i = 1, 2$, resp., increase at t are, resp.,

$$\left[\sqrt{N}B^N - X^{1,N}(t) + b\right]\mu^{01,N}a^N(t),$$

$$\left[\sqrt{N}B^N - X^{1,N}(t) + b\right]\mu^{02,N}\left(1 - a^N(t)\right).$$

The differences of the dominant arrival and departure terms for the two subclasses are, resp.,

$$\sqrt{N}\left[\lambda^{01,N} - B^N\mu^{01,N}a^N(t)\right], \tag{5.4a}$$

$$\sqrt{N}\left[\lambda^{02,N} - B^N\mu^{02,N}\left(1 - a^N(t)\right)\right]. \tag{5.4b}$$

For the case of Section 2, the analogs of these terms have the value zero.

Let $\epsilon > 0$ be small. If, for large N, $a^N(t) \notin [\bar{a} - \epsilon, \bar{a} + \epsilon]$, then (5.4a) implies that there is a large force (of the order of $\sqrt{N}$) returning it to this interval. Similarly, (5.4b) implies that $(1 - \alpha^N(t))$ must be very close to $(1 - \bar{a})$. The contribution of the non dominant terms is relatively small in comparison.

The degeneracy situation is similar if there are more than 2 subclasses.

Many interesting variations of the multiple subclass problem can be analyzed. For example, we might wish to alter the above formulation to allow each of the 0i subclasses a different fraction of the available bandwidth. More concretely, suppose that there are positive numbers k_i such that for each unit of bandwidth allocated to a customer of subclass 01, we allocate k_2/k_1 units of

bandwidth to each customer of subclass 02. Then the dominant term in the conditional mean instantaneous rate at which $D^{0i,N}(\cdot)$ increases at t is

$$\mu^{0i,N} k_i X^{0i,N}(t) \frac{B^N N}{\sqrt{N}(k_1 X^{01,N}(t) + k_2 X^{02,N}(t))}. \tag{5.5}$$

Redefine $a^N(t)$:

$$a^N(t) = \frac{k_1 X^{01,N}(t)}{\sum_j k_i X^{0i,N}(t)}. \tag{5.6}$$

Then, the difference between the dominant input and output terms is (5.4), but with the new value of $a^N(t)$ used. Thus, we see that the new value of $a^N(t)$ is very close to $\bar{a}$ for large N.

Note that the weighted fair-share for different types of BE customers in the above equations is the one defined by the ATM forum for sharing bandwidth among ABR users (see the Section I.3 in Appendix I "Implementation Examples on ABR Service" [1]).

6 Multicast: The Limit Dynamical Equations and The Discounted Cost Function

Now, consider the case where there are two channels. There are three classes of customers. Class 0 is as in Section 2, but must be transmitted simultaneously and with the same instantaneous rate on both channels. Class i, $i = 1, 2$, is to be transmitted on channel i only. We make the natural analogs of the assumptions of Sections 2 and 3, defining $A^{i,N}(\cdot), D^{i,N}(\cdot), X^{i,N}(\cdot), i = 0, 1, 2$, etc., analogously to what was done there. Analogously to (2.1), the capacity of channel i is assumed to be

$$C_N^i = N \sum_i \frac{\lambda^{i,N}}{\mu^{i,N}} + b_i \sqrt{N}, \quad b_i > 0. \tag{6.1}$$

Any number of channels and 0 subclasses could also be used, with an arbitrary assignment of the subclasses to the channels.

At time t, the bandwidth available for class 0 customers on channel i is

$$N \frac{\lambda^{0,N}}{\mu^{0,N}} + b^i \sqrt{N} - X^{i,N}(t) \sqrt{N}.$$

Thus, the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is determined by the channel with the largest available bandwidth and (analo-

gously to what was done in Section 2) is

$$\frac{\mu^{0,N}\sqrt{N}X^{0,N}(t)}{\sqrt{N}}\times$$
$$\min\left[\frac{N\frac{\lambda^{0,N}}{\mu^{0,N}}+b_1\sqrt{N}-X^{1,N}(t)\sqrt{N}}{\sqrt{N}X^{0,N}(t)},\ \frac{N\frac{\lambda^{0,N}}{\mu^{0,N}}+b_2\sqrt{N}-X^{2,N}(t)\sqrt{N}}{\sqrt{N}X^{0,N}(t)}\right]I^{0,N}(t). \tag{6.2}$$

This equals

$$\left[\sqrt{N}\lambda^{0,N}+\mu^{0,N}g_2(X^N(t))\right]I^{0,N}(t), \tag{6.3}$$

where we define

$$g_2(x)=\min\left[b_1-X^{1,N}(t),\ b_2-X^{2,N}(t)\right]. \tag{6.4}$$

Thus analogously to what was done in Section 2, we can write

$$D^{0,N}(t)=\lambda^{0,N}\sqrt{N}t+\int_0^t g_2(X^N(s))ds+\tilde{D}^{0,N}(t)-Y^{0,N}(t),$$

where the Doob-Meyer increasing process associated with the martingale is

$$<\tilde{D}^{0,N}>(t)=\int_0^t\left[\lambda^{0,N}+\frac{\mu^{0,N}}{\sqrt{N}}g_2(X^N(s))\right]I^{0,N}(s)ds.$$

The analog of (2.11) is

$$\begin{aligned}X^{0,N}(t)=X^{0,N}(0)+W^{0,N}(S^{0,N}(t))-\mu^{0,N}\int_0^t g_2(X^N(s))ds\\ -\tilde{D}^{0,N}(t)+Y^{0,N}(t)-F^{0,N}(t),\end{aligned} \tag{6.6a}$$

and for $i=1,2$,

$$\begin{aligned}X^{i,N}(t)=X^{i,N}(0)+W^{i,N}(S^{i,N}(t))-\mu^{i,N}\int_0^t X^{i,N}(s)ds\\ -\tilde{D}^{i,N}(t)-F^{i,N}(t)-U^{i,N}(t),\end{aligned} \tag{6.6b}$$

where the $Y^{0,N}(t)$ term compensates for the fact that there are no departures of class 0 customers at time t if $X^{0,N}(t)=0$, and $U^{i,N}(t)$ compensates for the class i arrivals lost due to a full system (when the entire channel is occupied by class i customers).

The discounted cost function is still (3.1), but now the sum has three terms. The analysis given in Section 3 holds here in the same way and the limit equations are

$$X^0(t)=X^0(0)-\mu_0\int_0^t g_2(X(s))ds+B^0(t)+Y^0(t)-F^0(t), \tag{6.7a}$$

$$X^i(t) = X^i(0) - \mu_i \int_0^t X^i(s)ds + B^i(t) - F^i(t), i = 1, 2, \tag{6.7b}$$

where the $B^i(\cdot)$ are mutually independent Wiener processes with variance parameters $\lambda_i(\sigma_i^2 + 1)$. The associated cost function is (3.3), and we still have

$$V_\beta^N(X^N(0)) \to V_\beta(X(0)), \tag{6.8}$$

if $X^N(0) \Rightarrow X(0)$.

The model of Section 4. Now, suppose that each class 0 customer can use at most a bandwidth C_0. Then define $X^{0,N}(t)$ as in Section 4, and let $B_+^0 < \infty$. The development is a combination of those of Sections 3 and 4. Now, the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is obtained as the minimum of three terms, depending on whether the available capacities in channels 1, 2, or C_0 is the limiting factor. It is

$$\begin{aligned}&\frac{\mu^{0,N}}{\sqrt{N}}[\#\text{ of class 0 in system at } t]\\ &\min\left[\frac{\text{avail BW in ch 1 at } t}{\#\text{ of class 0 in system at } t}, \frac{\text{avail BW in ch 2 at } t}{\#\text{ of class 0 in system at } t}, C_0\right] I^{0,N}(t).\end{aligned} \tag{6.9}$$

Define

$$g_3(x) = \min\left[b_1 - x^1, b_2 - x^2, C_0 x^0\right]. \tag{6.10}$$

Then (6.9) can be written as

$$\left[\sqrt{N}\lambda^{0,N} + \mu^{0,N} g_3(X^N(t))\right] I^{0,N}(t).$$

All of the previous results continue to hold with $g_3(\cdot)$ replacing $g_2(\cdot)$.

7 The Ergodic Cost Function: The Basic Model: Bounded Control Rate

For concreteness, we work with the system model and assumptions of Sections 2 and 3, although all of the results hold for all of the other models. In this section, we will suppose that the controls $F^{1,N}(\cdot)$ and $F^1(\cdot)$ (for the limit system) have bounded derivatives in the sense used in Theorem 3.3, as follows: There is a constant R, which can be as large as we wish, such that $\dot{F}^1(t) \le R$ for all t, and for all $t, s > 0$, $F^{1,N}(t+s) - F^{1,N}(t) \le Rs + 1/\sqrt{N}$. Thus, the maximum "rate" of refusing admission to class 1 customers is bounded by $\sqrt{N}R$.

The reasonableness of the bounded derivative assumption is also seen from the form of the limit equation (3.2), which (informally) suggests that one loses very little by bounding the derivative of $F^1(\cdot)$. Furthermore, it is completely borne out by our numerical data. The next section shows that we can make this assumption in the proofs with no loss of generality.

Although useful in applications, the mathematical reason for the assumption of bounded control "derivatives" concerns the mathematics of the ergodic cost problem. Little is known about the ergodic cost problem for the limit system when the control functions are arbitrary right continuous functions. But a great deal is known when they have uniformly bounded derivatives. In that case, for the current non degenerate model (3.2), there is an optimal feedback control which is time independent and the optimal value $\bar{\gamma}_R(x)$ (defined below) does not depend on x. More importantly, for our purposes, for any $\epsilon > 0$, there is an $\epsilon-$optimal time independent feedback control $u^\epsilon(\cdot)$ such that $u^\epsilon(\cdot) = \dot{F}^{\epsilon,1}(\cdot)$ is arbitrarily smooth, and under which there is a unique stationary measure. The $F^{\epsilon,1}(\cdot)$ plays the role of the $F^{1,\epsilon}(\cdot)$ in Theorem 3.2. The basic convergence results are quite technical. They are in [20] for the unconstrained (no reflecting boundaries) problem, with extensions to the constrained problem being in [22, 23]. Indeed, under our basic setup, the needed convergence results can be obtained from [23] by appropriate identification of terms.

Define the cost functions

$$\mathcal{C}^N(X^N(0), T, F^{1,N}) = \int_0^T k(X^{0,N}(s))ds + \sum_i c_i F^{i,N}(T),$$

$$\bar{\gamma}_R^N(X^N(0), T) = \inf_{F^{1,N}} E\mathcal{C}(X^N(0), T, F^{1,N})/T,$$

$$\bar{\gamma}_R^N(X^N(0)) = \limsup_T \bar{\gamma}_R^N(X^N(0), T).$$

For the limit system, define the analogous quantities, with the N dropped. If there is no rate R restriction, we drop the subscript R. We also suppose that (with little loss of generality)

$$\sup_N E|X^{1,N}(0)|^2 < \infty. \tag{7.1}$$

In Section 3, it was shown that, for the discounted cost problem and large enough R, we can get as close to optimality as we wish. The proof in Theorem 3.3 used the fact that the discounting implied that we need concern ourselves only with a finite time interval. The proof is more subtle for the ergodic cost problem, and is given in the next section.

The results in [23] will apply if we have tightness of the doubly indexed (both t and N are indices now) set of processes

$$\left\{X^N(t+\cdot), B^N(t+\cdot) - B^N(t), F^{1,N}(t+\cdot) - F^{1,N}(t); N, t\right\}. \tag{7.2}$$

The tightness holds for the set of $F^{1,N}(t+\cdot) - F^{1,N}(t)$ processes by the assumption on boundedness of the derivative. The set $\{W^{i,N}(S^{i,N}(t+\cdot)) - W^{i,N}(S^{i,N}(t); N, t\}$ is tight due to the independence properties of the interarrival intervals, (2.2), the weak convergence of the $S^{i,N}(t+\cdot) - S^{i,N}(t)$, as $N \to \infty$ and for any sequence t, and the use of the criterion (2.14), (2.15).

A standard Liapunov function argument (using the Liapunov function $|X^{1,N}|^2$) and the "$R-$derivative" restrictions on the controls can be used to prove directly that

$$\sup_{t,N,F^{1,N}} E\left|X^{1,N}(t)\right|^2 < \infty. \tag{7.3}$$

Here, the sup is over the $F^{1,N}$ which satisfy the $R-$derivative restriction. Then, tightness can be shown for the set of $\tilde{D}^{i,N}(t+\cdot) - \tilde{D}^{i,N}(t)$ processes by a direct application of the criterion (2.14), (2.15) and the use of (7.3) to bound the expectation of the Doob-Meyer processes associated with $\tilde{D}^{i,N}(t+\cdot) - \tilde{D}^{i,N}(t)$. The proof of the tightness of the doubly indexed sequence $\{X^{i,N}(t+\cdot); N, t\}$ is then the same as the proof of tightness of $\{X^{i,N}(\cdot), t\}$ where the initial conditions vary over a tight set (the $X^{0,N}(t)$ are bounded by B_0^+).

Given the tightness of (7.2) and the non degeneracy of the limit system (3.2) (the set of driving Wiener processes is non degenerate, in fact the components are mutually independent with positive variances), the following results follow by a direct application of the results and ideas in [23].

$$\bar{\gamma}_R^N(X^N(0), T) \to \bar{\gamma}_R, \tag{7.4}$$

as $T \to \infty$ and $N \to \infty$ in any way at all, where $\bar{\gamma}_R$ is the infimum of the costs for the limit system over controls with derivatives bounded by R, and it does not depend on the initial condition.

Furthermore, for any $\epsilon > 0$,

$$\lim_{N,T} P\left\{\frac{\mathcal{C}^N(X^N(0), T, F^{1,N})}{T} \le \bar{\gamma}_R - \epsilon\right\} = 0, \tag{7.5}$$

where N, T can go to their limits in any way at all, and $F^{1,N}(\cdot)$ is an arbitrary sequence of controls. There is a converse to (7.5), which says that a good control for the limit system is a good control for the physical system. Given $\epsilon > 0$, let $F^{\epsilon,1}(\cdot)$ be an $\epsilon-$optimal control with smooth derivative $u^\epsilon(\cdot)$ and an adaptation $F^{\epsilon,1,N}(\cdot)$ to the physical system such that

$$\lim_{N,T} P\left\{\frac{\mathcal{C}^N(X^N(0), T, F^{\epsilon,1,N})}{T} \ge \bar{\gamma}_R + 2\epsilon\right\} = 0, \tag{7.6}$$

The results of the next section imply that we can replace $\bar{\gamma}_R$ in (7.5) and (7.6) by $\bar{\gamma}$

The control $u^\epsilon(\cdot)$ can be adapted for use on the physical system in many ways, for large N. For example, by rejecting an arrival of class 1 at t with probability (conditioned on the past system data) $u^\epsilon(X^N(t))/[\lambda^{1,N}\sqrt{N}]$. Alternatively, we need not have the rejection choices being random, provided that $\sqrt{N}$ times the number rejected when the state is "near" x converges to $u^\epsilon(x)$ as $N \to \infty$.

Comments on the controls. Numerical data show that the derivative $\dot{F}(t) = u(t)$ of the optimal control takes either the value R or zero, with the regions

separated by a piecewise linear or nearly linear switching curve. One applies this control to the physical system as in the last paragraph. [This procedure is asymptotically equivalent to rejecting all arrivals when the state is above the switching curve.] Equation (7.6) holds for such discontinuous controls as well. This is important in applications since such controls are easily implemented. Numerical data show that the switching curves converge nicely to piecewise smooth (or even linear) curves as $R \to \infty$. (7.6) holds for this curve a well. Then we reject all arrivals of class 1 when the state is above the switching curve. Analogous remarks hold for the discounted cost problem.

Comment. Note that both (7.5) and (7.6) deal with pathwise average costs, not with average costs. Since any application is a single realization, the convergence of pathwise average costs is more important than the convergence of expectations. The inequalities (7.5) and (7.6) say that for large N, the optimal controls for the physical problem are (asymptotically) only negligibly better than the use of a nice almost optimal control for the limit system.

Finally, we simply note without further comment that the methods in [20, 22, 23] can be adapted to prove that $\lim_{N,\beta\to 0} \beta V_\beta^N(X^N(0)) = \bar{\gamma}$.

8 The Ergodic Cost Problem: The Basic Model: Arbitrary Controls

Now, return to the problem of bounded derivative controls. We will show that we can approximate any optimal or nearly optimal control by a control with which has a "derivative" bounded by R, for large enough R. In particular:

Theorem 8.1.

$$\lim_R [\bar{\gamma} - \bar{\gamma}_R] = 0. \tag{8.1a}$$

$$\lim_R \limsup_N \left[\bar{\gamma}^N(x) - \bar{\gamma}_R^N(x)\right] = 0. \tag{8.1b}$$

Proof. A detailed outline of the steps will be given. Unlike as in Theorem 3.3, we can not restrict ourselves to a finite interval. We need to show that for any $\delta > 0$, there is $R_\delta < \infty$ such that there are $\delta-$optimal controls for both the physical and the limit system with bounded rate R_δ.

The development proceeds in several steps. The steps will be outlined (informally to save space) for the physical system. The details are a little simpler for the limit system.

1. Given $\epsilon > 0$, show that there is a $B_\epsilon < \infty$ such that the optimal cost will change by no more than ϵ if we do not reject when $X^{1,N}(t) < -B_\epsilon$.

2. Let $\epsilon > 0$. Allowing only controls which do not reject if $X^{1,N}(t) < -B$, for some given $0 < B < \infty$, show that there is $K_\epsilon < \infty$ such that if we further restrict the controls such that the increments $F^{1,N}(n+1) - F^{1,N}(n)$ are at most K_ϵ for all n, N, then the optimal cost will change by no more than ϵ.

3. Let $\epsilon > 0$. Allowing only with controls satisfying the restrictions of the first two steps for some finite B, K, show that there is $R_\epsilon < \infty$ such that the optimal cost will change by at most ϵ if we further restrict the controls to have maximum "derivative" R_ϵ.

Step 1 is the least difficult to accept even without a proof, since it is quite reasonable that there is a $B < \infty$ such that an optimal or nearly optimal control that would not reject if $X^{1,N}(t) \leq -B$. The proof is a formalization of the following idea. Given a control $F^{1,N}(\cdot)$ and a $B > 0$, define another control $F_B^{1,N}(t) \leq F^{1,N}(t)$ where $F_B^{1,N}(t)$ is as close as possible to $F^{1,N}(t)$, but acts only when $X^{1,N}(t) \geq -B$. For large B, the change in $X^{0,N}(\cdot)$ is slight. To save space, we concentrate on the outline for the other steps.

Thus, we start by supposing that there is $0 < B < \infty$ such that there are no rejections if $X^{1,N}(t) \leq -B$. We will show that, given $\epsilon > 0$, there is $K_\epsilon < \infty$, such that we lose less than ϵ in the cost if we restrict the control to satisfy $F^{1,N}(n+1) - F^{1,N}(n) \leq K_\epsilon$ for all N, n, ω.

Owing to the fact that we do not reject if $X^{1,N}(t) \leq -B$, a Liapunov function argument can be used to get that there is $C < \infty$ such that

$$\sup_{N,t,F^{1,N}} E\left|X^{1,N}(t)\right|^2 \leq C. \tag{8.2}$$

Also, the same $-B$ restriction and (8.2) can be used to show that

$$\sup_{n,N,F^{1,N}} E\left[F^{1,N}(n+1) - F^{1,N}(n)\right]^2 < \infty. \tag{8.3}$$

The sup in (8.2) is over all controls satisfying our $-B$ restriction. The proof of (8.3) computes a worst case on each interval, which is a control taking $X^{1,N}(n)$ satisfying only (8.2) to $-B$ as quickly as possible, then keeping it there until the end of the interval, and repeating on the next interval, etc. The uniform mean square boundedness of the part due to keeping $X^{1,N}(\cdot)$ at $-B$ on $[n, n+1]$ follows from the reflection mapping and the mean square bounds on the martingales driving (2.11b). For the reflection mapping and the Lipschitz continuity of the reflection term as a function of the driving processes, see [8],[7, Proposition 2.1].

Given any $F^{1,N}(\cdot)$ satisfying our restriction, we proceed to approximate it by bounding the increments by K. The approximation will be denoted by $F_K^{1,N}(\cdot)$, and the associated processes denoted by $X_K^N(\cdot)$. Define $F_K^{1,N}(\cdot)$ such that it satisfies the restriction $F_K^{1,N}(n+1) - F_K^{1,N}(n) \leq K$, it is no greater than $F^{1,N}(\cdot)$, and tries to keep $X_K^{1,N}(\cdot)$ as close as possible to $X^{1,N}(\cdot)$.

We always have $X_K^{1,N}(t) \geq X^{1,N}(t)$, hence $X_K^{0,N}(t) \geq X^{0,N}(t)$. It is not hard to see that, for large enough K, $X_K^{1,N}(\cdot)$ will repeatedly catch up to and equal $X^{1,N}(\cdot)$. We can decompose time into successive intervals where $X^{1,N}(t) < X_K^{1,N}(t)$, and where $X^{1,N}(t) = X_K^{1,N}(t)$. The key to the proof of step 2 is the observation that, as $K \to \infty$, a larger percentage of time will be taken up by the latter intervals. More precisely, for any $T_0 < \infty$, it can be shown that

$$\lim_K \limsup_N \sup_{t,F^{1,N}} P\left\{X^{1,N}(s) \neq X_K^{1,N}(s) \text{ for some } s \in [t, t+T_0]\right\} = 0. \tag{8.4}$$

(8.4) follows from the observations made before it. Choose the control $F_K^{1,N}(\cdot)$ as described. Then, starting at time $t-k$ for large k, the probability that $X_K^{1,N}(\cdot)$ catches up to $X^{1,N}(\cdot)$ and equals it on $[t, t+T_0]$ goes to unity as $K \to \infty$.

Note that if $X_K^{0,N}(t) = 0$, then $X^{0,N}(t) = 0$. We next bound the "return times" to the boundary $x^0 = 0$. Indeed, it can be shown that

$$\lim_{T\to\infty} \limsup_N \sup_{t, F_K^{1,N}, K} \sup_\omega P\left\{X_K^{0,N}(t+s) \neq 0, \text{ for some } s \leq T \middle| \text{data to } t\right\} = 0. \tag{8.5}$$

This can be shown by a weak convergence argument, using the fact that it holds for the limit process, as follows. The worst case for proving (8.5) is where there is no control since the control only decreases $X^{0,N}(t)$. Thus, suppose that there are $\rho > 0$, t_n and $N_n \to \infty, T_n \to \infty$ such that (no control)

$$\limsup_n \sup_\omega P\left\{X^{0,N_n}(t_n+s) \neq 0, \text{ for some } s \leq T_n \middle| \text{data to } t_n\right\} \geq \rho. \tag{8.6a}$$

We will show a contradiction to (8.6a). Actually, it is more direct to show that the assertion

$$\limsup_n \sup_\omega P\left\{Y^{0,N_n}(t_n+T_n) - Y^{0,N_n}(t_n) = 0 \middle| \text{data to } t_n\right\} \geq \rho. \tag{8.6b}$$

is false. The falsity of (8.6b) implies the falsity of (8.6a).

A Liapaunov function argument using (7.1) and the fact that there is no control can be used to prove that

$$\sup_n E|X^{1,N_n}(t_n)|^2 < \infty. \tag{8.7}$$

Now, extract a weakly convergent subsequence of $X^{N_n}(t_n+\cdot)$, and note that its limit $X(\cdot)$ satisfies (3.2). The distribution of $X^1(0)$ depends on the selected convergent subsequence. But, owing to (8.7), $E|X^1(0)|^2$ is bounded uniformly in the selected convergent subsequence. Using this last fact, the properties of (3.2) and the weak convergence now imply that (8.6b) cannot hold unless $\rho = 0$. Now note that (a key point), if $X_K^{0,N}(t) = 0$ and $X^{1,N}(t) = X_K^{1,N}(t)$, then the two processes start again at t with equal initial values.

The above results imply the following. For any $\delta > 0$, with a probability arbitrarily close to one, the fraction of time that $|X^{0,N}(t) - X_K^{0,N}(t)| \geq \delta$ on any time interval goes to zero as $K \to \infty$, uniformly in the time interval and in (large) N. This implies that the change in the $k(\cdot)$ part of the cost can be made as small as desired by making K large enough. By construction, $F^{1,N}(t) \geq F_K^{1,N}(t)$, hence the control cost is no greater for the approximating control. We omit the details of the fact that

$$E[F_K^{0,N}(T) - F^{0,N}(T)]/T$$

can be made as small as desired by making K large. But it can be proved by a weak convergence argument and the facts established above.

Thus for large enough K, we lose as little as desired by restricting $F^{1,N}(\cdot)$ such that $F^{1,N}(n+1) - F^{1,N}(n) \leq K$ for all N, n. This completes step 2.

Now, we turn to step 3 and make a few comments concerning the $k(\cdot)$ component of the cost. Given a control $F^{1,N}(\cdot)$ satisfying the restrictions of steps 1 and 2 (with constants B and K, resp.), find a suitable approximation with a bounded "derivative". Let R denote the derivative bound. Define a control $F_R^{1,N}(\cdot)$ with derivative bounded by R, such that $F_R^{1,N}(t) \leq F^{1,N}(t)$, but where the associated process $X_R^{1,N}(\cdot)$ is allowed to catch up with $X^{1,N}(\cdot)$ when possible. It will catch up repeatedly, for large enough R. This is because the maximum number of rejects on any time interval of unit length is $2K/\sqrt{N}$. Since K is bounded and R large, except for an arbitrarily small time subinterval the number of rejects on any time interval $[n, n+1]$ can be made as close as desired to what is needed, uniformly in N, n. Additionally, $\left|X^{1,N}(t) - X_R^{1,N}(t)\right|$ is uniformly (in N, t) bounded. Thus, the values of $X^{1,N}(t)$ and $X_R^{1,N}(t)$ will be arbitrarily close when $X_R^{0,N}(t)$ (hence, $X^{0,N}(t)$) hits zero, or very shortly thereafter (at most a time $K/R + O(1/\sqrt{N})$ later).

Note that the approximation problem is more subtle than in step 2, since we cannot guarantee that $X_R^{1,N}(\cdot)$ will equal $X^{1,N}(\cdot)$ on longer and longer intervals. [E.g., if $F^{1,N}(\cdot)$ jumps periodically, or if the limit is singular with respect to Lebesgue measure.]

The following properties can be proved. First, by a Liapunov function argument, it can be shown that

$$\limsup_N \sup_{t, R, F_R^{1,N}} E\left|X_R^{1,N}(t)\right|^2 < \infty.$$

Using this, it can be shown that

$$\limsup_N \sup_{n, R, F_R^{1,N}} E \sup_{n \leq s \leq n+1} \left|X_R^{1,N}(s)\right|^2 < \infty, \tag{8.8}$$

with a similar estimate holding for the $X^{1,N}(\cdot)$. The above comments imply that

$$\lim_R \limsup_N \sup_\tau E \sup_{t \leq T} \int_\tau^{\tau+t} \left|X^{1,N}(s) - X_R^{1,N}(s)\right| ds = 0, \tag{8.9}$$

for any $T < \infty$, and where τ are stopping times. Now use an argument based on recurrence to $X^{0,N}(\cdot)$ to zero analogously to what was done in in step 2 to get that the $k(\cdot)-$costs are close for large R. Obviously, the component of the cost due to $F_R^{1,N}(\cdot)$ is no greater than that due to $F^{1,N}(\cdot)$. Again, by a weak convergence argument, it can be shown the overflow costs also converge, and the details are omitted.

9 Data

Some typical data is given in the tables below. The cost function is $c_0 EX^0(1) + EF^1(1) + 5EF^0(1)$, all stationary values. $B_+^0 = 6.4$, and larger values made little

difference. The individual components of the cost function are tabulated, and we write $EF^0(1) = 0$ if it is less than 10^{-4}. The fraction of lost class 1 customers is $EF^1(1)/[\lambda_1\sqrt{N}]$, so it depends on N. The tables indicate the potential tradeoffs between the time gained for class 0 and the lost class 1 customers.

Table A: $\mu_1 = .5, \mu_0 = 1, \lambda_1 = 1, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 1, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	%savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	.555	0	.0032	na	0	0	0
$c_0 = 5$	.267	.489	0	52%	4.89	1.55	.489
$c_0 = 10$	.184	1.01	0	67%	1.01	.319	.101

Table B: $\mu_1 = .25, \mu_0 = .25, \lambda_1 = .5, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 1, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	%savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	1.57	0	.002	na	0	0	0
$c_0 = 5$	.424	1.463	0	73%	1.463	.386	.1463
$c_0 = 10$	.299	2.34	0	81%	2.34	.740	.234

Table C: $\mu_1 = 1, \mu_0 = 1, \lambda_1 = 1, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 3, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	%savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	.853	0	.004	na	0	0	0
$c_0 = 5$	.557	.750	0	35%	.750	.237	.0750
$c_0 = 10$	.415	1.77	0	51%	1.77	.316	.177

In the above examples and in all other cases that we tested numerically, a considerable saving in the global performance is obtained. The price payed for this saving is the rejection of class 1 customers. However, the fraction of rejected class 1 customers is acceptable for large N. As is seen in the tables, it is of the order of 1% for $N = 1000$, and less than 0.5% for $N = 10000$. We thus conclude that for large systems operating at a heavy traffic regime, we may gain considerably in overall performance of the system at a the cost of rejection a very small fraction of GP calls.

References

[1] The ATM Forum Technical Committee, *Traffic Management Specification*, Version 4.0, af-tm-0056, April 1996.

[2] E. Altman, 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, F. Boccara, J. Bolot, P. Nain, P. Brown, D. Collange and C. Fenzy, "Analysis of the TCP/IP Flow Control Mechanism in High-Speed Wide-Area Networks", European Trans. on Telecommunications, **10** pp125–134, 1999

[4] E. Altman, A. Orda and N. Shimkin, "Bandwidth allocation for Guaranteed versus Best Effort service categories" *in Proc. of IEEE Infocom, San Francisco, Ca., March 29-Apr 2, 1998* pp617–624

[5] J-C. Bolot and A. Vega-Garcma, "Control mechanisms for packet audio in the Internet," Proceeding of the *IEEE Infocom'96.*, San Francisco, CA, pp. 232-239, April 1996.

[6] P. Billingsley. *Convergence of Probability Measures.* John Wiley, New York, 1968.

[7] H. Chen and W. Whitt, "Diffusion approximations for open queueing networks with service interruption", *Queueing Systems, Theory and Appl.* **13**, pp335–359, 1993.

[8] P. Dupuis and H. Ishii, "On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications" *Stochastics and Stochastics Rep.*, **35**, pp.31–62, 1991.

[9] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley, New York, 1986.

[10] A. I. Elwalid and D. Mitra, "Effective Bandwidth of general Markovian traffic sources and admission control of high speed networks", *IEEE/ACM Trans. on Networking*, **1**, pp. 329-343, 1993.

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

[12] E. Gelenber, X. Mang and R Onvural, "Diffusion based statistical call admission in ATM networks", *Performance Evaliation*, **27 & 28**, pp. 411-436, 1996.

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

[14] M. Grossglauser and D. Tse, "A Framework for Robust Measurement-Based Admission Control", *IEEE/ACM Trans. on Networking* **7**, pp293–309, 1999.

[15] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in High-speed networks", *IEEE J. on Selected Areas in Communications* **9**, 1991

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

[17] V. Jacobson. "Congestion avoidance and control". *ACM SIGCOMM 88*, pages 273-288, 1988.

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

[19] 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.

[20] H.J. Kushner. Optimality conditions for the average cost per unit time problem with a diffusion model. *SIAM J. Control Optim.*, 16:330–346, 1978.

[21] H.J. Kushner *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory* MIT Press, Cambridge, MA, 1984

[22] H.J. Kushner. Control of trunk line systems in heavy traffic. *SIAM J. Control Optim.*, 33:765–803, 1995.

[23] H.J. Kushner. Heavy traffic analysis of controlled multiplexing systems. *Queueing Systems, Theory and Appl.*, 28:79–107, 1998.

[24] H.J. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, Berlin and New York, 1992.

[25] H.J. Kushner, D. Jarvis, and J. Yang. Controlled and optimally controlled multiplexing systems: A numerical exploration. *Queueing Systems, Theory and Appl.*, 20:255–291, 1995.

[26] H.J. Kushner and L.F. Martins. Numerical methods for stochastic singular control problems. *SIAM J. Control Optim.*, 29:1443–1475, 1991.

[27] H.J. Kushner and L.F. Martins. Heavy traffic analysis of a data transmission system with many independent sources. *SIAM J. Appl. Math.*, 53:1095–1122, 1993.

[28] H.J. Kushner and L.F. Martins. Heavy traffic analysis of a controlled multi class queueing network via weak convergence theory. *SIAM J. on Contral and Optimization*, 34:1781–1797, 1996.

[29] H.J. Kushner and K.M. Ramachandran. Optimal and approximately optimal control policies for queues in heavy traffic. *SIAM J. Control Optim.*, 27:1293–1318, 1989.

[30] H.J. Kushner and J. Yang. Numerical methods for controlled routing in large trunk line systems via stochastic control theory. *ORSA J. Computing*, 6:300–316, 1994.

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

[32] Z. Liu, P. Nain and D. Towsley, " Exponential bounds with application to call admission." Journal of the ACM, Vol. 44, 1997, pp366–394

[33] A. Mandlebaum and G. Pats "State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits," Annals of Applied Probability, Vol. 8, pp. 569-646, 1998.

[34] 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.

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

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

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