Providing bandwidth guarantees over a best-effort network: call-admission and pricing*

Costas A. Courcoubetis
Department of Informatics,
Athens University of Economics and Business,
47A Evelpidon Str. & 33 Lefkados Str., GR 113 62, Athens, Greece.
E-mail: courcou@aueb.gr

Antonis Dimakis
Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley,
Berkeley, CA 94720-1770, USA.
E-mail: dimakis@eecs.berkeley.edu

Martin I. Reiman
Bell Laboratories, Lucent Technologies,
Murray Hill, NJ 07974, USA.
E-mail: marty@research.bell-labs.com

Abstract— **This paper introduces a framework for answering questions regarding the conditions on the network load that allow a best-effort network like the Internet to support connections of given duration that require a certain quality of service. Such quality of service is expressed in terms of the percentage of time the bandwidth allocated to a connection may drop below a certain level or the maximum allowable delay in placing the call through the network waiting for more favorable loading conditions. The call-acceptance conditions, which depend on the behavior of the system over the lifetime of accepted calls, are thus based on transient models for the congestion (instead of looking at the average behavior) and attempt to exploit the time-scales of the fluctuations of the number of connections competing for bandwidth. Extensions of the model consider the case of dynamic pricing which allows connections that pay more to get larger shares of the bandwidth, and investigate the trade-off between quality of service, the size of the acceptance region, and the charge to be paid by the connection. Under this framework we introduce an option contract that reduces the risk of quality disruption, if a user has a fixed budget at his disposal, and calculate its price. One potential use of this methodology is towards developing a simple admission control mechanism for placing voice calls through an IP network, where the decisions can be taken by edge devices.**

I. Introduction

A serious criticism of the existing Internet protocols is their inability to support services that require minimum bandwidth guarantees. The Internet as it currently exists provides a simple but robust best-effort service where connections obtain a share of the available bandwidth that depends on the number of connections that are active and compete for network resources. In this paper we examine the conditions on the network load under which a value-added service involving minimum bandwidth guarantees can be supported for a given time period over such a basic best effort service. As a good example, one can think of the value-added service as being internet telephony, which requires a minimum bandwidth defined by the coding standards for voice.

The key idea in our approach is to consider the amount of bandwidth that a new connection will receive over its duration, which depends on the transient behavior of the network. The basic problem we solve is to provide ways to calculate, for a given time window and a state of congestion (number of active connections), the percentage of the time during which the bandwidth the best-effort network allocates to a new active connection will be above a minimum level. Interestingly enough, providing some flexibility for the time the call must be placed can increase the chance for the call to be accepted, and hence delaying the placement of the call can be a reasonable strategy. Our methodology allows us to develop a tractable way to compute the admission control policy that takes into consideration such broader quality-of-service definitions that include the time duration over which the quality must be guaranteed, the delay in call-setup, and the fraction of time the bandwidth requirements must hold.

Our approach considers a number of different congestion models that model the network at a connection level rather than at a packet level, and in this paper are applied to single link. The simplest model is the case of a link of fixed capacity that is shared by simple best-effort (elastic) connections and by quality-seeking (inelastic) connections, where capacity is allocated on an equal basis to all connections. Best-effort connections arrive as a Poisson process, have exponential duration, and receive an equal share of the bandwidth. Note that our model focuses on connection dynamics rather than the packet scale. Whereas packet dynamics have been shown to exhibit self-similar behavior, i.e., persisting burstiness across time-scales that differ by many orders of magnitude, connections initiated by human users are well modeled by Poisson processes with perhaps time-varying arrival rates [6]. Other models (discussed in the Appendix) include the case of a link where connections stay longer when bandwidth is more scarce, and the case where best-effort connections are generated by on-off users that model web browsing, where a user alternates between “thinking” and requesting a file transfer.

The problem we solve is to determine for a given state of the link (number of active connections) and for a given quality-of-service as defined previously, whether to accept or reject a quality-seeking call. In this paper the above admission control rule is defined for the case where the majority of the traffic is of the elastic type, and hence the transients of the occupancy process of the link are defined by the process of the elastic calls.

* This work was partly supported by EU’s Fifth Framework project “Market Managed Multiservice Internet” - M3I (RTD No IST-1999-11329).

The idea is that the amount of bandwidth a connection receives varies with time, and although this is a random process there is some predictability to the way that this takes place. Lets say that at this time the amount of bandwidth a connection obtains is very small. Will this remain so or it will drastically change in the near future? And in which direction will it move? These are the questions we answer. Consider first the case where the average load is low enough and on the average, the connection should receive an amount of bandwidth m larger than the minimum amount f required by the connection. Then if currently the number of connections is higher than normal and hence the share of the bandwidth below m, we anticipate that soon many connections will terminate and the share of the bandwidth will approach m, hence it may be reasonable to delay the connection. Similarly, if the system is lightly loaded but the average load is higher than the one required for supporting the bandwidth f, this extra available bandwidth will soon vanish. But how soon? This is important if these time scales are comparable to the duration of the quality-seeking connection. Our approach allows for the calculation of the above time scales and provides the tools to answer the above questions with reasonable accuracy (for the models we consider) using appropriate asymptotics.

The next important issue is the capability of a connection to get more bandwidth when paying more. There is a substantial amount of recent research in this area [8], [1], that suggest mechanisms for connections which by paying more can get a proportionally larger share of the bandwidth. The idea is of a simple network that sends to its edges congestion information (in a way that generalizes the concept of packet loss in TCP), and serves as an indication of the rate of charge that the network charges the traffic streams. The sources at the edges decide at any given moment the rate of transmission based on the rate of charge received, in a way to maximize their net benefit. In this framework, one can think of more general rate control algorithms than TCP which, by operating at the edges of the network, will optimize the overall economic efficiency of the system. In our paper we consider such possibilities, without dealing with implementation details. In particular we consider the case where a connection that is willing to pay $\$w$/s gets w times more bandwidth than a connection that pays $1/s (the typical best-effort connection in our case). In this price-sensitive context, we investigate issues related to the optimal spending of a fixed budget, and the conditions for arbitrage if a similar service is offered by another network where quality and prices are fixed (like the PSTN).

An interesting approach that has many common points with ours is the one in [2], [3]. The idea is that a call should decide whether to enter the network based on the current congestion level (and hence the current price), where congestion information along a certain route is signalled to the edges of the network by marking packets. One can analyze the fixed-point of such an interaction between offered load and congestion signals, and solve for the steady-state of the system. Our approach is more refined since it allows the study of the transients. Also we do not explicitly model congestion at a packet level, but in terms of the decrease of the bandwidth share. On the other hand, as it currently stands, we deal with a simpler network case where a single link is the bottleneck.

The paper is organized as follows. In Section II we present the basic model for the link and the bandwidth allocation. In Section III we analyze the system using asymptotics that capture the important aspects when the size of the system is large. In Section IV we derive the shape of the admissible region as a function of the various parameters (quality of service, call duration, willingness to pay more, etc.). In Section V we introduce an option contract that reduces the risk of quality disruption, if a user has a fixed budget at his disposal, and calculate its price. Two other models that are based on our basic model are discussed in the Appendix. Finally, we end the paper with some concluding remarks and suggestions for extensions.

II. The Model

The system we consider has a single link of bandwidth C shared by two types of services, best-effort data (elastic calls) and bandwidth sensitive calls (inelastic calls) that have similar requirements. Best-effort data calls arrive as a Poisson process of rate ν and inelastic calls arrive as an independent Poisson process of rate κ. We assume that all call holding times are i.i.d. exponentially distributed with mean μ^{-1} and all calls share equally the available bandwidth. This last assumption models to some extent a best-effort network like the Internet in the case of a single bottleneck link and connections using TCP with similar round trip times.

Call durations in the above model are not affected by the load of the network. We relax this assumption in the two models discussed in the Appendix. The results of this paper are obtained in the context of the above basic model but can be extended in the context of the other two models.

Let N_t denote the number of calls in progress at time t, and X_t the bandwidth available for each call. (Note that, under our assumptions, once admitted, all calls behave the same. This assumption is relaxed below in Section IV-B when we allow inelastic calls to obtain more bandwidth than best-effort calls by paying more.) Then

$$X_t = \frac{C}{N_t}\,. \tag{1}$$

In the absence of any admission control the process $\{N_t, t \geq 0\}$ corresponds to the number of customers in service in an M/M/∞ system with arrival rate $\nu + \kappa$ and service rate (per server) μ. The stationary distribution of this process is Poisson with mean $(\nu+\kappa)/\mu$. The stationary distribution of $\{N_t, t \geq 0\}$ immediately yields the stationary distribution of $\{X_t, t \geq 0\}$ via equation (1).

The most reasonable admission control is a threshold policy: For some K, accept an inelastic call if and only if the current number of calls in progress is less than K. Under this admission control policy the process $\{N_t, t \geq 0\}$ is again a birth-death process, whose stationary distribution can be calculated straightforwardly. The issue then becomes one of determining a good value for K.

Recall that the purpose of admission control is to provide acceptable quality of service to the inelastic calls. The quality of service experienced by an inelastic call depends on the bandwidth allocated to it throughout its lifetime in the system. Thus, in order to answer questions related to call admission controls, we need to understand the system evolution following an observed system state, not just the steady state behavior. In the next section, we introduce fluid and diffusion limits for N_t and X_t that allow us to study the transient behavior of these processes.

III. Asymptotics

We consider a sequence of systems, indexed by C, with $C \to \infty$. Such an asymptotic regime is certainly well motivated by real communication networks, where bandwidths continue to grow larger. The asymptotics we obtain are immediately translatable into an approximation for a system with a fixed (large) value of C. We keep μ fixed while we let $\nu = C\theta$ with $0 < \theta < \infty$. In addition we assume that the requests for best-effort service dominate the link as compared to requests for inelastic service, so that $\kappa/\sqrt{C} \to 0$ as $C \to \infty$.

A. Number of calls in progress

We first consider the number of calls in progress. Let

$$\bar{N}_t^{(C)} = \frac{N_t^{(C)}}{C} . \qquad (2)$$

Assume that $\bar{N}_0^{(C)} \to \bar{N}_0$ a.s. (almost surely), as $C \to \infty$, where $\bar{N}_0 > 0$ is a constant. Then by [4] or [5] $\bar{N}_t^{(C)} \to \bar{N}_t$ a.s., uniformly on $[0,T]$ for $0 < T < \infty$, where $\bar{N} = \{\bar{N}_t, t \geq 0\}$ is the unique solution, given $\bar{N}_0$, to the ordinary differential equation

$$\frac{d\bar{N}_t}{dt} = \theta - \mu\bar{N}_t , \quad t \geq 0 ,$$

which yields

$$\bar{N}_t = \bar{N}_0 + \left(\frac{\theta}{\mu} - \bar{N}_0\right)\left(1 - \mathrm{e}^{-\mu t}\right) , \quad t \geq 0 . \qquad (3)$$

For any $\bar{N}_0 > 0$, $\bar{N}_t \to \theta/\mu \equiv \bar{N}_\infty$ as $t \to \infty$. If $\bar{N}_0 = \bar{N}_\infty$, then $\bar{N}_t = \bar{N}_\infty$ for all $t \geq 0$.

Let

$$\hat{N}_t^{(C)} = \frac{N_t^{(C)} - C\bar{N}_t}{\sqrt{C}} = \sqrt{C}\left[\bar{N}_t^{(C)} - \bar{N}_t\right] . \qquad (4)$$

Assume that $\hat{N}_0^{(C)} \xrightarrow{d} \hat{N}_0$. (Where $\xrightarrow{d}$ denotes convergence in distribution.) Then by [4] or [5] $\hat{N}_t^{(C)} \xrightarrow{d} \hat{N}_t$, where $\hat{N} = \{\hat{N}_t, t \geq 0\}$ is the unique solution, given $\hat{N}_0$, to the stochastic differential equation

$$d\hat{N}_t = -\mu\hat{N}_t dt + \sqrt{\theta + \mu\bar{N}_t} dB_t ,$$

and $\{B_t, t \geq 0\}$ is a standard (0 drift, unit variance) Brownian motion process. If $\bar{N}_0 = \bar{N}_\infty$ then $\hat{N}$ is an Ornstein-Uhlenbeck diffusion process.

Combining (2) and the convergence of $\bar{N}_t^{(C)}$ to $\bar{N}_t$ motivates using the fluid approximation $C\bar{N}_t$ for $N_t^{(C)}$. Similarly, combining (4) and the convergence of $\hat{N}_t^{(C)}$ to $\hat{N}_t$ motivates $C\bar{N}_t + \sqrt{C}\hat{N}_t$ as a diffusion approximation of $N_t^{(C)}$.

B. Bandwidth available for each call

We next consider the bandwidth available for each call. Let

$$\bar{X}_t^{(C)} = \frac{1}{\bar{N}_t^{(C)}} = \frac{C}{C\bar{N}_t^{(C)}} = \frac{C}{N_t^{(C)}} = X_t^{(C)} .$$

If $\bar{N}_0^{(C)} \to \bar{N}_0$, with $\bar{N}_0 > 0$, then $X_0^{(C)} \to \bar{X}_0 \equiv \frac{1}{\bar{N}_0}$. It is immediate from (3) that if $\bar{N}_0 > 0$, then $\bar{N}_t > 0$ for $t \geq 0$, so $\bar{X}_t^{(C)} \to \bar{X}_t = \frac{1}{\bar{N}_t}$, $t \geq 0$.

Let $\hat{X}_t^{(C)} = \sqrt{C}\left[X_t^{(C)} - \bar{X}_t\right]$. Then

$$\begin{aligned} \hat{X}_t^{(C)} &= \sqrt{C}\left[X_t^{(C)} - \bar{X}_t\right] = \sqrt{C}\left[\bar{X}_t^{(C)} - \bar{X}_t\right] \\ &= \sqrt{C}\left[\frac{1}{\bar{N}_t^{(C)}} - \frac{1}{\bar{N}_t}\right] = \sqrt{C}\frac{\bar{N}_t - \bar{N}_t^{(C)}}{\bar{N}_t\bar{N}_t^{(C)}} \\ &= -\frac{\hat{N}_t^{(C)}}{\bar{N}_t\bar{N}_t^{(C)}} \xrightarrow{d} -\frac{\hat{N}_t}{\bar{N}_t^2} . \end{aligned}$$

C. Acceptable Bandwidth Threshold

Assume that the minimum bandwidth required for the inelastic calls is f, i.e., if $X_t \geq f$ then callers are receiving acceptable quality. Under our assumption that the users are sharing the link equally the last inequality can be expressed as

$$N_t^{(C)} \leq C/f \equiv N^* . \qquad (5)$$

Consider a system without admission control, in steady state. As already pointed out, without any control the steady state distribution for the total number of calls in progress follows from that of the M/M/∞ system, i.e., it is Poisson with mean $(\nu + \kappa)/\mu = C\theta/\mu + o(\sqrt{C})$. The probability that an incoming call will find the system in an overloaded state ($N_t > N^*$) can thus be easily calculated. In order to keep this probability within an acceptable level (typically below, e.g., 10^{-3}) we would like to know the values that $\nu + \kappa$ may take. We set $C = 1000, \mu = 1$ and determine $\mathbb{P}(N_t > N^*)$ as depicted in Fig. 1. Observe that when $C\theta/\mu$ is close to N^* the probability changes abruptly from almost zero to almost one.

Let us consider the limit case where $C \to \infty$, then $\mathbb{P}(\bar{N}_\infty^{(C)} > 1/f) \to 1$ or 0, if $\theta/\mu > 1/f$ or $\theta/\mu \leq 1/f$ respectively. The diffusion approximation attains the same zero-one limit, except at $\theta/\mu = 1/f$, where the above probability converges to $1/2$. The zero-one behavior of the diffusion with $\theta/\mu \neq 1/f$ is to be expected since, in this case, $|N^* - C\bar{N}_\infty^{(C)}|$ is $O(C)$. This would require $\hat{N}_\infty^{(C)}$ to be $O(\sqrt{C})$ to bridge the gap, which cannot happen because $\hat{N}_\cdot^{(C)}$ is an $O(1)$ stochastic process. This situation motivates a "more sensitive" choice of the minimum bandwidth requirement f. In particular we choose $f^{(C)}$ so that

$$\frac{C}{f^{(C)}} = \frac{C\theta}{\mu} + \gamma\sqrt{C}, \qquad (6)$$

where γ is a constant. The easiest way to interpret the scaling in (6) is to note that, given the original data C, ν, μ, and f, if we choose $\theta = \nu/C$ and $f^{(C)} = f$, then (6) implies that

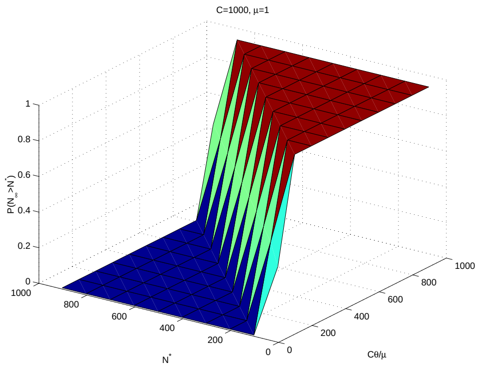

Fig. 1. Probability that the number $N_t^{(C)}$ of users is greater than the maximum number N^* allowed, in order that the available bandwidth to be enough (f). When $C\theta/\mu$ is close to N^*, probability changes abruptly from almost zero to almost one.

$\gamma = \sqrt{C}\left(\frac{1}{f} - \frac{\theta}{\mu}\right)$. This choice affects our diffusion analysis, but not our fluid analysis. With $\bar{N}_0 = \theta/\mu$, the steady state probability that $N_t^{(C)} > N^*$ is approximated by the steady state probability that $\hat{N}_t > \gamma$, which is $1 - \Phi\left(\gamma\sqrt{\frac{\mu}{\theta}}\right)$, where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-z^2/2}dz$.

IV. Call Admission Controls

Connections that compete for bandwidth have different weights that may also vary with time, denoted by w_t, and receive service from a GPS server that models the link. An interpretation for w_t is the rate of charge a user is incuring; users which are being charged more should receive a proportionally larger share of the bandwidth. Since we concentrate on a single bottleneck link, bandwidth allocations produced by various notions of fairness coincide, e.g., weighted max-min fairness [7] and proportional fairness [8]. We scale the values of the above weights by assuming the best-effort users request connections with $w_t = 1$.

There has been considerable research of how a user picks his weight w_t in order for certain global efficiency criteria to be met. In general, weights are produced by some kind of market mechanism, such as smart markets [9] or tatonnement processes [10], where they correspond to bids or charges per unit of time reflecting network congestion. The idea is that users monitor the amount of bandwidth X_t they receive for a given value of w_t (rate of charge), and determine the current price per unit bandwidth w_t/X_t. Then, taking into account their utility for bandwidth, they increase or decrease the value of their willingness to pay, w_t, in order to increase or decrease the amount of purchased bandwidth. At the equilibrium, such a strategy results in maximizing the total user utility (social welfare maximum). In this paper we simply assume that the system allows inelastic users to declare their willingness to pay w_t, and get a proportional share of the bandwidth ($w_t = 1$ for best-effort).

An inelastic user can share bandwidth in many ways according to the way his weight w_t varies with time. We consider the following cases:

1. $w_t = 1$, i.e., he is treated as a best-effort user. This is the case where no pricing algorithm is implemented and all users are treated equally.
2. $w_t = w > 1$, i.e., constant weight throughout the connection, giving him greater share than the best-effort, but this share depends on the overall congestion.
3. w_t varying with t, in a way that the bandwidth received at any point in time is exactly f, the minimum bandwidth required.

The above cases may fit to different transport services, e.g., TCP in 1, Diff-Serv in 2, CBR in 3, or any scheme employing complete priority over best-effort users. In the next sections we give call admission controls for inelastic users that depend on system evolution after an observed state, for the three aforementioned ways that their weight w_t varies.

A. *Case 1:* $\mathbf{w_t = 1}$

A.1 Fluid analysis

Given a number of $\bar{N}_0$ calls at $t = 0$, the system relaxes to $\bar{N}_\infty = \theta/\mu$ as the time passes. If $\bar{N}_0 < 1/f < \theta/\mu$ then $\bar{N}_t < 1/f$ for small t, and $\bar{N}_t > 1/f$ for large t. We would like to calculate the time for the system to reach the N^* boundary, i.e., the T_{relax} such that $\bar{N}_{T_{\text{relax}}} = 1/f$. Using (3), we have

$$T_{\text{relax}} = \frac{1}{\mu}\ln\left(\frac{\frac{\theta}{\mu} - \bar{N}_0}{\frac{\theta}{\mu} - \frac{1}{f}}\right). \qquad (7)$$

This relation can be elaborated into in a simple decision rule in the case that the best-effort link is shared mainly by elastic calls; inelastic calls come once in a while. At the instant that an inelastic call arrives, a simple rule is to check if the time to relax to $1/f$ (after the call is accepted) is larger than the length of the call. That is, an inelastic call is accepted (routed through the best-effort network) if $T \leq T_{\text{relax}}$, where T is the call holding time. Using this constraint for T in (7) one obtains that the call is accepted if

$$\bar{N}_0 \leq \frac{\theta}{\mu} - \left(\frac{\theta}{\mu} - \frac{1}{f}\right)e^{\mu T}. \qquad (8)$$

Thus, if $T \geq \mu^{-1}\left[\ln(\theta/\mu) - \ln(\theta/\mu - 1/f)\right]$, no inelastic calls can ever be accepted.

So far we have assumed that $\bar{N}_0 < 1/f < \theta/\mu$. There are three other cases to consider: (i) $1/f > \theta/\mu$ and $\bar{N}_0 < 1/f$: $\bar{N}_t$ will never reach $1/f$, so calls can always be accepted; (ii) $1/f < \bar{N}_0 < \theta/\mu$: $N_t > 1/f$ for all t, so calls can never be accepted; (iii) $1/f > \theta/\mu$ and $\bar{N}_0 > 1/f$: $\bar{N}_t > 1/f$ for small t and $\bar{N}_t < 1/f$ for large t, so calls with long enough duration can be accepted, if we are willing to accept bandwidth under the minimum, f, initially. This occurs because the system starts at $t = 0$ with too many calls and then relaxes towards steady state, where the bandwidth requirement can be met.

All of the above are true at the fluid scale, i.e., within o(C) as $C \to \infty$. A refinement to the above analysis, on the order of O$(\sqrt{C})$ can be given when statistical quality guarantees are provided, as in the next section.

A.2 Diffusion approximation

Given the duration (or an estimate) T of the call and $N_0^{(C)} = n$, the quality received by the best-effort link may be assessed by the fraction of time that (5) is violated during the length of the entire call. This can be written as

$$\begin{aligned} &\tfrac{1}{T}\mathbb{E}\left[\int_0^T 1(N_t^{(C)} > N^*)dt \,\middle|\, N_0^{(C)} = n\right] = \\ &\tfrac{1}{T}\int_0^T \mathbb{P}\left(N_t^{(C)} > N^* \,\middle|\, N_0^{(C)} = n\right) dt \,. \end{aligned} \tag{9}$$

In Section III-A it was indicated that as $C \to \infty$, $\hat{N}_t^{(C)}$ converges in distribution to a diffusion process with drift $-\mu x$ and time dependent diffusion coefficient $\theta + \mu\bar{N}_t$.

Suppose that $\bar{N}_0 = \theta/\mu$. Then $\theta + \mu\bar{N}_t = 2\theta$, for all $t \geq 0$, and $\hat{N}$ is an Ornstein-Uhlenbeck process. The transition distribution $\mathrm{P}_t(\cdot|x)$ (where $\mathrm{P}_t(z|x) = \mathrm{P}(\hat{N}_t \leq z|\hat{N}_0 = x)$) of this process is equal to a Gaussian cdf centered at $x\exp(-\mu t)$ with variance $\theta(1-\exp(-2\mu t))/\mu$. Equation (9) can thus be approximated for large C by

$$\int_0^T \left[1 - \mathrm{P}_t\left(\frac{N^* - C\bar{N}_t}{\sqrt{C}}\,\middle|\, x\right)\right] dt = \int_0^T \left[1 - \mathrm{P}_t(\gamma|\, x)\right] dt \,, \tag{10}$$

where $x = (n - C\bar{N}_0)/\sqrt{C}$ and $\gamma = \sqrt{C}\left(\frac{1}{f} - \frac{\theta}{\mu}\right)$. $\Big($Recall from Section III-C that for the diffusion analysis we assume $f^{(C)}$ is given by equation (6). Solving equation (6) with C fixed yields $\gamma = \sqrt{C}\left(\frac{1}{f} - \frac{\theta}{\mu}\right)$.$\Big)$

Now suppose we set an upper bound α on the fraction of time that (5) is violated during $[0, T]$. With γ, α, and T fixed, let

$$x^* \equiv \max\left\{x : T^{-1}\int_0^T \left[1 - \mathrm{P}_t(\gamma|x)\right]\, dt \leq \alpha\right\} \,.$$

Then the maximum allowable number of users at $t = 0$ is approximated (for large C) by $C\frac{\theta}{\mu} + \sqrt{C}x^* \equiv M^{(C,T,\alpha)}$.

Under this call admission control, a call is accepted if $N_0^{(C)} \leq M^{(C,T,\alpha)}$.

Denote by α_∞ the stationary fraction of time that (5) is violated (as $C \to \infty$), i.e. $1 - P_\infty(\gamma|x) = 1 - \Phi\left(\gamma\sqrt{\frac{\mu}{\theta}}\right)$. Then if $\alpha > \alpha_\infty$, calls with long enough duration can be accepted, while if $\alpha < \alpha_\infty$ shorter calls are favored. This effect is illustrated in Figs. 2 and 3. In Fig. 2 we examine $M^{(C,T,\alpha)}$ as a function of T with $C = 1000$ for 3 different values of α: $\alpha = 3.0182 \times 10^{-4} = \alpha_\infty$, $\alpha = 2.5 \times 10^{-4} < \alpha_\infty$, and $\alpha = 10^{-3} > \alpha_\infty$. We also take $\theta = .85$, $\mu = 1$, and $N^* = 950$. This yields (using (5) and (6)) $\gamma = \sqrt{10}$. If $\alpha = \alpha_\infty$, call admission decisions are independent of call holding times. If $\alpha > \alpha_\infty$, the acceptance threshold on the number of users tends to ∞, while if $\alpha < \alpha_\infty$, only short enough calls may be admitted. In Fig. 3 we examine $M^{(C,T,\alpha)}$ as a function of T with $C = 1000$ and $\alpha = 10^{-3}$ for 3 different values of θ: $\theta = .8$, $\theta = .85$, and $\theta = .86$. In all cases $\mu = 1$ and $N^* = 950$. This yields respective values for γ of $1.5\sqrt{10}$, $\sqrt{10}$, and $.9\sqrt{10}$. Keeping level of quality α fixed (10^{-3} in the example), the link may operate in one of two regimes depending on link load: (i) long enough calls can always be accepted (e.g., for $\theta = .85$) or (ii) short calls are favored ($\theta = .86$). Even small variations of θ can make the link switch between these two regimes.

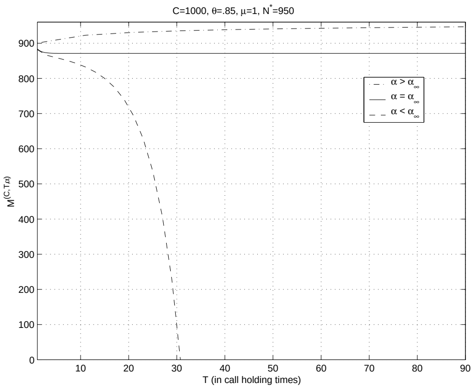

Fig. 2. Call acceptance threshold $M^{(C,T,\alpha)}$ for different quality levels: $\alpha = 3.0182 \times 10^{-4} = \alpha_\infty, \alpha = 2.5 \times 10^{-4} < \alpha_\infty$, and $\alpha = 10^{-3} > \alpha_\infty$.

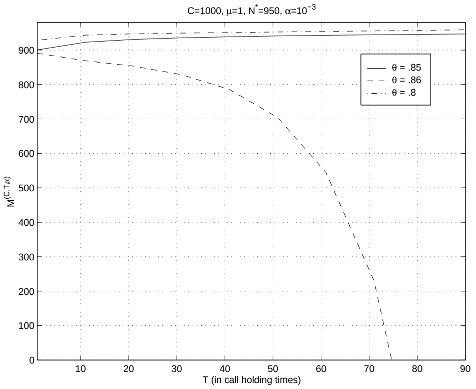

Fig. 3. Call acceptance threshold $M^{(C,T,\alpha)}$ for different call arrival rates θ.

It is worth noting that, because only the voice calls are subject to admission control, and they are a vanishingly small fraction of the total traffic, the admission control does not affect the dynamics of the limit process. Admission control does, however, affect the blocking probability of voice calls. The voice call blocking probability at time t is $\mathbb{P}(N_t^{(C)} > M^{(C,T,\alpha)})$. The diffusion approximation to this quantity is $\mathbb{P}(\hat{N}_t > x^*)$. If the process is in steady state, $\mathbb{P}(\hat{N}_t > x^*) = 1 - \Phi\left(x^*\sqrt{\frac{\mu}{\theta}}\right)$.

Figure 4 illustrates the call blocking probability for various call durations (T). In a lightly loaded link, longer calls have more chances to be accepted This is because if in the steady state enough bandwidth is available, and shortage occurs in the beginning of the call, then if the call stays for long a time, it will obtain its bandwidth requirements during the specified fraction of its duration. On the other hand under heavy load, long calls are always denied service and only short enough calls can be accepted.

B. *Case 2:* $\mathbf{w_t = w > 1}$

With $w > 1$ inelastic calls receive more bandwidth than best-effort calls, and we need to keep track of how many of each are in the system to determine exactly how much bandwidth each receives. Fortunately, due to our assumption that $\kappa/\sqrt{C} \to 0$, the number of inelastic calls in the system is $o(\sqrt{C})$ and does not affect either the fluid or diffusion asymptotics. In particular, $X_t^{(C)} = wC/N_t^{(C)} + o(C^{-1/2})$, and this $o(C^{-1/2})$ term does

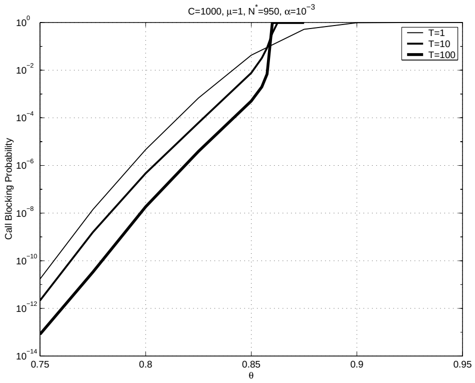

Fig. 4. Achieved call blocking probability for different call durations (T).

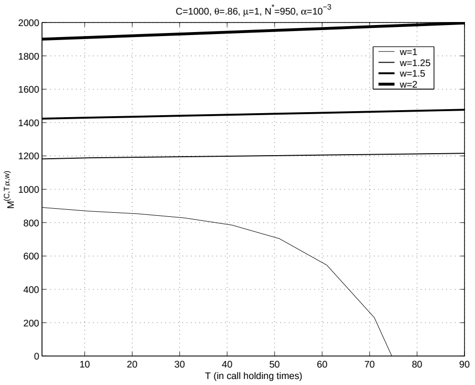

Fig. 5. Call acceptance threshold $M^{(C,T,\alpha,w)}$ for different choices of weight w at call arrival.

not affect either the fluid limit or the diffusion limit. It is also clear that, by the same argument, it is not necessary that all inelastic calls behave in the same manner. Thus, each can have its own f and w.

B.1 Fluid analysis

Essentially the same analysis in IV-A.1 can be used on this case, and the call admission control now is:

$$\bar{N}_0 \leq \frac{\theta}{\mu} - \left(\frac{\theta}{\mu} - \frac{w}{f}\right) \mathrm{e}^{\mu T} . \tag{11}$$

B.2 Diffusion approximation

The probability that the instantaneous available bandwidth to a user having weight w, will drop below $f^{(C)}$, is $\mathbb{P}(wC/N_t^{(C)} < f^{(C)}) = \mathbb{P}(N_t^{(C)} > wN^*)$, which leads to an expression for the mean time that the bandwidth drops below $f^{(C)}$, similar to (10). A similar reasoning to that in Section IV-A.2, leads to a quantity $M^{(C,T,\alpha,w)}$ which is depicted in Fig. 5 for various choices of w. Each curve in Fig. 5 gives raise to an acceptance region. An arriving user will be positioned onto the (N_0, T) plane according to the link load (N_0) and his holding time (T). His declared weight specifies a $(M^{(C,T,\alpha,w)}, T)$ curve on that plane, which determines whether the user is accepted, or not, if he is positioned below the curve, or not, respectively. Thus, the call admission control is

$$N_0^{(C)} \leq M^{(C,T,\alpha,w)} . \tag{12}$$

Conversely, one can compute the minimum weight that one has to declare in order to attain a fraction α of time spent by w/N_t below $f^{(C)}$ during $[0, T]$; this is depicted in Fig. 6.

Suppose now that weights w reflect user's willingness to pay for a unit of flow per unit time (see [10]) and he is given the option to connect to a circuit-switched network at a price of w_* per unit of time. Then, he will certainly choose to connect to the circuit-switched network when the best-effort link determines that it cannot serve him at a lower price.

The corresponding $(M^{(C,T,\alpha,w_*)}, T)$ curve is the "maximal" acceptance region. That is, if the state of the network N_0 and the declared call holding time T specify a point over that curve, then the best-effort link cannot serve the user at a price lower than w_*.

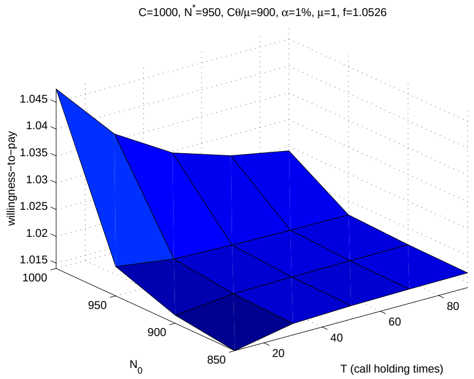

Fig. 6. The smallest choice of willingness-to-pay (weight w) so that a call that lasts time T and requests quality level α, is accepted, when, at call arrival instance, N_0 best-effort users are being served.

From now on, w_t will reflect willingness-to-pay for a unit of flow per unit time unless otherwise stated.

C. *Case 3: Constant bandwidth share*

By letting the user declare (or assigning him) a willingness-to-pay at arrival that is constant in time, it may be the case that there are times that the received bandwidth drops well under f. Thus, at those times the received bandwidth is totally useless to him. At other times, he may receive bandwidth well above f, whose slack is again useless. Thus, although at certain times he does not use the bandwidth that the network gives him, he is charged for it a proportional price.

The network itself, operating under competition, should link charges to actual usage as accurately as possible. Thus, it has the incentive to assign a bandwidth share close to f. By doing that, the user now is charged for the usage of f, which is approximately $fN_t^{(C)}/C$. This corresponds to a constant adaptation of the user's willingness-to-pay in order to maintain a share of f throughout the duration of the call.

It is obvious here that a user can maintain any bandwidth share (up to C) throughout his connection by continuously adjusting his weights. So network load clearly cannot be the only constraint, given that no other inelastic users are using the network at the same time. If w_t expresses willingness-to-pay then certainly a user might want to bound the values that w_t might take. So if a circuit-switched network exists, which offers the inelastic service for w_* per unit of time then a constraint arises imposed by competition.

Assuming a pricing scheme like the above, what is the "competitive" acceptance region for the best-effort link? Competition requires that a user connecting to a best-effort link should not be charged an amount exceeding the amount charged by the competitor network, i.e., the circuit-switched network in our case. This can be expressed as the requirement that

$$\int_0^T w_t dt \le w_* T \ , \quad \text{or} \quad f \int_0^T \bar{N}_t^{(C)} dt \le w_* T \ . \qquad (13)$$

C.1 Fluid analysis

If we take $C \to \infty$, then we obtain

$$\bar{N}_0 \le \frac{\left(w_* - f\frac{\theta}{\mu}\right)T}{\frac{f}{\mu}\left(1 - \mathrm{e}^{-\mu T}\right)} + \frac{\theta}{\mu} \ .$$

For $T \to \infty$, the right hand side tends to $-\infty$, θ/μ, or $+\infty$, if $w_* < f\theta/\mu$, $w_* = f\theta/\mu$, or $w_* > f\theta/\mu$ respectively.

An increase in utilization can be achieved if users tolerate "bad quality" for a fraction of time up to α. Since f is the minimum bandwidth that they can operate on, "bad quality" can mean "temporarily out of service", thus they should not pay during these periods of time. If a user arrives at a time when the prices are high enough then the best-effort link may not be able to serve him for less than $w_* T$. If instead, the user is tolerant as we previously described, then the network might be able to serve him for less than $w_* T$, since he would not have to pay for an initial period up to αT. Similarly, he would not have to pay for a period before the end of his call, if the prices are high at the end of his call.

Recall that price per unit of time at time t is given by $f\bar{N}_t$, so prices go up or down according to whether $\bar{N}_0 < \theta/\mu$, or $\bar{N}_0 > \theta/\mu$ respectively.

In the first case,

$$\int_0^{(1-\alpha)T} f\bar{N}_t dt \le w_*(1-\alpha)T \ .$$

This expression is identical to (13) for a call of duration $(1-\alpha)T$.

In the second case,

$$\int_{\alpha T}^{T} f\bar{N}_t dt \le w_*(1-\alpha)T \ , \quad \text{or}$$

$$\bar{N}_0 \le \mathrm{e}^{\mu\alpha T} \frac{\left(w_* - f\frac{\theta}{\mu}\right)(1-\alpha)T}{\frac{f}{\mu}\left(1 - \mathrm{e}^{-\mu(1-\alpha)T}\right)} + \frac{\theta}{\mu} \ . \qquad (14)$$

C.2 Diffusion approximation

The diffusion case is a bit more involved. Using $w_t^{(C)} = f^{(C)}\,\frac{N_t^{(C)}}{C}$ we can write

$$w_t^{(C)} = \frac{C\bar{N}_t + \sqrt{C}\hat{N}_t^{(C)}}{C\frac{\theta}{\mu} + \gamma\sqrt{C}} \ .$$

As usual, we assume that $\bar{N}_0 = \frac{\theta}{\mu}$, so that $\bar{N}_t = \frac{\theta}{\mu}$. We can thus write

$$w_t^{(C)} = 1 + \frac{1}{\sqrt{C}}\frac{\mu}{\theta}\left(\hat{N}_t - \gamma\right) \ . \qquad (15)$$

Recall that $\mathbb{E}[\hat{N}_t|\hat{N}_0] = \hat{N}_0 e^{-\mu t}$. Thus the condition $\mathbb{E}\left[\int_0^T w_t^{(C)} dt \middle| \hat{N}_0\right] \le w_*^{(C)} T$ is, approximately (for large C),

$$\hat{N}_0 \le \theta\sqrt{C}\,\frac{w_*^{(C)} - 1 + \frac{\mu\gamma}{\theta\sqrt{C}}}{1 - e^{-\mu T}} \ .$$

If we write $w_*^{(C)} = 1 + \delta/\sqrt{C}$, this becomes

$$\hat{N}_0 \le \frac{\theta\delta + \mu\gamma}{1 - e^{-\mu T}} \ .$$

V. Option contracts

In practice an inelastic user will not be able to change w_t infinitely fast as to attain the desired bandwidth amount. Furthermore there is a risk of exceeding a fixed budget. In these cases a user might want to reduce his exposure to risk by specifying an upper bound on the values that w_t may take.

Financial derivatives can be used to hedge risk in these cases. Similar to [11], we construct an option contract that gives the user the right to buy bandwidth share at a prescribed "strike price" (denote it by $w^{(C)}$) anytime during the length of a connection. As in [11], we interpret this option as a series of European call options (cf. [12]) integrated over time.

Due to the fact that bandwidth is a perishable commodity, there seems to be no possibility of developing a value of this option based on arbitrage arguments. We thus take a direct approach. Using (15), and writing $w_t^{(C)} = 1 + \hat{w}_t^{(C)}/\sqrt{C}$, we have $\hat{w}_t^{(C)} = \frac{\mu}{\theta}\left(\hat{N}_t^{(C)} - \gamma\right)$. Now, writing the strike price as $w^{(C)} = 1 + \tilde{w}/\sqrt{C}$, we obtain the expected value of the option to the caller to be

$$V_x^{(C)} = \frac{1}{\sqrt{C}}\mathbb{E}\left[\int_0^T \left(\hat{w}_t^{(C)} - \tilde{w}^{(C)}\right)^+ dt \middle| \hat{w}_0^{(C)} = x\right] \ .$$

Although a non-risk-seeking third party would have no incentive to sell these options at a price of $V_x^{(C)}$, one can argue that the network provider does have an incentive, because it entices risk averse customers to use its guaranteed quality service. (One could also argue that, with no transaction costs, a profit seeking third party with sufficient capital would have an incentive to sell these options at a price of $V_x^{(C)} + \epsilon$ for any $\epsilon > 0$.)

Let $\mathcal{N}(a,b)$ denote a normally distributed random variable with mean a and variance b. Recall that, conditioned on $\hat{N}_0 = x$, $\hat{N}_t$ is normally distributed with mean $xe^{-\mu t}$ and variance $\frac{\theta}{\mu}\left[1 - e^{-2\mu t}\right]$. Define $\hat{V}_x^{(C)} = \sqrt{C}\,V_x^{(C)}$. Then $\hat{V}_x^{(C)} \to \hat{V}_x$ as $C \to \infty$ for $-\infty < x < \infty$. We can then write

$$\begin{aligned}
\hat{V}_x &= \int_0^T \mathbb{E}\left[(\hat{w}_t - \tilde{w})^+ \middle| \hat{w}_0 = x\right] dt \\
&= \int_0^T \mathbb{E}\left(\frac{\mu}{\theta}\mathcal{N}\left(\left[\frac{\theta x}{\mu} + \gamma\right]e^{-\mu t},\ \frac{\theta}{\mu}\left[1 - e^{-2\mu t}\right]\right)\right. \\
&\qquad \left. - \frac{\mu\gamma}{\theta} - w\right)^+ dt \ .
\end{aligned}$$

Taking into consideration the fixed price w^* and the option instrument introduced above, we are led into a simple call admission control in a way similar to the previous section: a call is accepted in the best-effort network if the cumulative charge (option price plus maximum per unit of time charge) is less than the corresponding charge of a fixed price alternative network, i.e.

$$V_x^{(C)} + \int_0^T \mathbb{E}\left[w_t^{(C)} \wedge w^{(C)} \middle| \hat{w}_0^{(C)} = x \right] dt \le w_*^{(C)} T \, .$$

Note that

$$w_t^{(C)} = w_t^{(C)} \wedge w^{(C)} + \left[w_t^{(C)} - w^{(C)} \right]^+ \, ,$$

so that

$$\begin{aligned} V_x^{(C)} &+ \int_0^T \mathbb{E}\left[w_t^{(C)} \wedge w^{(C)} \mid \hat{w}_0^{(C)} = x \right] dt \\ &= \mathbb{E}\left[\int_0^T w_t^{(C)} dt \mid \hat{w}_0^{(C)} = x \right] \, . \end{aligned}$$

Thus this admission control is identical to that of the previous section.

VI. Conclusions

In this paper we have presented a new approach for reasoning whether or not to accept a connection that requires some minimum quality of service over a simple best-effort network where connections get an equal share of the bandwidth. The contribution of the paper is to relate the above issue with the condition of the current state of the network rather than its average behavior. We show that call durations are crucial (if calls that seek quality have infinite duration, then the relevant analysis is based on steady state), and that there is an interesting relation with pricing. We view the work in this paper as being a first step in the direction of constructing state-dependent price-sensitive call-admission controls. We also show how financial derivatives can be used in order to reduce the risk inherent in congestion pricing and hide the time-scales mismatch between end-to-end rate control algorithms and rapid fluctuation of congestion prices, yielding robust call admission controls.

We believe that the analysis can be further extended to include the network case where more than one link can constrain the bandwidth along a route. Although this is theoretically interesting, we believe that we should validate our results first in an actual network like the Internet by measuring the actual bandwidth allocation process and see whether the time scales are close to the ones predicted in our models. Another interesting approach would be to fit the parameters of our model in the actual situation and use it in order to predict its future evolution.

There are several directions for further research. One is towards allowing the inelastic traffic to be of a substantial percentage, and hence influence the transient analysis. Another is towards modeling the effect of UDP traffic. When UDP traffic is substantial, the available capacity for TCP flows will be $C_t = C - U_t$, where U_t is the amount of capacity taken up by UDP. An interesting problem is the description of the U_t process in a way that allows a similar analysis as in this paper to carry through.

Another basis on which to make a call acceptance decision, using the *same* model, is the distribution of the first passage time of $N_t^{(C)}$ to N^*, assuming that $N_0^{(C)} < N^*$. Approximating $N_t^{(C)}$ as before by $C\bar{N}_t + \sqrt{C}\hat{N}_t$, and assuming that $\bar{N}_0 = \theta/\mu$, $N_t^{(C)} = N^*$ is equivalent to $\hat{N}_t = \gamma$. Let

$$T_\gamma = \inf\{t \ge 0 : \hat{N}_t = \gamma\} \, .$$

We can then base the call acceptance decision on $\mathbb{E}\left[T_\gamma \middle| \hat{N}_0 = x \right]$ or $\mathbb{P}\left(T_\gamma \le T \middle| \hat{N}_0 = x \right)$. For example, we may accept a call only if $\mathbb{P}\left(T_\gamma \le T \middle| \hat{N}_0 = x \right) \le \alpha$, for some α. The Laplace transform of T_α is available, see e.g. [13] and references contained therein.

Appendix

I. Two Other Models

There is another model for best effort data calls, where each such call has an i.i.d. exponentially distributed "file size" with mean μ^{-1}. There is a maximum access rate r at which the file can be served (we assume that $f < r$). Thus the service rate for each file is the minimum of r and X_t, which we write as $r \wedge X_t$, and the call departure rate is $\mu N_t (r \wedge X_t) = \mu r (N_t \wedge \frac{C}{r})$. The process $N_\cdot$ is that of an $M/M/n$ queue with $n = \frac{C}{r}$. (This is only true if C/r is an integer. If C/r is not an integer there is a "fractional server"; this has no effect on our asymptotics.) The asymptotics for this process are known, but are not as explicit as for the $M/M/\infty$ queue. In particular, the results of [5] can be used to show that $\bar{N}^{(C)} \to \bar{N}$ a.s., where $\bar{N}$ is the unique solution to the ordinary differential equation

$$\frac{d\bar{N}_t}{dt} = \theta - \mu \left(r\bar{N}_t \wedge 1 \right) \, , \tag{16}$$

and $\hat{N}^{(C)} \xrightarrow{d} \hat{N}$, where $\hat{N}$ is the unique solution to the stochastic differential equation

$$\begin{aligned} d\hat{N}_t &= \mu 1_{\{r\bar{N}_t \le 1\}} \hat{N}_t^- - \mu 1_{\{r\bar{N}_t < 1\}} \hat{N}_t^+ \\ &\quad + \sqrt{\theta + \mu \left(r\bar{N}_t \wedge 1 \right)} \, dB_t \, , \end{aligned} \tag{17}$$

where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$.

A straightforward analysis of (16) shows that, if $\theta > \mu$ then $\bar{N}_t \to \infty$ as $t \to \infty$, while if $\theta < \mu$ then $\bar{N}_t \to \bar{N}_\infty = \theta/\mu r$. (If $\theta = \mu$ then $\bar{N}_t = \bar{N}_0$ for $\bar{N}_0 > 1/r$, and $\bar{N}_\infty = 1/r$ for $\bar{N}_0 \le 1/r$.) Thus, either the system is unstable or access limited (in which case the link is not fully utilized). This does not seem to be reasonable. The problem is that this model is flawed: It lacks any kind of "self-limiting" mechanism that would allow users to fully utilize the link while cutting back on their usage when their throughput deteriorates.

A third model is motivated by best effort web browsing. We assume that there is a finite population K of web browsing users, and let $K = \eta C$, so that $K \to \infty$ as $C \to \infty$. Web browsing users can be in one of two states: thinking or waiting. In the thinking state users are deciding what to click on next. Each thinking user exits the thinking state after an independent (of other users and his/her own past behaviors) exponentially distributed time with mean λ^{-1}. After clicking, users enter the waiting state, waiting to receive the response from the network. Here we assume, as in the second model, that the service rate for each waiting user is $r \wedge X_t$, and the total departure rate from the waiting state is $\mu r\left(N_t \wedge C/r\right)$, where N_t denotes the number of waiting users. We let M_t denote the number of users in the thinking state, and define $\bar{M}_t^{(C)} = M_t^{(C)}/C$. The results of [5] can be used to show that $\left(\bar{N}^{(C)}, \bar{M}^{(C)}\right) \rightarrow \left(\bar{N}, \bar{M}\right)$ a.s., where $\left(\bar{N}, \bar{M}\right)$ is the unique solution, given $\left(\bar{N}_0, \bar{M}_0\right)$ with $\bar{N}_0 + \bar{M}_0 = \eta$, to the system of differential equations

$$\frac{d\bar{N}_t}{dt} = \lambda\bar{M}_t - \mu\left(r\bar{N}_t \wedge 1\right) \tag{18a}$$

$$\frac{d\bar{M}_t}{dt} = \mu\left(r\bar{N}_t \wedge 1\right) - \lambda\bar{M}_t\ . \tag{18b}$$

A straightforward analysis of (18a) shows that $\left(\bar{N}_t, \bar{M}_t\right) \rightarrow \left(\bar{N}_\infty, \bar{M}_\infty\right)$ as $t \rightarrow \infty$. If $r[\eta - \mu/\lambda] \geq 1$, then $\bar{N}_\infty = \eta - \mu/\lambda \geq r^{-1}$ and $\bar{M}_\infty = \mu/\lambda$. If $r[\eta - \mu/\lambda] \leq 1$, then $\bar{N}_\infty = \lambda\eta/[\lambda + \mu r] \leq r^{-1}$ and $\bar{M}_\infty = \mu r\eta/[\lambda + \mu r]$.

Next let $\hat{M}_t^{(C)} = \sqrt{C}\left(\bar{M}_t^{(C)} - \bar{M}_t\right)$. The results of [5] can be used to show that $\left(\hat{N}^{(C)}, \hat{M}^{(C)}\right) \xrightarrow{d} (\hat{N}, \hat{M})$, where $\hat{N}$ is the unique solution to the stochastic differential equation

$$\begin{aligned} d\hat{N}_t &= \mu 1_{\{r\bar{N}_t \leq 1\}}\hat{N}_t^- - \mu 1_{\{r\bar{N}_t < 1\}}\hat{N}_t^+ \\ &\quad + \sqrt{\lambda\bar{M}_t + \mu(r\bar{N}_t \wedge 1)}\, dB_t\ , \end{aligned} \tag{19}$$

and $\hat{M}_t = -\hat{N}_t$.

Suppose that $\bar{N}_0 = \bar{N}_\infty$. With $\bar{N}_\infty \leq r^{-1}$ (which occurs if $r[\eta - \mu/\lambda] \leq 1$) the system is access limited, and $\bar{X}_t \geq r > f$. In this case inelastic calls always receive sufficient bandwidth and can always be accepted. With $\bar{N}_\infty > r^{-1}$ (which occurs if $r[\eta - \mu/\lambda] > 1$), we can rewrite (19) as

$$d\hat{N}_t = -\mu\hat{N}_t + \sqrt{2\mu}\, dB_t\ .$$

In this case $\hat{N}$ is an Ornstein-Uhlenbeck process, so the analysis of Section IV can be used.

References

[1] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, pp. 1969–1985, 1999, http://www.statslab.cam.ac.uk/˜ frank/evol.html.

[2] R. J. Gibbens and F. P. Kelly, "Distributed connection acceptance control for a connectionless network," in *Proc. of ITC 16*, P. Key and D. Smith, Eds. June 1999, pp. 941–925, Elsevier.

[3] F. P. Kelly, P. B. Key, and S. Zachary, "Distributed admission control," *IEEE JSAC*, 2000, unpublished.

[4] T. G. Kurtz, "Strong approximation theorems for density dependent Markov chains," *Stoch. Proc. Appl.*, vol. 6, pp. 223–240, 1978.

[5] A. Mandelbaum, W. A. Massey, and M. I. Reiman, "Strong approximations for Markovian service networks," *Queueing Systems*, vol. 30, pp. 149–201, 1998.

[6] V. Paxson, and S. Floyd, "Wide-Area Traffic: the Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, pp. 226–244, June 1995.

[7] B. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1987.

[8] F. P. Kelly, "Charging and rate control for elastic traffic," *Europ. Trans. Telecom.*, vol. 8, pp. 33–37, 1997.

[9] J. K. MacKie-Mason and H. R. Varian, "Pricing the Internet," in *Public Access to the Internet*, B. Kahin and J. Keller, Eds. Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[10] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *Journal of the Operational Research Society*, vol. 49, 1998.

[11] N. Semret and A. A. Lazar, "Spot and derivative markets in admission control," in *Proc. of ITC 16*, P. Key and D. Smith, Eds. June 1999, pp. 925–941, Elsevier.

[12] J. C. Hull, *Options, Futures, and other Derivatives*, Prentice-Hall, 1997.

[13] L. M. Ricciardi and S. Sato, "First-passage time density and moments of the Ornstein-Uhlenbeck process," *J. Appl. Prob.*, vol. 25, pp. 43–57, 1988.