A Decision-Theoretic Approach to Call Admission Control in ATM Networks

Richard J. Gibbens, Frank P. Kelly, and Peter B. Key

(Invited Paper)

Abstract—**This paper describes a simple and robust ATM call admission control, and develops the theoretical background for its analysis. Acceptance decisions are based on whether the current load is less than a precalculated threshold, and Bayesian decision theory provides the framework for the choice of thresholds. This methodology allows an explicit treatment of the trade-off between cell loss and call rejection, and of the consequences of estimation error. Further topics discussed include the robustness of the control to departures from model assumptions, its performance relative to a control possessing precise knowledge of all unknown parameters, the relationship between leaky bucket depths and buffer requirements, and the treatment of multiple call types.**

I. Introduction

Asynchronous transfer mode (ATM) is the recommended transfer mode for the introduction of broadband services, capable of integrating services as diverse as telemetry and broadcast TV. (ATM divides all information into short, fixed length cells of 53 octets, comprising a 5-octet header and a 48-octet payload which allows simple and fast hardware switching.) Integrating all services onto a common platform brings a number of benefits, one of which is increased efficiency.

One aspect of efficiency is statistical multiplexing: Information to be transmitted usually varies in time, with peaks and troughs, and if we can fit different calls together in such a way that the peaks do not coincide, then we can carry more calls than by purely allocating capacity according to peak requirements. Typical examples are provided by video, where the rate may depend upon the scene being viewed, and traffic interconnecting local area networks, which is bursty in nature. In fact, one could argue that all information is inherently variable, and it is only the limitations of technology that created constant-rate encoders. For example, the most ubiquitous telecommunications service is telephony, which is usually encoded as a constant bit-rate stream, yet silences exist within speech which can be exploited: Indeed this is done on international links to pack more calls.

Statistical multiplexing offers large potential gains, yet the nature of future traffic is largely unknown. We are unsure what the dominant service in a future integrated broadband network will be; applications are being developed all the time, and the applications can use different coding schemes. On the other hand, ATM has to be capable of offering a very high quality of service. For instance, one of the key quality of service parameters is the cell loss ratio, and values of 1 in 10^9 or better are discussed [41]. How is it possible to guarantee such a high quality of service if we don't know what the traffic will look like, and where in applications such as video the information rate can depend on what is being viewed?

This dilemma has led some to shrink back from statistical multiplexing, and argue that a peak rate should be allocated to each connection. Others have argued that the only way to achieve worthwhile gains is to tightly constrain (police) each source, thus solving the characterization problem by forcing sources to fit into a certain mold. Still others, influenced by computing technologies, advocate feedback controls, which throttle back the source if the network gets congested.

We take a different view. Peak rate allocation solves the problem by ignoring it—treating all sources as though they are constant bit rate—which may be very inefficient. Tightly policing a source potentially requires more of customers than they know, and could have the effect of requiring equipment to be designed according to the policing policy of a particular network. Feedback controls are inappropriate for real-time services, and can result in a very different network structure. Instead, we argue that with a minimum of information, coupled with observation of the network, it is possible to achieve worthwhile statistical multiplexing gains, at the same time meeting stringent quality of service targets.

A major ingredient of our approach is the time-scale decomposition, introduced by Hui [21]. ATM is a connection-oriented transfer mode. If we consider a particular resource, then calls will be set up for a certain time (perhaps minutes) and then cleared down. On a shorter time-scale information is sent in bursts (the burst-scale), interspersed with silences, and on a cell-scale cells may be transmitted at the line rate.

On the cell-scale the peak rate of a source is limited by hardware constraints, and declaration of a *peak cell rate* is mandatory in current ATM standards [39], [40]. These standards allow some variation about the peak rate, the *cell delay variation*, to account for physical layer overheads and distortion introduced into the cell stream by multiplexing with other streams. The standards [39], [40], [18] describe how declarations of peak cell rate and cell delay variation are to be policed by a leaky bucket. Our model assumes cell-scale buffers at each resource, large enough to interleave

Manuscript received, September 30, 1994; revised April 1, 1995. This work is supported in part by a Royal Society University Research Fellowship and in part by the EPSRC under Grant GR/J371896.

R. J. Gibbens and F. P. Kelly are with the Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K.

P. B. Key is with BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.

IEEE Log Number 9412659.

0733–8716/95$04.00 © 1995 IEEE

simultaneous cell arrivals from different streams and to cope with cell delay variation.

At the burst level, efficient statistical multiplexing requires some implicit or explicit estimation of the mean rate of a source. We claim that it is possible to estimate this quantity in such a way that quality of service is guaranteed. Simple estimation will not do—underestimating the mean may cause us to admit too many sources, thus compromising the quality of service standard. Bayesian decision theory allows us to quantify the damage of misestimation, and also incorporate prior knowledge into the model.

In our approach call acceptance decisions are based on a simple threshold rule: An offered call is accepted if the current load is less than a precalculated threshold. The threshold implements an implicit robust estimation procedure, and the decision-theoretic framework facilitates the essential trade-off between the benefits of accepting a call (earning revenue, customer satisfaction) and the disadvantages (threatening quality of service targets).

The problem of call admission control has received considerable attention in the literature: We note in particular the approaches of [4], [20], [33], [35] and the recent work reported in [26]. The approach taken in this paper, developing on [1], [2], [6], [8], [19], [27], [30] differs in that we eschew any attempt to police mean rates, or to gain benefit from burst-scale queueing. For obvious statistical reasons a long-term mean (over, say, the life of a call) cannot be policed efficiently. Similarly the tail behavior of a burst-scale queue is too sensitive to the characteristics of burst lengths to permit robust statistical multiplexing over time. Instead we focus on a very simple scheme, which aims to extract maximum benefit from statistical multiplexing over different calls without requiring detailed knowledge of their mean rates.

The organization of the paper is as follows. In Section II we develop the stochastic process describing the number of calls in progress. This involves thresholds which determine whether or not an offered call is accepted, and in Section II we review various rationales for the choice of these thresholds. We shall find that Bayesian decision theory provides a coherent and general framework within which the several trade-offs involved may be effected. In Section IV we investigate the robustness of our threshold scheme to departures from the model assumptions. In Section V we review the performance of our scheme, and find that it compares favorably with the performance achieved by a scheme which has precise knowledge of all unknown parameters. In Section VI we outline the interrelationship between the parameters of our cell-scale and burst-scale analyses. Finally, in Section VII, we briefly indicate how our approach extends to the case of multiple call types.

II. The Call Process

In Sections II-A and II-B we describe our basic model for the number of calls in progress. This model assumes calls are "on" or "off" and have unit peak rate, and that the resource is unbuffered and of capacity C. Later, in Section VI, we relate these assumptions to a detailed model of the cell level: In particular, we relate the parameter C to the leak rate and bucket depth of leaky bucket policers and the transmission speed and buffer size at a switch. In Section II-C we collect some numerical and analytical observations which aid calculation with our model of the call process.

A. The Basic Model

Suppose that

$$S_n(t) = X_1(t) + X_2(t) + \cdots + X_n(t) \qquad (1)$$

where $X_i(t)$, for distinct values of i and t, are independent, identically distributed random variables with

$$P\{X_i(t) = 1\} = p, \qquad P\{X_i(t) = 0\} = 1 - p. \qquad (2)$$

We interpret $X_i(t)$ as the load produced by call i at time t, and the superposition $S_n(t)$ as the instantaneous load on the resource at time t. We shall call p the *burstiness* parameter: Note that $1/p$ measures the peak to mean ratio of the load produced by a call. We suppose for the moment that the resource is unbuffered and of capacity C: The proportion of cells (or load) lost is then

$$L(n;p) = \frac{M(n;p)}{np} \qquad (3)$$

where

$$\begin{aligned} M(n;p) &= E(S_n - C)^+ && (4)\\ &= \sum_{m=1}^{n-C} mP\{S_n = C + m\}. && (5)\end{aligned}$$

The quantities L and M may be calculated, since under our assumptions S_n has a binomial distribution $B(n, p)$.

The schemes we shall consider have the following basic form: When a call is offered it is accepted if the current load, S_n, is less than $s(n)$. Here n is the number of calls currently in progress, and the vector $\mathbf{s} = (s(n), n = 0, 1, \cdots)$ defines the call admission scheme. In Section II-B we shall discuss the choice of the vector $\mathbf{s}$: Here we analyze the consequences for the stochastic process describing the number of calls in progress of a given vector $\mathbf{s}$. Assume that calls arrive as a Poisson process of rate ν, and that holding times of accepted calls are independent, exponentially distributed random variables with unit mean. (These and our other assumptions will be discussed in Section IV.) Then n will be a birth and death process with transition rates

$$q(n, n-1) = n, \qquad q(n, n+1) = \nu a(n) \qquad (6)$$

where $a(n)$, the acceptance probability when n calls are carried, is given by

$$a(n) = \sum_{i=0}^{s(n)-1} \binom{n}{i} p^i (1-p)^{n-i}. \qquad (7)$$

The stationary distribution for this birth and death process is

$$\pi(n) \propto \frac{\nu^n}{n!} \prod_{r=0}^{n-1} a(r) \qquad (8)$$

where the constant of proportionality is determined by the requirement that $\pi(n), n = 0, 1, \cdots$ sums to unity. Call $\lambda = \nu p$ the *offered load.* When we wish to emphasize the dependence of π on p and λ we shall write $\pi(n) = \pi(n; p, \lambda)$.

The overall cell loss rate is

$$M(p,\lambda) = \sum_{n=1}^{\infty} \pi(n;p,\lambda) M(n;p) \tag{9}$$

while the cell loss ratio is

$$L(p,\lambda) = \frac{M(p,\lambda)}{\sum\limits_{n=1}^{\infty} \pi(n;p,\lambda) np}. \tag{10}$$

The *call* loss probability is

$$E(p,\lambda) = \sum_{n=0}^{\infty} \pi(n;p,\lambda)(1 - a(n)) \tag{11}$$

and the *utilization* is

$$\begin{aligned} U(p,\lambda) &= \sum_{n=1}^{\infty} \pi(n;p,\lambda)(np - M(n;p)) & (12)\\ &= \lambda(1 - E(p,\lambda)) - M(p,\lambda). & (13) \end{aligned}$$

B. Backoff

The offered load to a resource has an important impact upon the performance of a call admission control. If the load is very high, then it may not be enough that a call admission control makes the correct decision on any single occasion with high probability: If calls are offered at a very high rate, the *rate* at which calls are admitted in error may become too large. A natural defence against high offered loads is the following backoff strategy, first described by Bean [1], [2].

Suppose that when an arriving call is rejected, no other calls are considered for acceptance until *after* a call currently in progress has ended. The embedded discrete time chain obtained by observing the system just after departure and acceptance epochs has transition probabilities

$$\begin{aligned} P(n,n+1) &= \frac{\nu a(n)}{\nu + n} & (14)\\ P(n,n-1) &= 1 - P(n,n+1). & (15) \end{aligned}$$

The time spent by the continuous time process in state n has mean

$$\frac{1}{\nu+n} + \frac{\nu}{\nu+n}(1 - a(n))\frac{1}{n} = \frac{1}{n}(1 - P(n,n+1)). \tag{16}$$

Thus, the stationary distribution for the continuous time process can be deduced [2] and is

$$\pi(n) \propto \begin{cases} \frac{1}{n} \prod\limits_{r=1}^{n-1} \frac{\nu a(r)}{r + \nu(1 - a(r))}, & n = 1, 2, \cdots \\ \frac{1}{\nu}, & n = 0. \end{cases} \tag{17}$$

Interestingly n is not Markov: To make it Markov append a 0 or 1 according as the system is awaiting an arrival or a departure. The augmented process (n, d) is Markov, with transition rates

$$\begin{aligned} q((n,0),(n+1,0)) &= \nu a(n) & (18)\\ q((n,0),(n,1)) &= \nu(1 - a(n)) & (19)\\ q((n,0),(n-1,0)) &= q((n,1),(n-1,0)) = n & (20) \end{aligned}$$

and stationary distribution

$$\begin{aligned} \pi(n,0) &= \pi(n)\frac{n}{n + \nu(1 - a(n))} \qquad n = 1, 2, \cdots & (21)\\ \pi(n,1) &= \pi(n)\frac{\nu(1 - a(n))}{n + \nu(1 - a(n))} \qquad n = 1, 2, \cdots & (22) \end{aligned}$$

with $\pi(0,0) = \pi(0)$ and $\pi(0,1) = 0$.

The effectiveness of the backoff strategy is well illustrated by its performance under very heavy traffic: Observe that the limiting case of the distribution (17), as $\nu \to \infty$, is

$$\pi(n) \propto \frac{1}{n} \prod_{r=n_{\min}}^{n-1} \frac{a(r)}{1 - a(r)}, \quad n \geq n_{\min} \tag{23}$$

where $n_{\min} = 1 + \max\{r : a(r) = 1\}$, a distribution with mode n^* where $a(n^*) \approx \frac{1}{2}$. This model, and a variation where a deterministic wait is imposed after any change in the number of calls before a new connection request is considered, are further discussed by Bean [1], [2].

The backoff strategy may lose a little efficiency when the offered load λ is known (for example, an optimized scheme with backoff may have a slightly lower utilization for the same cell loss ratio than an optimized scheme without backoff), but this seems to be more than outweighed by its robustness against unpredicted variations in λ. Of course there are many other methods of producing a delay following a rejection decision, or of otherwise limiting the offered load. We choose the backoff strategy as one that is easy to define and to analyze, and which captures the essential features of a good scheme. Except where explicitly indicated, we shall henceforth compute π from (17).

C. Computational Preliminaries

The Chernoff bound [3] for a binomial random variable is

$$P\{S_n > C\} = P\{S_n > na\} \leq e^{-nK(a,p)} \tag{24}$$

where

$$K(a,p) = a \log \frac{a}{p} + (1 - a) \log \frac{1 - a}{1 - p} \tag{25}$$

and $a = C/n$. This bound is often used as an approximation, the *large deviations* approximation, and is asymptotically tight, in that

$$\lim_{n \to \infty} \frac{1}{n} \log P\{S_n > C\} = -K(a,p). \tag{26}$$

The approximation is not used directly in our later calculations, but it provides a useful check and it gives some analytical insight into our results.

Later we shall find it convenient to have quantities such as the cell loss rate defined as continuous functions of a real

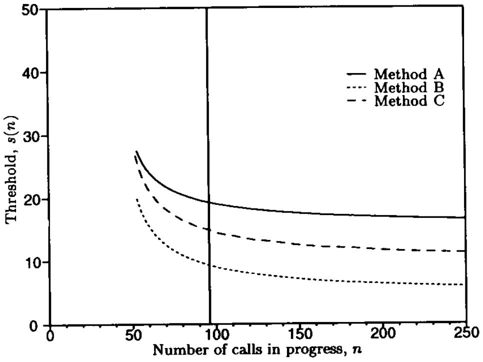

Fig. 1. Call acceptance boundaries: Introductory examples ($C = 50$). The region below or to the left of a curve indicates where a new call may be accepted.

vector s: This may be achieved by randomizing acceptance decisions. When $s(n)$ is nonintegral, extend (7) as follows

$$a(n) = \sum_{i=0}^{\lfloor s(n)\rfloor - 1} \binom{n}{i} p^i (1-p)^{n-i} + (s(n) - \lfloor s(n)\rfloor) \binom{n}{\lfloor s(n)\rfloor} p^{\lfloor s(n)\rfloor}(1-p)^{n-\lfloor s(n)\rfloor}. \tag{27}$$

The interpretation is straightforward: A call is accepted if $S_n < \lfloor s(n)\rfloor$, rejected if $S_n > \lfloor s(n)\rfloor$, and accepted with probability $s(n) - \lfloor s(n)\rfloor$ if $S_n = \lfloor s(n)\rfloor$.

III. Choice of Thresholds

In this section we discuss the choice of the vector s. There are many possibilities, some of them illustrated in Fig. 1. If the burstiness parameter p is known, then one possibility is to accept a call if and only if the proportion of cells lost $L(n; p)$ remains less than a predefined limit. This corresponds to a vertical line whose position depends upon p. Another possibility, available if the offered load λ is also known, is to accept a call if and only if the cell loss ratio $L(p, \lambda)$ remains less than a predefined limit. This corresponds to a vertical line whose position depends upon both p and λ. (The vertical line in Fig. 1 achieves a cell loss ratio of 10^{-10} when $(p, \lambda) = (0.25, 25)$.) Other possibilities, more appropriate when neither p nor λ are known with certainty, are shown labeled A, B and C, (where, to the left of the sections shown, $s(n) = +\infty$). A horizontal line corresponds to a call admission control which uses only the measured load S_n, and requires no knowledge of n.

There are various rationales for the choice of the vector s. The first approach we describe, in Section III-A, is a straightforward but naive attack on the problem of estimating a source type's effective bandwidth [20]–[22]. Essentially it assumes the peak rate of a source is known, and uses the current load to estimate the source burstiness. The difficulty with this approach is that it does not adequately deal with the uncertainty inherent in the resulting estimate of burstiness, and as a consequence cell loss ratios can become too high. It is possible to make the approach more conservative, by adding safety margins, but there is no clear criterion for the choice of these margins.

Section III-B describes a simple Bayesian approach. This assumes a prior distribution is available for the burstiness parameter. Different choices of prior can represent differing amounts of uncertainty about the parameter. Additionally information is available from measurements of load. The Bayesian formulation allows these two forms of information, prior and data based, to be integrated. A minimax variant is discussed in Section III-C.

In Section III-D we describe our preferred approach, a Bayesian decision-theoretic approach. This provides a coherent and general framework within which to combine prior knowledge and measurement data, and to trade off utilization and cell loss. In Section III-E we describe its application to the important case where we may have load measurements only.

A. A Naive Approach

From knowledge of n, the number of calls in progress, and $S_n(t)$, the instantaneous load at time t, can we learn something of the parameter p, and hence of whether or not an additional call should be accepted? Consider the following very simple scheme: Estimate p by $\widehat{p} = S_n/n$, and accept a newly offered call if and only if

$$L(n+1; \widehat{p}) < 10^{-\gamma}. \tag{28}$$

Thus the estimate $\widehat{p}$ is treated as an exact observation on p, and a newly offered call is admitted if, for this value of p, the quality of service guarantee can be met with $n+1$ accepted sources. The vector s, defined in Section II-A, is determined by

$$s(n) = \max\left\{s : L\left(n+1; \frac{s}{n}\right) < 10^{-\gamma}\right\} \tag{29}$$

and the cell loss ratio and utilization may be calculated from the earlier equations.

The basic scheme described above can be refined in several ways. For example, the estimate S_n/n might be improved by averaging over a longer period, and a more conservative estimate of p might be used. For example, since S_n/n has an approximate $N(p, p(1-p)/n)$ distribution

$$\widehat{p} = \frac{S_n}{n} + \alpha\sqrt{\frac{\frac{S_n}{n}\left(1 - \frac{S_n}{n}\right)}{n}} \tag{30}$$

estimates the mean of $\frac{S_n}{n}$ plus α standard deviations. The parameter $\alpha > 0$ is then a safety margin which reduces the proportion of cells lost. The curve labeled A in Fig. 1 illustrates this approach with the choice $C = 50$, $\gamma = 10$ and $\alpha = 1$.

B. A Simple Bayes Model

The approach in Section III-A leaves unclear how a control variable such as α should vary with parameters such as the capacity C or the desired cell loss ratio $10^{-\gamma}$. Next, we consider an approach which attempts to avoid arbitrary choices of control variables, through a Bayesian formulation.

We now suppose that the parameter p has a known prior distribution $f(p)$ over $p \in [0, 1]$. Then, conditional on the load $S_n(0) = s$ at time $t = 0$, we can calculate the posterior distribution $f(p \mid S_n(0) = s)$. For example, if $f(p)$ is the Beta distribution with parameters α and β

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1 - p)^{\beta - 1} \tag{31}$$

then $f(p \mid S_n(0) = s)$ is again a Beta distribution, with parameters $\alpha + s$, $\beta + n - s$ [5]. Conditional on the observation $S_n(0) = s$, the ratio of the expected number of cells lost to the expected number of cells offered, over the following short interval, from $n + 1$ calls in progress is

$$\begin{aligned} &\hat{L}(n+1, s) \\ &= \frac{E[(S_{n+1}(\epsilon) - C)^+ \mid S_n(0) = s]}{(n+1)E[X_1(\epsilon) \mid S_n(0) = s]} \qquad (32)\\ &= \frac{\int\limits_0^1 \sum\limits_{m=1}^{n+1-C} mP\{S_{n+1}(\epsilon) = C + m \mid p\} f(p \mid S_n(0) = s)dp}{(n+1)\int\limits_0^1 pf(p \mid S_n(0) = s)dp} \qquad (33) \end{aligned}$$

for $\epsilon > 0$. Consider now the following call acceptance strategy: Accept a new call if and only if

$$\hat{L}(n+1, s) \leq 10^{-\gamma}. \tag{34}$$

The rationale is that a new call is accepted when the estimated cell loss proportion appears satisfactory.

The above approach is extensively illustrated in [6], [19] and the case with uniform prior $f(p) = 1$, $\gamma = 10$, $C = 50$ is shown in Fig. 1 labeled B. A fully sequential Bayesian approach would repeatedly update the posterior distribution with each successive observation of load: The mean of the posterior distribution would converge to p while its variance would converge to zero. For simplicity and robustness we prefer not to adopt this approach: Rather, in Section III-D, we shall further develop our simple Bayes model within a decision-theoretic framework. We shall see in Section V that the loss of efficiency is minimal.

C. A Minimax Approach

Some analytical insight into the issues of this paper is given by the following simplified model [2], [6]. Suppose we replace the condition (34) by the condition

$$\max_{p \in [0,1]} P\{S_n(\epsilon) > C \quad \text{and} \quad S_n(0) < s \mid p\} \leq e^{-\xi}. \tag{35}$$

for $\epsilon > 0$. We thus control the probability that a measurement $S_n(0) = s$ will appear low enough to accept further calls, and yet the subsequent load $S_n(\epsilon)$ will be too large. In the simplified model the criterion is resource-based rather than stream-based, in that it is expressed in terms of the event that the resource is overloaded, rather than directly in terms of the proportion of cells from the offered stream that are lost. Otherwise the simplified model may be viewed as a variant of the Bayesian approach, where the prior is chosen to concentrate mass on the worse case value of p.

Using the large deviation approximation and bound of (24)–(26)

$$P\{S_n > C \mid p\} \leq e^{-nK(a,p)} \tag{36}$$

where $a = C/n$. Similarly

$$\begin{aligned} P\{S_n < s \mid p\} &= P\{n - S_n > n - s \mid p\} && (37)\\ &\leq e^{-nK(1-a, 1-p)} && (38)\\ &= e^{-nK(a,p)} && (39) \end{aligned}$$

where $a = s/n$. Thus since under our assumptions $S_n(0)$ and $S_n(\epsilon)$ are independent given p, (35) is implied by

$$\max_p \left[e^{-nK(\frac{C}{n}, p)} e^{-nK(\frac{s}{n}, p)}\right] \leq e^{-\xi}. \tag{40}$$

The maximum over p occurs where

$$p = \frac{C + s}{2n}. \tag{41}$$

Thus (35) is implied by

$$n\left[K\left(\frac{C}{n}, \frac{C+s}{2n}\right) + K\left(\frac{s}{n}, \frac{C+s}{2n}\right)\right] \geq \xi. \tag{42}$$

Observe that the maximizing p is halfway between the unbiased estimator s/n and the saturating level C/n: Overload in this simplified model is caused by the joint occurrence of a moderately low value of $S_n(0)$ and a moderately high value of $S_n(\epsilon)$, rather than an especially extreme value of either alone. This, the *estimation effect*, is an important insight: Errors in estimation can be a major cause of cell loss.

We have used some simplifying approximations in order to obtain the above analytical insight. It is, however, possible to define the minimax approach without such approximations, and with a stream-based criterion: We simply define $s(n)$ to the maximum value of s such that

$$\max_p \left[P\{S_n(0) < s \mid p\} \frac{E[(S_{n+1}(\epsilon) - C)^+ \mid p]}{(n+1)p}\right] \leq 10^{-\gamma}. \tag{43}$$

Thus if we consider a single call acceptance decision and the immediately following interval, this criterion limits, over all values of p, the expected cell loss ratio on the event that a call is accepted. The resulting s curve for the case $\gamma = 10$ is illustrated in Fig. 1 labeled C.

The approaches of Sections III-A, III-B, and III-C are concerned to bound cell loss over the period immediately following a call admission. For example, the maximizing p in (35) or (43) is recomputed for each distinct value of n. It seems unduly pessimistic to suppose that changes in p are this malevolent, and next we describe an approach which assumes that p, while unknown, is relatively stable.

D. A Decision-Theoretic Framework

Now suppose that we have a prior distribution $f(p, \lambda)$ for the parameters p and λ, and that we choose the control s to maximize

$$\begin{aligned} &\iint \left[\sum_{n=1}^{\infty} \pi(n; p, \lambda)(np - yM(n; p))\right] f(p, \lambda)dpd\lambda && (44)\\ &= \iint [U(p, \lambda) - (y-1)M(p, \lambda)] f(p, \lambda)dpd\lambda. && (45) \end{aligned}$$

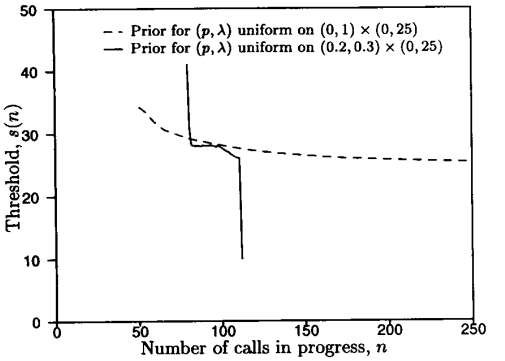

Fig. 2. Optimized s curves. As prior information on burstiness parameter p, becomes more precise, the optimal curve approaches a vertical line.

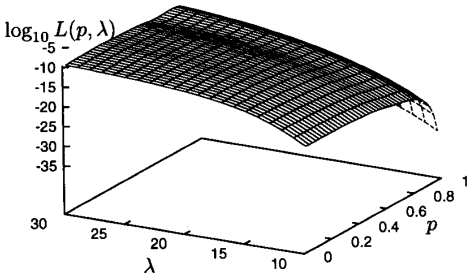

Fig. 3. Cell loss for s the dashed curve from Fig. 2. Over a wide range of offered load, λ, and burstiness parameter, p, the cell loss ratio is well controlled.

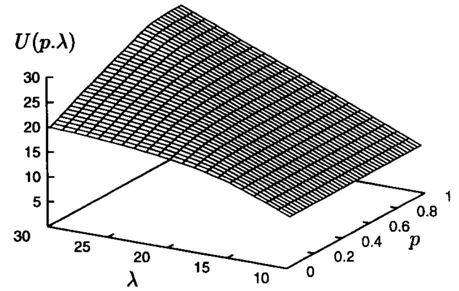

Fig. 4. Utilization for s the dashed curve from Fig. 2. Utilization necessarily drops below offered load when p is small and λ is large.

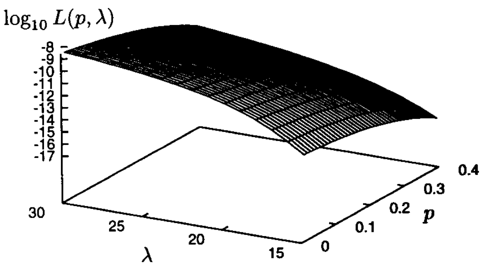

Fig. 5. Cell loss with load measurements only: s horizontal. Note that the region used is smaller than in Fig. 3.

Expression (44) is proportional to the expected reward per unit time if each offered cell attracts a reward of one unit while each lost cell incurs a penalty of y units. Thus the constant y measures our trade-off between utilization and cell loss.

Note that the optimization of (45) is just a maximization of expected utilization for a given expected cell loss rate, with $y - 1$ a Lagrange multiplier attached to the cell loss constraint. Thus y measures the *marginal* cell loss ratio: If a perturbation to the s curve allows additional carried traffic, then each additional offered cell has probability y^{-1} of being lost. In the classical theory of loss networks the use of a *marginal* loss rate in capacity expansion decisions is known as Moe's principle [11, pp. 216–210]. Although we do not discuss routing in this paper, we note in passing the importance of marginal rather than average loss rates in system optimal routing strategies [23, sec. 6].

The dashed line illustrated in Fig. 2 shows the form of the optimizing s curve for $C = 50$, $y = 10^9$ and $f(p, \lambda)$ uniform on $(0, 1) \times (0, 25)$. The cell loss ratio (10) and utilization (12) achieved by this s curve are illustrated in Fig. 3 and Fig. 4, respectively. Note that the cell loss ratio is well controlled over a wide range of values of p and λ: This property is not explicitly sought in the optimization procedure, but is a natural consequence of the form of the objective function (45) and the use of a uniform prior distribution. Bean [2], in his study of the case $\lambda = \infty$, shows that, when this is the objective, it is possible to choose the s curve so that the cell loss ratio is approximately constant over nearly all values of p, apart from a small interval close to $p = 1$.

The solid line in Fig. 2 illustrates the form of the optimal s curve for $f(p, \lambda)$ uniform on $(0.2, 0.3) \times (0, 25)$ (to the right of the section shown, $s(n) = 0$). As the prior information on the parameter p becomes more precise the optimal s curve approaches a vertical line. If p is known, then the load measurement $s(n)$ conveys no useful information, and the number of calls which may be safely admitted can be precomputed [21], [22]. Of course such an admission control is highly vulnerable to errors in the specification of p. Next we consider another extreme, where the s curve is a horizontal line, an extreme which is much more robust against misspecification of the parameter p.

E. Load Measurements Only

Now we consider a scheme where $s(n) = s$ for all n, so that call admission decisions are taken on the basis of the known peak rate and the instantaneous load, and without knowledge of the number of sources already connected. In Fig. 5 we illustrate the cell loss ratio for a scheme with $s = 26.83$ for $C = 50$, for various values of p and λ. This value of s optimizes the objective function (44) when $y = 10^9$ and $f(p, \lambda)$ places mass 1 on $(p, \lambda) = (0.2, 25)$: From Fig. 5 we see that the cell loss ratio is fairly well controlled for a range of (p, λ) values about (0.2,25): Compared with Fig. 3 the cell loss ratio falls away more quickly as p increases.

Recall that the cell loss ratio $L(p, \lambda)$ is defined, via equations (9) and (10), in terms of the weighted sum $M(p, \lambda)$. In

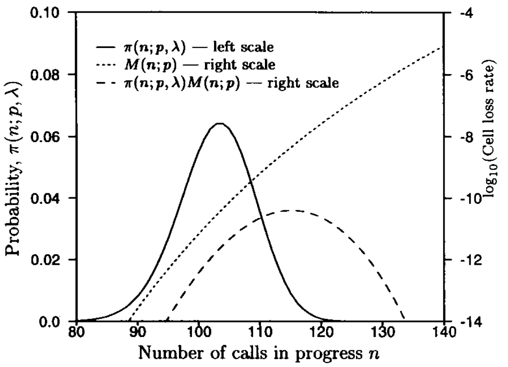

Fig. 6. Terms of $M(p,\lambda)$ for varying n. Larger values of $M(n;p)$ are associated with rapidly diminishing values of $\pi(n;p,\lambda)$.

Fig. 6 we show the magnitude of the various contributions to this sum, when $(p,\lambda) = (0.2, 25)$ and $L(p,\lambda) = 10^{-9.33}$. We note that the largest contribution occurs at $n = 115$, where $\pi(n;p,\lambda) = 0.009$ and $M(n;p) = 10^{-8.37}$. At $n = 131$, $\pi(n;p,\lambda) = 10^{-6.87}$ and $M(n;p) = 10^{-6.11}$, corresponding to a proportion of cells lost of $10^{-6.11}/131 \times 0.2 = 10^{-8.29}$. Thus, while the overall cell loss ratio is $10^{-9.3}$, for a fraction $10^{-6.9}$ of the time the proportion of cells lost is $10^{-8.3}$. Any call admission control which admits the possibility of estimation error may occasionally have too many calls in progress: Our methodology allows this effect to be quantified and assessed.

In Fig. 7 we illustrate how s varies with capacity C; for given p and C we plot the value s optimizing the objective function (44) when $y = 10^9$ and $f(p,\lambda)$ places unit mass on (p, C).

As p approaches zero the distribution π places probability mass on increasingly large values of n. This can complicate numerical calculations, but fortunately there are simple analytical relationships for the limit case, as $p \to 0$. In this limit the product np converges in distribution to a constant δ: Consideration of the mode of the distribution π fixes δ as the solution to

$$\lambda P_\delta(s) = \delta + \lambda(1 - P_\delta(s)) \qquad (46)$$

where $P_\delta(s)$ is constructed from the Poisson distribution function by

$$P_\delta(s) = e^{-\delta} \sum_{r=0}^{\lfloor s \rfloor - 1} \frac{\delta^r}{r!} + (s - \lfloor s \rfloor)\frac{\delta^{\lfloor s \rfloor}}{\lfloor s \rfloor !}. \qquad (47)$$

The cell loss rate is $E(S - C)^+$ where S has a Poisson distribution with mean δ and is thus

$$M(0,\lambda) = e^{-\delta} \sum_{m=1}^{\infty} m \frac{\delta^{C+m}}{(C+m)!}. \qquad (48)$$

The cell loss ratio is thus

$$L(0,\lambda) = \frac{M(0,\lambda)}{\delta} \qquad (49)$$

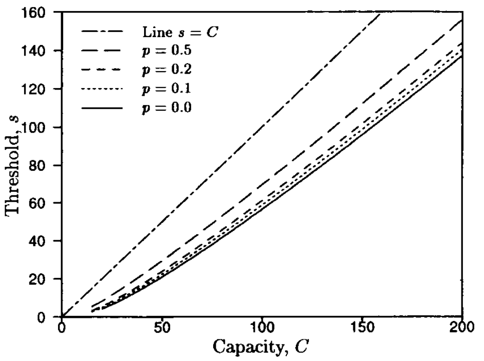

Fig. 7. Variation of s with C. The threshold level s decreases slowly with p.

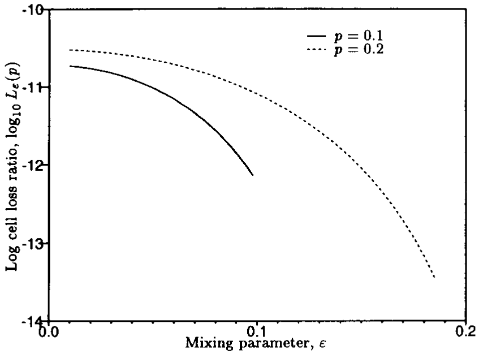

Fig. 8. Mixed call types: $L_\varepsilon(p)$. The assumption that calls have the same burstiness parameter, p, is conservative: It corresponds to the choice $\varepsilon = 0$.

and the utilization is

$$U(0,\lambda) = \delta(1 - L(0,\lambda)). \qquad (50)$$

These relations can be used to compute the $p = 0$ values in Fig. 3, 5, 7, and 11.

The above limit analysis complements the discussion of the estimation effect, in Section III-C. For p small and C large, typical values of n are larger, and thus many call admissions are needed to substantially increase the carried load. However for C smaller and p larger, relatively fewer call admissions may be enough to overload the resource, and we may expect the estimation effect to be more marked. See also Griffiths and Key [19], where the simple estimator (41) is related to a limit case, as $n \to \infty$, of the model of Section III-B.

We note here that a good starting point for the optimization of Section III-D can be constructed as follows. Choose the value $s(p)$ so that s $= (s(p), n = 0, 1, \cdots)$ optimizes (45), with a prior which concentrates mass on a point (p, λ). Let $n(p)$ be the mode of the distribution $(\pi(n), n = 0, 1, \cdots)$, under this scheme. Then define the curve s $= (s(n), n = 0, 1, \cdots)$ by $s(n) \geq n + 1$ for $0 \leq n < C$, and by $s(n) = s(p^*)$ where $n(p^*) = n$ for $n \geq C$.

To give some feel for the various levels of cell loss ratio, we remark that on a 150 Mb/s link, fully utilized, a cell loss ratio of 10^{-10} corresponds to losing about three cells per day:

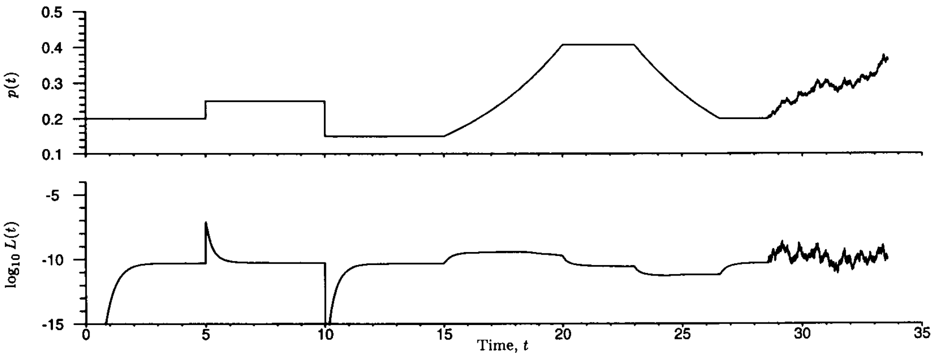

Fig. 9. Fluctuating burstiness: The effect of changes in $p(t)$ on the cell loss ratio $L(t)$.

A cell loss ratio of 10^{-13} corresponds to losing about one cell per year.

IV. Robustness

In this section we assume a control s is specified and study its behavior under departures from the model assumptions.

A. Mixed call types

Suppose that while n calls are in progress, $\lfloor \frac{n}{2} \rfloor$ of them have burstiness parameter $p - \varepsilon$, and $n - \lfloor \frac{n}{2} \rfloor$ have burstiness parameter $p + \varepsilon'_n$, where

$$\left\lfloor \frac{n}{2} \right\rfloor (p - \varepsilon) + \left(n - \left\lfloor \frac{n}{2} \right\rfloor\right)(p + \varepsilon'_n) = np. \tag{51}$$

Thus the expected proportion of calls active remains at p. The birth and death process n is altered, since

$$a_\varepsilon(n) = \sum_{i=0}^{s(n)-1} P_{n,\varepsilon}\{X = i\} \tag{52}$$

where

$$P_{n,\varepsilon}\{X = i\} = \sum_{k=0}^{i} g\left(\varepsilon, \left\lfloor \frac{n}{2} \right\rfloor, k\right) g\left(-\varepsilon'_n, n - \left\lfloor \frac{n}{2} \right\rfloor, i - k\right) \tag{53}$$

and

$$g(\eta, m, j) = \binom{m}{j} (p - \eta)^j (1 - p + \eta)^{m-j}. \tag{54}$$

Hence the altered stationary distribution $\pi_\varepsilon(n)$ can be calculated, using (17). The altered cell loss ratio can be determined using

$$M_\varepsilon(n, p) = \sum_{m=1}^{n-C} m P_{n,\varepsilon}\{X = C + m\} \tag{55}$$

and hence $L_\varepsilon(p)$, the altered cell loss ratio, can be calculated from (9) and (10).

Fig. 8 shows $L_\varepsilon(p)$ as a function of ε for various values of p where $C = 50$, $\lambda = 25$ and s is the dashed curve from Fig. 2. Observe that as ε increases, the cell loss ratio improves. This is not unexpected [1], [6], [19], [33]: The sum of a collection of independent, not necessarily identical, Bernoulli random variables has largest variance for given mean when the Bernoulli random variables are identically distributed. Nevertheless, the magnitude of the effect is quite striking.

It is perhaps even clearer that our use of a Bernoulli random variable for the load produced by a call is conservative: See [19] for a formalization of this remark in terms of the Chernoff bound (24).

B. Fluctuating burstiness

Next we investigate the sensitivity of our results to fluctuations over time in the burstiness parameter p. We suppose that $p = p(t)$, where $p(t) = 0$ for $t \leq 0$ so that the system starts empty at time $t = 0$. The subsequent evolution of the distribution $\pi(n, d; t)$ is given by the forward equations

$$\frac{\partial \pi(n, d; t)}{\partial t} = Q(t)\pi(n, d; t) \tag{56}$$

where the q-matrix $Q(t)$ is given through (18)–(20) and (27) in terms of $p(t)$. Thus the time dependent cell loss ratio, $L(t) = L(p, \lambda; t)$, can be calculated from (5) and (9)–(10).

We illustrate the case $C = 50$, $\lambda = 25$, and use the s shown as the dashed line in Fig. 2. Let us suppose that $p(t)$, $t \geq 0$ is given by the top graph of Fig. 9; then the cell loss ratio, calculated as described above, is given by the lower graph. The system starts empty, at time $t = 0$, and so initially $L(t)$ increases from zero. The equilibrium cell loss ratio for $p = 0.2$ is $10^{-10.28}$, and we see that $L(t)$ converges to this value in about two call holding times. At $t = 5$ the parameter $p(t)$ is suddenly increased to 0.25, and $L(t)$ increases to $10^{-7.14}$. After about one call holding time $L(t)$ has dropped to $10^{-10.25}$. Similarly, after $p(t)$ drops suddenly to 0.15, it takes one or two call holding times for near equilibrium to be restored. The exponential convergence to equilibrium, and the fact that it takes just a few call holding times to reach near equilibrium, are straightforward consequences of the simple structure of the Markov chain describing the number of calls in progress.

We have used sudden changes in the parameter $p(t)$ to illustrate the essential time constants of the system, rather than because such changes are likely to occur with real traffic. A more realistic fluctuation in p would be either of smaller magnitude, corresponding to changes in the nature of a single call, or continuous, corresponding to a gradual shift in the nature of many different calls. We next discuss such changes. Suppose, that from $t = 15$ to $t = 20$ the parameter $p(t)$ increases so that $p'(t) = 0.2p(t)$. Then we observe a slight, but not substantial, upward displacement in the cell loss ratio over the same time interval. Similarly as $p(t)$ decreases so that $p'(t) = -0.2p(t)$ over the interval beginning at $t = 23$ up until $p(t) = 0.2$, there is a slight downward displacement in the cell loss ratio.

The cell loss ratio is relatively insensitive to the level of the parameter p, but is affected by its rate of change and, in particular, the derivative of $\log p$. The final part of Fig. 9 illustrates the system's response when p is driven by the equation

$$d(\log p(t)) = 0.1(dW(t) + dt) \tag{57}$$

where $W(t)$ is a standard Brownian motion.

It is perhaps worth emphasising that Fig. 9 is calculated assuming all calls share the same parameter p: If fluctuations in the parameter p are independent between calls, with the *average* parameter p behaving as illustrated in Fig. 9, then the cell loss ratio will be improved (see Section IV-A). We conclude that the call admission control is able to respond flexibly and robustly to changes in the burstiness of calls.

C. Sensitivity to Traffic Processes

There has been much discussion recently about the traffic models appropriate for high-speed networks [32], [37] and part of our motivation has been to develop a call admission model that does not depend critically upon traffic assumptions. We note here that the bufferless burst-scale model, leading to (3), makes no assumptions about burst lengths, while our cell-scale queueing model, described in Section VI, makes minimal assumptions on the policing of peak rates. At the call-scale, the distribution (17) is relatively robust to departures from the Poisson and exponential assumptions: Indeed the simple model leading to the distribution (8) is, like the Erlang loss model, insensitive to the holding time distribution of accepted calls [12], [15], [23].

D. Separation of Time-Scales

Our analysis assumes a complete separation of time-scales between the call, burst and cell levels. The assumed separation between call and burst time-scales is embodied in our supposition in Section II-A that the random variables $X_i(t)$ are independent for distinct values of t, while the assumed separation between burst and cell time-scales will allow our stationary analysis at the cell level in Section VI.

The separation of burst and cell levels has been well studied [38]. The assumed separation between call and burst levels requires more discussion. Bean [1] adapts a weak convergence theorem of Hunt and Kurtz [17] to establish the separation as an asymptotic result, as the ratio of mean burst length to mean call holding time approaches zero. This limit result is reassuring, as too is the observation that given n and p a positive correlation between loads at distinct times makes less likely the dangerous combination identified in Section III-C of a low value of $S_n(0)$ and a high value of $S_n(\epsilon)$. There remains, however, one further point to explore. If successive load measurements are positively correlated, then under a very high offered load a sequence of calls may be admitted in quick succession as a consequence of a sequence of positively correlated load measurements. We note that there are several safeguards which will prevent this, and the simplicity of the basic model allows many of them to be readily analyzed. For example, Bean's analysis [1], [2] includes a deterministic wait after an admission before a new connection request is considered. Alternatively, we could analyze the model with $\nu = \infty$, and assume that several connections may be accepted together, as a result of a single load measurement. For example, if a measurement S_n leads to the acceptance of $s(n) - S_n$ calls, then the stationary distribution for the number of calls in progress becomes

$$\pi(n) \propto \frac{1}{n} \prod_{r=n_{\min}}^{n-1} \frac{P\{S_{r-1} < s(r-1) - 1\}}{P\{S_r \geq s(r) - 1\}}. \tag{58}$$

For comparison distribution (23) may be rewritten as

$$\pi(n) \propto \frac{1}{n} \prod_{r=n_{\min}}^{n-1} \frac{P\{S_r < s(r)\}}{P\{S_r \geq s(r)\}}. \tag{59}$$

The explicit forms (58) and (59) allow the effects of multiple acceptances to be readily assessed.

V. Performance

In this section we compare the performance of a scheme s with the best performance possible when the burstiness parameter p and the offered load λ are known.

A. Optimal Performance

If the parameters p and λ are known then the policy maximizing the utilization U for a given cell loss ratio L is of a very simple form. The policy is as follows: Accept an offered call if the number of calls in progress, n, is less than a critical value N; reject an offered call if $n > N$; and accept an offered call with probability η if $n = N$. The parameters N and η are chosen so that the cell loss ratio is exactly the desired level L. Thus the optimal scheme has stationary distribution (8) with

$$a(n) = \begin{cases} 1, & n < N \\ 0, & n > N \\ \eta, & n = N \end{cases} \tag{60}$$

where N and η depend upon p and λ.

B. Comparisons

How do the schemes s of Sections III-D and III-E compare with the optimal scheme? In Fig. 10 we plot the ratio of the utilizations achieved over different values of p and λ, for s the dashed line of Fig. 2. For each value of p and λ, the optimal scheme is chosen to achieve the *same* cell loss ratio (10) as is achieved at the point (p, λ) by the scheme s. This plot is necessarily bounded above by 100%, by the definition of the

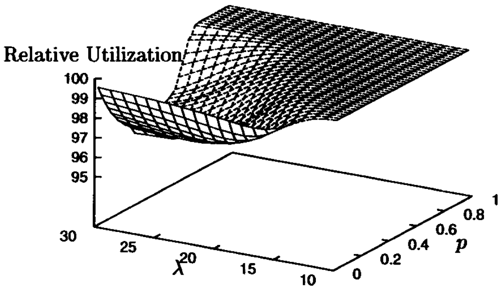

Fig. 10. Utilization relative to optimal scheme where the parameters p and λ are known ($C = 50$).

optimal scheme. The important point to note is that, relative to the optimal scheme, the scheme s looses no more than 3% utilization over a very wide range of values of p and λ.

We note that any scheme which uses repeated measurements in order to estimate p and λ could not perform better than the optimal scheme; hence, as mentioned at the end of Section III-B, the loss of efficiency through using but a single observation of load is minimal.

In Fig. 11 we plot the utilizations achieved by the schemes of Section III-E, with $\lambda = C$ and with s values as given in Fig. 7. We also plot the utilizations achieved by the optimal scheme, where for each value of p and $\lambda = C$ the optimal scheme is chosen to match the cell loss ratio of the corresponding scheme of Section III-E. The optimal schemes necessarily achieve a higher utilization, but we note that the dominance is not great. We have seen in Fig. 5 that the "load measurement only" schemes of Section III-E are relatively robust to variations in p and λ: Fig. 11 shows that they are also nearly as efficient as a scheme that knows p and λ precisely.

VI. Time Scales, Buffers and Cell Delay Variation

The development of earlier sections has assumed a very simple model of the relationship between the cell and burst levels: We have assumed, in (1) and (2), that the load produced by a call is either 0 or 1, In this section we explore the cell time-scale in more detail, and show how the essential parameter of the burst level, the capacity C, may be deduced from the parameters describing the cell level. We adopt a direct approach, motivated by ATM standards [39], [40], [18]. These standards require that each connection has to specify two parameters to the network, a peak rate and a cell delay variation (CDV) tolerance τ, which allows a specified variation about the peak rate. A tolerance of zero corresponds to no variation. The network can check that the connection is behaving by using a policer such as a simple leaky bucket, which has a specified leak rate of r cells per second and specified depth of b cells. The "bucket" is a counter which increments by 1 when a cell is admitted, and decreases at rate r when positive. Cells are discarded or delayed if the counter would otherwise increase beyond b. The CDV tolerance [39], [40] is then $\tau = (b - 1)/r$.

Suppose now we have a number of active connections which are multiplexed together in a switch, where the switch has a buffer of B cells, and a maximum service rate of R cells/s. This is a realistic model of common ATM switches which approximate to a nonblocking cross connect with output buffering [30]. Each connection is policed by a leaky bucket with parameters (r_i, b_i).

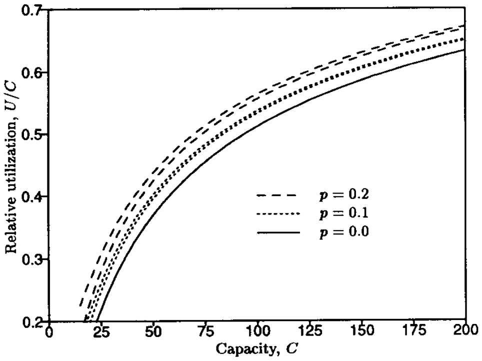

Fig. 11. Utilization relative to optimal scheme where the parameters p and λ are known. For each value of p the upper curve is the optimal scheme and the lower curve is the load only scheme. Note that the relative utilization's natural scale is 0%–100%, while our scale runs from 20% to 70%.

In order not to overload the buffer, the traffic intensity must be less than 1, say less than ρ, that is

$$\sum_i r_i \leq \rho R. \tag{61}$$

Now in a time interval of length t, the number of cells allowed through leaky bucket i is bounded above by

$$b_i + tr_i \tag{62}$$

so that the number of cells arriving at the buffer in an interval of length t is bounded by

$$\sum_i b_i + t\sum_i r_i. \tag{63}$$

But the queue length Q of the buffer at an arbitrary time is bounded [13] by

$$Q \leq \sup_{t>0}\left\{t\sum r_i + \sum_i b_i - Rt\right\} \tag{64}$$

$$\leq \sum_i b_i \tag{65}$$

using (61). Therefore no cells will be lost provided that

$$\sum_i b_i \leq B. \tag{66}$$

Equation (61) corresponds to peak rate allocation, where we use the contracted peak rate or bucket leak rate to decide whether to admit calls, subject to the CDV tolerances being within the limits of the buffer, as given by (66). Note that (66) is implied by (61), provided that

$$\frac{b_i}{r_i} \leq \frac{B}{\rho R} \tag{67}$$

a relation bounding the time to empty each bucket in terms of the time to empty the buffer.

Inequality (66) is a worst case bound, and can be overly pessimistic for large numbers of sources. In this case the $N * D/D/1$ or $\sum D_i/D/1$ analysis [38, sec. 6] is able to provide better bounds, under the assumption that distinct sources are independent. Suppose, for example, that we have 1000 sources, all with the same rate and with zero CDV tolerance ($b = 1$). Then the random phasing of these constant bit rate sources is described by an $N * D/D/1$ queue, whose analysis [38] shows that with buffer B of 150 cells the proportion of cells lost remains negligible (less than 10^{-12}, say) even when the traffic intensity is high ($\rho = 0.9$, say). If the traffic intensity is moderate ($\rho \leq 0.75$, say) then a buffer B of 50 cells will produce a negligible cell loss.

The $N * D/D/1$ analysis provides a bound for much larger CDV tolerances. Suppose, for example, that sources are each policed by a leaky bucket of depth b, and that each source emits through the leaky bucket a cluster of b cells at infinite rate, every b/r units of time. Then a buffer of size $B = 150b$ cells is large enough to cope with the random phasing of the clusters, even when the load is high ($\rho = 0.9$, say).

More generally, suppose that r, R, b, B are given. Then, we can calculate the number of sources C that can be simultaneously carried, by determining the traffic intensity $\rho = Cr/R$ at which the proportion of cells lost, determined by the $N * D/D/1$ analysis with $N = C$, is negligible (10^{-12}, say). Fig. 12 shows how C varies with the ratios R/r and B/b. Similarly the load S used in earlier sections can be defined rather simply on the cell-scale as A/b where A is the number of cells that arrive at the buffer B in b/r units of time.

The above discussion, and Fig. 12, indicate the importance of the buffer ratio B/b. To provide a reference for the absolute levels of buffers, we remark that a 150 Mb/s link between Europe and North America will have around 10 000 cells in flight along the link at any given instant. For a buffer substantially smaller than 10000 cells, the delay in passing through the buffer will be substantially smaller than the transatlantic delay.

In earlier sections we have assumed that if the load S_n is less than C, then no cells are lost, while if the load S_n is greater than C, then the excess, $(S_n - C)^+$, is lost. If C is chosen by the above analysis then when the load S_n is less than C the proportion of cells that are lost will be negligible (less than 10^{-12}), while when S_n exceeds C, a simple coupling argument shows that no more than $(S_n - C)^+$ of the load will be lost. Of course if any particular model is adopted for the cell scale, then a less conservative argument is available. For example, under the cell-scale model of this section, we could let $M(N)$ be the rate of cell loss from N connections, calculated from an $N * D/D/1$ analysis, and replace expression (4) by

$$M(n;p) = \sum_{N=0}^{n} \binom{n}{N} p^N (1-p)^{n-N} M(N). \qquad (68)$$

We have used the simpler expressions (3)–(5) in our earlier analysis, since the particular cell-scale model described in this section is not essential to that analysis.

This section has given one possible outline of the interrelationship between the parameters of the cell-scale and burst-scale analyses. There are, of course, many other cell-scale effects that are important. For example [8], [27] discusses how a slightly higher leak rate r than the agreed peak rate might be chosen so as not to penalize a source for unavoidable CDV introduced by an earlier multiplexing stage, and [8], [28] discuss the CDV introduced by many stages of multiplexing. For a detailed review of recent work on CDV, see [16]. If further constraints upon source behavior may be enforced or assumed, then other source models than that used in our worst-case analysis can be constructed [14], [7], [22], [25]. In particular, we expect that real data, obtained under a plausible policing and tariffing regime, will allow Fig. 12 to be recomputed in a less conservative manner. There has been much discussion recently of scheduling algorithms [31], [34], [36]. Here we simply note that our analysis assumes only that the cell-scale scheduling algorithm is work conserving.

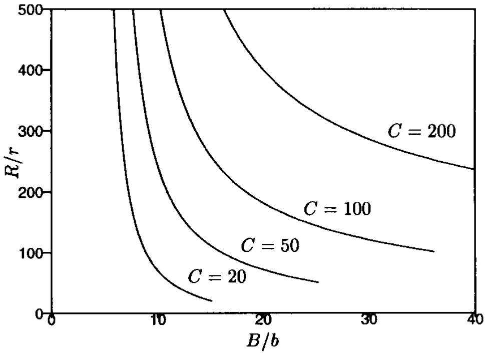

Fig. 12. Relationship between R/r and B/b for fixed C. There is a trade-off between buffer B and service rate R at the switch, as well as between the bucket depth b and leak rates r at each source.

VII. Multiple Call Types

In this section we briefly indicate, through two examples, that the schemes described earlier may be readily extended to deal with multiple call classes.

A. Aggregated Load Measurements

We have seen the effectiveness of schemes based on load mesurements only, in Section III-E. When distinct call classes have differing peak rates, it is important that load measurements for different classes be aggregated separately [6, ch. 4]. We now illustrate a simple and effective mechanism.

Suppose that

$$S(t) = \sum_{j=1}^{J} S_j(t), \qquad S_j(t) = \sum_{i=1}^{n_j(t)} X_{ji}(t) \qquad (69)$$

where $X_{ji}(t)$, for distinct values of i, j, and t, are independent random variables with

$$P\{X_{ji}(t) = h_j\} = p_j \qquad P\{X_{ji}(t) = 0\} = 1 - p_j. \quad (70)$$

Here, $X_{ji}(t)$ is the load produced by a call of class j at time t, and the calls of class j have peak rate h_j, and mean rate

TABLE I
CELL LOSS RATIO AND UTILIZATION FOR 3 CALL CLASSES ($C = 50$); NOTE THAT THE CELL LOSS RATIO IS WELL CONTROLLED WHATEVER THE COMPOSITION OF THE OFFERED LOAD

λ_1	λ_2	λ_3	Cell loss ratio $\log_{10}(L)$	Utilization U
25.000	0.000	0.000	-10.17	23.86
12.500	12.500	0.000	-11.95	16.19
0.000	25.000	0.000	-12.06	12.99
0.000	12.500	12.500	-10.00	8.39
0.000	0.000	25.000	-10.55	5.69
12.500	0.000	12.500	-10.88	8.42
6.250	9.375	9.375	-9.01	9.69
6.250	6.250	12.500	-10.33	8.41
9.375	6.250	9.375	-10.18	9.73
12.500	6.250	6.250	-9.97	11.67
9.375	9.375	6.250	-9.90	11.52
6.250	12.500	6.250	-9.89	11.30
8.333	8.333	8.333	-9.01	10.25

$m_j = p_j h_j$. The cell loss rate is

$$M(\mathbf{n}) = E(S - C)^+ \tag{71}$$

where $\mathbf{n} = (n_1, n_2, \cdots, n_J)$. Assume that calls of class j arrive as a Poisson process of rate ν_j. Let $\lambda_j = \nu_j m_j$, and call $\lambda = \sum_{j=1}^{J} \lambda_j$ the offered load. Assume, for simplicity, that the holding times of accepted calls are independent and exponentially distributed with unit mean. Suppose that a call arriving at time t is accepted if

$$\sum_{j=1}^{J} \alpha_j S_j(t) < 1. \tag{72}$$

Then $\mathbf{n}(t)$ is a multidimensional birth and death process, with transition rates which can be readily evaluated. Observe that if $\nu_j = 0$ for $j \neq k$ we recover the model of Section II-A, with C replaced by C/h_k. But how does the model perform in the presence of multiple call classes? In Table I we show the cell loss ratio and utilization achieved in a system with $\lambda = 25$, $C = 50$, $J = 3$, and call class parameters as follows

$$\begin{aligned} h_1 &= 1.0 & h_2 &= 2.0 & h_3 &= 4.0 \\ m_1 &= 0.5 & m_2 &= 0.5 & m_3 &= 0.5 \\ \alpha_1 &= 0.032 & \alpha_2 &= 0.08 & \alpha_3 &= 1.0. \end{aligned} \tag{73}$$

We note that the cell loss ratio remains under good control throughout the region.

By construction the *call* loss probability is identical for the various call classes. This might be appropriate if, for example, a tariff structure, such as that described in [24], has accounted for the different resource requirements of the different call classes. If, however, it is desired to differentiate call loss between different call classes then this can be achieved by a simple extension of the above acceptance rule designed to imitate trunk reservation [9], [10]: Accept an arriving call of class j if

$$\sum_{j=1}^{J} \alpha_j S_j(t) < 1 - \beta_j. \tag{74}$$

Note that the larger β_j, the lower the priority of call class j. Of course other simple priority mechanisms are possible, one of which we discuss next.

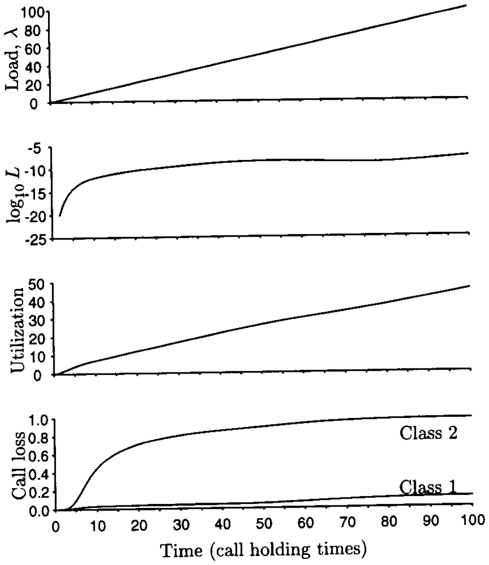

Fig. 13. System evolution with 2 call classes ($C = 100$). Class 1 has peak rate 1, while class 2 has peak rate 10. Differential backoff is used to give priority to class 1.

B. Differential Backoff

The introduction of backoff into the scheme in Section VII-A provides a further mechanism to differentiate between call classes. As an illustration consider a system with $C = 100$, $\lambda_1 = \lambda_2 = \lambda/2$ and two classes with parameters

$$\begin{aligned} h_1 &= 1.0 & h_2 &= 10.0 \\ m_1 &= 0.5 & m_2 &= 0.5 \\ \alpha_1 &= 0.01 & \alpha_2 &= 0.07. \end{aligned} \tag{75}$$

But suppose now that when a call of either class is rejected, the system refuses to consider for acceptance any call of class 2 for an exponentially distributed period with mean σ^{-1} (clearly several refinements are possible). In Fig. 13 we show the performance of such a scheme with $\sigma = 1.0$. Observe that as the offered load λ increases, the cell loss rate remains controlled, and the utilization steadily increases. The increase in utilization is achieved since, as the load increases, priority is given to calls with lower peak rate. Note how the call loss probability for calls of class 1 increases slowly, while that for calls of class 2 increases steeply. Of course many other refinements and variations are possible, and in particular there are many other forms of backoff. For example, when a call of either class is rejected, instead of locking out arriving calls of class 2, the system might require all calls of class 2 in progress to reduce their peak rate. The essential point is that robust and effective multitype call admission controls may be constructed using a weighted load criterion of the form (72).

VIII. CONCLUSION

This paper has described a family of simple and robust call admission controls, which are able to adapt to unknown and possibly varying mean rates while achieving stringent quality of service targets.

A separation of time-scales provides the framework for our analysis. Buffering caters for the cell delay variation, while a bufferless model is used on the burst-scale. Load measurements on the cell-scale are used to control call admission in such a way that strict cell loss requirements are met.

A call need only specify its peak rate and a CDV tolerance: Mean rates are implicitly and robustly estimated through the operation of simple threshold rules. Our analysis does not presume any particular choice of buffer size within the network: There might be small buffers, 150 cells, say, for certain real-time services, but if data services produce clumps of say 1000 cells, then buffers might be measured in thousands of cells. Similarly our examples have used tight quality of service constraints, aiming for cell loss rates of 1 in 10^9 or better, but the analysis could equally well have used less strict requirements, with increased multiplexing gains.

A major feature of our approach is the simplicity of the basic model. Such simplicity is, we believe, essential if a full understanding of the behavior of a call admission control is to be achieved.

REFERENCES

[1] N. G. Bean, "Statistical multiplexing in broadband communication networks," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 1993.

[2] ——, "Robust connection acceptance control for ATM networks with incomplete source information," *Ann. Op. Res.*, 1994.

[3] P. Billingsley, *Probability and Measure*, 2nd ed. New York: Wiley, 1986.

[4] C. Courcoubetis *et al.*, "Admission control and routing in ATM networks using inferences from measured buffer occupancy," *IEEE Trans. Commun.*, to be published.

[5] M. H. DeGroot, *Probability and Statistics*, 2nd ed. Boston: Addison-Wesley, 1986.

[6] R. J. Gibbens and F. P. Kelly, "ATM connection acceptance control, stage 2," Rep. prepared for British Telecommun., plc by Lyndewode Res. Ltd., Jan. 1991.

[7] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.

[8] R. J. Gibbens and F. P. Kelly, "Modeling cell delay variation in ATM networks," Rep. prepared for British Telecommun., plc by Lyndewode Res. Ltd., Aug. 1993.

[9] N. G. Bean, R. J. Gibbens, and S. Zachary, "The performance of single resource loss systems in multiservice networks," pp. 13–21 in [26].

[10] ——, "Asymptotic analysis of single resource loss systems in heavy traffic with applications to integrated networks," *Advances Appl. Probability*, Mar. 1995.

[11] E. Brockmeyer, H. L. Halstrom, and A. Jensen, *The Life and Works of A. K. Erlang*. Copenhagen, Denmark: Acad. Tech. Sci., 1948,

[12] D. Y. Burman, J. P. Lehoczky, and Y. Lim, "Insensitivity of blocking probabilities in a circuit switching network," *J. Appl. Probability*, vol. 21, pp. 850–859, 1984.

[13] R. L. Cruz, "A calculus for network delay, pt. I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, 1991.

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

[15] P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*. Berlin: Akademie-Verlag, 1981.

[16] A. Gravey and S. Blaabjerg, Eds., *Cell Delay Variation in ATM Networks*, 1994, Cost 242.

[17] P. J. Hunt and T. G. Kurtz, "Large loss networks," *Stochastic Processes and their Applications*, to be published.

[18] "ATM user-network interface specification," 1993, ATM Forum, Version 3.0.

[19] R. Griffiths and P. Key, "Adaptive call admission control in ATM networks," pp. 1089–98 in [26].

[20] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Select. Areas Commun.*, vol. 9, pp. 968–981, 1991.

[21] J. Y. Hui, "Resource allocation for broadband networks," *IEEE Select. Areas Commun.*, vol. 6, pp. 1598–1608, 1988.

[22] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, pp. 5–15, 1991.

[23] ——, "Loss networks," *Ann. Appl. Probability*, vol. 1, pp. 319–378, 1991.

[24] ——, "Tariffs and effective bandwidths in multiservice networks," pp. 401–410 in [26].

[25] D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, "Methods for performance evaluation of VBR video traffic models,"*IEEE/ACM Trans. Networking*, vol. 2, pp. 176–180, 1994.

[26] J. Labetoulle and J. W. Roberts, Eds., "The fundamental role of teletraffic in the evolution of telecommunnication networks," in *Proc. 14th Int. Teletraffic Cong.-ITC 14*. June 1994.

[27] F. P. Kelly, "Mathematical models of multiservice networks," in *Complex Stochastic Systems and Engineering*, D. M. Titterington, Ed. The Inst. Mathematics and its Applications: Oxford Univ. Press, 1994.

[28] F. P. Kelly and P. B. Key, "Dimensioning playout buffers from an ATM network," in *11th UK Teletraffic Symp.*, London, U.K., 1994.

[29] P. B. Key, "Admission control problems in telecommunications," in *Complex Stochastic Systems and Engineering*, D. M. Titterington, Ed. The Inst. of Mathematics and its Applications, Oxford Univ. Press, 1994.

[30] ——, "An introduction to ATM performance issues and modeling," in *IFIP Workshop TC6 Perform. Modeling, Eval. ATM Networks*, Bradford, U.K., July 1994.

[31] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 343–357, 1993.

[32] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *Comput. Commun. Rev.*, vol. 24, pp. 257–268, 1994.

[33] C. Rasmussen, J. H. Sorensen, K. S. Kvols, and S. B. Jacobsen, "Source-independent acceptance procedures in ATM networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 351–358, 1991.

[34] J. W. Roberts, "Virtual spacing for flexible traffic control," *Int. J. Analog, Digital Commun. Syst.*, Dec. 1994.

[35] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *IEEE Select. Areas Commun.*, vol. 9, pp. 982–989, 1991.

[36] S. Shenker, D. D. Clark, and L. Zhang, "A scheduling service model and a scheduling architecture for an integrated services packet network," to be published.

[37] W. Willinger, "Traffic modeling for high-speed networks: Theory versus practice," in *Stochastic Networks*, F. P. Kelly and R. J. Williams, Eds. Berlin: Springer, 1994.

[38] J. W. Roberts, Ed., *COST224 Perform. Eval., Design of Multiservice Networks*. Comm. Europe. Commun. Final Report., Oct. 1992.

[39] "ITU Recommendation I.121: Broadband aspects of ATM," 1991, Geneva, Switzerland.

[40] "ITU Recommendation I.371: Traffic control and congestion control in B-ISDN," Mar. 1994, Geneva, Switzerland.

[41] "CCITT IVS Baseline document, SGXVIII/8," June 1992, Geneva, Switzerland.

Richard J. Gibbens received the B.A. degree in mathematics, the Diploma in mathematical statistics, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1983, 1984, and 1988, respectively.

During 1988–1993 he worked as a Research Associate and since 1993 he has held a Royal Society University Research Fellowship in the Statistical Laboratory at the University of Cambridge. His research interests are in the area of mathematical modeling of telecommunication systems, especially the design of dynamic routing schemes. He is a coinventor of the Dynamic Alternative Routing (DAR) strategy.

Frank P. Kelly, for a photograph and biography please see page 936 of this issue.

Peter B. Key received the B.A. degree in mathematics from the University of Oxford, Oxford, U.K., in 1978, and the M.Sc. and Ph.D. degrees in statistics from the University of London, London, U.K., in 1979 and 1985.

From 1979 to 1982, he was a Research Assistant in the Statistics and Computer Science Department of Royal Holloway College, University of London. He joined BT in 1982, working in the field of teletraffic and performance evaluation, initially in the area of circuit-switched networks, where he was involved with the development of network analysis tools and techniques, and the introduction of the Dynamic Alternative Routing (DAR) strategy. After leading a mathematical service group, he became involved with network reliability. In 1992 he took over an ATM performance group. He currently leads a team looking at performance across the transport layers, attempting to look at end-to-end performance across networks and across layers, taking into account the requirements of applications.