Distributed Admission Control

Frank P. Kelly, Peter B. Key and Stan Zachary

Abstract— **This paper describes a framework for admission control for a packet-based network where the decisions are taken by edge devices or end-systems, rather than resources within the network. The decisions are based on the results of probe packets that the end-systems send through the network, and require only that resources apply a mark to packets in a way that is load dependent. One application example is the Internet, where marking information is fed back via an ECN bit, and we show how this approach allows a rich QoS framework for flows or streams. Our approach allows networks to be explicitly analysed, and consequently engineered.**

I. Introduction

The current Internet does not give any Quality of Service (QoS) guarantees. This is a particular problem for streaming applications. Typical applications are voice or video streams; however the problem applies much more generally to transfers of data or flows which have an inelastic component, that is which need some minimum level of bandwidth to function correctly. Current interest in labelling and switching flows increases the importance of Quality of Service questions.

Measurement based admission control (MBAC) research has shown that it possible to assure Quality of Service for varying heterogeneous traffic by inferring information from on-line measurements. ATM, a connection oriented transfer mode, was a natural setting for early work [1], [2], [3], [4] with variable bit-rate (VBR) traffic providing the impetus: VBR traffic is hard to characterise and introducing regulators, policers or shapers does not solve the problem since these also require some characterisation to sensibly set parameters. MBAC offered an attractive solution providing useful multiplexing gains with minimal assumptions.

The timescale decomposition of Hui [5] shed light on how to apply theory in practice: connections last for a period of time and generate periodic *bursts* of activity, where each burst of activity consists of a number of cells transmitted at the line rate. A natural timescale separation follows if just enough buffering is present in the network switches to absorb cell-scale effects (caused by phase effects of cells sent at the peak rate), with admission control limiting the probability of an excess number of bursts to an acceptably low level. In other words, the buffering in switches is sufficient to absorb the aggregate cell-delay variation.

Initially intended for real-time streaming applications, the ideas also apply to non-real time traffic ('data') provided the cell-scale and burst scales are suitably reinterpreted. MBAC ideas also found a home in the Internet community [6], [7], [8], and have been proposed as a way of limiting the number of flows or connections, where a flow can be a general transfer of data between a source and an endpoint or endpoints.

It is worthwhile recapping the lessons that have been learned from the measurement-based research. First, simple algorithms based solely on load measurements are generally robust. Secondly, there is a trade-off between connection blocking and packet loss: most algorithms can be tuned to improve one at the cost of the other. Thirdly, the timescale over which load is measured is important, but little is gained by building history into the inference process since the connection admission process integrates decisions.

A problem nevertheless remains with current MBAC work: how should decisions be made on an end-to-end basis? One option is to invoke signalling to pass messages between resources. However, we propose a different approach: let the connection decide whether or not it should enter the network. The connection has to infer information about aggregate load, for example by sending probe packets, and necessarily is in a poorer position to deduce state information about a resource than the resource itself. But, conversely, it is in a much better position to make timely end-to-end inferences along a probed path.

In our approach, we require information to be fed back to the end-systems associated with a potential connection. The framework is not tied to any particular implementation, however it is natural to think of the Internet, in which case the recent RFC on Explicit Congestion Notification, [9] would provide a mechanism for conveying information back. Elsewhere[10], [11] it is argued that such a framework allows a coherent treatment of pricing and QoS issues, in integrated networks carrying both streaming and data applications.

Gibbens and Kelly [12] and Turányi and Westberg [13] have looked at distributed connection admission control with an edge-device or broker acting as a gateway to determine whether to accept a connection or not, with the aim of keeping the experienced packet marking rate to an acceptable level. Thus the gateways act as aggregators and have access to aggregate information. We go one step further here, where information is only obtained through the (noisy) packet marking signal the end-system receives. Alternative mechanisms for conveying information to end-systems have been studied by Elek *et al.* [14], who use packet loss, and Bianchi *et al.* [15], who require the network to give lower priority to probe packets.

The organisation of this paper is as follows. In Section II we describe a model for probabilistic admission to a loss network. We develop methods for its analysis and approximation. Product form distributions and fixed point models

F.P. Kelly is with the University of Cambridge, Cambridge, CB2 1SB U.K.

P. B. Key is with Microsoft Research, Cambridge, CB2 3NH,U.K.

S. Zachary is with Heriot-Watt University, Edinburgh, EH14 4AS, U.K.

have long been used to model telecommunication networks (see for example, [16, page 203], [17], and, for a more recent discussion, [18]), and the implied costs and shadow prices derived from fixed point models have found extensive use in studies of design optimization and resource management for several forms of circuit-switched network (Key [19], Farago *et al.* [20], Mitra *et al.* [21], [22]). Introductions to the area are provided by Kelly [23] and Ross [24]. We show that these methods may be generalized to networks where call admission is probabilistic, and thus allow the analysis of networks where admission control is measurement-based and distributed.

In Sections II-A and II-B we develop a simple network model that admits a product form solution. The simple model makes strong assumptions on, for example, the homogeneity of bandwidth requirements, but allows exact calculations for various quantities of interest. The simple model leads directly, in Sections II-C, II-D and II-E, to fixed point approximations for networks carrying large numbers of connections, or for networks with diverse routing. The fixed point approximations are applicable more widely than the product form solutions, and in Sections II-F, II-G we describe how they emerge from the average dynamics of a large network, following the work of Hunt and Kurtz [25] and Zachary [26].

The connection level models of Section II uses a very simple abstraction of resource behaviour. In Section III we develop more detailed models of resources, showing how the abstraction used in Section II may arise from the packet level dynamics of the resource. We consider models where the connections generate bursts of activity, as well as models where connections behave more smoothly. A robust mechanism to detect approaching traffic overload, without the need for source traffic characterization, is provided by the Virtual Queue mechanism: a resource marks packets or not depending on the state a fictitious queue, of lower capacity than the real queue. In Section III-D we use the many sources asymptotic [27], [28], [29] to provide important insights into the relevant packet level timescales, and thus into the robustness of such mechanisms. Finally, in Section IV, we conclude.

II. Connection level network models

We now explore more fully how an end-system or user might decide whether or not to enter the system. In effect, this is admission control performed by the *user*. We assume that each arriving call request (where we use the term 'call' to represent a connection or flow) probes the network and is accepted, that is, decides to enter the network, with a probability which depends on the current load on resources. In this section we study a model for such distributed call acceptance control, in which the load is assumed to be generated by the calls themselves. We also assume independence of the packet level dynamics at each resource, *conditional on the current load on the network*. We first analyse a specific model in which we further assume homogeneity of bandwidth requirements across different call types, and also that, for a call to be accepted, each resource along its route must signal that it is uncongested. This leads to the product form acceptance probability (1) below, and so to the product form stationary distribution (3). In Section II-G, we discuss the analysis of more general models.

A. A simple model

Let J be a set of resources, and R a set of routes, where $r \in R$ identifies a subset of J. Calls requesting route r arrive as a Poisson stream of rate ν_r. Each such call probes the network and is accepted with probability

$$\bar{a}_r(m(n)) = \prod_{j \in r} a_j(m_j(n)). \tag{1}$$

Here $n = (n_r,\ r \in R)$ where n_r is the number of calls already in progress on each route r and $m(n) = (m_j(n),\ j \in J)$ where $m_j(n) = \sum_{r \in R: j \in r} n_r$ is the existing occupancy of each resource j. For each j, the function a_j is non-increasing and takes values in the interval $[0, 1]$. We can interpret $a_j(m_j)$ as the probability resource j marks none of a fixed number of probe packets used by a call requesting connection through resource j, when m_j is the number of calls already in progress through that resource. The product form acceptance probability (1) corresponds to the assumption of independence of the packet level dynamics at each resource, conditional on the current load, together with the requirement (in this specific model) that, for a call to be connected, it must be "accepted" by each resource along its route. As an example, we might take

$$a_j(m_j) = P\{X(m_j, p) < \theta_j\}, \tag{2}$$

for some probability p and threshold θ_j, where $X(m_j, p)$ is a binomial random variable—an example we explore in more detail in Sections III-A and III-B.

Suppose further, without loss of generality, that accepted calls have holding times with unit mean. Then the stationary distribution π of the vector n is given by

$$\pi(n) = \pi(0) \prod_{r \in R} \frac{\nu_r^{n_r}}{n_r!} \prod_{j \in J} \prod_{k=1}^{m_j(n)} a_j(k-1). \tag{3}$$

In the case of exponential holding times this is immediate from the reversibility of the Markov process $(n(t),\ t \geq 0)$ giving the number of calls in progress at each time t. More generally the result follows from the quasireversibility of this process [30].

Note that this is also the mathematical model which is appropriate to a traditional uncontrolled loss network with "hard" capacity constraints, in which each resource j has a fixed capacity C_j and $a_j(m_j) = 1$ if $m_j < C_j$, $a_j(m_j) = 0$ otherwise. It turns out that many of the results for such a loss network generalise to the present model. In particular, for a large network the stationary distribution above is concentrated on the same point as that for the traditional loss network with appropriately defined C_j—see Appendix 1.

B. The occupancy distribution

We now study the stationary distribution of the resource occupancy vector m, together with the corresponding call

acceptance probabilities. We give a recursion which enables the efficient and exact computation of these quantities—at least in networks of small to medium capacity—and which more generally provides insight into the behaviour of the model (see Section II-C).

For each m, define $\pi^*(m) = \sum_{n:\, m=m(n)} \pi(n)$ (where π is as given by (3)) to be the stationary probability that the total load on the network is given by the vector m. For each $r \in R$, define the vector $\delta_r = (\delta_{rs},\ s \in R)$ by $\delta_{rs} = 1$ if $s = r$, $\delta_{rs} = 0$ otherwise; define also the vector $e_r = m(\delta_r)$. It follows from (3) that, for each n and for each r,

$$\pi(n)n_r = \pi(n - \delta_r)\nu_r \bar{a}_r(m(n) - e_r), \tag{4}$$

where the function $\bar{a}_r$ is given by (1). (In the case of exponential holding times, when the process $(n(t),\ t \geq 0)$ is Markov and reversible, the equations (4) are the detailed balance equations for this process.) Now, for each m and for each $j \in J$, sum the equations (4) over those n such that $m(n) = m$ and over those r such that $j \in r$ to obtain, straightforwardly,

$$\pi^*(m)m_j = \sum_{r:\, j \in r} \nu_r \pi^*(m - e_r)\bar{a}_r(m - e_r). \tag{5}$$

The equations (5) enable the recursive determination of the stationary probabilities $\pi^*(m)$—for example, by induction on $|m| = \sum_{j \in J} m_j$. They form a natural generalisation of the Kaufman-Dziong-Roberts recursion for the direct computation of the exact resource-occupancy distribution in a traditional uncontrolled loss network (see [31]). There the ultimate goal is usually the computation of the stationary blocking, or equivalently the stationary acceptance, probabilities. For the present model, the stationary probability that a call of type r is accepted is given by $\sum_m \pi^*(m)\bar{a}_r(m)$.

C. A simple fixed point approximation

Simplifications arise in large networks, often formalised for loss networks by a familiar limiting regime in which arrival rates and capacities are allowed to grow in proportion to some scale parameter N—see Appendix 1 for a formal result in the context of the current model. Thus suppose that both the arrival rates ν_r, $r \in R$, and the capacities of the resources $j \in J$ (effectively defined by the functions a_j) are large. We develop a simple fixed point approximation identifying those regions in which the stationary distributions of n and $m(n)$ are concentrated, and derive the corresponding call acceptance probabilities.

The stationary distribution π of n is given by (3). Since, for each j, $a_j(m)$ is decreasing in m, the function $\log \pi$ is concave, and so $\log \pi(n)$ is maximised by some unique $n = \bar{n}$, which, by our above assumptions, will again be large. Now take logarithms in (3), and use Stirling's approximation $(\log n! = \frac{1}{2}\log 2\pi n + n \log n - n + O(n^{-1}))$. Assume that the functions a_j are sufficiently smooth to permit partial differentiation of $\log \pi(n)$ with respect to each n_r (treating the latter as continuous). We then obtain that $\bar{n}$ is given, to a good approximation in a large network, by

$$\bar{n}_r = \nu_r \bar{a}_r(m(\bar{n})), \quad r \in R. \tag{6}$$

Further, for each j, by summing these equations over r such that $j \in r$, we obtain

$$m_j(\bar{n}) = \sum_{r:j \in r} \nu_r \bar{a}_r(m(\bar{n})), \quad j \in J. \tag{7}$$

The equations (6) and (7) form sets of fixed point equations for the vectors $\bar{n}$ and $m(\bar{n})$ respectively. (Note that the equations (7) may also be obtained directly from (5) by making the reasonable assumptions that $\pi^*(m)$ varies smoothly in m and is maximised, to a sufficiently good approximation, by $m(\bar{n})$.) Now the stationary distribution of m typically declines more sharply away from its mode than does that of n—see [32] and [33] for the analogous results in the case of a traditional uncontrolled loss network, where $a_j(m_j) \in \{0, 1\}$. Further, the size of the set J is usually less than that of the set R. Hence if, for example, recursive substitution is used to solve either the equations (6) or (7) then the latter set may generally be expected to be easier to solve, after which $\bar{n}$ may be obtained from (6).

Again as for of the traditional loss network above, we may show that, for a large network, $\pi(n)$ does decrease sufficiently fast as n moves away from $\bar{n}$ that the stationary distributions π and π^* are effectively concentrated in small neighbourhoods of $\bar{n}$ and $m(\bar{n})$ respectively. For each call type r, the corresponding stationary acceptance probability is clearly given by $\mathbb{E}_\pi(n_r)/\nu_r$, where $\mathbb{E}_\pi(n_r)$ is the expectation of n_r under the stationary distribution π. The concentration of π in a small neighbourhood of $\bar{n}$ ensures that $\mathbb{E}_\pi(n_r)$ is close to $\bar{n}_r$ and so, from (6) and to a good approximation in a large network, the stationary acceptance probability for a call of type r is given by $\bar{a}_r(m(\bar{n}))$. It now follows also from (1) that the *stationary* acceptance probabilities are, approximately, *as if* each resource j accepts calls with probability $a_j(m_j(\bar{n}))$, independently of other resources. We refine these approximations in Section II-D.

Note also that if, for each j, we define $C_j = m_j(\bar{n})$, then comparison of (1), (6) and (7) with the equations of Theorem 2.1 of [32] shows that $\bar{n}$ is also the point on which is concentrated the stationary distribution of n in the corresponding (traditional) uncontrolled loss network in which calls are accepted subject only to a hard capacity constraint C_j on each resource j. We make a further important connection in Appendix 1.

D. A refined fixed point approximation

We now give a refinement of the approximation of the previous section. For each resource j, let $A_j(\rho_j, \theta_j)$ be the stationary acceptance probability when that resource is offered a single Poisson stream of call traffic of rate ρ_j and operates with a threshold parameter θ_j. (We show in Section III-B how the parameter θ_j arises in the binomial model: more generally we use this parameter as a proxy for the capacity of resource j.) Then $A_j(\rho_j, \theta_j)$ is readily computed exactly via the relation $\rho_j A_j(\rho_j, \theta_j) = \sum_{n \geq 0} n\pi_j(n)$

where, from (3),

$$\pi_j(n) = \pi_j(0)\frac{\rho_j^n}{n!}\prod_{k=1}^{n} a_j(k-1), \quad \sum_{n\geq 0}\pi_j(n) = 1. \qquad (8)$$

We treat the network call acceptance probabilities as if, under stationarity, the network resources accept calls independently of each other—as was shown to be true, to a good approximation, in the previous section. However, we use the result above to refine the resource acceptance probabilities. We thus associate with each resource j a stationary acceptance probability A_j given by the solution of the fixed point equations

$$A_j = A_j(\rho_j, \theta_j) \quad j \in J$$

where

$$\rho_j = \sum_{r:j\in r} \nu_r \prod_{i\in r-\{j\}} A_i \quad j \in J.$$

Thus the traffic offered to each resource is viewed as a sum of Poisson streams from those routes using it, and each of these streams has a rate thinned by the acceptance probabilities for the other resources along the corresponding route. The stationary acceptance probability for a call of type r is then taken to be $\prod_{j:j\in r} A_j$. This approximation becomes exact in the case of a single-resource network. For an uncontrolled loss network with hard constraints, it is the well-known Erlang fixed-point, or reduced load, approximation. Conditions for this approximation to be accurate include the case of diverse routing, where different resources within the network have only a small proportion of their load arising from the same calls, as well as the large network regime considered in Appendix 1.

The fixed point equations above have a unique solution, identified by the solution of the problem

$$\text{Minimize} \qquad \sum_r \nu_r \exp\left(-\sum_{j\in r} y_i\right) + \sum_j \int_0^{y_j} U_j(z)\,dz$$
$$\text{over } y_j \geq 0,\ j \in J,$$

where $U_j(y) = \rho_j A_j(\rho_j, \theta_j)$ is the mean utilization at resource j, when ρ_j is the solution of $A_j(\rho_j, \theta_j) = e^{-y}$. This follows, as in [32], since the earlier condition that $a_j(\cdot)$ is monotone decreasing ensures that the stationary acceptance probability of resource j is monotone decreasing, and the mean utilization of resource j is monotone increasing, in the offered traffic (by a coupling argument), and hence $U_j(y)$ is increasing. The function displayed above is then strictly convex, and the stationarity conditions identifying the unique minimum are just the fixed point equations.

E. Optimization of routing and capacity

How should calls be routed or capacity allocated so as to improve the performance of the network? For example, there may be a number of routes r that carry traffic between the same two end points, and we might be interested in varying the amounts of traffic ν_r offered to each of these routes. Or we might be interested in how to allocate additional capacity over the resources of the network. What is the effect on the performance of the network of changes in the parameters ν or θ? To make some progress with these issues, let us suppose that a call carried on route r generates an expected revenue w_r (or, equivalently, interpret w_r as the cost of losing a call on route r). Extend the definition of the functions $A_j(\rho_j, \theta_j)$ to non-integral values of θ_j in such a manner that the functions have continuous derivatives. Let $A = (A_j, j = 1, 2, \ldots, J)$ be the unique solution to the fixed point equations of Section II-D. To emphasize its dependence on the parameter vectors ν and θ, write $A = A(\nu; \theta)$. Under the fixed point approximation the rate of return from the network is given by

$$W(\nu;\theta) = \sum_r w_r\lambda_r \qquad (9)$$

where

$$\lambda_r = \nu_r(1 - L_r), \quad 1 - L_r = \prod_j A_j \qquad (10)$$

and $A = A(\nu; \theta)$. Thus L_r, λ_r are the stationary loss probability and carried traffic respectively on route r, as calculated from the approximation. Let

$$\xi_j = -A_j(\rho_j, \theta_j)^{-2}\frac{\partial A_j(\rho_j, \theta_j)}{\partial \rho_j}$$

and let

$$\eta_j = -A_j(\rho_j, \theta_j)\frac{\partial A_j(\rho_j, \theta_j)}{\partial \theta_j}\left(\frac{\partial A_j(\rho_j, \theta_j)}{\partial \rho_j}\right)^{-1}.$$

Then, by proceeding as in [34], we may calculate

$$\frac{d}{d\nu_r}W(\nu;\theta) = (1 - L_r)s_r \qquad (11)$$

and

$$\frac{d}{d\theta_j}W(\nu;\theta) = \eta_j c_j, \qquad (12)$$

where $s = (s_r, r \in R)$, $c = (c_j, j \in J)$ are the unique solution to the linear equations

$$s_r = w_r - \sum_{j\in r} c_j \qquad (13)$$

$$c_j = \xi_j \sum_{r:j\in r} \lambda_r(s_r + c_j). \qquad (14)$$

We can interpret s_r as the *surplus value* of a call on route r: if such a call is accepted it will earn w_r directly, but at an *implied cost* of c_j for each circuit used from link j. The implied costs c measure the expected knock-on effects of accepting a call upon later arrivals at the network, and, through the derivative (12), they give information on the effect of increasing the capacity of resources. Note that the implied costs c_j are derived from the stationary acceptance probabilities A_j, and so represent information integrated over many call holding times. This is in contrast to the probabilities a_j, which fluctuate with the number of calls in progress.

F. Dynamics

The shadow prices described in Section II-E are useful for certain forms of longer-term static optimization: for example, for the allocation of traffic across routes, or for capacity expansion decisions. In this section we explore how the connection-level dynamics of the network may also be interpreted in terms of an implicit, and different, optimization problem. Consider the system of differential equations

$$\frac{d}{dt}x_r(t) = \nu_r \prod_{j\in r} a_j \left(\sum_{s:j\in s} x_s(t) \right) - x_r(t) \qquad (15)$$

for $r \in R$. We may motivate the system (15) as describing average dynamics, with $x(t) = n(t)$, in a large network. These ideas are formalized in [25]—see also Appendix 1 for a limiting result. In Appendix 3 it is shown that the strictly concave function

$$\begin{aligned} \mathcal{U}(x) &= \sum_{r\in R} (x_r \log \nu_r - x_r \log x_r + x_r) + \\ &\quad \sum_{j\in J} \int_0^{\sum_{s:j\in s} x_s} \log a_j(y) dy \end{aligned} \qquad (16)$$

is a Lyapunov function for the system of differential equations (15). The unique value $x = \bar{x}$ maximizing $\mathcal{U}(x)$ is a stable point of the system, to which all trajectories converge. Further, under the identification $x(t) = n(t)$ above, $\bar{x}$ is the fixed point $\bar{n}$ given by the solution of the equations (6). (This follows on noting that $\bar{x}$ is given by setting the right hand side of the equations (15) equal to zero.) This result establishes a stability property of the network dynamics under the given call acceptance strategy—a result which need not be true for all control strategies. Further, for other models (for example, those in which different call types have different bandwidth requirements) and for other call acceptance strategies, consideration of network dynamics is often the only way to derive equilibrium behaviour—see the discussion of Section II-G.

We can view the connection-level dynamics of the system as implicitly attempting to choose the flows $x_r, r \in R$, to maximize an aggregate utility made up of a benefit to users on route r of $x_r \log \nu_r - x_r \log x_r + x_r$ for each route $r \in R$, less a cost to the network of the negative of the final term in expression (16), providing a connection with the economic framework of [10], [35], [11].

G. More general models

It is natural to consider also more general models—for example, those incorporating heterogeneity of bandwidth requirements or more general call acceptance strategies. An example of the latter occurs when the condition for a call to enter the network is that total number of probe packets marked by all resources along its route should not exceed a given number. For such models we do not in general have the simple product form stationary distribution (3). Observe, however, that we may still consider average dynamics in a large network. The analogue of the differential equations (15) is here

$$\frac{d}{dt}x_r(t) = \nu_r \bar{a}_r(x(t)) - x_r(t) \qquad (17)$$

for $r \in R$, where $x(t) = (x_r(t), r \in R)$ and $x_r(t)$ may again be identified with the number of calls of type r in progress at time t. Again ν_r is the arrival rate for calls of each type r, while $\bar{a}_r(x)$ is the corresponding acceptance probability when the state of the system is given by x. In the case where, for a call of type r to be accepted, each resource along its route must signal that it is uncongested, then the function $\bar{a}_r$ factorizes as before—this again follows from the assumed conditional independence of packet level dynamics at each resource. For an interpretation of (17) in terms of a functional law of large numbers for a traditional loss network, again see [25].

Fixed points of the network are given by setting the right hand side of the equations (17) equal to zero, providing a generalization of the fixed point (6). Equations (17) also permit an analysis of the stability of fixed points, although in general it is difficult to find Lyapunov functions to establish global stability. However, extensive analysis of the corresponding models for traditional loss networks suggests that, in all but the most badly controlled systems, there will be a single stable fixed point, to which all trajectories of the dynamics converge. This fixed point then determines stationary loads and stationary call acceptance probabilities. Thus, for large networks at least, we are able to deal with design and optimization issues.

The approach of [10], [35], [11] envisages using marks not just for inelastic connections but also for adaptive streams and for short transfers. For such heterogeneous mixes of traffic an analysis at the level of detail of Sections II-A, II-B is unattainable, but models of the form (15), (17) may well be tractable, and a simple example is given in [36].

III. Packet level resource models

In this section we develop several more detailed representations of resources, modelling behaviour at the burst and packet level. We consider the behaviour of an isolated resource conditional upon its current load, and calculate acceptance probabilities $a(\cdot)$ which can then be integrated into a connection model using the results of Section II. Since we are considering an isolated resource, we drop j from the notation and set $m = n$, and $\pi = \pi^*$.

The first example, the binomial model used extensively in [3], allows us to explore numerically and analytically some of the consequences of a particular form for the function $a(\cdot)$, and to consider the conditions under which timescale separation might take place. The second example, an M/M/1 queue, allows us to obtain straightforward analytical forms for various quantities concerned with system sizing. These two examples assume particular source traffic characterizations, where it is reasonably easy to understand the behaviour of the resource at the packet level. A robust mechanism to detect approaching traffic overload, without the need for source traffic characterization or even

call homogeneity, is provided by the Virtual Queue mechanism, previously used in [10], [12]. In Section III-D we use the insight provided by the many sources asymptotic into the behaviour of this mechanism.

A. Origins of the binomial model

The binomial model (2) arises naturally as follows. Suppose that connections last for a period of time and generate periodic bursts of activity, where each burst of activity consists of packets transmitted at a constant rate. Then the binomial model would result from a timescale separation where bursts are long in comparison with queue dynamics at resources, but short in comparison with the periods between connection arrivals and departures. A resource would then see a stream of packets whose rate was given by the number of active bursts and various simple marking mechanisms would either mark all packets or none, depending on whether the number of active bursts was above or below a threshold.

How reasonable is the modelling assumption of a timescale separation? An example may help clarify this question. Suppose that connections have an average holding time of about 200 seconds, and alternate between 'on' and 'off' periods each of mean 500 milliseconds, with packets emitted in a periodic fashion, every 25 milliseconds, during an 'on' period. (These numbers are chosen to be broadly comparable with those used in studies of both ATM and IP networks—see [37, Section 8.14], [38].) If m is the number of connections carried through a resource, then the number of bursts, or 'on' periods, in progress is about $m/2$. The number of connections in progress changes about every $100/m$ seconds, while the number of bursts in progress changes about every $1/(2m)$ seconds. Fluctuations in the number of bursts thus occur a few hundred times faster than fluctuations in the number of connections carried. At the packet level, our stylized model allows the resource to deduce the number of active bursts by counting the number of packets arriving over a measurement interval of length 25 milliseconds; and the number of packets arriving in a measurement interval of length $t \leq 25$ milliseconds, is binomially distributed, with parameters m and $t/50$. The resource may improve its estimate of the number of connections by counting packets over a longer measurement interval, but in [3], Section V, it is shown that the potential for improved performance is minimal.

Thus we may conclude that for this example and with a small enough measurement interval ($t \leq 25$ milliseconds, $100/m$ seconds) there is a clear separation of timescales, and that successive acceptance decisions at a resource are, conditional on the conection load, approximately independent. Further, if different resources within the network have, with high probability, only a small proportion of their load arising from the same connections (the diverse routing condition familiar for loss networks), then the acceptance decisions at different resources are again, conditional on current load, approximately independent. These observations motivated the assumption, made in Section II, of conditional independence of acceptance decisions (or more generally marking behaviour) at each resource and for each arriving call request.

B. Analysis of the binomial model

The binomial model is defined by acceptance probabilities given by

$$a(m) = P\{X(m,p) < \theta\} \tag{18}$$

for some probability p and threshold θ, where $X(m,p)$ is a binomial random variable. This model arises, for example, as in the previous section. Here, in a given measurement interval, each of the m connections already carried by the resource sends a packet with probability p; the random variable $X(m,p)$ is then the total count of such packets.

In [3], this model was studied for a single resource, where the problem was to estimate p, the peak to mean ratio of a connection, in a robust way. The trade-off between call blocking and cell-loss was made explicit through a Bayesian decision-theoretic framework. Both the case where the resource knows the number of connections m, and the case where the resource has only aggregate load measurements, were considered: it was noted that lack of knowledge of the number of connections did *not* degrade the performance greatly. In the current framework, the number of connections m is unknown to both the resource and to end-systems.

It is possible to exactly mirror the approach of [3]: apply priors to the offered load as well as the burstiness parameter p, and push through a Bayesian analysis to determine optimal thresholds θ. However it is more natural to see the marking strategy as under the network's control, whilst the decisions are the responsibility of the end-systems or users and the main focus of this paper. One of the key-insights of the previous work was to show that θ is a robust control with respect to varying arrival rates.

Now, for each m, $X(m,p)$ has mean mp and variance $mp(1-p)$. Hence nearly all its distribution is concentrated within, say, 3 standard deviations of mp. Thus $a(m) = P\{X(m,p) < \theta\}$ tends rapidly to 0 or 1 outside the region $m \in (\theta/p) \pm (3\sqrt{(1-p)\theta/p}) + o(\sqrt{\theta})$. (In Appendix 2 a Chernoff bound is used to further explore the fall off in $a(m)$ above θ/p.)

As an illustration, consider an isolated resource system with $p = 0.5$. The *rejection* probabilities $1 - a(m)$ are shown in Figure 1(a), the solid line corresponding to $\theta = 10$, and the broken line corresponding to $\theta = 20$[1]. Note that both curves are 's-shaped', and, as shown above, the curves flatten out with increasing θ. Suppose now that calls are offered at rate $\nu = 50$: the cumulative probability distributions for the stationary distribution π of the number of calls in progress (given by (8)) are shown in Figure 1(b). The corresponding probability density functions $\pi(m)$ are shown in Figure 2(a). Figure 2(b) shows the density functions in the case $\nu = 100$, illustrating the

[1] A threshold of $\theta = 15$ corresponds to an optimal choice for a system with capacity 25, $p = 0.5$, $\nu = 50$ and packet loss target 10^{-3} or less, using the methodology of [3].

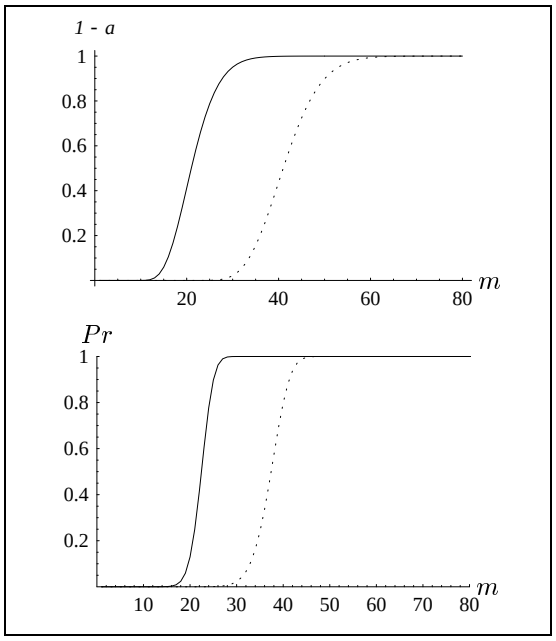

Fig. 1. (a) Rejection probabilities (upper figure) for thresholds of $\theta = 10$ (solid line), and $\theta = 20$ (broken line). (b) Cumulative occupancy distributions with offered rate $\nu = 50$.

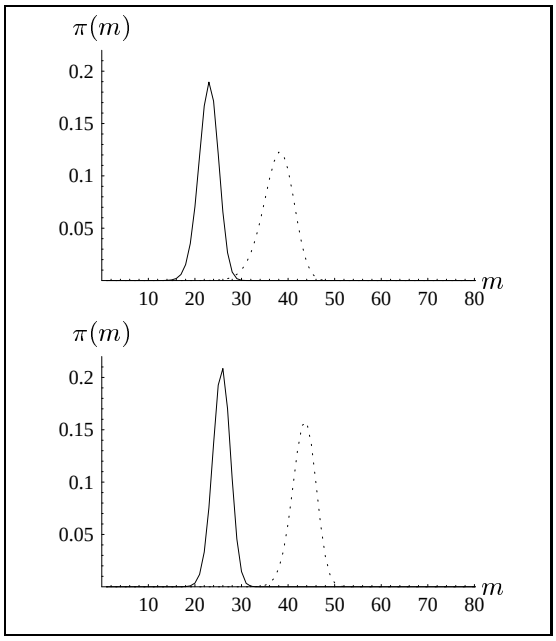

Fig. 2. (a) Occupancy density functions $\pi(m)$ for $\nu = 50$, $\theta = 10, 20$. (b) $\pi(m)$ for high offered load $\nu = 100$, $\theta = 10, 20$.

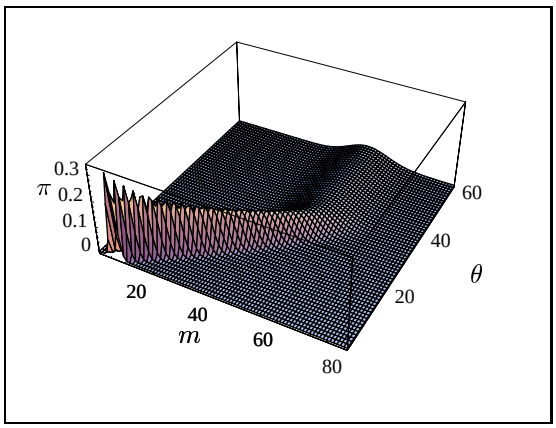

Fig. 3. Density Functions $\pi(m)$ for varying θ from 1 to 60, $\nu = 50$, $p = 0.5$

robustness against offered load. In effect, to reach a high number of calls in progress the offered load would have to be extremely high.

The admission strategy thus has the effect of truncating the number of calls in progress, largely independently of the offered load. This is illustrated graphically in Figure 3, showing how the stationary distribution π gradually spreads out and approaches an unconstrained Poisson distribution as the threshold θ increases.

Of course, the number of calls m admitted is very sensitive to p, since it is concentrated in the region θ/p, but the expected load generated, $\mathbb{E}(mp)$ is not (see [3] for a fuller discussion).

C. A simple continuous-time packet model

In this section we explore a contrasting packet level resource model, and look in detail at the marking strategy. To make progress, we assume the resource is modelled by an M/M/1 queue[2], with traffic intensity

$$\rho = m\mu \tag{19}$$

where m is the number of connections, and μ is the load imposed by a single connection. For example we might assume exponential packet sizes, and a Poisson packet generation process per stream.

As a motivational example, suppose streams generate packets at the rate of 40 pps, for nominal 200 byte packets, which would correspond to PSTN quality speech with simple PCM encoding. Thus with capacities of a small leased line, legacy LAN or Backbone taken to be 2, 10 or 600 Mb/s respectively, the loads produced by a stream are of the order $\mu_s = \frac{1}{30}$, $\mu_{LAN} = \frac{1}{150}$, $\mu_B = \frac{1}{9000}$. The reciprocals $1/\mu$ can thus be regarded as a measure of system size, or how many streams of a particular type can be fitted onto a system. If the bit-rate of streams is an order of magnitude larger, corresponding to video for example, then the effective system size becomes much smaller, with $1/\mu$ given by 3, 15 or 900 respectively.

An acceptance strategy is the following: accept a call if none of a sequence of M probe packets is marked. If the packets are marked with probability $P(\rho)$, then the acceptance probabilities for $M = 1$ are given by

$$a(m) = 1 - P(m\mu). \tag{20}$$

For M greater than 1, marking will in general be positively correlated between packets unless the probe packets are sent at intervals long compared to the critical timescale of

[2]We could, with some loss of tractability, use an M/D/1 or N*D/D/1 model to link in with Section III-B

the queue. In the latter case the approximation

$$a(m) \approx (1 - P(m\mu))^M . \tag{21}$$

may be reasonable, and in general we would expect $a(m)$ to lie between the values given in expressions (20) and (21).

A simple marking strategy is to mark all packets when the queue exceeds some threshold K, much as RED does [39]. However, any threshold marking scheme acts as an integral controller, since it is based on queue length, and hence is slow to react. One way around this is to use a Virtual Queue marking strategy [12], where we run a queue at a reduced service rate (and usually with a reduced buffer) and mark as though the demand is offered to the Virtual Queue, for example when a threshold is exceeded in the Virtual Queue. The intuitive idea, which we clarify later, is that the Virtual Queue anticipates the onset of congestion.

Let the rate and buffer size of the Virtual Queue be c_v, b_v, with $c_v \leq c$, $b_v \leq b$, where c, b are the rate and buffer size of the real queue. How should the parameters c_v, b_v be chosen? To make progress with this question, suppose that a packet arriving to find K or more packets already in the real queue incurs damage (for example, incurs unacceptable delay). Then from the stationary distribution of an M/M/1 queue, $p_K(\rho) = \rho^K$ is the probability that a packet incurs such damage, and the rate at which packets arrive to incur damage is $\rho p_K(\rho)$. The expected impact of an additional packet (the *shadow price* discussed in [10]) is just

$$\frac{d}{d\rho}\rho p_K(\rho) = (K+1)\rho^K. \tag{22}$$

Now suppose that the Virtual Queue serves at a fraction $\kappa < 1$ of the real queue rate, so that $\kappa = c_v/c$, and suppose further that arriving packets are marked if the fictitious queue has K or more packets already present. Then the marking probability is $p_K(\rho/\kappa)$, and equating this to the expression (22) gives

$$\kappa = (K+1)^{-1/K} . \tag{23}$$

There is a trade off between utilisation and threshold. For the Virtual Queue typical values of K, κ are shown in Table I. Notice that the rate reduction κ is more dramatic for small K, where the virtual queue has a rate less than 80% of the rate of the real queue. Table I also gives values for load ρ for the real queue to give a marking probability of $p_K(\rho/\kappa) = 0.2$ or 0.5. Note that this loading is relatively insensitive to the marking probability.

We have shown how given a K, κ can be chosen. How might K be fixed? This is, in our view, an important area for further research. For a real time service, K might bound the acceptable per hop delay, and so might scale with link capacity. Table I shows the trade-off between utilisation and delay: the last line ($K = 1000$) illustrates how little we gain by having an extreme value of K. A typical router in the current Internet might have buffering per output link equivalent to between 60 and 240KB, equating to 300

TABLE I

VIRTUAL QUEUE THRESHOLDS AND LOADS

K	κ	$\rho, p = 0.2$	$\rho, p = 0.5$
5	.699	.506	.608
10	.787	.670	.734
20	.859	.792	.830
50	.924	.895	.912
100	.955	.940	.948
1000	.993	.992	.992

to 1200 packets in our nominal units units, corresponding to a maximum delay of 30ms for a backbone link. We would advocate a smaller value for the threshold K. More generally, this marking scheme applies to a heterogeneous network where adaptive and non-adaptive traffic is mixed together. For adaptive traffic, it is important to keep K small to avoid oscilliatory behaviour.

With this Virtual Queue marking, for $M = 1$ the acceptance probabilities are given by

$$a(m) = \max\left(0, 1 - \left(\frac{m\mu}{\kappa}\right)^K\right) \tag{24}$$

$$= \max\left(0, 1 - (K+1)(m\,\mu)^K\right). \tag{25}$$

Figure 4 is analogous to Figure 1, but using Virtual Queue marking strategies with $K = 5$ and $K = 10$. As before, Figure 4(a) shows the rejection probabilities $1 - a(m)$ and Figure 4(b) the cumulative occupancy distribution. There are some differences: the acceptance curve $1 - a$ is now convex rather than s-shaped below $a = 0$ for small K, which makes the π distribution less symmetric (because of harder truncation). Figure 5 is analogous to Figure 2, giving the stationary occupancy distribution $\pi(m)$ for differing loads.

The figures are remarkably similar to those of Section III-B, despite the different marking behaviour and assumptions. Note that for this example the key parameters are K and μ. The parameter $1/\mu$ is effectively the system size—because the load of the system cannot go above 1, and limits the maximum number of connections.

From Table I, for $K = 10$, $\kappa = 0.79$ and we aim to keep the system loading below 79%. Hence, the effective capacity of the system is $C = 0.79/\mu$. For a particular value of offered load ν, we can compare the stationary rejection probability $\mathbb{E}(1 - a(m))$ under our proposal with that obtained if we can count connections, and limit their number to C'. The rejection probability under the latter scheme is $Erl(\nu, C')$, Erlang's formula for the blocking of Poisson traffic rate ν offered to C' circuits. Table II gives the rejection probabilities and percentage capacity reduction $100(C - C')/C$ if C' is chosen such that $Erl(\nu, C') = \mathbb{E}(1-a(m))$. The Erlang limit is in some sense the best that be accomplished, and Table II shows an upper bound on the capacity savings if connections could be counted.

We have said little about how the user might choose M,

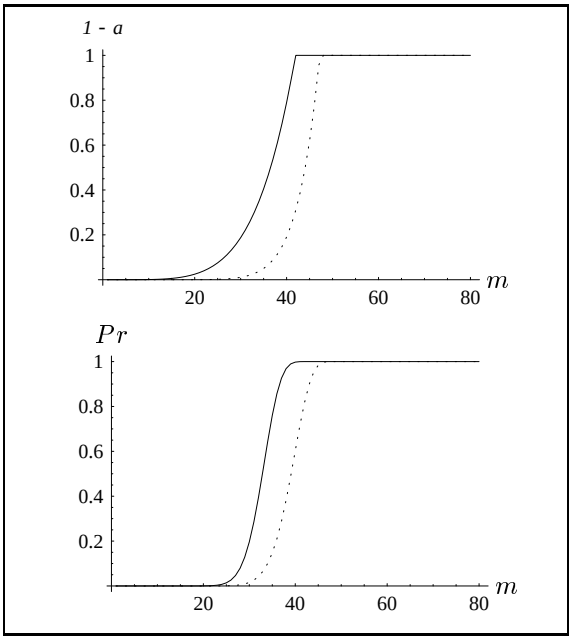

Fig. 4. (a) Rejection probabilities, and (b) Cumulative dccupancy distribution for $\nu = 50$, $1/\mu = 60$, Virtual Queue marking with $K = 5, 10$.

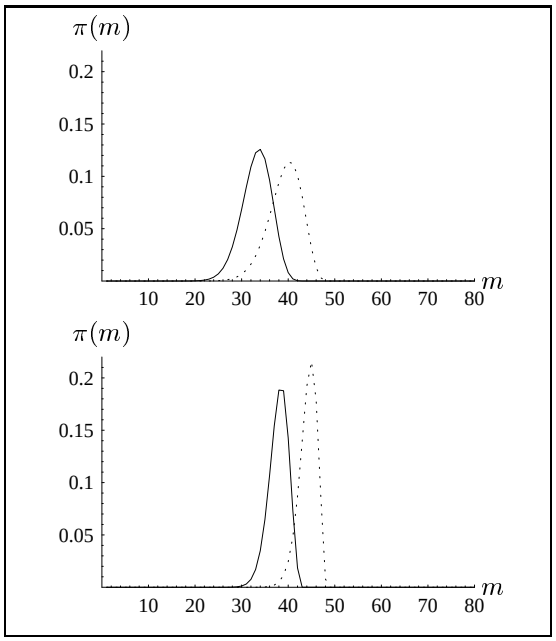

Fig. 5. (a) Occupancy density function $\pi(m)$ for $1/\mu = 60$, Virtual Queue marking with $K = 5, 10$, $\nu = 50$; (b) Occupancy density $\pi(m)$ at high offered load $\nu = 100$.

TABLE II

REJECTION PROBABILITIES FOR $K = 10$, $M = 1$, $1/\mu = 60$ AND CAPACITY SAVINGS WITH A COUNTING SCHEME.

ν	$\mathbb{E}(1 - a(m))$	Capacity Savings
20	0.001	25%
30	.024	17%
40	.106	13%
50	.215	8%
60	.313	6%

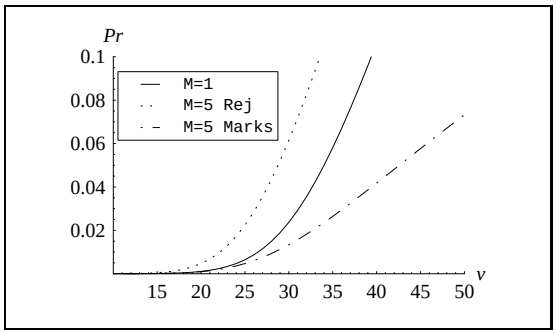

Fig. 6. Expected rejection and expected marking probabilities against varying ν, with $K = 10$, $\mu = 1/60$ for $M = 1$ (solid line), and $M = 5$ (broken lines)

the number of probe packets to offer to the network. Figure 6 shows stationary rejection probabilities, $\mathbb{E}(1 - a(m))$, and expected marking probabilities, $\mathbb{E}(P(m\mu))$, for $M = 5$ (the broken lines) compared to $M = 1$ (solid line) as ν is varied, computed using relations (21) and (20) respectively. For $M = 1$, the expected rejection probability is the same as the expected marking rate, shown as the solid line. For $M = 5$, the rejection probability is increased, while the marking rate is reduced (lower dashed line) compared to $M = 1$. In effect, a higher value of M protects against higher marking rates: if more probe packets are used by all potential connections, then the network stabilises with a lower packet marking rate. Varying M between different types of connections will produce service differentiation.

D. Time scales for Virtual Queues

It is interesting to compare the binomial model of Sections III-A and III-B with the M/M/1 model of Section III-C, where arrivals at the queue form a Poisson process of rate proportional to the number of connections. The workload arrival process at the resource under the latter model exhibits short-range order, and the resource is able to learn about its load relatively quickly, on the packet timescale. Under the binomial model, in contrast, arrivals at the queue exhibit a more complex structure. Over small time intervals, less than the packet spacing within a burst, the packet arrival process exhibits *negative* correlations. But over longer periods the packet arrival process exhibits long-range order, varying slowly as the number of active bursts changes. The resource may well mark packets when the number of connections is quite low, or fail to mark packets when the number of connections is quite high, since its marking strategy depends on the random number of active bursts. In general, the timescales at which queue overload occurs depend subtly upon the statistical properties of individual streams and the amount of traffic aggregation, as well as upon the capacity and buffer size of the resource, and it would not be appropriate to use a fixed measurement interval, as in Section III-A, for heterogeneous mixes of traffic or for poorly characterized sources. In this section we shall explore the robustness of marking mechanisms based on Virtual Queues.

Suppose that the resource can serve packets at rate c

and has a buffer size b. Let n be the number of connections, and let $P(c,b,n)$ be the proportion of packets that are lost. For simplicity of exposition we assume the connections are all of the same type: the argument is identical with multiple traffic types, at the cost of a more elaborate notation. Suppose a connection generates a workload for the resource of $X[0,t]$ in time $[0,t]$: assume X has stationary increments, and that the workloads generated by different connections are independent. Then the many sources asymptotic regime [27], [28], [29] shows that for large systems

$$\log P(c,b,n) \approx \sup_t \inf_s \{stn\alpha(s,t) - s(b+ct)\}. \qquad (26)$$

where

$$\alpha(s,t) = \frac{1}{st}\log \mathbb{E}\left[e^{sX[0,t]}\right] \qquad 0 < s,t < \infty \qquad (27)$$

is the effective bandwidth of an individual source. Let (s^*,t^*) be an extremal pair for Equation (26): then t^* is the *critical timescale* of the resource, giving the most likely timescale on which the buffer threshold is exceeded [40]. As an example, suppose the workload generated by a connection is Gaussian, with

$$X[0,t] = \lambda t + Z(t) \qquad (28)$$

where $Z(t)$ is normally distributed with zero mean. If

$$\mathrm{var}Z(t) = \sigma^2 t^{2H} \qquad (29)$$

then the process represents fractional Brownian motion with Hurst parameter H, and long-range dependence if $H > \frac{1}{2}$. For this model [41]

$$t^* = \frac{b}{c-n\lambda}\frac{H}{1-H}. \qquad (30)$$

Thus, for this example, the critical time scale t^* is linear in the buffer size b, and increases with the Hurst parameter H.

While the formalization of the result (26) requires a limiting regime with small overload probabilities, the time scale t^* identified by the theory is useful in a much wider range of settings. For example, in the case where traffic is Gaussian, t^* is precisely the value t which maximizes the probability that the workload arriving in a period of length t will exceed the maximal capacity of the resource to accept input, $ct+b$.

Write $t^* = t^*(c,b,n)$ to emphasise the dependence of the critical time scale on the capacity and buffer of the queue. Now suppose that marking is determined by a Virtual Queue, of reduced rate κc and reduced buffer size κb. Then observe, from relation (30) and more generally from relation (26), that

$$t^*(\kappa c, \kappa b, \kappa n) = t^*(c,b,n). \qquad (31)$$

Thus the critical time scale for the Virtual Queue is the *same* as it would be for the real queue if the load on the real queue were a factor $1/\kappa$ larger. The Virtual Queue thus provides an early warning of overload, and implicitly and robustly tunes the time period over which overload is detected to the statistical properties of the traffic sources.

IV. Concluding remarks

The framework we have presented is general, and can be analysed in detail. We have assumed the marking rate is determined by the number of connections alone. More generally, the marking rate can depend on a weighted sum $m_j = \sum_{r:j\in r}\alpha_r n_r$ reflecting differing bandwidth requirements. Differing user behaviour naturally translates into user dependent responses to marks: for example the user or system may adopt a more complicated acceptance strategy, such as deciding to enter if no more than a specified number of packets is marked, where the number is greater than 1. For these extensions the product form solution breaks down, without certain side conditions that essentially preserve reversibility. However, fixed-point approximations in the spirit of Section II can still be constructed, and complete networks analysed.

We have said little about why a connection should chose to react in the way described in the paper. If the connection is an application running on the end-system, then the behaviour could be embedded in a protocol stack, with behaviour mandated, much as the behaviour of TCP connections is under the control of the operating system. Different types of connection (voice, video, streamed data etc) could have different mandated reactions to the receipt of probe packets. Viewed in this light, our scheme can represent a lightweight signalling system with soft guarantees, which could be used in an Intranet for example, sharing bandwidth within a Virtual Path. An alternative and more controversial scenario would see the marked packets as representing some cost or charge to the user, which might represent real money or a distributed mint. In this case, the probing phase enables the user to form an estimate of the likely charge of a connection. Some possible user reactions are investigated in [42]. This is a natural framework for heterogeneity, where different users can have different strategies for choosing whether or not to enter the network.

References

[1] T. R. Griffiths and P. B. Key, "Adaptive call admission control in ATM networks," in *The fundatmental role of Teletraffic Engineering in the Evolution of Telecommunication Networks: Proceedings of the 14th International Teletraffic Congress* (J. Labetoulle and J. Roberts, eds.), vol. 1b, International Teletraffic Congress, Elsevier, 1994.

[2] P. B. Key, "Connection admission control in ATM networks," *BT Technology Journal*, vol. 13, July 1995.

[3] R. J. Gibbens, F. P. Kelly, and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE Journal on Selected Areas in Communications, special issue on Advances in the Fundamentals of Networking*, vol. 13, no. 6, pp. 1101–1114, 1995.

[4] R. J. Gibbens and F. P. Kelly, "Measurement-based connection admission control," in *Teletraffic Contributions for the Information Age, Proceedings ITC16* (V. Ramaswami and P. Wirth, eds.), vol. 2a, pp. 879–888, Elsevier, June 1997.

[5] J. Y. Hui, "Resource allocation for broadband networks," *IEEE Journal on Selected Areas in Communication*, vol. 6, pp. 1598–1608, 1988.

[6] S. Floyd, "Comments on measurement-based admissions control for controlled-load services." http://www.aciri.org/floyd/admit.html, July 1996.

[7] S. Jamin and S. Shenker, "Measurement-based admission control algorithms for controlled-load service: a structural examination," Tech. Report CSE-TR-333-97, University of Michigan, April 1997. http://irl.eecs.umich.edu/jamin/papers/.

[8] M. Grossglauser and D. N. C. Tse, "A framework for robust measurement-based admission control," *IEEE / ACM Transactions on Networking*, vol. 7, pp. 293–309, 1999.
[9] K. Ramakrishnan and S. Floyd, "A proposal to add explicit congestion notification ECN to IP," RFC 2481, IETF, Jan. 1999. ftp://ftp.isi.edu/in-notes/rfc2481.txt.
[10] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, pp. 1969–1985, 1999. www.statslab.cam.ac.uk/∼frank/evol.html.
[11] P. B. Key, D. R. McAuley, P. Barham, and K. Laevens, "Congestion pricing for congestion avoidance," Microsoft Research Technical Report MSR-TR-99-15, MSR, 1999. http://research.microsoft.com/pubs/.
[12] R. J. Gibbens and F. P. Kelly, "Distributed connection acceptance control for a connectionless network," in *Teletraffic Engineering in a Competitive World, Proceedings ITC16* (P. Key and D. Smith, eds.), pp. 941–952, Elsevier, June 1999.
[13] Z.R. Turányi and L. Westberg, "Load control: lightweight provisioning of Internet resources." www.ericsson.co.hu/ethzrt/ 1999.
[14] V. Elek, G. Karlsson and R Rönngren, "Admission control based on end-to-end measurements." INFOCOM 2000. www.comnet.technion.ac.il/infocom2000
[15] G. Bianchi, A. Capone and C. Petrioli, "Throughput analysis of end-to-end measurement based admission control in IP." INFOCOM 2000. www.comnet.technion.ac.il/infocom2000
[16] E. Brockmeyer, H. L. Halstrom, and A. Jensen, *The Life and Works of A.K. Erlang*. Copenhagen: Academy of Technical Sciences, 1948.
[17] R. I. Wilkinson, "Theory for toll traffic engineering in the USA," *Bell System Technical Journal*, vol. 35, pp. 421–513, 1956.
[18] A. G. Greenberg and R. Srikant, "Computational techniques for accurate performance evaluation in multirate, multihop communication networks," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 253–277, 1997.
[19] P. B. Key, "Implied cost methodology and software tools for a fully connected network with DAR and trunk reservation," *British Telecom Technology Journal*, vol. 6, pp. 52–65, 1988.
[20] A. Farago, S. Blaabjerg, L. Ast, G. Gordos, and T. Henk, "A new degree of freedom in ATM network dimensioning: optimizing the logical configuration," *IEEE Journal Selected Areas Communications*, vol. 13, pp. 1199–1206, 1995.
[21] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "ATM network design and optimization: a multirate loss network framework," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 531–543, 1996.
[22] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "Virtual Private Networks: joint resource allocation and routing design," in *INFOCOM*, IEEE, 1999.
[23] F. P. Kelly, "Loss networks," *Annals of Applied Probability*, vol. 1, pp. 319-378, 1991.
[24] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Berlin: Springer, 1995.
[25] P. J. Hunt and T. G. Kurtz, "Large loss networks," *Stochastic Processes and their Applications*, vol. 53, pp. 363–378, 1994.
[26] S. Zachary, "Dynamics of large uncontrolled loss networks." *J. App. Prob.*, vol. 37, 2000, http://www.ma.hw.ac.uk/˜stan/papers/
[27] D. D. Botvich and N. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.
[28] C. Courcoubetis and R. Weber, "Buffer overflow asymptotics for a switch handling many traffic sources," *Journal of Applied Probability*, vol. 33, pp. 886–903, 1996.
[29] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis, "Application of the many sources asymptotic and effective bandwidths to traffic engineering." To appear in Telecommunication Systems. http://www.ics.forth.gr/netgroup/msa, 1999.
[30] F. P. Kelly, *Reversibility and stochastic networks*. Chichester: Wiley, 1979.
[31] Z. Dziong and J. W. Roberts, "Congestion probabilities in a circuit-switched integrated services network," *Performance Evaluation*, vol. 7, pp. 267–284, 1987.
[32] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability*, vol. 18, pp. 473–505, 1986.
[33] S. Zachary, "On blocking in loss networks," *Advances in Applied Probability*, vol. 23, pp. 355–372, 1991.
[34] F. P. Kelly, "Routing and capacity allocation in networks with trunk reservation," *Mathematics of Operations Research*, vol. 15, 1990.
[35] P. B. Key and D. R. McAuley, "Differential QoS and pricing in networks: where flow control meets game theory." *IEE Proc Software*, vol. 146, pp. 39–43, 1999.
[36] F.P. Kelly, "Models for a self-managed Internet", *Phil. Trans. R. Soc. Lond. A* 2000. www.statslab.cam.ac.uk/∼frank/smi.html
[37] J. W. Roberts, ed., *COST224 Performance evaluation and design of multiservice networks*, Commission of the European Communities, Oct. 1992. Final Report.
[38] S. Bajaj, L. Breslau, and S. Shenker, "Is service priority useful in networks?," 1998. http://www.cs.wpi.edu/∼sigmet98/bajaj.ps.
[39] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993. http://www-nrg.ee.lbl.gov/floyd/red.html.
[40] D. Wischik, "Sample path large deviations for queues with many inputs." Submitted to Annals of Applied Probability, 1999.
[41] B. K. Ryu and A. Elwalid, "The importance of the long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities," in *Proceedings ACM SIGCOMM*, (Stanford), 1996.
[42] P. B. Key and L. Massoulié, "User policies in a network implementing congestion pricing," in *Workshop on Internet Service Quality Economics*, (MIT), December 1999. http://www.marengoresearch.com/isqe/massoulie_fnl.pdf.

Appendix 1

Consider the simple model of Section II-A. Suppose that the network is large, and that, for each resource j, the function $a_j(\cdot)$ decreases rapidly from 1 to 0 in the neighbourhood of some "capacity" C_j. This is so in the case of each of the models studied in Section III. For example, in the binomial model of Section III-B we have $C_j = \theta_j/p$—see the discussion there. We show here that, to a good approximation, the stationary acceptance probabilities and the corresponding fixed point $\bar{n}$ of Section II-C are the same as for a traditional uncontrolled loss network with hard capacity constraints C_j, $j \in J$. These are therefore independent of the more detailed structure of the functions $a_j(\cdot)$. We also discuss briefly limiting dynamics.

To make these ideas more formal, consider a sequence of networks, indexed by some scale parameter N. All members of the sequence have a common *incidence matrix* $\alpha = (\alpha_{jr},\ j \in J, r \in R)$, where $\alpha_{jr} = 1$ if $j \in r$ and $\alpha_{jr} = 0$ otherwise: we assume the matrix α is of full rank. The N^{th} member of the sequence has arrival rates $\nu_r^N = N\nu_r$, $r \in R$, and acceptance functions $a_j^N(\cdot)$ satisfying

$$\lim_{N\to\infty} a_j^N(N(C_j - \epsilon)) = 1, \qquad \lim_{N\to\infty} a_j^N(N(C_j + \epsilon)) = 0, \tag{32}$$

for all $\epsilon > 0$; define also $\bar{n}^N$ to be the corresponding solution of the fixed point equations (6) of Section II-C.

Let $(A_j,\ j \in J)$ be the unique solution in $A_j \in (0,1]$, $j \in J$, of the equations

$$\sum_{r:j\in r} \nu_r \prod_{i\in r} A_i \le C_j, \qquad j \in J, \tag{33}$$

$$\sum_{r:j\in r} \nu_r \prod_{i\in r} A_i = C_j \qquad \text{if } A_j < 1, \qquad j \in J \tag{34}$$

(the uniqueness here follows from Theorem 2.1 of [32]). For each r, define also $x_r = \nu_r \prod_{j\in r} A_j$.

Theorem 1: Under the above limiting regime,

$$\lim_{N\to\infty} a_j^N(m_j(\bar{n}^N)) = A_j, \qquad j \in J, \tag{35}$$

$$\lim_{N\to\infty} \frac{\bar{n}^N}{N} = x_r, \qquad r \in R. \tag{36}$$

Proof: From (3), (6) and (7) we have, for the N^{th} member of the sequence,

$$\frac{1}{N} m_j(\bar{n}^N) = \sum_{r:j\in r} \nu_r \prod_{i\in r} a_i^N(m_i(\bar{n}^N)), \qquad j \in J, \tag{37}$$

$$\frac{1}{N} \bar{n}_r^N = \nu_r \prod_{j\in r} a_j^N(m_j(\bar{n}^N)), \qquad r \in R. \tag{38}$$

Now consider any subsequence in which $(a_j^N(m_j(\bar{n}^N)),\ j \in J)$, is convergent. Then, for all j, $\lim_{N\to\infty} m_j(\bar{n}^N)/N \le C_j$ (for otherwise, from (32), $\lim_{N\to\infty} a_j^N(m_j(\bar{n}^N)) = 0$, leading to a contradiction for the j^{th} member of (37)). Further, for j such that $\lim_{N\to\infty} m_j(\bar{n}^N)/N < C_j$, it follows from (32) that $\lim_{N\to\infty} a_j^N(m_j(\bar{n}^N)) = 1$. It now follows from the uniqueness of $(A_j,\ j \in J)$ above that, in the entire sequence, the result (35) holds, and (36) is now immediate from (38). ■

Finally, we may show as in [32] that, under the above limiting regime, the fixed point approximation of Section II-C for the acceptance probabilities becomes asymptotically exact. It then follows from Theorem 1 that the limiting acceptance probability for a call of type r is $\prod_{j\in r} A_j$.

We mention also the limiting dynamics $(x(t),\ t \ge 0)$, where $x(t) = n(t)/N$, under the above regime. These are as given by (15), but since here, in the limit, each function $a_j(\cdot)$ jumps from 1 to 0 as its argument passes through C_j, it is necessary to use stochastic averaging as described in [25]. Using the techniques of [26], we may show that the appropriate Lyapunov function for the limiting dynamics is again given by the function $\mathcal{U}$ defined by (16), where here we take the integrand $\log a_j(y) = 1$ for all y. This is as might be expected: the limiting dynamics $x(\cdot)$ are of course constrained to satisfy $\sum_{r:j\in r} x_r(t) \le C_j$ for all j.

Appendix 2

How rapidly does the product of acceptance probabilities in (3) decay, when given by the binomial probabilities (2)? Well, if $mp > \theta$,

$$a(m) \le \exp\left\{-\left((m-\theta)\log\frac{m-\theta}{m(1-p)} + \theta\log\frac{\theta}{mp}\right)\right\}$$

by a Chernoff bound on the Binomial distribution. Hence

$$\begin{aligned}\prod_{k=\theta/p}^{n+1} a(k-1) \le\ & \\ & \exp\left\{-\int_{\theta/p}^{n}\left((x-\theta)\log\frac{x-\theta}{x(1-p)} + \theta\log\frac{\theta}{xp}\right)dx\right\} \\ =\ & \exp\left\{-\frac{1}{2}\left[(n-\theta)^2\log\frac{(n-\theta)p}{\theta(1-p)} - \right.\right. \\ & \left.\left. n^2\log\frac{np}{\theta} + \theta\left(n-\frac{\theta}{p}\right)\right]\right\},\end{aligned}$$

indicating a rather rapid decay in the product form above $n = \theta/p$. Similarly, there is a rapid approach to unity of the product $\prod_{k=1}^{n} a(k-1)$ below $n = \theta/p$.

Appendix 3

We prove the result of Section II-F. Assume $\nu_r > 0$ for $r \in R$, and that $a_j(y)$ is a non-negative continuous decreasing function of y, not identically zero, for $j \in J$. Then we may establish the following result.

Theorem 2: The function (16) is a Lyapunov function for the system of differential equations (15). The unique value x maximizing $\mathcal{U}(x)$ is a stable point of the system, to which all trajectories converge.

Proof: The assumptions on $\nu_r, r \in R$, and $a_j, j \in J$, ensure that $\mathcal{U}(x)$ is strictly concave on the positive orthant with an interior maximum; the maximizing value of x is thus unique. Observe that

$$\frac{\partial}{\partial x_r}\mathcal{U}(x) = \log\frac{\nu_r}{x_r} + \sum_{j\in r}\log a_j\Big(\sum_{s:j\in s} x_s\Big); \tag{39}$$

setting these derivatives to zero identifies the maximum. Further

$$\frac{d}{dt}\mathcal{U}(x(t)) \quad = \quad \sum_{r\in R}\frac{\partial\mathcal{U}}{\partial x_r}\cdot\frac{d}{dt}x_r(t).$$

Next note that expression (39) necessarily has the same sign as expression (15) for each $r \in R$, establishing that $\mathcal{U}(x(t))$ is strictly increasing with t, unless $x(t) = x$, the unique x maximizing $\mathcal{U}(x)$. The function $\mathcal{U}(x)$ is thus a Lyapunov function for the system (15), and the theorem follows. ■