10.1098/rsta.2000.0651

THE ROYAL SOCIETY

Models for a self-managed Internet

By Frank P. Kelly

*Centre for Mathematical Sciences, University of Cambridge,
Wilberforce Road, Cambridge CB3 0WB, UK*

This paper uses a variety of mathematical models to explore some of the consequences of rapidly growing communications capacity for the evolution of the Internet. It argues that queueing delays may become small in comparison with propagation delays, and that differentiation between traffic classes within the network may become redundant. Instead, a simple packet network may be able to support an arbitrarily differentiated and constantly evolving set of services, by conveying information on incipient congestion to intelligent end-nodes, which themselves determine what should be their demands on the packet network.

Keywords: stability; shadow price; fairness; explicit congestion notification; queues; short transfers

1. Introduction

When a communication network becomes congested, a trade-off must be made between traffic that is carried and traffic that is not. In telephone networks the traditional mechanism has been a call admission control, which blocks a newly arriving call if any of the resources that would be needed by the call are congested. The implementation of this mechanism has generally required a parallel signalling network, to monitor resource loads and make admission decisions. In contrast, congestion in the current Internet causes traffic through a congested resource to suffer packet loss, and this in turn causes at least some end-systems to reduce their load on that resource.

In a traditional telephone network, the load of a call is well defined, and users tolerate occasional blocking in return for a guaranteed bandwidth for accepted calls. The statistical properties of packet streams carrying voice, together with the limit on load achieved by the call admission control, allow queues within the network to be kept short and quality of service for accepted calls to be assured.

In the Internet, the load imposed by end-systems is much less predictable, and the bandwidth achieved by a user fluctuates as a consequence of the behaviour of other users. Buffers, important since the early days of store-and-forward communication networks (Kleinrock 1964), are used to smooth statistical fluctuations in demand for scarce transmission capacity.

Might it be possible to design a network so that it can carry existing loads on telephone networks and the Internet, as well as new forms of traffic? A problem is that some current Internet traffic needs large buffers, causing unacceptable queueing delays for real-time applications such as telephony, a problem compounded by the current use of packet loss and delay as the default mechanism for deterring demand. Attempts to solve the problem are generally based on a small number of

© 2000 The Royal Society

service classes with well-defined quality and prices, delivered by a network using queueing mechanisms that treat differently packets belonging to different classes. By giving higher priority to some packets, and using large buffers for others, it may be possible to carry real-time applications over the same transmission capacity as other less-delay-sensitive traffic. But this approach has drawbacks. In particular, the development of asynchronous transfer mode (ATM) traffic classes (International Telecommunications Union 1996) has illustrated some of the difficulties of defining service categories and incentive-compatible pricing schemes (Songhurst 1999) before the applications that might use the categories have been invented or have become widespread.

A body of work is now emerging that takes a radically different approach to differentiated services (see, for example, Crowcroft & Oechslin 1998; Gibbens & Kelly 1999*a*; Key & McAuley 1999). Its premise is that a simple packet network may be able to support an arbitrarily differentiated set of services by conveying information on congestion from the network to intelligent end-nodes, which themselves determine what should be their demands on the packet network. There would then be no need for large buffers or priority queues within the network, or for connection acceptance control at the border of the network. This paper is intended to provide a brief overview of some of the mathematical models that have been found useful in the exploration of this approach.

The organization of the paper is as follows. In § 2 we explore the impact of rapidly growing communications capacity upon queueing delays. Under various scaling regimes we argue that queueing delays become small in comparison with propagation delays. This point is hardly controversial, and yet its consequences are potentially profound. One consequence is that attempts to differentiate between service classes within the network using discriminatory queueing disciplines may become redundant. Another consequence is that a very simple network mechanism, the setting of just a single bit to mark some packets (Ramakrishnan & Jain 1990; Floyd 1994), may be enough to implement a *smart market* (MacKie-Mason & Varian 1994) for the efficient allocation of resources; since, as end-systems see more packets per round-trip time, any randomness associated with whether or not a particular packet is marked becomes less relevant than the proportion of packets marked.

If market mechanisms are able to broadly align supply and demand for communications capacity, then rather simple models of queueing behaviour may be enough to understand the dynamics of various load control algorithms. In § 3 we review a tractable mathematical model of a network carrying adaptive traffic, from end-systems able to adjust their rates to available bandwidth. Stability, at least on a time-scale of seconds, is established by the explicit construction of a Lyapunov function. Stability on shorter time-scales, comparable with round-trip times in the network, is considered in § 4. Simple queueing models are used to establish sufficient conditions for local stability, in terms of relationships between the gain parameters of rate or window control algorithms and the marking strategies of resources. A striking observation is that the lower the buffer level at which marking occurs, or the higher the statistical variability of traffic at the packet level, the *fewer* the possibilities for lag-induced oscillatory behaviour.

Gibbens & Kelly (1999*b*), Turányi & Westberg (1999) and Kelly *et al.* (2000) have described how packet marking at congested resources may be used as the basis for distributed admission control of non-adaptive applications such as traditional telephony, without the need for a parallel signalling network. In § 5 we explore the behaviour of competing aggregates of adaptive traffic and non-adaptive traffic subject to distributed admission control. In § 6 we consider short transfers, where a file may have completed its transmission within a round-trip time. There is no possibility for adaptive control on such a short time-scale, and the effect of short transfers is to place a random background load upon the network. A simple queueing model indicates that this background load may well *improve* the stability of the adaptive traffic with which resources are shared.

The marks discussed in this paper give shadow prices at the finest possible granularity; many choices are possible about the level of aggregation at which the marks are reflected as costs or prices to economic agents. For example, the marks could allow the dispersal of charging operations such as metering, accounting and billing to customer systems; Briscoe *et al.* (2000) argue that such an architecture gives simplicity and scalability, without sacrificing commercial flexibility or security. Or the marks could be charges to third-party software running on customer systems, software which undertakes the risks associated with uncertain network loads and user behaviour (Semret & Lazar 1999; Key 1999), and presents an economic agent with pricing choices tailored for that agent.

2. Scalings for queueing delay

In this section we study queueing delay under various scaling regimes where traffic and capacity grow in line with one another. We shall find that, in two of the three regimes considered, queueing delays decrease.

Let $X[s,t]$ be the amount of work that arrives at a queue in the time-interval $[s,t]$. If the queue has a service rate C and no work is lost, then the amount of work buffered in the queue at time 0 is

$$Q = \sup_{u \geqslant 0}\{X[-u,0] - Cu\}. \tag{2.1}$$

Assume that the service rate C is adequate, so that Q is defined as a proper random variable.

Now suppose that the input to the queue in the time-interval $[s,t]$ becomes

$$\sum_{i=1}^{c} aX_i[bs,bt], \tag{2.2}$$

where the random processes $X_i[s,t]$, $i = 1,2,\dots,c$, are independent and each distributed as $X[s,t]$. Let $Q(a,b,c)$ label the random variable defined by the formula (2.1) when $X[s,t]$ is replaced by expression (2.2) and C is replaced by $abcC$. Thus $Q(a,b,c)$ describes the queue length in a system where three forms of scaling have been applied: the original stream $X[s,t]$ has been increased in volume by a factor a, speeded up by a factor b, and c streams have been multiplexed. The mean amount of work arriving at the queue has been increased by a factor abc, as has the service rate at the queue. Let

$$\tau(a,b,c) = Q(a,b,c)/(abcC),$$

the queueing delay under the first-in-first-out queueing discipline.

The impact of the first two forms of scaling on queueing delays is straightforward:

$$\tau(a,b,c)=\tau(1,b,c), \qquad \tau(a,b,c)=b^{-1}\tau(a,1,c). \tag{2.3}$$

Certainly, $\tau(a,b,c)\leqslant\tau(a,b,1)$, but the precise impact of the multiplexing factor c depends on the statistical properties of the streams. An illuminating example, illustrating the impact of traffic variability on several scales, is provided by a self-similar traffic model (Willinger *et al.* 1996). If the increments of $X_i[s,t]$ are stationary, with $X_i[0,t]\sim N(\lambda t,\sigma^2 t^{2H})$, corresponding to fractional Gaussian input with Hurst parameter $H\in(0,1)$, then

$$\tau(a,b,c)=c^{-1/(2-2H)}\tau(a,b,1) \tag{2.4}$$

(Norros 1994). Observe that short-range order, the case $H=0.5$, gives a reduction of queueing time by a factor c; larger values of H, corresponding to increasing degrees of long-range order, give even larger reductions.

We have defined $\tau(a,b,c)$ to be the queueing delay under the first-in-first-out discipline, but the relationships (2.3)–(2.4) hold similarly for other important time periods that can be expressed in terms of the queue length process, such as the busy period of the queue. Detailed investigations of queueing time-scales, using real traffic traces, are described in Courcoubetis *et al.* (1999) and Gibbens & Teh (1999), and some of the implications for queueing networks are developed in Wischik (1999*a*).

We conclude from relations (2.3)–(2.4) that the volume scaling parameter a does not impact on queueing delay, but the speed and multiplexing parameters b and c both cause substantial reductions in queueing delay. In contrast, propagation delays depend on the speed of light and are unaffected by these scalings. These observations motivate our later models, in which queueing delays and busy periods are assumed small in comparison with propagation delays, and in which packets are served anonymously by a single queue at each resource.

The scalings of this section assume that capacity is adequate: in an overloaded network it is certainly possible to differentiate between traffic classes by using queueing mechanisms. To assure adequate capacity requires suitable capacity provisioning, to deal with growth in demand measured over days, weeks or longer (cf. Gibbens *et al.*, this issue); we do not consider this topic here. It also requires load control, to deal with fluctuations over seconds or less. In the following sections we consider load control for various types of traffic.

3. Rate control of adaptive traffic

In this section we outline a tractable mathematical model of a network carrying rate-adaptive traffic, following the development of Kelly *et al.* (1998); related approaches are described in Golestani & Bhattacharyya (1998) and Low & Lapsley (1999). The algorithm described is closely related to the window-based Congestion Avoidance algorithm of Jacobson (1988) for file transfers, but is also intended to model rate-adaptive real-time applications.

Consider a network with a set J of *resources*. Let a *route* r identify a non-empty subset of J, and write $j\in r$ to indicate that route r passes through resource j. Let R be the set of possible routes, and suppose that route r carries a flow of rate x_r for each $r\in R$. Suppose that as a resource becomes more heavily loaded it generates feedback signals intended to indicate congestion to the end-systems or users responsible for routes passing through that resource.

How might the end-systems react? Consider the system of differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}x_r(t) = \kappa_r\biggl(w_r - x_r(t)\sum_{j\in r}\mu_j(t)\biggr), \tag{3.1}$$

for $r \in R$, where

$$\mu_j(t) = p_j\biggl(\sum_{r:j\in r} x_r(t)\biggr), \tag{3.2}$$

for $j \in J$. We interpret relations (3.1), (3.2) as follows. We suppose that resource j marks a proportion $p_j(y)$ of packets with a feedback signal when the total flow through resource j is y; and that user r views each feedback signal as a congestion indication requiring some reduction in the flow x_r. Then equation (3.1) corresponds to a rate control algorithm for user r that comprises two components: a steady increase at rate proportional to w_r, and a steady decrease at a rate proportional to the stream of congestion-indication signals received.

It is shown in Kelly *et al.* (1998) that

$$\mathcal{U}(x) = \sum_{r\in R} w_r \log x_r - \sum_{j\in J}\int_0^{\sum_{r:j\in r} x_r} p_j(z)\,\mathrm{d}z$$

is a Lyapunov function for the system of differential equations (3.1), (3.2), and it is deduced that the unique value x maximizing $\mathcal{U}(x)$ is a stable point of the system, to which all trajectories converge. The variable $\mu_j(t)$ has an interpretation as the implied *shadow price* per unit of flow through resource j at time t, and, at the stable point,

$$x_r = \frac{w_r}{\sum_{j\in r}\mu_j}. \tag{3.3}$$

The weights $(w_r, r \in R)$ determine the share of scarce resources obtained by different flows, and the rates x given by equation (3.3) have an interpretation in terms of a *weighted proportional fairness* criterion (Kelly 1997; Crowcroft & Oechslin 1998).

In the current Internet, the rate at which a source sends packets is often controlled by the Transmission Control Protocol (TCP) of the Internet, implemented as software on end-systems (Jacobson 1988). When a resource within the network becomes overloaded, one or more packets are lost; loss of a packet is taken as an indication of congestion, the destination informs the source, and the source slows down. The source then gradually increases its sending rate until it again receives an indication of congestion. In the future, resources may also have the ability to indicate congestion by marking packets, using an Explicit Congestion Notification bit (Floyd 1994), and current questions concern how packets might be marked and how TCP might be adapted to react to marked packets. Equation (3.1) describes a form of linear increase and multiplicative decrease similar (differences are discussed in Key *et al.* (1999) and Kelly (2000)) to that used in Jacobson's (1988) Congestion Avoidance algorithm, but designed to react less severely to congestion indication signals.

Observe that if several flows $r(1), r(2), \dots, r(n)$ use an identical set of resources, r, say, and share the same gain parameter $\kappa_{r(1)} = \cdots \kappa_{r(n)} = \kappa_r$, then the behaviour of the aggregate,

$$x_r(t) = \sum_{i=1}^{n} x_{r(i)}(t), \tag{3.4}$$

may be studied by simply removing the labels $r(1), r(2), \dots, r(n)$ from the set R, replacing them by the aggregate label r, and, using the aggregate weight,

$$w_r = \sum_{i=1}^{n} w_{r(i)}. \tag{3.5}$$

One consequence of this important scaling property is that the differential equations (3.1), (3.2) may be used to study the behaviour of the network at various levels of aggregation. In the next section we consider forms of packet-level behaviour that lead to these equations, with each route r corresponding to an individual flow, while in § 4 we use the equations to study the behaviour of large aggregates in competition with other forms of traffic.

4. Packet-scale behaviour

In this section we describe how the differential equations (3.1), (3.2) might arise naturally from the detailed packet-level behaviour of end-system and resources, and consider stability over time-scales comparable with round-trip times in the network. The impact of time-lags on the stability of equations (3.1), (3.2) has been considered in the network context by Kelly *et al.* (1998), Tan (1999) and Johari & Tan (2000). Our emphasis here is on how the packet-level behaviour of marking strategies determines the functions p_j, $j \in J$, appearing in these treatments.

We begin by noting the important *self-clocking* feature of Jacobson's (1988) algorithm: the sender uses an acknowledgement from the receiver to prompt a step forward, and this produces a key dependence on the round-trip time T of the connection. In more detail, TCP maintains a window of transmitted but not yet acknowledged packets; the rate x and the window size cwnd satisfy the approximate relation cwnd $= xT$. Each positive acknowledgement increases the window size cwnd by $1/\text{cwnd}$; each congestion indication halves the window size.

Consider a variant constructed by incrementing cwnd by

$$\bar{\kappa}\left(\frac{\bar{w}}{\text{cwnd}} - f\right),$$

per acknowledgement, where $f = 1$ or 0 according to whether the packet acknowledged was marked or not. Since the time between update steps is about T/cwnd, the expected change in the rate x per unit time is approximately

$$\frac{\bar{\kappa}((\bar{w}/\text{cwnd}) - p)/T}{T/\text{cwnd}} = \kappa(w - xp),$$

where $\kappa = \bar{\kappa}/T$, $w = \bar{w}/T$ and p is the probability of a mark. This expression corresponds with the form of linear increase and multiplicative decrease described by equation (3.1), where the probability a packet is marked somewhere along its route is approximated by the sum of the marking probabilities at the separate resources along that route.

The model (3.1), (3.2) ignored round-trip times within the network, in order to explore broader aspects of dynamical behaviour. Next we consider the impact of round-trip times on stability, for a very simple example.

Consider a collection of streams all using a single scarce resource. Let the round-trip time be T for each connection, and suppose connections share the same gain parameter κ. Then equations (3.1), (3.2) become, aggregating all connections into a total flow x and taking the time-lag into account:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \kappa(w - x(t-T)p(x(t-T))). \tag{4.1}$$

The unique equilibrium point of this system does not depend on the round-trip time T, but its stability does. To explore this issue, we first recall some facts about the linear retarded equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = -\alpha u(t-T), \tag{4.2}$$

where $\alpha > 0$. Solutions to equation (4.2) converge to zero as t increases if $\alpha < \pi/2T$, and the convergence is non-oscillatory if $\alpha < 1/\mathrm{e}T$ (see Hale (1977) and Johari & Tan (2000) for a discussion of the multi-dimensional network generalization).

Let x be the equilibrium point of the system (4.1), let $x(t) = x + u(t)$, and write p, p' for the values of the functions $p(\cdot)$, $p'(\cdot)$ at x. Observe that x, p are related by $xp(x) = w$; we assume that the capacity of the resource C is adequate, i.e. $w < C$. Then, linearizing the system (4.1) about x, we obtain equation (4.2) with $\alpha = \kappa(p + xp')$. Hence, the equilibrium point of the differential equation (4.1) is stable, and the local convergence is non-oscillatory, if

$$\kappa T(p + xp') < \mathrm{e}^{-1}; \tag{4.3}$$

stability alone is assured if condition (4.3) is satisfied with e^{-1} replaced by $\pi/2$.

Next we explore some simple forms for the function $p(x)$. Suppose that the workload arriving at the resource over a time-period τ is Gaussian, with mean $x\tau$ and variance $x\tau\sigma^2$, and that a packet is marked if when it arrives the workload already present in the queue is larger than a threshold level B. Then, from the stationary distribution for a reflected Brownian motion (Harrison 1985),

$$p(x) = \exp\left\{\frac{-2B(C-x)}{x\sigma^2}\right\}, \tag{4.4}$$

and the condition (4.3) for a non-oscillatory stable equilibrium becomes

$$\kappa T\left(1 + \frac{2BC}{x\sigma^2}\right)p(x) < \frac{1}{\mathrm{e}}. \tag{4.5}$$

The left-hand side of relation (4.5) is increasing in w $(= xp(x))$, and so the relation is satisfied for any $w < C$ if

$$\kappa T\left(1 + \frac{2B}{\sigma^2}\right) < \frac{1}{\mathrm{e}}. \tag{4.6}$$

Note that as the threshold level B increases, or as the variability of traffic at the packet level σ^2 decreases, the greater the possibilities for lag-induced oscillatory behaviour. The reason is straightforward: increasing B or decreasing σ^2 causes p' to increase. This increased sensitivity of the resource's load response may compromise stability, unless there is a corresponding decrease in κT, the sensitivity of response of end-systems to marks. (Recall that for the self-clocking window control algorithm described earlier in this section, $\kappa T = \bar{\kappa}$ is the window size decrement made by an end-system in response to a marked packet.) The magnitudes of κ, p' also affect the variance about the equilibrium point in the presence of noise, and speed of convergence (Kelly *et al.* 1998): broadly, smaller values of κ or larger values of p' lessen the random fluctuations of rates at equilibrium, while larger values of κ or larger values of p' increase the speed with which changes in parameters such as w may be tracked.

Many variants of the above packet marking algorithm can be analysed. For example, suppose we mark a packet with probability $1 - \exp(-qW)$ if it arrives to find a workload of W already present (the Random Early Marking proposal of Lapsley & Low (1999), a variant of the Random Early Detection proposal of Floyd & Jacobson (1993); see also Floyd (1994)). Then the probability that a packet is marked is readily deduced to be

$$p(x) = \frac{qx\sigma^2}{qx\sigma^2 + 2(C - x)}, \tag{4.7}$$

and a simple calculation shows that condition (4.3) is satisfied for any $w < C$ if

$$\kappa T\left(1 + \frac{2}{q\sigma^2}\right) < \frac{1}{\mathrm{e}}. \tag{4.8}$$

Observe the correspondence between q^{-1} in relation (4.8) and the threshold level B in the earlier relation (4.6): indeed the second algorithm corresponds to randomly resetting the threshold level upon each arrival, according to an exponential distribution with mean q^{-1}.

A suggestion of Gibbens & Kelly (1999*a*) is that a virtual buffer be maintained of finite size B, and that from the time of a virtual buffer overflow to the end of the virtual buffer's busy period, all packets leaving the real queue be marked. Thus the virtual buffer's contents evolve as if overflow is lost, while the real buffer may or may not have loss. For our Gaussian traffic model, the rate at which workload overflows the virtual buffer is

$$L(x, C) = (C - x)\left(\exp\left\{\frac{2B(C - x)}{x\sigma^2}\right\} - 1\right)^{-1}$$

(Harrison 1985) and the proportion of workload marked is given by

$$p(x) = -\frac{\mathrm{d}}{\mathrm{d}C}L(x, C).$$

Virtual loss from the virtual buffer causes the function $p(x)$ to increase more slowly as x approaches C, and this reduces the maximum value of the stability factor $p + xp'$ appearing in relation (4.3) (see figure 1).

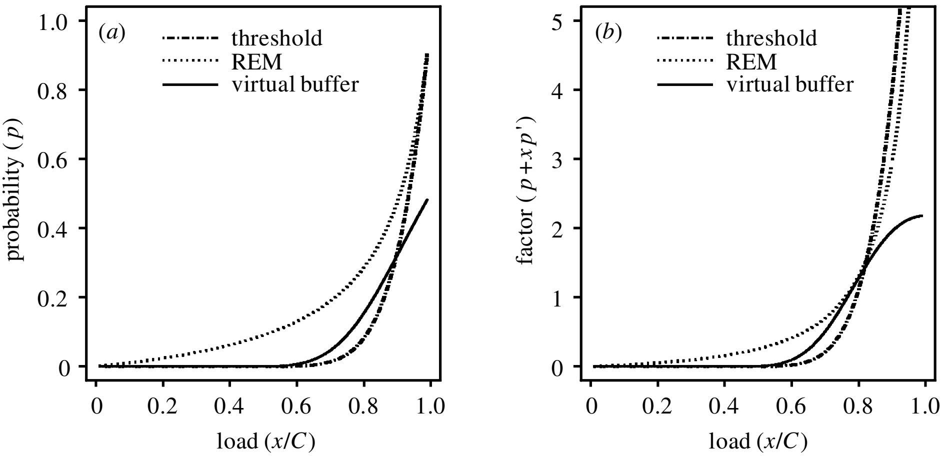

Figure 1. Marking probabilities (*a*) and stability factors (*b*) for the threshold, Random Early Marking and virtual buffer algorithms, for $B/\sigma^2 = 5, q = B^{-1}$.

Motivations for different marking strategies, and hence for different functions $p(x)$, are many and varied (in addition to the references above and the seminal paper of Ramakrishnan & Jain (1990), the interested reader should see Wischik (1999*b*)). But, in general, the higher the buffer level at which marking occurs, the greater the possibility of lag-induced oscillatory behaviour. An additional effect becomes apparent if the buffer level at which marking occurs is so high that the time taken by the queue to build up becomes comparable with round-trip times. In this case it becomes necessary to model instantaneous queue size as well as source rates in any dynamical representation. Bolot & Shankar (1990), Fendick *et al.* (1992) and Bonomi *et al.* (1995) have considered deterministic models of this form, establishing the local instability of the equilibrium point whatever gain parameters are used. Fendick *et al.* (1992) provide a careful analysis of the limit cycle behaviour, showing that the queue necessarily hits zero in each cycle.

The approach described in this paper, appropriate when marking occurs at much lower buffer levels, treats the stochastic effects present on the queueing time-scale as essential features of system behaviour, which are averaged out over round-trip times. The specific functions $p(x)$ used in this section help us understand the impact of threshold levels and traffic variability at the packet level, but their precise forms are not essential for the results of the previous section, where it is enough that $p(x)$ be smoothly increasing. Provided packet marking algorithms do not destabilize the feedback loops created by adaptive applications, the differential equations of the last section should represent the dynamical behaviour of the network on time-scales longer than a few round-trip times.

5. Distributed admission control

Gibbens & Kelly (1999*b*), Turányi & Westberg (1999) and Kelly *et al.* (2000) have described how packet marking at congested resources allows resource allocation decisions to be distributed to the edges of networks or to end-systems, and have developed various models for the resulting distributed admission control. In this section we consider a network carrying rate-adaptive real-time traffic, of the form discussed in § 3, and non-adaptive traffic which is subject to distributed admission control. Our aim is to explore the behaviour of competing aggregates of adaptive and non-adaptive traffic.

Suppose that routes $s \in S$ carry non-adaptive traffic, described as follows. Calls wishing to use route s arrive at the network in a stream of rate ν_s. When a call arrives a set of m_s probe packets is transmitted along route s, and the call is rejected and lost if any of these probe packets is marked. Otherwise the call is accepted and produces a flow of unit mean rate for a call holding period of unit mean. Let $y_s(t)$ represent the aggregate flow along route s. Consider the equations

$$\frac{\mathrm{d}}{\mathrm{d}t} y_s(t) = \nu_s \Big(1 - m_s \sum_{j \in s} \mu_j(t) \Big) - y_s(t), \tag{5.1}$$

for $s \in S$, where

$$\mu_j(t) = p_j \Big(\sum_{r : j \in r} x_r(t) + \sum_{s : j \in s} y_s(t) \Big), \tag{5.2}$$

for $j \in J$, together with equation (3.1) for $r \in R$. Equation (5.2) gives the proportion of packets marked by resource j, and this depends on the total of both adaptive and non-adaptive traffic. As before, equation (3.1) describes adaptive traffic on route r, although now x_r, w_r correspond to an aggregate of flows, as described by equations (3.4), (3.5). Equation (5.1), describing the aggregate non-adaptive traffic on route s, corresponds to an assumption that marking probabilities are small and approximately independent for each of the m_s probe packets of a call attempting route s. The equations treat the average flow along a route as evolving smoothly on a time-scale comparable with a call holding time, and correspond to a scaling regime where the number of calls carried on routes is large, a regime studied in detail by Zachary (2000).

A straightforward exercise in partial differentiation yields that the strictly concave function

$$\mathcal{U}(x, y) = \sum_{r \in R} w_r \log x_r + \sum_{s \in S} \frac{1}{m_s} \Big(y_s - \frac{y_s^2}{2\nu_s} \Big) - \sum_{j \in J} \int_0^{\sum_{r : j \in r} x_r + \sum_{s : j \in s} y_s} p_j(z)\, \mathrm{d}z \tag{5.3}$$

is a Lyapunov function for the system of differential equations (3.1), (5.1), (5.2); hence the unique value (x, y) maximizing $\mathcal{U}(x, y)$ is a stable point of the system, to which all trajectories converge. If calls have flow rates and holding times that depend on s, then the natural generalization of equation (5.1) again has an associated Lyapunov function of the form (5.3): the parameter ν_s is the product of the arrival rate, the flow rate and the mean call holding time for calls of type s, i.e. the offered load on route s measured in Erlangs.

Thus the system implicitly trades off the benefits of the two types of traffic, as represented by the first two terms of expression (5.3). For example, if the load w_r of adaptive traffic on route r gradually increases, then the stable point (x, y) will gradually shift, and in particular x_r will gradually increase, as expression (5.3) places more weight on the term $w_r \log x_r$. Or, if the offered load ν_s of non-adaptive traffic on route s gradually increases, then y_s will gradually increase.

In this section equations (3.1), (5.1), (5.2) are used to represent aggregates of traffic. At a finer level of detail, the number of calls on route s will fluctuate randomly, with y_s corresponding to a mean value (Kelly *et al.* 2000). The impact of non-adaptive traffic on the form of the functions $p_j, j \in J$, will be treated in the next section.

For file transfers, rather than rate-adaptive real-time traffic, different models are appropriate, since the transfer time and the average rate will be inversely proportional, rather than, as in the model above, independent. Gibbens & Kelly (1999*a*) discuss algorithms for file transfers similar to Jacobson's (1988) Slow Start algorithm, which rapidly increase the transfer rate in the absence of congestion, but back-off otherwise. Key & Massoulié (1999) develop a network model for mixtures of file transfers and rate-adaptive real-time traffic in which file transfers are either done at maximal speed or not at all.

6. Short transfers

At the start of a file transfer the congestion window maintained by TCP increases exponentially, doubling every round-trip time, until congestion is detected (the Slow Start algorithm of Jacobson (1988)), whereupon TCP switches to the Congestion Avoidance behaviour sketched in § 4. Short files may well have completed their transfer before congestion is detected. The performance of a short transfer could clearly be improved by increasing the initial congestion window size, and the rate of exponential growth could also be varied, either up or down. Crowcroft & Oechslin (1998) discuss how a different rate of growth could be implemented; Key & Massoulié (1999) develop a theoretical framework for the choice of the rate of growth, where the choice is made by a risk-averse end-system possessing a prior distribution for the marking probability. Similarly a choice of initial window might reasonably depend upon an end-system's prior distribution, based on past experience (over preceding hours or days) of transfers from a given location.

Feedback from the network, in the form of marked packets, would then affect the choice of initial window and rate of growth of a short transfers, but only on a long time-scale corresponding to the accumulation of information from prior experience of the network. On shorter time-scales, comparable with round-trip times, the effect of short transfers will be to place an uncontrolled and random background load upon the network. If this background load on a resource over a time-period τ is Gaussian, with mean $u\tau$ and variance $u\tau v^2$, then equation (4.4) for threshold marking should be replaced by

$$p(x) = \exp\left\{ \frac{-2B(C-x-u)}{x\sigma^2 + uv^2} \right\}, \tag{6.1}$$

and condition (4.3) for a non-oscillatory stable equilibrium is satisfied for any $w < C - u$ if

$$\kappa T\left(1 + \frac{2B(C-u)}{(C-u)\sigma^2 + uv^2}\right) < \frac{1}{\mathrm{e}}.$$

Alternatively, for the Random Early Marking algorithm, equation (4.7) becomes

$$p(x) = \frac{q(x\sigma^2 + uv^2)}{q(x\sigma^2 + uv^2) + 2(C-x-u)}$$

and condition (4.8) becomes

$$\kappa T\left(1+\frac{2(C-u)}{q((C-u)\sigma^2+uv^2)}\right)<\frac{1}{\mathrm{e}}.$$

Thus we observe two major effects of the background load: the condition $w < C$ for adequate capacity becomes the more onerous condition $u + w < C$; but, provided this condition is met, the further condition for the equilibrium to be stable and non-oscillatory is more easily met the *larger* the infinitesimal variance v^2 of the background load. Additional variability caused by short transfers lessens the sensitivity of the resource's load response, and this improves the stability of the feedback loops created by adaptive traffic.

7. Conclusion

A consequence of rapidly growing communications capacity may be that it becomes feasible to align supply and demand, and that queueing delays become small in comparison with propagation delays. In such circumstances, a simple packet network may be able to support an arbitrarily differentiated and constantly evolving set of services, by conveying information on incipient congestion to intelligent end-nodes which themselves determine what should be their demands on the packet network. This paper has outlined work on the stability of such a self-managed network, with illustrative results for the cases of adaptive and non-adaptive traffic, and short transfers.

I am grateful to the organizers of the meeting, and to the EPSRC for its support with computing facilities under grant no. GR/M09551.

References

Bolot, J.-C. & Shankar, A. U. 1990 Dynamic behavior of rate-based flow control mechanisms. *ACM Comp. Commun. Rev.* **20**, 35–49.

Bonomi, F., Mitra, D. & Seery, J. B. 1995 Adaptive algorithms for feedback-based flow control in high-speed wide-area networks. *IEEE J. Selected Areas Commun.* **13**, 1267–1283.

Briscoe, B., Rizzo, M., Tassel, J. & Damianakis, K. 2000 Lightweight policing and charging for packet networks. In *3rd IEEE Conf. on Open Architectures and Network Programming.* (See http://www.labs.bt.com/people/briscorj/.)

Courcoubetis, C., Siris, V. A. & Stamoulis, G. D. 1999 Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecom. Systems* **12**, 167–191.

Crowcroft, J. & Oechslin, P. 1998 Differentiated end-to-end Internet services using a weighted proportionally fair sharing TCP. *ACM Comp. Commun. Rev.* **28**, 53–67.

Fendick, K. W., Rodrigues, M. A. & Weiss, A. 1992 Analysis of a rate-based feedback control strategy for long-haul data transport. *Performance Eval.* **16**, 67–84.

Floyd, S. 1994 TCP and explicit congestion notification. *ACM Comp. Commun. Rev.* **24**, 10–23. (See http://www.aciri.org/floyd/ecn.html.)

Floyd, S. & Jacobson, V. 1993 Random Early Detection gateways for congestion avoidance. *IEEE/ACM Trans. Networking* **1**, 397–413. (See ftp://ftp.ee.lbl.gov/papers/early.pdf.)

Gibbens, R. J. & Kelly, F. P. 1999*a* Resource pricing and the evolution of congestion control. *Automatica* **35** 1969–1985. (See http://www.statslab.cam.ac.uk/~frank/evol.html.)

Gibbens, R. J. & Kelly, F. P. 1999*b* Distributed connection acceptance control for a connectionless network. In *Proc. 16th Int. Teletraffic Congr.* (ed. P. B. Key & D. G. Smith), pp. 941–952. Elsevier. (See http://www.statslab.cam.ac.uk/~frank/dcac.html.)

Gibbens, R. J. & Teh, Y. C. 1999 Critical time and space scales for statistical multiplexing in multiservice networks. In *Proc. 16th Int. Teletraffic Congr.* (ed. P. B. Key & D. G. Smith), pp. 87–96. Elsevier.

Golestani, S. J. & Bhattacharyya, S. 1998 A class of end-to-end congestion control algorithms for the Internet. In *Proc. 6th Int. Conf. on Network Protocols.* (See http://www.bell-labs.com/user/golestani/.)

Hale, J. 1977 *Theory of functional differential equations*. Springer.

Harrison, J. M. 1985 *Brownian motion and stochastic flow systems*. New York: Krieger.

International Telecommunications Union 1996 Recommendation I371: traffic control and congestion control in B-ISDN. Geneva.

Jacobson, V. 1988 Congestion avoidance and control. In *Proc. ACM SIGCOMM '88*, pp. 314–329. (See ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z.)

Johari, R. & Tan, D. K. H. 2000 End-to-end congestion control for the Internet: delays and stability. Statistical Laboratory Research Report 2000-2, University of Cambridge.

Kelly, F. P. 1997 Charging and rate control for elastic traffic. *Eur. Trans. Telecom.* **8**, 33–37. (See http://www.statslab.cam.ac.uk/~frank.)

Kelly, F. P. 2000 Mathematical modelling of the Internet. In *Proc. 4th Int. Congr. on Industrial and Applied Mathematics.*

Kelly, F. P., Maulloo, A. K. & Tan, D. K. H. 1998 Rate control in communication networks: shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.* **49**, 237–252.

Kelly, F. P., Key, P. B. & Zachary, S. 2000 Distributed admission control. *IEEE J. Selected Areas Commun.* **18**.

Key, P. 1999 Service differentiation: congestion pricing, brokers and bandwidth futures. In *Proc. 9th Int. Workshop on Network and Operating Systems Support for Digital Audio and Video.* (See http://www.nossdav.org.)

Key, P. & McAuley, D. 1999 Differential QoS and pricing in networks: where flow control meets game theory. *IEE Proc. Software* **146**, 39–43.

Key, P. & Massoulié, L. 1999 User policies in a network implementing congestion pricing. In *Workshop on Internet Service Quality Economics, MIT 1999.* (See http://research.microsoft.com/research/network/disgame.htm.)

Key, P., McAuley, D., Barham, P. & Laevens, K. 1999 Congestion pricing for congestion avoidance. Microsoft Research report MSR-TR-99-15. (See http://research.microsoft.com/pubs/.)

Kleinrock, L. 1964 *Communication nets: stochastic message flow and delay.* New York: McGraw-Hill.

Lapsley, D. E. & Low, S. H. 1999 Random Early Marking: an optimisation approach to Internet congestion control. In *Proc. IEEE ICON '99, Brisbane, Australia.*

Low, S. H. & Lapsley, D. E. 1999 Optimization flow control. I. Basic algorithm and convergence. *IEEE/ACM Trans. Networking.* (See http://www.ee.mu.oz.au/staff/slow/.)

MacKie-Mason, J. K. & Varian, H. R. 1994 Pricing the Internet. In *Public access to the Internet* (ed. B. Kahin & J. Keller). Englewood Cliffs, NJ: Prentice-Hall.

Norros, I. 1994 A storage model with self-similar input. *Queueing Systems* **16**, 387–396.

Ramakrishnan, K. K. & Jain, R. 1990 A binary feedback scheme for congestion avoidance in computer networks. *ACM Trans. Comp. Systems* **8**, 158–181.

Semret, N. & Lazar, A. A. 1999 Spot and derivative markets in admission control. In *Proc. 16th Int. Teletraffic Congr.* (ed. P. B. Key & D. G. Smith), pp. 757–766. Elsevier.

Songhurst, D. J. (ed) 1999 *Charging communication networks: from theory to practice.* Elsevier.

Tan, D. K. H. 1999 Mathematical models of rate control for communication networks. PhD thesis, University of Cambridge. (See http://www.statslab.cam.ac.uk/~dkht2/phd.html.)

Turányi, Z. R. & Westberg, L. 1999 Load control: lightweight provisioning of Internet resources. (See http://www.ericsson.co.hu/ethzrt/.)

Willinger, W., Taqqu, M. S. & Erramilli, A. 1996 A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. In *Stochastic networks: theory and applications* (ed. F. P. Kelly, S. Zachary & I. Ziedins), pp. 339–366. Royal Statistical Society Lecture Notes Series, vol. 4. Oxford University Press.

Wischik, D. 1999*a* The output of a switch, or, effective bandwidths for networks. *Queueing Systems* **32**, 383–396. (See http://www.statslab.cam.ac.uk/~djw1005.)

Wischik, D. 1999*b* How to mark fairly. In *Workshop on Internet Service Quality Economics, MIT 1999*.

Zachary, S. 2000 Dynamics of large uncontrolled loss networks. *J. Appl. Prob.* **37**. (See http://www.ma.hw.ac.uk/~stan/papers.) (In the press.)