Stability and fairness of explicit congestion control with small buffers

Frank Kelly
Statistical Laboratory
University of Cambridge
Cambridge, U.K.
fpk1@cam.ac.uk

Gaurav Raina
Statistical Laboratory
University of Cambridge
Cambridge, U.K.
G.Raina.99@cantab.net

Thomas Voice
Statistical Laboratory
University of Cambridge
Cambridge, U.K.
T.D.Voice.99@cantab.net

ABSTRACT

Rate control protocols that utilise explicit feedback from routers are able to achieve fast convergence to an equilibrium which approximates processor-sharing on a single bottleneck link, and hence such protocols allow short flows to complete quickly. For a network, however, processor-sharing is not uniquely defined but corresponds with a choice of fairness criteria, and proportional fairness has a reasonable claim to be the network generalization of processor-sharing.

In this paper, we develop a variant of RCP (rate control protocol) that achieves α-fairness when buffers are small, including proportional fairness as the case $\alpha = 1$. At the level of theoretical abstraction treated, our model incorporates a general network topology, and heterogeneous propagation delays. For our variant of the RCP algorithm, we establish a simple decentralized sufficient condition for local stability.

An outstanding question for explicit congestion control is whether the presence of feedback based on queue size is helpful or not, given the presence of feedback based on rate mismatch. We show that, for the variant of RCP considered here, feedback based on queue size may cause the queue to be *less* accurately controlled.

A further outstanding question for explicit congestion control is the scale of the step-change in rate that is necessary at a resource to accommodate a new flow. We show that, for the variant of RCP considered here, this can be estimated from the aggregate flow through the resource, without knowledge of individual flow rates.

Categories and Subject Descriptors

C.2.2 [**Computer-Communication Networks**]: Network Protocols—*congestion control*

General Terms

Modeling of communication networks, Theory

Keywords

Internet, rate control

1. INTRODUCTION

There is currently considerable interest in explicit congestion control protocols, which use a field in each packet to convey relatively precise information on congestion from resources to endpoints. These protocols contrast with TCP and its various enhancements, where endpoints implicitly estimate congestion from noisy information, essentially the single bit of feedback provided by a dropped or marked packet. Examples of explicit congestion control protocols include XCP [10] and RCP [6].

We shall discuss in detail the rate control protocol RCP, which uses explicit feedback from routers to allow fast convergence to an equilibrium where a bottleneck link is shared equally over flows [6]. RCP approximates the processor-sharing queueing discipline when there is a single bottleneck link, and hence allows short flows to complete quickly [4, 5]. For the processor-sharing discipline at a single bottleneck link, the mean time to transfer a file is proportional to the size of the file, and is insensitive to the distribution of file sizes [2, 5].

For a network with multiple constrained resources there exist several possible generalizations of the processor-sharing discipline. A conveniently parameterized family, that of the α-fair rate allocations, was introduced in [15]. The parameter α lies in the range $(0, \infty)$, and the cases $\alpha \to 0$, $\alpha = 1$ and $\alpha \to \infty$ correspond respectively to an allocation which achieves maximum throughput, is *proportionally fair* or is *max-min fair* [15, 21].

Max-min is the fairness criterion commonly envisaged in connection with RCP, but it is not the only possibility. Proportional fairness, in particular, has a claim to be the natural network generalization of processor-sharing, with a growing literature showing that it has exact or approximate insensitivity properties [14, 20] and important efficiency and robustness properties [3, 12].

Buffer sizing is an important issue for RCP, as for other protocols. If links are run close to capacity, then buffers need to be large, so that new flows can be given a high starting rate. But if links are run with some spare capacity, then this may be sufficient to cope with new flows, and allow buffers to be small.

In previous work the local stability of RCP has been studied for a single resource with a large buffer [1] and for a network with small buffers under the max-min criterion [25]. In this paper we model a variant of RCP which achieves α-fairness when buffers are small, and we establish, in Section 3, a simple decentralized sufficient condition for the local stability of this algorithm. In Sections 4 and 5 we show that, with small buffers, feedback based on queue size has a major impact on equilibrium utilization. In Section 6 we study whether the presence of feedback based on queue size is helpful or not, given the presence of feedback based on rate mismatch: we show that it may cause the queue to be *less* accurately controlled.

Finally, under weighted proportional fairness, we explore the scale of the step-change in rate necessary at a resource to accommodate a new flow. We show that it can be estimated from the aggregate flow through the resource, and without knowledge of individual flow rates. In Section 7 we illustrate the effectiveness of a step-change algorithm, which we introduce, under various traffic scenarios.

2. MODEL OF RCP

RCP updates its estimate of a fair rate through a single bottleneck link from observations of the spare capacity at the link and the queue size, as described by the equation [1, 4, 5]

$$\frac{d}{dt}R(t) = \frac{R(t)}{C\overline{T}}\left(a(C - y(t)) - \beta\frac{q(t)}{\overline{T}}\right) \tag{1}$$

where

$$y(t) = \sum_s R(t - T_s) \tag{2}$$

and

$$\begin{aligned}\frac{d}{dt}q(t) &= [y(t) - C] \qquad q(t) > 0\\ &= [y(t) - C]^+ \qquad q(t) = 0,\end{aligned} \tag{3}$$

using the notation $x^+ = \max(0, x)$. Here $R(t)$ is the rate being updated by the router and advertised to endpoints, C is the link capacity, $y(t)$ is aggregate load at the link, $q(t)$ is the queue size, T_s is the round-trip time of flow s, and $\overline{T}$ is the average round-trip time, over the flows present.

The key relation (1) contains two forms of feedback - a term based on the rate mismatch $C - y(t)$, and a term based on the instantaneous queue size $q(t)$.

Sufficient conditions for local stability of the system (1-3) about its equilibrium point were derived in [1], using results for a switched linear control system with a time delay. The analysis explicitly models the discontinuity in system dynamics that occurs as the queue becomes empty. The sufficient conditions, on the non-negative dimensionless constants a and β, take the form

$$a < \frac{\pi}{2} \tag{4}$$

and $\beta < f(a)$ where f is a positive function that depends on $\overline{T}$.

3. SMALL BUFFER MODEL

With small buffers and large rates the queue size fluctuations are very fast – so fast that it is impossible to control the queue size (Figure 1 is an illustration we shall discuss in Section 5). Instead, as described in [18, 27], protocols act to control the *distribution* of queue size. On the time-scale relevant for convergence of the system (1-3), it is then the *mean* queue size that is important.

This produces a simplification of the key relation (1): the instantaneous queue size, $q(t)$, can be replaced by its mean. This simplification of the treatment of queue size allows us to obtain a model that remains tractable even for a general network topology.

Next we describe our network model of RCP with small buffers. It is a multiple resource generalization of the system (1-3), under the simplification of the queueing term described above.

We shall consider a network with a set J of *resources*. A *route* r will be identified with a non-empty subset of J, and we shall write $j \in r$ to indicate that route r passes through resource j. Let R be the set of possible routes.

For each j, r such that $j \in r$, let T_{rj} be the propagation delay from the source of flow on route r to the resource j, and let T_{jr} be the return delay from resource j to the source. Then

$$T_{rj} + T_{jr} = T_r \quad j \in r, r \in R, \tag{5}$$

where T_r is the round-trip propagation delay on route r: the identity (5) is a direct consequence of the end-to-end nature of the signalling mechanism, whereby congestion on a route is conveyed via a field in the packets to the destination, which then informs the source. We assume queueing delays are negligible in comparison with propagation delays - this is consistent with our assumption of small buffers.

Our small buffer RCP variant is modelled by the system of differential equations

$$\frac{d}{dt}R_j(t) = \frac{aR_j(t)}{C_j\overline{T}_j(t)}\left(C_j - y_j(t) - b_jC_jp_j(y_j(t))\right) \tag{6}$$

where

$$y_j(t) = \sum_{r:j\in r} x_r(t - T_{rj}) \tag{7}$$

is the aggregate load at link j, $p_j(y_j)$ is the mean queue size at link j when the load there is y_j, and

$$\overline{T}_j(t) = \frac{\sum_{r:j\in r} x_r(t)T_r}{\sum_{r:j\in r} x_r(t)} \tag{8}$$

is the average round-trip time of packets passing through resource j. We suppose the flow rate x_r is given by

$$x_r(t) = \left(\sum_{j\in r} R_j(t - T_{jr})^{-\alpha}\right)^{-1/\alpha}. \tag{9}$$

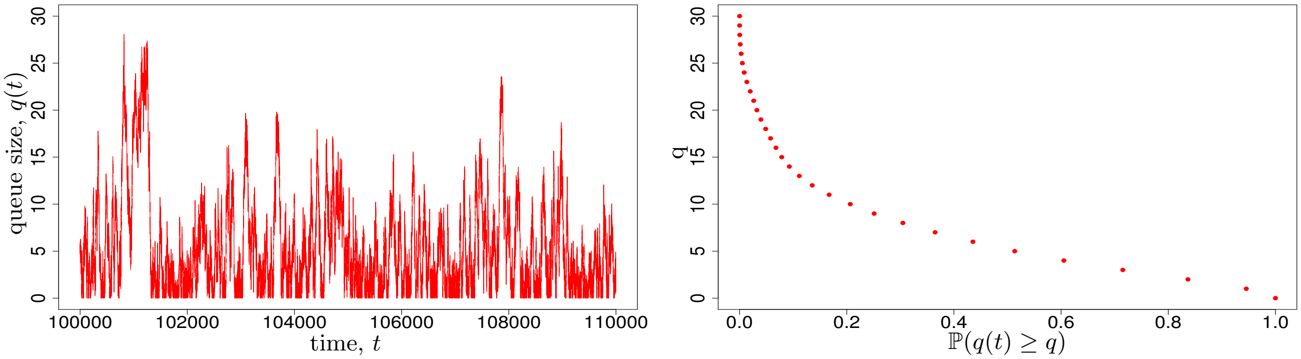
Observe that $R_j(t - T_{jr})$ gives the flow rate on a route r which passes through resource j alone. Observe also, as $\alpha \to \infty$, the expression (9) approaches $\min_{j\in r}(R_j(t - T_{jr}))$, corresponding to max-min fairness. In general, the flows at equilibrium will be α-fair [11, 15], as we describe in Appendix II.

Note that for bounded values of α the computation (9) can be performed as follows. If a packet is served by link j at time t, $R_j(t)^{-\alpha}$ is added to the field in the packet containing the indication of congestion. When an acknowledgement is returned to its source, the acknowledgement feedbacks the sum, and the source sets its flow rate equal to the returning feedback to the power of $-1/\alpha$.

A simple approximation for the mean queue size is as follows. Suppose that the workload arriving at resource j over a time period τ is Gaussian, with mean $y_j\tau$ and variance $y_j\tau\sigma_j^2$. Then the workload present at the queue is a reflected Brownian motion [8], with mean under its stationary distribution of

$$p_j(y_j) = \frac{y_j\sigma_j^2}{2(C_j - y_j)}. \tag{10}$$

The parameter σ_j^2 represents the variability of link j's traffic at a packet level. Its units depend on how the queue size is measured: for example, packets if packets are of constant size, or Kilobits otherwise.



(a) Evolution of the queue over a single round-trip time.

(b) Empirical distribution of queue size within one round-trip time.

Figure 1: Illustration of the small buffer variant of RCP.

At the equilibrium point $y = (y_j, j \in J)$ for the dynamical system (6-10) we have

$$C_j - y_j = b_j C_j p_j(y_j). \tag{11}$$

From equations (10-11) it follows that at the equilibrium point

$$p'_j(y_j) = \frac{1}{b_j y_j}, \tag{12}$$

a relation that will be key to our analysis of stability.

In Appendix II we establish, using an important early result of Vinnicombe [22], that a sufficient condition for the dynamical system (6-10) to be locally stable about its equilibrium point is that

$$a < \frac{\pi}{4}. \tag{13}$$

It is noteworthy that this simple decentralized sufficient condition places *no* restriction on the parameters $b_j, j \in J$, provided our modelling assumption of small buffers is satisfied.

4. DISCUSSION

The parameter a is the same as in the original model (1) of RCP. But the parameter b_j is a rescaled version of β,

$$b_j = \frac{\beta}{a C_j \overline{T}_j}, \tag{14}$$

and its units are the reciprocal of the units in which the queue size is measured.

The parameter a controls the speed of convergence at each resource, while the parameter b_j controls the utilization of resource j at the equilibrium point. From (10-11) we can deduce that the utilization of resource j is

$$\rho_j \equiv \frac{y_j}{C_j} = 1 - \sigma_j \left(\frac{b_j}{2} \cdot \frac{y_j}{C_j} \right)^{1/2}$$

and hence that

$$\begin{aligned} \rho_j &= \left(\left(1 + \frac{\sigma_j^2 b_j}{8} \right)^{1/2} - \left(\frac{\sigma_j^2 b_j}{8} \right)^{1/2} \right)^2 \\ &= 1 - \sigma_j \left(\frac{b_j}{2} \right)^{1/2} + O(\sigma_j^2 b_j). \end{aligned} \tag{15}$$

For example, if $\sigma_j = 1$, corresponding to Poisson arrivals of packets of constant size, then a value of $b_j = 0.022$ produces a utilization of 90%. Figure 2 plots the function (15), under the label 'Gaussian analysis' and shows how utilization decreases as b_j increases.

It is important to note that setting the parameter b_j to control utilization produces a very different scaling for β from that of [1], as a consequence of the presence of the bandwidth-delay product $C_j \overline{T}_j$ in the relation (14). In particular, if the bandwidth-delay product $C_j \overline{T}_j$ is large, then the values we consider for β are much larger than those considered in [1].

If the parameters b_j are all set to zero, and the algorithm uses as C_j not the actual capacity of resource j, but instead a target, or virtual, capacity of say 90% of the actual capacity, then this too will achieve an equilibrium utilization of 90%. In this case the equivalent sufficient condition for local stability is

$$a < \frac{\pi}{2} \tag{16}$$

(cf. [23], [11] Section 6.2, [24]). An outstanding question, to be considered more fully in Section 6, is whether it is advantageous to set the parameters b_j to be positive, or equivalently whether to include a queueing term in the definition of the protocol. We note here that although the presence of a queueing term is associated with a smaller choice for the parameter a – note the factor two difference between conditions (13) and (16) – nevertheless, close to the equilibrium the local responsiveness is comparable, since the queueing term contributes roughly the same feedback as the term measuring rate mismatch. Below equilibrium, the $b = 0$ case is more responsive (up to a factor of 2); above equilibrium, the $b > 0$ case is more responsive (how much more responsive depends on the buffer size).

We remark that, in the taxonomy of [11], we are considering fair dual algorithms rather than delay-based dual algorithms [13, 16], and this is important for the form of the sufficient conditions (13) and (16).

5. ILLUSTRATION

Next we illustrate our small buffer variant of the RCP algorithm with a simple packet level simulation. The network simulated has a single resource, of capacity one packet per unit time [1], and 100 sources that each produce Poisson traffic. Poisson traffic is simulated by randomly drawing the interval time between packet transmission from an exponential distribution of parameter equal to flow rate. The round-trip time is 10000 units of time. Assuming a packet size of 1000 bytes, this would translate into a service rate of 100Mbytes/s, and a round-trip time of 100ms, or a service rate of 1 Gbyte/s and a round-trip time of 10ms. The RCP parameters take the values $a = 0.5$ and $\beta = 100$. Thus $b = \beta/(aCT) = 0.02$ packets. Further details of the simulation are given in Appendix I.

Figure 1(a) shows the evolution of the queue size in one round-trip time. Note that the queue size fluctuates rapidly within a round-trip time, frequently reflecting from zero. Figure 1(b) shows the empirical distribution of the queue size over the same single round-trip time - it is calculated from the sample path shown in Figure 1(a).

Figure 2 plots the utilization observed in the simulations for the case where 100 sources each send Poisson traffic, under the label '100 Poisson sources'. For comparison we also plot the relation (15) obtained from our earlier analysis with $\sigma = 1$, labelled 'Gaussian analysis'. Two features of the simulated results are notable. First, the variability of the utilization, measured over one round-trip time. This is to be expected, since there remains variability in the empirical distribution of queue size, Figure 1(b). This source of variability decreases as the bandwidth-delay product CT increases. Second, apart from this variability, the utilization is rather well represented by relation (15). Further simulations, not described here, show the match become closer and closer as the bandwidth-delay product CT increases.

Our differential equations describe the system behaviour at the macroscopic level, where flows are described by rates. At the packet, or microscopic level, there is choice on how the sources may regulate their flow, in response to the feedback that they get from the network. Sources that send approximately Poisson traffic might be expected to lend themselves especially well to our approach, since the superposition of independent Poisson streams is a Poisson stream, and the number of streams superimposed does not affect the statistical characteristics of the superposition other than through the rate, which we model explicitly. Furthermore, for a constant rate Poisson arrival stream of constant size packets, i.e. an $M/D/1$ queue, the exact mean queue size is known, and indeed matches relation (15) with $\sigma = 1$ [19]. Thus the rather good match between the utilization and relation (15) is to be expected for Poisson sources.

Next we illustrate an example where each source sends a near periodic stream of traffic, with period the inverse of the source's rate. Figure 2 plots the utilization observed in the simulations under the label '100 periodic sources'. The simulated data show variability, as we expect, but now lie *above* the Gaussian analysis. Again an exact analysis of a special case is able to provide insight. A superposition of periodic streams produces queueing behaviour which has been studied extensively [7, 19]. The ND/D/1 queue, as it is termed, locks into a repeating pattern of busy periods. Over time intervals small in comparison with the period of a source, the queueing behaviour induced is comparable with that induced by a Poisson stream. But over longer periods the arrival pattern has less variability than a Poisson stream. This will lead to a lower expected queue size and hence a higher utilization for any given value of b.

We have simulated periodic sources through a single congested resource since this seems likely to be an extreme case, but a fuller exploration of the accuracy and robustness of our model is not attempted here.

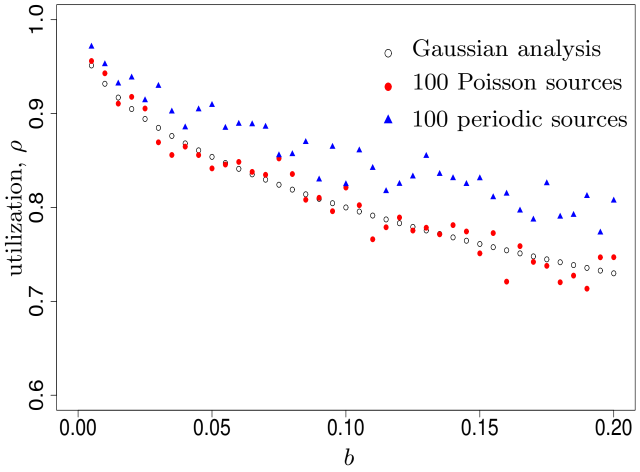


Figure 2: Utilization, ρ, measured over one round-trip time, for different values of the parameter b, with 100 RCP sources sending either Poisson, or periodic, traffic.

6. IS QUEUE FEEDBACK HELPFUL?

In this Section we address the question raised in Section 4: should we include feedback based on queue size, or should we instead set all the parameters b_j to zero?

We first describe our simulation set up. The network simulated has a single resource of capacity one packet per unit time and 100 sources that each produce Poisson traffic. The round-trip times that are simulated are in the range of 100 to $100,000$ units of time. In Section 4 it was highlighted that by removing feedback based on queue size, we can double the value of the parameter a in the sufficient condition for local stability. So, in all our simulations, when we included feedback based on queue size we set $a = 0.5$. When the queue term was excluded from the feedback, i.e. $b = 0$, we set $a = 1$ and replaced C with γC for some $\gamma < 1$. The simulations were started close to equilibrium.

Figures 3 and 4 show the comparison between theory and the simulation results, when the round-trip times are in the range of $1,000$ to $100,000$ units of time. Figure 3 represents the case where the queue term was present in the RCP definition. In Figure 4, where the queue term is absent, we replace C with γC for $\gamma \in [0.7, \cdots, 0.90]$ in the protocol definition.

We first note that when the round-trip time is in the region of $100,000$ there is excellent agreement between theory and simulations in both Figures 3 and 4. So, in this regime,

[1] At the resource the buffer size was 30 packets, and no packets were lost in our simulations. The buffer size would be important for behaviour away from equilibrium.

based on local stability analysis we are unable to distinguish between the two different design choices. This provides motivation for analysis which goes beyond local stability. The reader is referred to [26] which analyses some nonlinear properties of the RCP dynamical system, with and without the queue term, in a single resource setting where the conclusions tend to favour a system where the queue term is absent.

In both Figures 3 and 4 note that as one reduces the round-trip time from $100,000$ to $1,000$ time units we observe greater variability in utilization. If one reduces the round-trip time further, say down to 100 time units, then queueing delays can start to become comparable to physical transmission delays. In such a regime our small buffer assumption - that queueing delays are negligible in comparison to propagation delays - breaks down. This is a regime where, in control theoretic parlance, the queue is acting as an integrator on approximately the same time scale as the round-trip time of a congestion control algorithm. Models aiming to capture this regime have been analysed previously in the literature: for example, for RCP see [1] and for TCP see [9, 17]; all of whom employ different styles of analysis from each other.

We resort to simulations to develop our understanding of this regime with our variant of RCP. To achieve 90% utilization in our small buffer model we need to set $b = 0.02$. Now recall the relationship between b, the small buffer rescaled parameter, and the original RCP model parameter β: $baCT = \beta$. So $a = 0.5$, $C = 1$, $T = 100$ and $b = 0.02$ yields $\beta = 1$. Stability charts in [1] suggest that the choice $\beta = 1$ and $a = 0.5$ lies outside their provably safe stability region for a large range of round-trip times. And indeed we observed deterministic instabilities in our simulations: see Figure 5(a). To aim for a fixed utilization we can also set $b = 0$ and target a virtual capacity; say 90% of the actual capacity. Without the queue term in the RCP definition, the congestion controller is reacting only to rate mismatch, and with a round-trip time of 100 time units we did not observe any deterministic instabilities: see Figure 5(b). In this regime, the presence of the queue term in the definition of the RCP protocol causes the queue to be *less* accurately controlled.

All the previous experiments were conducted in a static scenario: fixed number of long lived flows, sending traffic, in equilibrium. We now describe a more dynamic setting. Consider a link, targeting 90% utilization with 100 flows and a round-trip time of 1000 time units, which suddenly has a 20% increase in load. As motivation, consider the failure of a parallel link with similar characteristics where 20% of the load is instantaneously transferred to the link under consideration.

We explore this scenario via a simulation. For this experiment, see Figure 6 for the evolution of the queue and rate for the cases with and without feedback based on queue size. The scenario when the queue size is included in the feedback has no clear advantage: the queue appears to have periodic spikes, and the rate seems to remain in a quasi-periodic state, even after 20 round-trip times. A full and comprehensive comparison would study behaviour away from equilibrium, and this is not attempted here.

This section leads us to conclude that, for our small buffer variant of RCP, there is no clear case that feedback based on queue size is helpful and some evidence that it is harmful. Whether or not alternative queue statistics, such as a

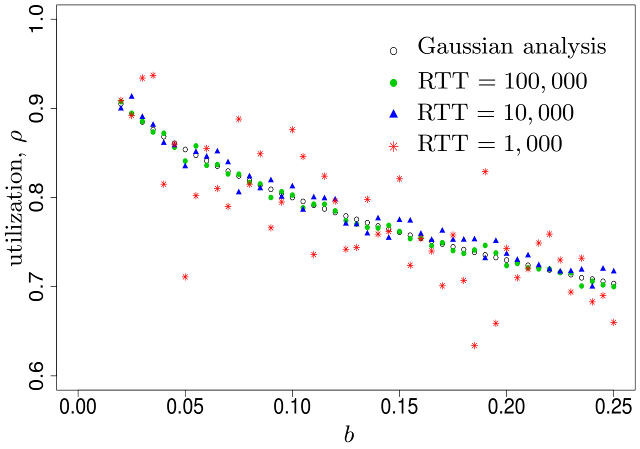


Figure 3: Utilization, ρ, measured over one round-trip time, for different values of the parameter b with 100 RCP sources sending Poisson traffic.

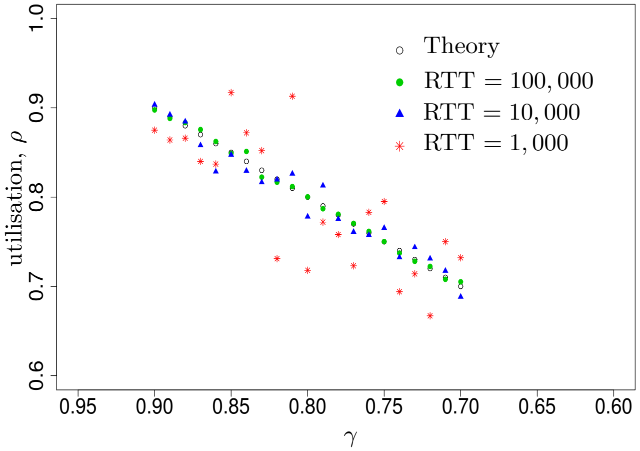


Figure 4: Utilization, ρ, measured over one round-trip time, for different values of the parameter γ with 100 RCP sources sending Poisson traffic.

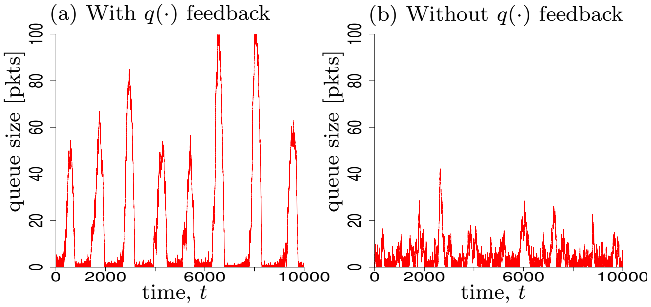
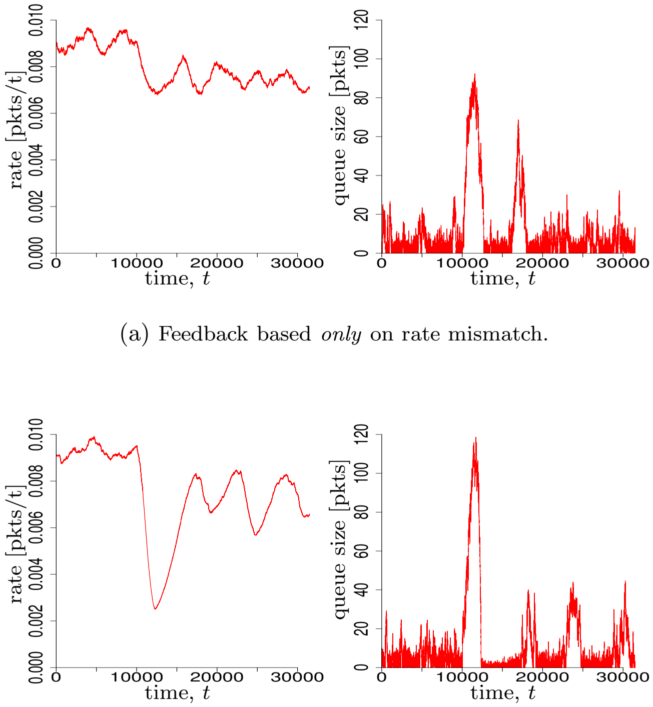


Figure 5: Traces from a packet-level simulation of a single bottleneck link with 100 RCP sources, round-trip time of 100, and a target link utilization of 90%.



(a) Feedback based *only* on rate mismatch.

(b) Feedback based on rate mismatch *and* queue size.

Figure 6: Illustrating the impact of a 20% increase in load after a 100 RCP flows, with a round-trip time of 1000, are in equilibrium.

smoothed estimate of average queue size, could improve algorithm performance remains an open question.

7. NEW FLOWS

When a new flow starts, it learns, after one round-trip time, of its starting rate. In this section we explore what should be the impact on a resource's estimate of R_j when it learns of a new flow about to start.

7.1 Step-change algorithm

We consider the case where the flow rate is set to

$$x_r(t) = w_r \left(\sum_{j \in r} R_j(t - T_{jr})^{-1} \right)^{-1} \tag{17}$$

which will produce weighted proportional fairness [11] at equilibrium, with weight w_r for flow r. Condition (13) remains sufficient for local stability (Appendix II).

In equilibrium, the aggregate flow through resource j is y_j, the unique value such that the right hand side of (6) is zero. When a new flow, r, begins transmitting, if $j \in r$, this will disrupt the equilibrium by increasing y_j to $y_j + x_r$. Thus, in order to maintain equilibrium, whenever a flow, r, begins R_j needs to be decreased, for all j with $j \in r$.

According to (7)

$$y_j = \sum_{r:j \in r} w_r \left(\sum_{k \in r} R_k^{-1} \right)^{-1}$$

and so the sensitivity of y_j to changes in the rate R_j is readily deduced to be

$$\frac{\partial y_j}{\partial R_j} = \frac{y_j \overline{x}_j}{R_j^2} \tag{18}$$

where

$$\overline{x}_j = \frac{\sum_{r:j \in r} x_r \left(\sum_{k \in r} R_k^{-1} \right)^{-1}}{\sum_{r:j \in r} x_r}.$$

This $\overline{x}_j$ is the average, over all packets passing through resource j, of the unweighted fair share on the route of a packet.

Suppose now that when a new flow r, of weight w_r, begins, it sends a request packet through each resource j on its route, and suppose each resource j, on observation of this packet, immediately makes a step-change in R_j to a new value

$$R_j^{new} = R_j \cdot \frac{y_j}{y_j + w_r R_j}. \tag{19}$$

The purpose of the reduction is to make room at the resource for the new flow. Although a step-change in R_j will take time to work through the network, the scale of the change anticipated in traffic from existing flows can be estimated from (18) and (19) as

$$(R_j - R_j^{new}) \cdot \frac{\partial y_j}{\partial R_j} = w_r \overline{x}_j \cdot \frac{y_j}{y_j + w_r R_j}.$$

Thus the reduction aimed for from existing flows is of the right scale to allow one extra flow at the average of the w_r-weighted fair share through resource j. Note that this is achieved without knowledge at the resource of the individual flow rates through it, $(x_r, r : j \in r)$: only knowledge of their equilibrium aggregate y_j is used in expression (19), and y_j may be determined from the parameters C_j and b_j as in (11). If the new flow r has a large target weight it could be initialized via a sequence of increments in w_r. Each increment could then be advertised to the resources which then react as though it was the request of a new flow, with weight equal to that increase. For example, for flows with integer value target weights a new flow could be initialized via a series of increments at the rate of size 1 per round-trip time.

7.2 Impact of a sudden change

In this subsection we briefly analyse the robustness of the admission control process based on the above step-change algorithm against large, and sudden, increases in the number of flows.

Consider the case where the network consists of a single link j with equilibrium flow rate y_j. If there are n identical flows, then at equilibrium $R_j = y_j/n$. When a new flow begins, the step-change (19) is performed and R_j becomes $R_j^{new} = y_j/(n+1)$. Thus, equilibrium is maintained. Now suppose that m new flows begin at the same time. Once the m flows have begun, R_j should approach $y_j/(n+m)$. However, each new flow's request for bandwidth will be received one at a time. Thus, the new flows will be given rates $y_j/(n+1), y_j/(n+2), \ldots, y_j/(n+m)$. So, when the new flows start transmitting, after one round-trip time, the new aggregate rate through j, y_j^{new} will approximately be

$$y_j^{new} \approx n \frac{y_j}{n+m} + \int_n^{n+m} \frac{y_j}{u} du.$$

If we let $\epsilon = m/n$, we have

$$y_j^{new} \approx y_j \left(\frac{1}{1+\epsilon} + \log(1+\epsilon) \right). \qquad (20)$$

Thus, for the admission control process to be able to cope when the load is increased by a proportion ϵ, we simply require y_j^{new} to be less than the capacity of link j. Direct calculation shows that if the equilibrium value of y_j is equal to 90% of capacity, then (20) allows an increase in the number of flows of up to 66%. Furthermore, if at equilibrium y_j is equal to 80% of capacity, then the increase in the number of flows can be as high as 120% without y_j^{new} exceeding the capacity of the link.

Although the above analysis and discussion concerns a single link, it does provide a simple rule of thumb guideline for choosing parameters such as b_j or C_j. If one takes ϵ to be the largest plausible increase in load that the network should be able to withstand, then from (20), one can calculate the value of y_j which gives y_j^{new} equal to capacity. This value of y_j can then be used to choose b_j or C_j, using the equilibrium relationship $C_j - y_j = b_j C_j p_j(y_j)$.

7.3 Illustration

We first recapitulate the processes involved in admitting a new flow into an RCP network. A new flow first transmits a request packet through the network. The links, on detecting the arrival of the request packet, perform the step-change algorithm to make room at the respective resources for the new flow. After one round-trip time the source of the flow receives back acknowledgement of the request packet, and starts transmitting at the rate that is conveyed back. This procedure allows a new flow to reach near equilibrium within *one* round-trip time. We now illustrate, via some simulations, the admission control procedure for dealing with newly arriving flows.

7.3.1 Some experiments on a toy network

Consider a toy network, depicted in Figure 7, consisting of 5 links labelled A, B, C, D and X where the links have a capacity of $1, 10, 1, 10$ and 20 packets per unit time, respectively. The physical transmission delays on links A, B and X are 100 time units and on links C and D are 1000 time units. In our illustration we do not include feedback based on queue size in the RCP definition. The target utilization is 90% for each of the links. In our experiments, links A, B, C and D each start with 20 flows operating in equilibrium. Each flow uses link X and one of links A, B, C or D.

In the first scenario we have a 50% increase in flows, i.e. on each of the links A, B, C and D we have 10 new flows that arrive and request to enter the network. So, for example, a request packet originating from flows entering link A, would first go through link A and then link X before returning back to the source. We then repeat the scenario with a 100% increase in flows.

In Figure 8 the necessary step-change required to accommodate the new flows is clearly visible on link C. Also, at approximately 1100 time units after the step-change in rate we observe a spike in the evolution of the queue in link C: 1100 time units is, of course, the sum of the physical propagation delays along links C and X. We can clearly observe two distinct step-changes on link X: first reacting to the flows that originated from links A and B, and then reacting to the flows that started from links C and D.

Figure 9 shows the scenario when we have a 100% increase in flows. The step-change in rate, and then the spike in evolution of the queue, are both again visible; this time, as expected, they are more pronounced.

The above scenarios are limited, but, with the analysis of Section 7.2, they do illustrate the effectiveness of the step-change algorithm introduced in Section 7.1.

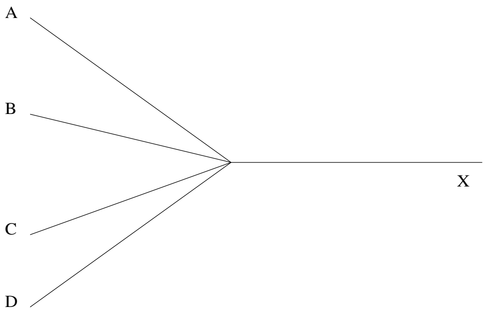
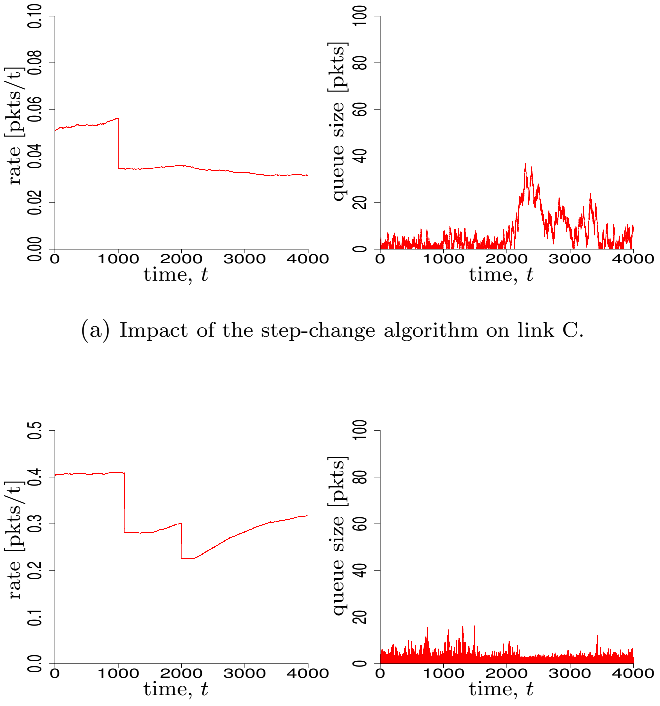


Figure 7: Toy network used, in packet-level simulations, to illustrate the process of admitting new flows into a RCP network.



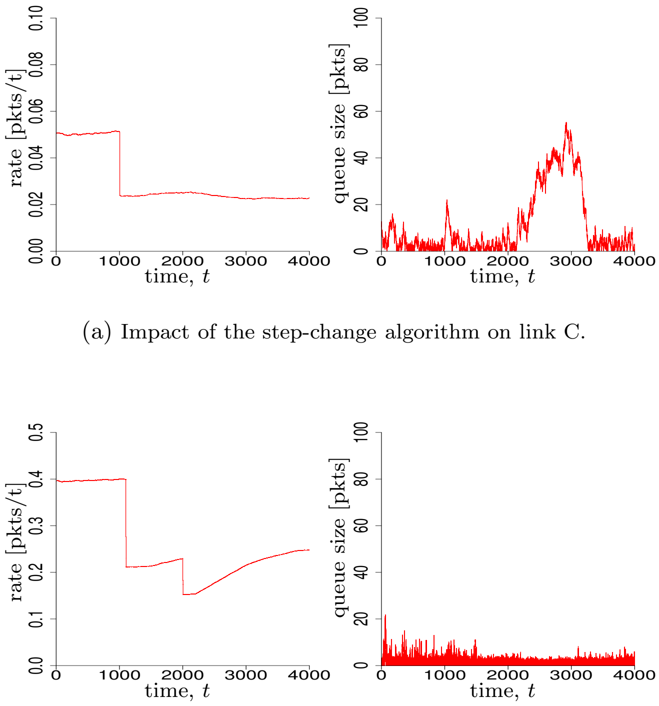
(a) Impact of the step-change algorithm on link C.

(b) Impact of the step-change algorithm on Link X.

Figure 8: Illustration of a scenario with a 50% increase in flows which instantaneously request to be admitted into the network depicted in Figure 7.

8. CONTRIBUTIONS

In this paper we used a fluid-flow model to analyse the local stability of a variant of RCP that achieves α-fairness



(a) Impact of the step-change algorithm on link C.

(b) Impact of the step-change algorithm on Link X.

Figure 9: Illustration of a scenario with a 100% increase in flows which instantaneously request to be admitted into the network depicted in Figure 7.

when buffers are small. We exhibited a simple decentralized sufficient condition for local stability of the algorithm, where our small buffer model incorporates a general network topology with heterogeneous propagation delays. Additionally, we show that feedback based on queue size may not be helpful, given feedback based on rate mismatch.

In the case of weighted proportional fairness, we also specify the scale of the step-change in rate necessary at a resource to accommodate a new flow, and show that it can be estimated from the aggregate flow through the resource, and without knowledge of individual flow rates.

Packet-level simulations served to illustrate some of the analysis in this paper.

9. APPENDIX I: SIMULATION

The figures bearing observations or traces from packet-level simulations were produced using a discrete event simulator of packet flows in RCP networks. We outline the simulation process for a single link which would generalize in a natural way, as outlined in the paper, for a network with multiple resources.

The links are modelled as FIFO queues, with internal feedback variables which evolve according to a discrete approximation of (1). The sources are modelled either as N time-varying Poisson sources or N periodic sources.

The link has an internal variable, $R(t)$, which represents the fair rate through the link for a flow unconstrained elsewhere. If a packet arrives or leaves a link at time t, and the previous time such an event occurred was $t - \delta t$, then $R(t)$ updates according to

$$\begin{aligned}\log(R(t)) &= \log(R(t-\delta t)) \\ &+ \frac{1}{CT}\left(a(C\delta t - I(t-\delta t,t)) - \beta\frac{q(t-)}{T}\delta t\right)\end{aligned}$$

where a, β are positive constants, C is the capacity of the link, T is the common round-trip time, $q(t-)$ is the queue size immediately before the event at time t and $I(t-\delta t, t)$ is the number of packet arrivals in the interval $[t-\delta t, t)$. The queue size is not necessarily integral - a partially served packet contributes only its remaining service time; $q(t-)$, so defined, is often termed the *virtual waiting time* [19].

This is our discrete approximation to the differential equation (1). The discrete approximation also reduces to equation (6) if we identify $p(y)$ with the mean value of $q(t)$, and relate b and β by equation (14).

Following (9), if a packet is served by a link at time t, $R(t)^{-\alpha}$ is added to that packet's congestion feedback variable. When an acknowledgement is returned to its source, the source sets its flow rate equal to the returning feedback to the power of $-1/\alpha$. When the RCP sources are Poisson, the remaining time until next packet transmission is simply recalculated as an exponential random variable with parameter equal to the new flow rate. For a network with a single resource, this corresponds to each source sending a Poisson stream at the latest rate $R(t)$ to be received from the link. When an RCP source is periodic, it sends a stream of packets with period $R(t)^{-1}$.

The observations plotted in Figures 1-4 were obtained over one round-trip time, after the simulation had been running for ten round-trip times starting from near equilibrium. The traces plotted in Figures 5-6 were for a network with a single resource. The sample traces plotted in Figures 8-9 illustrate, on the toy network depicted in Figure 7, the effectiveness of the step-change algorithm in admitting new flows into a network, in a few scenarios.

10. APPENDIX II: STABILITY

In this appendix we shall derive conditions for the local stability of the system of delayed differential equations (6-8), (10), (17). We shall assume that the $|J| \times |R|$ connectivity matrix A, which has entry $A_{jr} = 1$ if $j \in r$ and $A_{jr} = 0$ otherwise, has full row rank. This is a common, and weak, assumption [11, 21].

First we establish that the equations (6-8), (17) have a unique equilibrium. We shall assume that $p_j(\cdot)$ is an increasing function, for $j \in J$, as it is for the special case (10): hence there is a unique value of $y_j(t)$, call it Y_j, such that the derivative (6) is zero. Let $Y = (Y_j, j \in J)$. Given Y, consider the problem of choosing $x = (x_r, r \in R)$ in order to

$$\begin{aligned}&\text{maximize} \quad \sum_{r\in R} w_r U(x_r) \\ &\text{over} \quad\quad Ax \le Y, \quad x \ge 0,\end{aligned} \qquad (21)$$

where $\alpha > 0$ and

$$\begin{aligned}U(x) &= \frac{x^{1-\alpha}}{1-\alpha} \quad \alpha \neq 1 \\ &= \log(x) \quad \alpha = 1.\end{aligned}$$

The unique solution to this strictly convex optimization problem is called a weighted α-fair rate allocation, or, if $w_r = 1$,

$r \in R$, an α-fair rate allocation [11, 15, 21]. We can identify the stationary version

$$x_r = w_r \left(\sum_{j \in r} R_j^{-\alpha} \right)^{-1/\alpha}$$

of the form (17) with the unique optimum to the problem (21): $(R_j^{-\alpha}, j \in J)$ is simply the vector of Lagrange multipliers for the constraints $Ax \leq Y$. Since A is of full row rank, this vector is unique.

Next, we linearize the system (6-8), (10), (17) about its unique equilibrium. Let R_j denote the equilibrium value of $R_j(t)$ for each $j \in J$, and let x_r be the equilibrium value of $x_r(t)$ for each $r \in R$. Taking $R_j(t) = R_j + R_j v_j(t)$, for all $j \in J$, we get the following linearized version

$$\dot{v}_j(t) = -\frac{a_j(Y_j + C_j)}{C_j Y_j \overline{T}_j} \sum_{r:j \in r} \frac{x_r^{\alpha+1}}{w_r^\alpha} \sum_{l \in r} R_l^{-\alpha} v_l(t - T_{lr} - T_{rj}). \tag{22}$$

We have used the result $(Y_j + C_j)/Y_j = 1 + b_j C_j p_j'(Y_j)$, from (12), to reduce to this form.

Let us define

$$z_r(t) = x_r T_r \sum_{j \in r} R_j^{-\alpha} v_j(t - T_{jr}) \tag{23}$$

for each $r \in R$. Then from (22) we get

$$\dot{v}_j(t) = -\frac{a_j(Y_j + C_j)}{C_j Y_j \overline{T}_j} \sum_{r:j \in r} \frac{x_r^\alpha}{w_r^\alpha T_r} z_r(t - T_{rj}) \tag{24}$$

and,

$$\dot{z}_r(t) = -x_r T_r \sum_{j \in r} R_j^{-\alpha} \frac{a_j(Y_j + C_j)}{C_j Y_j \overline{T}_j} \sum_{s:j \in s} \frac{x_s^\alpha}{w_s^\alpha T_s} z_s(t - T_{sj} - T_{jr}). \tag{25}$$

If (25) is exponentially stable, then, from (24), $\dot{v}(t)$ must tend to 0 exponentially and so $v(t)$ must tend to a limit. However, $z(t) \to 0$ and the connectivity matrix has full row rank, and so, from (23), we must have $v(t) \to 0$.

To find conditions for the exponential stability of (25), we turn to control theory. Let us overload notation and write $z(\omega)$ for the Laplace transform of $z(t)$. A natural control loop version of (25) is

$$z(\omega) = X(\omega) P(\omega) K(\omega) (w(\omega) - z(\omega)), \tag{26}$$

where $X(\omega)$, $P(\omega)$ and $K(\omega)$ are matrix functions, defined below, and $w(\omega)$ represents the input into the control loop. We define $X(\omega)$ and $K(\omega)$ to be diagonal matrices with entries

$$X_{r,r}(\omega) = T_r e^{-T_r \omega} x_r^{1-\alpha} w_r^\alpha, \quad K_{r,r}(\omega) = \frac{1}{T_r \omega}.$$

The matrix $P(\omega)$ has entries

$$P_{r,s}(\omega) = e^{\omega(T_{rj} - T_{sj})} \frac{x_r^\alpha x_s^\alpha}{w_r^\alpha w_s^\alpha} \sum_{j \in r \cap s} R_j^{-\alpha} \frac{a_j(Y_j + C_j)}{C_j Y_j \overline{T}_j},$$

and thus satisfies $P^T(-\omega) = P(\omega)$. Theorem 1 of [22] implies that (26) is asymptotically stable, and so (25) is exponentially stable, if the maximum absolute row sum norm of $P(i\theta)X(0)$ is less than $\pi/2$ for all real θ. For any real θ, the maximum absolute row sum norm of $P(i\theta)X(0)$ is given by

$$\|P(i\theta)X(0)\|_\infty = \max_{r \in R} \frac{x_r^\alpha}{w_r^\alpha} \sum_{j \in r} R_j^{-\alpha} \frac{a_j(Y_j + C_j)}{C_j Y_j \overline{T}_j} \sum_{s:j \in s} x_s T_s$$

$$\leq \max_{r \in R} \left(\sum_{l \in r} R_l^{-\alpha} \right)^{-1} \sum_{j \in r} R_j^{-\alpha} \max_{l \in J} a_l \frac{Y_l + C_l}{C_l} \leq 2 \max_{j \in J} a_j.$$

Thus, if, for all $j \in J$, $a_j < \pi/4$, then the system of delayed differential equations (6-8), (10), (17) is locally stable about its unique equilibrium point.

11. ACKNOWLEDGEMENTS

The first author is grateful for discussions of RCP with Nandita Dukkipati and Nick McKeown. The research of the authors is supported in part by EPSRC grant GR/S86266/01.

12. REFERENCES

[1] H. Balakrishnan, N. Dukkipati, N. McKeown and C. Tomlin. Stability analysis of explicit congestion control protocols. *IEEE Communications Letters*, 11(10): 823–825, 2007.

[2] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié and J.W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. *Proceedings of ACM Sigcomm, 2001.*

[3] T. Bonald, L. Massoulié, A. Proutière and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems*, 53: 65–84, 2006.

[4] N. Dukkipati and N. McKeown. Why flow-completion time is the right metric for congestion control. *Computer Communication Review*, 36(1): 59–62, 2006.

[5] N. Dukkipati, M. Kobayashi, R. Zhang-Shen and N. McKeown. Processor sharing flows in the Internet. *Thirteenth International Workshop on Quality of Service (IWQoS)*, Passau, Germany, June 2005.

[6] N. Dukkipati, N. McKeown and A.G. Fraser. RCP-AC: Congestion control to make flows complete quickly in any environment. High-Speed Networking Workshop: The Terabits Challenge (In Conjunction with IEEE Infocom 2006), Barcelona, Spain, April 2006.

[7] B. Hajek. A queue with periodic arrivals and constant service rate. In F.P. Kelly (ed.) *Probability, Statistics and Optimisation: a Tribute to Peter Whittle*. Wiley, Chichester. 1994, 147–157.

[8] J.M. Harrison. *Brownian Motion and Stochastic Flow Systems*. Krieger, 1985.

[9] C.V. Hollot, V. Misra, D. Towsley and W. Gong. Analysis and design of controllers for AQM routers supporting TCP flows. *IEEE/ACM Transactions on Automatic Control*, 47(6):945-959, 2002.

[10] D. Katabi, M. Handley and C. Rohrs. Internet congestion control for future high bandwidth-delay product environments. *Proceedings of ACM Sigcomm*, 2002.

[11] F. Kelly. Fairness and stability of end-to-end congestion control. *European Journal of Control*, 9:159–176, 2003.

[12] J.-Y. Le Boudec and B. Radunovic. Rate performance objectives of multihop wireless networks. *IEEE Transactions on Mobile Computing*, 3:334–349, 2004.

[13] S.H. Low and D.E. Lapsley. Optimization flow control.I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7:861-874, 1999.

[14] L. Massoulié. Structural properties of proportional fairness: stability and insensitivity. *The Annals of Applied Probability*, 17(3):809–839, 2007.

[15] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8:556–567, 2000.

[16] F. Paganini, Z. Wang, J.C. Doyle and S.H. Low. Congestion control for high performance, stability and fairness in general networks. *IEEE/ACM Transactions on Networking*, 13:43-56, 2005.

[17] G. Raina and D. Wischik. Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. *Proc. EuroNGI Next Generation Internet Networks*, Rome, Italy, April 2005.

[18] G. Raina, D. Towsley and D. Wischik. Part II: Control theory for buffer sizing. *Computer Communication Review*, 35(3): 79–82, 2005.

[19] J.W. Roberts (ed.) (1992). *Performance Evaluation and Design of Multiservice Networks.* Office for Official Publications of the European Communities, Luxembourg.

[20] J. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. *ITC specialists seminar*, Yokohama, 1998.

[21] R. Srikant. *The Mathematics of Internet Congestion Control.* Birkhauser, 2004.

[22] G. Vinnicombe. On the stability of end-to-end congestion control for the Internet. Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.398. 2000.

[23] G. Vinnicombe. On the stability of networks operating TCP-like congestion control. *Proceedings of IFAC World Congress*, Barcelona, Spain 2002.

[24] T. Voice. Stability of multi-path dual congestion control algorithms. *IEEE/ACM Transactions on Networking*, 15: 1231–1239, 2007.

[25] T. Voice and G. Raina. (2007). Stability analysis of a max-min fair Rate Control Protocol (RCP) in a small buffer regime. Preprint.

[26] T. Voice and G. Raina. (2008). Rate Control Protocol (RCP): global stability and local Hopf bifurcation analysis. Preprint.

[27] D. Wischik and N. McKeown. Part I: Buffer sizes for core routers. *Computer Communication Review*, 35(3): 75–78, 2005.