PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2008) **366**, 1955–1971
doi:10.1098/rsta.2008.0021
Published online 6 March 2008

Control of communication networks: welfare maximization and multipath transfers

By Peter B. Key[1,*] and Laurent Massoulié[2]

[1]*Microsoft Research Ltd, 7 J J Thomson Avenue, Cambridge CB3 0FB, UK*
[2]*Thomson Paris Research Lab, 46 Quai A. Le Gallo, 92648 Boulogne, France*

We discuss control strategies for communication networks such as the Internet. We advocate the goal of welfare maximization as a paradigm for network resource allocation. We explore the application of this paradigm to the case of parallel network paths. We show that welfare maximization requires active balancing across paths by data sources, and potentially requires implementation of novel transport protocols. However, the only requirement from the underlying 'network layer' is to expose the marginal congestion cost of network paths to the 'transport layer'. We further illustrate the versatility of the corresponding layered architecture by describing transport protocols with the following properties: they welfare maximization, each communication may use an arbitrary collection of paths, where paths may be from an overlay, and paths may be combined in series and parallel. We conclude by commenting on incentives, pricing and open problems.

Keywords: welfare maximization; marginal congestion cost; transport control; overlays; multipath routing; congestion control

1. Introduction

The Internet has grown from a small network connecting a few nodes to be the dominant global communications network. The current Internet is an example of a fast packet network, and we shall take it as both an exemplar and a motivation for describing problems of controlling communication networks. In particular, we address problems of control in such a network, where the objective is to provide some form of service differentiation or quality of service (QoS) to different users or applications in the face of limited resources, and we want to optimize the performance in some sense.

The problem is a topical one since the current Internet essentially provides a 'best-effort' service to all users, with no clear mechanism for providing any form of end-to-end service differentiation. This is the subject of much current debate and research within the Internet community, often centred around clean-slate approaches or next generation architectures. This is because the Internet has failed to evolve architecturally. For instance, the two core protocols of the Internet, the Internet protocol (IP) and the transmission control protocol (TCP), have changed little over the last two decades. The reasons for this are many and

[*] Author for correspondence (peter.key@microsoft.com).

One contribution of 16 to a Discussion Meeting Issue 'Networks: modelling and control'.

varied, and include strong standards with a commitment to backward compatibility, and diverse ownership. These, coupled with an absence of a link between quality and price militates against any incentive or reward structure for unilateral innovation or evolution (e.g. Laskowski & Chuang 2006), which makes creating new end-to-end forms of differentiation difficult.

When there is excess demand, and limited resource, some form of control is needed. Different applications react differently to being given reduced resources or starved of resources. For example, a real-time media stream may be unable to decode meaningful content for the user to view, whereas a file transfer will be delayed. Our approach is general, and the resources could include wired or wireless network capacity, buffering in a switch or even processing capacity. However, we shall use capacity or bandwidth as a motivating example, with the rate that an application receives, the primary performance indicator. Other measures, such as latency or delay, are important from an application's perspective, but there are good reasons to see within-network queuing delay decreasing as networks evolve, leaving latency solely determined by path length and physics. Indeed, Kelly (2000) provides scaling regimes where queuing delays decrease, and others (such as Raina & Wischik 2005; Enachescu *et al.* 2006) have argued that buffers in the networks should be 'small'.

In this paper, we focus on the performance of the so-called 'elastic-traffic'. We first advocate welfare maximization as an objective for network control (§2). Our main argument is that it maximizes the so-called *schedulable region*, characterizing the load that the network can carry. Interestingly, for single-path data transfers, the current Internet TCP protocol achieves a particular type of welfare maximization.

We then address how to achieve this objective when multiple network paths connecting a data source to its destination can be used for individual data transfers (§3). We show that coordination of flow control along such multiple paths is required. We also explain why and how the current Internet transport protocol needs to be modified to allow suitable coordination.

We then discuss path sampling techniques for achieving welfare maximization at a macroscopic level, while ensuring that individual data transfers proceed along a small number of paths (§3*c*). We further introduce flow control protocols that combine network paths not only in parallel but also in series (§4). Finally, we comment on the architectural aspects of our proposals and related charging issues (§5).

2. Modelling and control of the Internet

TCP is the dominant transport protocol in the current Internet, and implements congestion control using a window-based flow control. Several authors (Mathis *et al.* 1997; Padhye *et al.* 2000) have proposed models for the rate that TCP allocates to a particular connection, say of class s. They express the achieved rate as a function of both the packet loss probability p_s along the corresponding network path, and the round-trip time (RTT) delay T_s along the same network path. The so-called *square-root formula* states that TCP Reno, when in congestion avoidance mode, gives a rate x_s equal to $1/(T_s\sqrt{p_s})$, when expressed in suitable units.

As Kunniyur & Srikant (2000) observe, this can be interpreted as saying that TCP implicitly solves a utility maximization problem, where the objective is to maximize the net utility (utility minus cost), with incurred cost the path loss rate $p_s x_s$ and with utility function

$$U_s(x) = -\frac{1}{T_s^2 x_s}. \tag{2.1}$$

The maximum is achieved when $U_s'(x_s) = p_s$, which does indeed recover the square-root formula for x_s. In other words, the predominant flow control protocol in the current Internet can be thought of as implementing a utility maximization. See Kelly (2000) for a more complete discussion.

We now describe a general framework for bandwidth allocations in a data network proposed by Kelly *et al.* (1998). We classify data flows into different types $s \in \mathcal{S}$. Let n_s denote the number of data flows of type s, and let x_s denote the data rate assigned to each type s flow, for all $s \in \mathcal{S}$. Then, the vector of the desired data rates $\{x_s\}_{s\in\mathcal{S}}$ is chosen to maximize the welfare function

$$W(\{x_s\}_{s\in\mathcal{S}}) := \sum_{s\in\mathcal{S}} n_s U_s(x_s) - \Gamma(\{n_s x_s\}_{s\in\mathcal{S}}), \tag{2.2}$$

where U_s is a strictly concave, non-decreasing utility function and Γ is a convex, non-decreasing cost function, representing the overall network cost of having aggregate bandwidth $n_s x_s$ for each flow type $s \in \mathcal{S}$. These are reasonable assumptions: concave utility captures the elastic nature of data traffic, while convex costs reflect congestion or capacity costs. This framework can be related to flow control in the current Internet, where flow types are associated with a pair of IP source and destination addresses.

For simplicity, we shall assume that the welfare function has one finite local maximum. Since welfare is a strictly concave function, this must be its unique global maximum. For the ease of exposition, we also assume that the functions U_s, $s \in \mathcal{S}$ and Γ are differentiable.

It is possible to design simple, decentralized rate adaptation schemes for send rates x_s, under which welfare increases over time and converges to the unique welfare maximizing allocation vector. Indeed, one such scheme that has these properties is the continuous rate adaptation algorithm

$$\frac{\mathrm{d}}{\mathrm{d}t} x_s(t) = \kappa_s \left[U_s'(x_s) - p_s \right],$$

where

$$p_s := \frac{\partial}{\partial z_s} \Gamma(z) \Big|_{z=\{n_s x_s\}}$$

is the so-called marginal cost for bandwidth of type s, and κ_s is some positive gain parameter (see Srikant (2003) and Voice (2006) for a detailed treatment).

As a specific example of network cost, let Γ be the sum of costs specific to links within the network

$$\Gamma(\{n_s x_s\}_{s\in\mathcal{S}}) = \sum_{\ell} \Gamma_\ell(y_\ell),$$

where the cost of link ℓ is some (convex, increasing) function Γ_ℓ of the aggregate rate y_ℓ through that link $y_\ell = \sum_{s\in\mathcal{S}:\ell\in s} n_s x_s$, and where $\ell \in s$ means that link ℓ is part of path s. Under these assumptions, the marginal cost p_s is the sum of the link prices $p_s = \sum_{\ell:\ell\in s} \Gamma'_\ell(y_\ell) \overset{\text{def}}{=} \sum_{\ell:\ell\in s} p_\ell$. One specific link cost function of interest consists in setting p_ℓ to be packet loss probability in a single server queue with capacity C and finite buffer size B (the so-called M/M/1/B queuing system), an approximate model for current drop-tail routers.

We now argue that utility maximization is the natural framework for addressing questions of fairness and design for rate control algorithms.

Firstly, this objective is achievable in practice: we saw that TCP performs such a maximization. It can be modified to mimic the behaviour of the previous dynamical systems to capture different utility functions (e.g. *scalable TCP* in Kelly 2003); the packet loss signal to which TCP reacts could be replaced by some measure of the path marginal cost, to capture different cost functions.

Secondly, if the utility functions U_s capture accurately the utility of data rates for type s flows, and the function Γ which determines the marginal costs or 'prices' p_s reflects accurately the network costs, then the above framework is in line with the microeconomic theory, with its corresponding emphasis on welfare maximization.

The previous argument applies to a scenario where data sources are long-lived, and the collection of active flows is essentially static. This stands in sharp contrast with the current usage of the Internet. There, data transfers have a limited duration, which typically depends on the achieved transmission rate for transfers of files that have an intrinsic volume.

We now describe the models of control and the network performance in a dynamic setting. Such models are important for addressing performance, and have a validity independent of utility maximization.

We assume that for each type s, there are n_s file transfers. Type s file transfers are created at the instants of a rate ν_s Poisson process, and the corresponding file size is exponentially distributed with parameter μ_s. Once initiated, transfers take place (there is no admission control or rejection policy) and obtain a welfare maximizing rate allocation x_s such that the vector $\{x_s\}_{s\in\mathcal{S}}$ maximizes (2.2). This corresponds to the best-effort service model paradigm, as implemented in the current Internet: data transfers are always accepted and proceed at the maximal rate that the standard transport protocol will enable.

To analyse the performance, we consider deterministic evolution equations, which correspond to the mean drift of the variables of interest in the original stochastic model for n_s. Such deterministic evolutions can be interpreted as accurate approximations of the original dynamics under some 'law of large numbers' scaling, where both arrival rates and network capacities are large, and n_s are the rescaled quantities. Specifically, we consider the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} n_s = \nu_s - \mu_s n_s x_s(n), \quad s \in \mathcal{S}. \tag{2.3}$$

We let dom(Γ) denote the set of rate vectors $\boldsymbol{z}$ for which $\Gamma(z)$ is finite, and then introduce the following stability condition:

$$\rho \in \operatorname{int}(\operatorname{dom}(\Gamma)), \tag{2.4}$$

$$\exists \delta \in (0,\infty)^{\mathcal{S}} \quad \text{such that} \quad U'_s(\delta_s) > \Gamma'_s(\rho+\delta), \quad s \in \mathcal{S}, \tag{2.5}$$

where Γ'_s denotes the sth partial derivative of Γ, and $\boldsymbol{\rho}$ is the vector of loads ρ_s for each file transfer type s, defined as $\rho_s = \nu_s/\mu_s$. The conditions are natural extensions of more familiar load constraints. For example, in the special case where there is a single resource having capacity C, with Γ the penalty function $\Gamma(z)=0$ if $z \le C$, and $+\infty$ otherwise, then the conditions are equivalent to $\sum_r \rho_s < C$. See Bonald & Massoulié (2001) for a discussion of the connection between this stability condition and fairness.

The following result is a consequence of a more general result proved in Key & Massoulié (2006), where real-time (fixed-duration) traffic is mixed with file transfers.

Theorem 2.1. *Under the conditions* (*2.4*) *and* (*2.5*), *the differential equations* (*2.3*) *have a unique invariant point* $\hat{\boldsymbol{n}}$ *characterized by*

$$U'_s\left(\frac{\rho_s}{\hat{n}_s}\right) = \Gamma'_s(\rho), \quad s \in \mathcal{S}. \tag{2.6}$$

Moreover, it is possible to show that all trajectories of the differential equations (2.3) converge to the equilibrium point.

In the case where the cost function Γ is constant on its domain dom(Γ), this result implies that the original stochastic system is ergodic, and hence file transfer times remain finite, whenever the vector of loads $\boldsymbol{\rho}$ lies in the interior of dom(Γ). In other words, the schedulable region of the network coincides with dom(Γ). This is the largest possible region, and thus welfare maximizing rate assignments lead to a maximal schedulable region (e.g. Massoulié 2007, for more details).

3. Multipath routing and congestion control

Multipath routing, where a flow can choose or use several paths rather than one, provides performance and reliability benefits over single-path routing. Reliability is improved since failures are less likely to disconnect a sender from a receiver when multiple paths exist between them. Performance is improved since higher data rates can potentially be achieved over several paths than over a single path. Multipath routing can also alleviate limitations of current architectures, where routing choices are dictated by network providers rather than end-users.

In this section, we investigate how to achieve the full benefits of multipath routing.

(*a*) *Uncoordinated parallel routing*

For uncoordinated parallel routing, we assume that each flow makes use of the underlying transport layer for each of its available paths, and that the transport layer assigns a rate according to some utility maximization principle, independently among paths used by the flow. This models traffic sources using several parallel standard TCP connections available for data transfer, and corresponds to the simplest implementation of multipath transfers within the current Internet architecture.

Let us assume that the transport layer solves

$$\text{maximize} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}(s)} N_s U_r(x_r) - \Gamma(N\boldsymbol{x}), \tag{3.1}$$

over $x_r \ge 0$, where s denotes a type of flow; $\mathcal{R}(s)$ denotes the collection of paths that type s flows can use; and N_s is the total number of type s flows. Without loss of generality, we assume that the path sets $\mathcal{R}(s)$ are disjoint. The utility on each

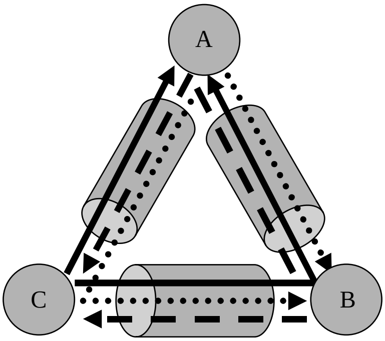

Figure 1. Example network where the use of parallel uncoordinated connections is inefficient.

path may represent a TCP controller in which case the utility functions depend on specific characteristics of the path, namely the corresponding RTT.

Such uncoordinated multipath transfers can be inefficient. To illustrate this, we focus on the performance with dynamic arrivals of file transfer requests, for a particular topology, shown in figure 1. Specifically, there are three types of flows, corresponding to all possible source–destination pairs (such as A–B). We denote the type of flow corresponding to a particular source–destination pair with the letter of the opposite vertex, e.g. type C flows correspond to the source–destination pair A–B. Flows between a pair of nodes can use both the direct one-hop route and the two-hop route (e.g. routes A–B and A–C–B for type C flows). We further assume sharp capacity constraints and unit capacity links, hence the cost Γ is zero unless some capacity constraint is violated, in which case it is infinite.

Consider the case of symmetric loads, where $\nu_i \equiv \nu$ denotes the arrival rate of type i flow transfer requests, $\mu_i \equiv \mu$ is the parameter of the assumed exponential distribution for the volume of such transfers and $\rho_i := \nu_i / \mu_i$ is the load contributed by such flows. The fluid dynamics corresponding to this process are given by

$$\frac{\mathrm{d}}{\mathrm{d}t} n_i = \nu - \mu_i n_i x_i(\boldsymbol{n}), \quad i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}. \tag{3.2}$$

If we assume that the utility functions associated with each path all coincide with $U(x) := x^{1-\alpha}/(1-\alpha)$ (a plausible model of TCP allocations if $\alpha = 2$ and RTTs are all equal), then it is possible to show that there is a solution of the fluid equations which converges to $n = 0$ provided $\rho < \rho^*$ where

$$\rho^* := \frac{1 + 2^{-1/\alpha}}{1 + 2^{1-1/\alpha}}. \tag{3.3}$$

A more detailed treatment of this example can be found in Key & Massoulié (2006), where the following is proved.

Theorem 3.1. *For the triangle network with symmetric offered loads ρ and path utility functions $U(x) = x^{1-\alpha}/(1-\alpha)$, the solution $\boldsymbol{n}(t)$ to the system of differential equations (3.2) diverges to infinity whenever $\rho > \rho^*$. In particular, when $\alpha = 2$, the system is unstable provided $\rho > 1/\sqrt{2} \approx 0.71$.*

Conversely, the solution $\boldsymbol{n}(t)$ decreases to zero in time at most $\theta[n_{\mathrm{A}}(0) + n_{\mathrm{B}}(0) + n_{\mathrm{C}}(0)]$ for a suitable constant $\theta > 0$ whenever $\rho < \rho^*$.

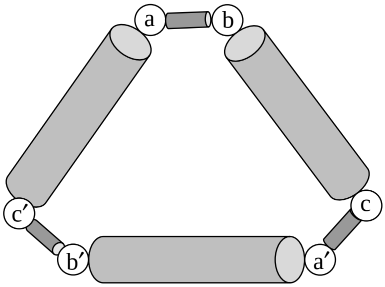

Figure 2. Network with alternation of fat and long with short and thin links.

It is possible to show that ρ^* is indeed the exact capacity of the triangle network under symmetric loads.

Let us anticipate the discussion of coordinated multipath control. As will be discussed in §3*c*, if we consider a suitably *coordinated* control, then the corresponding trajectories converge to zero, under the following condition:

$$\rho_i + \rho_j < 2, \quad i,j \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}, \quad i \neq j. \tag{3.4}$$

Consequently, the original stochastic system is ergodic under these conditions, which are essentially optimal. In particular, for symmetric loads this condition becomes $\rho < 1$.

In other words, the schedulable region is smaller than that for coordinated routing, and indeed approximately 30% less traffic can be carried when uncoordinated routing is used instead of a coordinated controller.

(*b*) *Path bias and suboptimal coordination*

The previous example showed that some form of coordination between paths is needed to reap the full benefits of multipath routing. One simple form of coordination is the following: for any flow of traffic offered access to several paths, choose the path for which the achieved transfer rate is largest. This could be implemented at the application layer, on top of the TCP transport layer for each path, by ‘feeding’ paths on the basis of their achieved rates.

We now show that this kind of adaptation is suboptimal when the underlying transport layer introduces biases between distinct paths. Specifically, we assume that the rate allocated by the transport layer to each path is the result of some utility maximization, where the utility function relative to path r is given by the utility function $U_r(x)$ defined in (2.1), reflecting the behaviour of TCP Reno.

We consider the hexagonal network of figure 2, which has long fat links, each having RTT T and capacity C, and short, thin links with capacity c and RTT τ where $c < C$ and $T > \tau$. Assume that type a users transfer data to a', and can either use a ‘long-fat’ path via c', b', comprising two long links and one short link (l–s–l), or a ‘short-thin’ route via b, c, which has two short links and one long route (s–l–s). Similarly, assume type b and c users transfer data to b' and c', respectively, and can each have a similar choice of short-thin or long-fat links.

We shall suppose that there are N' users of each type, who can use n_p parallel connections, and let $N = N' n_p$ denote the total number of connections users of each type make. Now let each user independently and selfishly seek to use the set of paths that maximize their rate. In other words, they are able to choose the path sets $\mathcal{R}(s)$ and the number of connections per paths subject to the constraint, $N_s=\sum_{r\in\mathcal{R}(s)}N_r$, and they seek the routes that maximize the achieved throughput $\sum_r N_r x_r$.

We now show that inefficient equilibria can exist, when using the TCP utility functions described above. Consider the case when all connections are via short–long–short links (s–l–s), having an RTT of $T+2\tau$. Then, there will be $2N$ connections on each short link, and N on each long link. To fully use each thin link, each connection will receive a rate of $c/(2N)$; for this to happen, the Lagrange multipliers p must satisfy

$$2p = U'_{s\text{–}l\text{–}s}(c/2N) = \left[(T+2\tau)\frac{c}{2N}\right]^{-2}.$$

Now, on the l–s–l route, with RTT $2T+\tau$, the path Lagrange multiplier is p, rather than $2p$, and the rate per connection is

$$x=\frac{1}{(2T+\tau)\sqrt{p}}=\sqrt{2}\frac{T+2\tau}{2T+\tau}\frac{c}{2N}.$$

This path will not be chosen, and hence s–l–s is indeed a Nash equilibrium provided

$$\sqrt{2}\frac{T+2\tau}{2T+\tau}<1.$$

But in this case, the throughput is *half* of what it would be if all connections were l–s–l.

As shown in Key *et al.* (2006), if the RTT bias is removed so that utility functions U_r are independent of path RTT, then the Nash equilibrium is efficient and solves the global social welfare problem (equations (3.5) and (3.6)) that we now introduce.

(*c*) *Multipath routing and coordinated congestion control*

With joint multipath routing and coordinated congestion control, for fixed numbers N_s of users of type s, the optimal rates for class s users solve the welfare maximization problem

$$\text{maximize} \sum_s N_s U_s\left(\sum_{r\in\mathcal{R}(s)} x_r\right) - \Gamma(N\boldsymbol{x}), \tag{3.5}$$

$$\text{over}\quad x_r\geq 0,\quad r\in\mathcal{R}. \tag{3.6}$$

Now there is a single utility per *user s* and $x_s=\sum_{r\in\mathcal{R}(s)}x_r$ is the allocation a user receives aggregated across paths. This coordinated problem is strong Lagrangian, hence has a unique optimum and its solution characterized by the Kuhn–Tucker conditions

$$U'_s\left(\sum_{r\in\mathcal{R}(s)} x_r\right)\leq \Gamma'_r(N\boldsymbol{x}),\quad U'_s\left(\sum_{r\in\mathcal{R}(s)} x_r\right)<\Gamma'_r(N\boldsymbol{x})\Rightarrow x_r=0, \tag{3.7}$$

and, as a consequence, the allocation for user s only puts a non-zero allocation on paths r whose price p_r is equal to the minimum price across possible routes. There may be only one or several such lowest cost paths. In other words, for a user s there is a critical value price p_s such that the prices on any path s uses equal p_s, and on possible paths r that s can use but does not (i.e. allocates zero to these paths), the prices must be higher than p_s.

Theorem 2.1 still applies, hence with dynamic arrivals, such coordinated controllers ensure that the schedulable region is maximized. Moreover (see Massoulié & Key 2006), the equilibrium cost of network operation is also minimized, independently of which utility functions are used for each traffic class.

Distributed rate control algorithms can be constructed for all of the above optimization problems, e.g. Kelly & Voice (2005) and Han *et al.* (2006).

(*d*) *Path selection and randomization*

We now assume that there is a large set of possible paths that a particular source–destination pair could use, but that each end-user uses only a small number of paths at any moment in time, while choosing to alter these paths over time as better ones become available. This models the practical limitations of opening a large number of potential connections, imposed by the cost of keeping state and the difficulty of assessing path quality on a large number of paths.

Assume that class s users can use concurrent paths from a collection c, where $c \subset \mathcal{R}(s)$, and denote by $\mathcal{C}(s)$ the family of all such path collections that are allowed. For definiteness, think of $\mathcal{C}(s)$ as the collection of all subsets of $\mathcal{R}(s)$ of size b. Denote by N_c the number of users with associated set of connections equal to c. When the number of class s users equals N_s, we have the relation

$$\sum_{c \in \mathcal{C}(s)} N_c = N_s, \quad s \in \mathcal{S}. \tag{3.8}$$

The allocation to a class s user is then given as the solution to the optimization

$$\text{maximize} \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}(s)} N_c U_s\left(\sum_{r \in c} x_{c,r}\right) - \Gamma(N\boldsymbol{x}), \tag{3.9}$$

over $x_{c,r} \geq 0$ where $N\boldsymbol{x}$ can be written as the vector of aggregate loads (X_r) where $X_r = \sum_{c:r \in c} N_c x_{c,r}$.

Now assume that a user has some current route set c, and is offered a new route set c', at some fixed rate $\lambda_{cc'}$, where this rate is on a fast time-scale compared with the time-scale of arrivals and departures. The new route set is accepted provided the *net benefit* the user receives from the new route set is higher than that of the current route set. The users of type s can explore the entire space provided that for all s, for any $r \in \mathcal{R}(s)$ and any set $c \in \mathcal{C}(s)$, there is some c' such that $r \in c'$ and $\lambda_{cc'} > 0$.

We denote by $x_c = \sum_{r \in c} x_{c,r}$ the aggregate data rate obtained by users streaming along routes $r \in c$, where $x_{c,r}$ is the sending rate along route r. We assume the following form of rate adaptation (see Kelly *et al.* 1998):

$$\frac{\mathrm{d}}{\mathrm{d}t} x_{c,r} = \kappa_{c,r}\left[U'_{s(c)}(x_c) - \partial_r \Gamma(X)\right] + \mu_{c,r}, \tag{3.10}$$

where the term $\mu_{c,r}$ is non-negative, satisfies $\mu_{c,r} x_{c,r} \equiv 0$ and is meant to ensure non-negativity of $x_{c,r}$, while $\kappa_{c,r}$ is a positive gain parameter.

The net benefit per unit time for type s users streaming along routes r in some set c, denoted by B_c, is given by

$$B_c = U_s(x_c) - \sum_{r \in c} x_{c,r} U_s'(x_s).$$

Users swap from set c to c' when the set is offered provided $B_{c'} > B_c$. For any strictly concave, continuously differentiable function U, this is equivalent to choosing path sets c with higher *rate*, x_c.

When there is a large population of users, we can model the evolution of users by deterministic mean-field equations

$$\dot{N}_c = \sum_{c'} N_{c'} \lambda_{c'c} \phi(B_c - B_{c'}) - \sum_{c'} N_c \lambda_{cc'} \phi(B_{c'} - B_c), \tag{3.11}$$

for some continuous Lipschitz function $\phi \rightarrow [0,1]$ that is equal to 0 on $\mathbb{R}_-$ and positive on $(0, \infty)$.

We can then show that the described rate adaptation scheme and route selection policy do indeed maximize the welfare, in that they maximize (3.9), and moreover solve the global optimization problem (3.5) and (3.6). We shall assume that utility functions U_s and the penalty function Γ are continuously differentiable on their domain, the former are strictly concave increasing, the latter convex increasing and that $U_s'(x) \rightarrow 0$ as $x \rightarrow \infty$. We have the following (Key *et al.* 2007).

Proposition 3.2. *Any absolutely continuous solution* $(N_c, x_{c,r})$ *to the system of ordinary differential equations* (*ODEs*) *defined by* (*3.10*) *and* (*3.11*) *converges to the set of maximizers of the welfare function*

$$\mathcal{W}(X, N) := \sum_{s \in \mathcal{S}} \sum_{c \subset \mathcal{R}(s)} N_c U_s(X_c) - \Gamma(X), \tag{3.12}$$

under the constraints (*3.8*). *The corresponding equilibrium rates* (x_r) *are solutions of the coordinated welfare maximization problem* (*3.5*) *and* (*3.6*).

This is a powerful result, which says that path reselection provides global allocations that coincide with the optimal allocations that would arise if users had simultaneous access to the full route set $\mathcal{R}(s)$. Reselection can be interpreted as a repacking algorithm that operates on a time-scale dictated by λ. Of course, there are practical issues that need to be addressed in directly implementing these ideas; care needs to be taken in deciding how long to try alternate paths before assessing the throughput (quality) and in allowing for statistical effects in such sampling. For flows that are 'short', there may be little benefit in repeated resampling.

What happens for parallel or uncoordinated controllers? It turns out that the same results apply *provided* there is no RTT bias in the controllers, unlike current TCP. This is another argument for removing the current RTT bias in TCP. If parallel controllers use a fixed number of paths b, for example $b=2$, then to give equivalence with the coordinated controller we need to replace the utility function $U_s(x)$ for the coordinated controller by $x \rightarrow b U_s(x/b)$.

Note that several peer-to-peer (P2P) applications such as BitTorrent effectively implement receiver-driven multipaths. Transfers take place using TCP, where throughput depends on the RTT; hence, this is an example of a multipath uncoordinated congestion controller, which uses a greedy path selection algorithm to find best paths.

4. Combinations of paths in series and parallel

We have just seen that suitably designed congestion controllers can combine paths in parallel and maximize utility. We now show how to achieve such a maximization when multiple paths can be used in series as well as in parallel. For instance, a source node S could send data to a destination node D via a relay node R by combining the IP paths S–R and R–D in series. Considering the two paths separately rather than as one combined S–R–D path corresponds to breaking or splitting the flow at R. This is an appropriate model when R is a node in an overlay network or is a middle-box or proxy.

The use of proxies is particularly appropriate for wireless networks, where there is an incentive to break the TCP control loop into parts to achieve better throughput, and proxies are used in practice in mobile networks. Split TCP breaks a TCP connection into two or more 'legs' in series, and has been widely studied (e.g. Balakrishnan *et al.* 1995; Kopparty *et al.* 2002).

We now discuss flow controllers that seek to maximize the overall welfare for source–destination flows by exploiting arbitrary network paths, generalizing the notion of coordinated controllers.

Assume that each flow f, with source $s(f)$ and destination $d(f)$, can use a collection of paths $\pi \in \Pi(f)$, where each path p is identified by its source and destination. A path π with source node $s(\pi)=i$ and destination node $d(\pi)=j$ is also denoted (ij).

We denote by x_{ij}^{f} the rate sent for flow f along path (ij), and by W_i^f the amount of data buffered at i for flow f, with the convention that

$$W_{s(f)}^{f} = W_{d(f)}^{f} = 0.$$

In this context, a natural utility maximization problem is the following:

$$\text{maximize} \sum_{f} U_f(x^f) - \Gamma(y), \tag{4.1}$$

$$\text{over} \quad x^f \geq 0, \quad x_{ij}^{f} \geq 0, \quad (ij) \in \Pi(f), \tag{4.2}$$

$$\text{under} \quad x^f = \sum_{j:(s(f)j)\in \Pi(f)} x_{s(f)j}^{f}, \tag{4.3}$$

$$i \notin \{s(f), d(f)\} \Rightarrow \sum_{j} x_{ij}^{f} = \sum_{j} x_{ji}^{f}, \tag{4.4}$$

$$y_{ij} = \sum_{f} x_{ij}^{f}. \tag{4.5}$$

The penultimate constraint specifies that the path rates x_{ij}^{f} satisfy flow conservation at intermediate, relay nodes, in other words they define a *flow* from $s(f)$ to $d(f)$.

Now consider the following rate adaptation scheme for rates x_{ij}^f: for each path $(ij) \in \Pi(f)$, we set

$$\dot{x}_{ij}^f = \kappa_{ij}^f \left[1_{i=s(f)} U_f' \left(\sum_{k:(ik)\in\Pi(f)} x_{ik}^f \right) - p_{ij} + \phi\left(W_i^f\right) - \phi\left(W_j^f\right) \right], \tag{4.6}$$

where ϕ is some continuous, strictly increasing function, such that $\phi(0)=0$. In the above, κ_{ij}^f is a positive gain parameter and p_{ij} denotes the marginal cost of sending along path (ij), that is

$$p_{ij} = \frac{\partial}{\partial y_{ij}} \Gamma(y), \tag{4.7}$$

where y is the vector of path rates and Γ is the network cost function.

Note that the adaptation rule (4.6) could require some x_{ij}^f to remain positive, while there are no data to forward from node i to node j at that particular time. This can be addressed in two different ways. One could send dummy packets instead of actual data packets, while still using path (ij) at rate x_{ij}^f. However, we prefer to follow an alternative approach and send at rate y_{ij}^f, where

$$y_{ij}^f = \beta_i^f x_{ij}^f, \tag{4.8}$$

where the adjustment variable β_i^f is such that

$$\beta_i^f \in [0,1], \quad W_i^f > 0 \Rightarrow \beta_i^f = 1, \quad W_i^f = 0 \Rightarrow \beta_i^f = \min\left(1, \frac{\sum_j y_{ji}^f}{\sum_j x_{ji}^f}\right). \tag{4.9}$$

Therefore, the vector $\boldsymbol{y}$ of path rates used in the definition of marginal costs in (4.7) is given by

$$y_{ij} = \sum_f y_{ij}^f. \tag{4.10}$$

Finally, for each i, we have

$$\dot{W}_i^f = \sum_{j:(ji)\in\Pi(f)} y_{ji}^f - \sum_{j:(ij)\in\Pi(f)} y_{ij}^f. \tag{4.11}$$

One technical point has been overlooked in the above description. The quantities x_{ij}^f should remain non-negative, a property that is not guaranteed for solutions of the ODEs (4.6)–(4.11). This can be addressed by adding to the r.h.s. of (4.6) a time-dependent, non-negative term u_{ij}^f such that $u_{ij}^f = 0$ if $x_{ij}^f > 0$.

We now verify that the stationary points of (4.6)–(4.11) are solutions of (4.1)–(4.3). First, setting $\dot{W}_i^f$, $i \notin \{s(f), d(f)\}$, to zero gives

$$\sum_j y_{ji}^f = \sum_j y_{ij}^f, \quad W_i^f > 0 \Rightarrow y_{ij}^f = x_{ij}^f. \tag{4.12}$$

Setting $\dot{x}^f_{ij}$ to zero yields

$$\left.\begin{aligned} &\mathbf{1}_{i=s(f)}\,U'_f(x^f)+\phi\Big(W^f_i\Big)-p_{ij}-\phi\Big(W^f_j\Big)\le 0,\\ &x^f_{ij}>0\Rightarrow \mathbf{1}_{i=s(f)}\,U'_f(x^f)+\phi\Big(W^f_i\Big)-p_{ij}-\phi\Big(W^f_j\Big)=0.\end{aligned}\right\}\qquad(4.13)$$

By (4.12), nodes i at which $\sum_j y^f_{ij} < \sum_j x^f_{ij}$ are such that $W^f_i = 0$. In view of (4.13), at such nodes i, it holds that either $x^f_{ij}=0$ or $W^f_j = p_{ij} = 0$ for all outgoing paths $(ij)\in \Pi(f)$.

Note that, necessarily, the equilibrium quantities y^f_{ij} define a flow from $s(f)$ to $d(f)$. By (4.13), at paths (ij), $i\neq s(f)$, at which x^f_{ij}, and hence y^f_{ij} is positive, it holds that $\phi(W^f_i)=\phi(W^f_j)+p_{ij}$. Thus, at each concatenation of paths $(s(f)i_1)$, $(i_1 i_2)$, …, $(i_m d(f))$ along which all corresponding rates x^f_{ij} are positive, by the previous argument it holds that

$$U'_f(x^f)=p_{s(f)i_1}+p_{i_1 i_2}+\cdots+p_{i_m d(f)}.$$

Furthermore, by the second part of (4.13), any concatenation of paths $(s(f)i_1)$, …, $(i_m d(f))$ along which some rate x^f_{ij} equals zero is such that

$$U'_f(x^f)\le p_{s(f)i_1}+p_{i_1 i_2}+\cdots+p_{i_m d(f)}.$$

These last two properties coincide with the Kuhn–Tucker optimality condition for the welfare maximization problem (4.1) and (4.2). Therefore, any equilibrium state for (4.6)–(4.11) provides a solution to this maximization problem.

We believe that a stronger result holds, namely that the above dynamical system converges asymptotically to solutions of this maximization problem. Lyapunov function techniques described in Voice (2006), in the work of Srikant (2003) and in Georgiadis *et al.* (2006) could plausibly be adopted to the present framework.

The algorithm (4.6) is a type of back-pressure algorithm, where transformed delays $\phi(W^f_i)$ are used to summarize or communicate downstream prices. The function ϕ allows scaling of the data queues, although the choice of ϕ has implications for the stability and the choice of gain parameter in the presence of a delayed feedback. In practice, the data queues could be implemented at the application layer of a node, or be thought of as virtual queues, used to summarize downstream prices that may be signalled out of band to upstream nodes.

5. Architecture, pricing and incentives

In this section, we briefly discuss some of the practical aspects surrounding our ideas.

To implement multipath in the current Internet, some alternative paths need to be provided to the end-users. If the end-user is an edge gateway, or a node in a commercial network, then there may already be two paths to different Internet service providers (ISPs) or different ASs (autonomous systems), perhaps provided for resilience. For a domestic subscriber, there is typically only one path to the Internet, and, without source routing, the path to a destination is specified by the user's ISPs and transit ISPs. Yet the *status quo* is changing: the advent of wireless mesh networks and the emergence of wireless for last-hop access facilitates multihoming, with different access points to the Internet, while there may be incentives for ISPs to offer alternate routes. The advent of IPv6 allows different addresses to be used for different interfaces for the same host, and both multipath and addressing are currently under discussion in the Internet Engineering Task Force (IETF).

At the present time, path choice is determined by the ISPs along a path. The current pricing structure makes the link between pricing and quality unclear: ISPs essentially charge for connectivity (to domestic customers) or on the basis of 95 percentile charging for large customers connected by large fibre-optic pipes, with little relation to quality. However, if ISPs could charge users that are not their direct customers for carrying transit traffic, then there is the potential to create a market where ISPs advertise paths or themselves to end-users, and receive income for carrying transit traffic. Moreover, the effects of competition lay the foundations for linking price with quality delivered and received. In terms of architecture, this could be done via ‘stepping-stone’ routers, described in Key *et al.* (2007), which could be advertised by a type of domain name system service; the user could then choose which stepping-stone router to use. If advertised by an ISP, then the ISP would forward the traffic to the destination and the ISP might choose to advertise a set of such routers.

The question of how to recover costs or charge end-users is vexed, and does not directly fit into the current charging models for the Internet. However, related schemes such as these occur in the telephony world: calling cards and other schemes allow indirect choice, through implicit choice of backbone or international carrier.

Congestion pricing is attractive economically and theoretically, and for fast-packet networks can create a version of the ‘smart-market’, first proposed by MacKie-Mason & Varian (1995), and discussed in Gibbens & Kelly (1999). Prices that reflect the congestion price of the resources used are fed back to users who have an incentive to behave rationally and react to such prices, while network providers have an incentive to upgrade resources to match user demand.

However, this assumes price signals can be fed back to the users, perhaps via an intelligent packet marking scheme. Intelligent packet marking has been advocated as a means of more intelligently signalling congestion than the current practice of discarding packets. The simplest scheme would convey such information via a single bit, such as the explicit congestion notification bit (Floyd & Fall 1999), yet even this has not been adopted in the Internet, owing to issues of backward compatibility and needing both hosts and end-systems to behave correctly when the bit has been set. Briscoe *et al.* (2005) have proposed using an edge-based deployment of congestion notification, which may enable partial deployment in managed networks. However, the lack of progress on implementing such a price-neutral mechanism illustrates some of the technical barriers to implementing congestion pricing, let alone philosophical or economic barriers.

One of the areas currently under discussion in Next Generation Internet is charging and economics. Users appear to have a strong preference for predictable charges, such as flat rate charging, despite the fact that most users would benefit from more flexible pricing schemes (typically the majority of traffic is created by a minority of users, reflecting quoted 80–20 behaviour, symptomatic of the heavy-tailed traffic nature of Internet traffic). The challenge is whether more usage-, quality- or congestion-based charging is desirable, and whether it can be combined with simple tariffs. This is a contentious issue where strong differences of opinion exist. It is interesting to note that the mobile phone market in the UK provides examples of stratified pricing, where different flat rates are linked to usage.

6. Concluding remarks

We have shown that a utility maximization framework is a natural one for addressing questions of resource allocation for elastic traffic, when combined with stochastic demand. In this framework, path 'prices' reflect congestion costs along a path. In the current Internet, such prices are signalled via packet drops, and the use of a single, dominant transport protocol, TCP, owes more to social pressure from peer groups and standards bodies than any incentive-compatible behaviour on the part of users. The lack of evolution of the Internet has also meant that route choice is determined by network providers, with the end-users having little say in the matter.

However, the current situation can be exploited to advantage by smart algorithms at the edge or end-nodes, when using overlay functionality or when an evolution of the current structure allows stepping-stone routers without recourse to overlays. In particular, we have used the example of combined multipath routing and congestion control at the end-nodes to show such smart solutions produce versatile utility-maximizing allocations, which have nice properties of maximizing the schedulable region, and minimizing network cost. In fact, some current P2P applications already implement variants of these controllers. Such controllers can work with current TCP; however, the RTT bias inherent in TCP can lead to network inefficiencies.

The use of flexible routing combined with smart congestion controllers can overcome many of the current architectural limitations of the current Internet. The one application type that is not completely accommodated in this framework is real-time streaming or real-time applications in general. Firstly, the suitability of utility-maximizing allocations for real-time flows is not as clear as for elastic traffic, although one possible integration is described in Key *et al.* (2004). Secondly, it is not possible to give hard QoS guarantees over a best-effort network. Thirdly, it is not easy to justify implementing rate-adaptive real-time streaming, since there is not much incentive for an application to limit its rate. At the time of writing, the volume of real-time traffic is small, when compared with data traffic, and some argue that this fact if coupled with a degree of over-provisioning in the network is sufficient to provide soft guarantees that are acceptable in practice.

An alternative scenario to the current *status quo* is one where some form of pricing is used to relate path prices to real prices, via congestion prices. This brings advantages in terms of controllability and incentive compatibility, and allows for the QoS differentiation for real-time traffic. Whether the hurdles of accounting, billing and diverse ownership can be overcome, or are worth overcoming, is a moot point. We have suggested a possible evolution path that may allow some intermediate form of pricing linked to quality to be introduced, provided transit ISPs can charge end-hosts for transit traffic.

We gladly acknowledge the contributions of Alan Bain, Frank Kelly and Don Towsley to several of the ideas in this paper.

References

Balakrishnan, H., Seshan, S., Amir, E. & Katz. R. H. 1995 Improving TCP/IP performance over wireless networks. In *Proc. 1st ACM Conf. on Mobile Computing and Networking, Berkeley, CA, November 1995*. New York, NY: ACM.

Bonald, T. & Massoulié, L. 2001 Impact of fairness on Internet performance. *ACM Sigmetrics Perf. Eval. Rev.* **29**, 82–91.

Briscoe, B., Jacquet, A., Cairano-Gilfedder, C. D., Salvatori, A., Soppera, A. & Koyabe, M. 2005 Policing congestion response in an Internet work using re-feedback. *ACM Comput. Commun. Rev.* **35**, 277–288. (doi:10.1145/1090191.1080124)

Enachescu, M., Ganjali, Y., Goel, A., McKeown, N. & Roughgarden, T. 2006 Routers with very small buffers. In *Proc. IEEE Infocom, Barcelona, Spain, April 2006*. Piscataway, NJ: IEEE Press. (doi:10.1109/INFOCOM.2006.240)

Floyd, S. & Fall, K. 1999 Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Trans. Netw.* **7**, 458–472. (doi:10.1109/90.793002)

Georgiadis, L., Neely, M. J. & Tassiulas, L. 2006 *Resource allocation and cross-layer control in wireless networks*. Hanover, MA: Now Publishers.

Gibbens, R. J. & Kelly, F. P. 1999 Resource pricing and the evolution of congestion control. *Automatica* **35**, 1969–1985. (doi:10.1016/S0005-1098(99)00135-1)

Han, H., Shakkottai, S., Hollot, C., Srikant, R. & Towsley, D. 2006 Multi-path TCP: a joint congestion control and routing scheme to exploit path diversity in the Internet. *IEEE/ACM Trans. Netw.* **14**, 1260–1271. (doi:10.1109/TNET.2006.886738)

Kelly, F. P. 2000 Models for a self-managed Internet. *Phil. Trans. R. Soc. A* **358**, 2335–2348. (doi:10.1098/rsta.2000.0651)

Kelly, F. P. & Voice, T. 2005 Stability of end-to-end algorithms for joint routing and rate control. *ACM Comput. Commun. Rev.* **35**, 5–12. (doi:10.1145/1064413.1064415)

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. (doi:10.1057/palgrave.jors.2600523)

Kelly, T. 2003 Scalable TCP: improving performance in highspeed wide area networks. *ACM Comput. Commun. Rev.* **33**, 83–91.

Key, P. & Massoulié, L. 2006 Fluid models of integrated traffic and multipath routing. *Queueing Syst.* **53**, 85–98. (doi:10.1007/s11134-006-7588-6)

Key, P., Massoulié, L., Bain, A. & Kelly, F. 2004 Fair Internet traffic integration: network flow models and analysis. *Ann. Telecommun.* **59**, 1338–1352.

Key, P., Massoulié, L. & Towsley, D. 2006 Combining multipath routing and congestion control for robustness. In *IEEE and CISS 2006, 40th Conf. on Information Sciences and Systems, Princeton, NJ, March 2006*. Piscataway, NJ: IEEE Press. (doi:10.1109/CISS.2006.286490)

Key, P., Massoulié, L. & Towsley, D. 2007 Path selection and multipath congestion control. In *Proc. IEEE Infocom 2007 Anchorage*, AL, *April 2007*, pp. 143–151. Piscataway, NJ: IEEE Press. (doi:10.1109/INFOCOM.2007.25)

Kopparty, S., Krishnamurthy, S. V., Faloutsos, M. & Tripathi, S. K. 2002 Split TCP for mobile ad hoc networks. In *Proc. IEEE Globecom 2002*, pp. 138–142. Piscataway, NJ: IEEE Press.

Kunniyur, S. & Srikant, R. 2000 End-to-end congestion control schemes: utility functions, random losses and ECN marks. In *Proc. IEEE Infocom 2000*, pp. 689–702. Piscataway, NJ: IEEE Press. (doi:10.1109/TNET.2003.818183)

Laskowski, P. & Chuang, J. 2006 Network monitors and contracting systems: competition and innovation. *ACM Sigcomm Comput. Commun. Rev.* **36**, 183–194. (doi:10.1145/1151659.1159935)

MacKie-Mason, J. K. & Varian, H. 1995 Pricing congestible resources. *IEEE J. Selected Areas Commun.* **13**, 1141–1149. (doi:10.1109/49.414634)

Massoulié, L. 2007 Structural properties of proportional fairness: stability and insensitivity. *Ann. Appl. Probab.* **17**, 809–839. (doi:10.1214/105051606000000907)

Massoulié, L. & Key, P. 2006 Schedulable regions and equilibrium cost for multipath flow control: the benefits of coordination. In *IEEE and CISS 2006, 40th Conf. on Information Sciences and Systems, Princeton, NJ, March 2006*. Piscataway, NJ: IEEE Press.

Mathis, M., Semke, J., Mahdavi, J. & Ott, T. 1997 The macroscopic behavior of the TCP congestion avoidance algorithm. *ACM Comput. Commun. Rev.* **27**, 67–82. (doi:10.1145/263932.264023)

Padhye, J., Firoiu, V., Towsley, D. & Kurose, J. 2000 Modeling TCP Reno performance: a simple model and its empirical validation. *IEEE/ACM Trans. Netw.* **8**, 133–145. (doi:10.1109/90.842137)

Raina, G. & Wischik, D. J. 2005 Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. In *Proc. EuroNGI Conference, Rome, Italy, April 2005*. Piscataway, NJ: IEEE Press.

Srikant, R. 2003 *The mathematics of Internet congestion control*. Cambridge, MA: Birkhauser.

Voice, T. 2006 Stability of congestion control algorithms with multi-path routing and linear stochastic modelling of congestion control. PhD thesis, Cambridge University.