Optimizing communication networks: a layered control architecture

By Peter Key and Laurent Massoulié

Microsoft Research, Cambridge UK and Thomson Research, Paris, France

We discuss control strategies for communication networks such as the Internet. We advocate a welfare maximization paradigm as the goal of network resource allocation. We examine ways of achieving this objective when communications can use 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 achieve welfare maximization when each communication may use an arbitrary collection of paths, available in an overlay, combined in series and parallel. We conclude by commenting on incentive and pricing dimensions, and open problems related to point-to-multipoint, broadcast or multicast communications.

Keywords: Welfare maximization, marginal congestion cost, layered architecture, transport control, overlays, multipath

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 optimizing 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 application 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 the subject of much current debate and research within the Internet community, often centered around clean-slate approaches or next generation architectures (such as the FIND and GENI initiatives in the US, and in the EU, the FIRE programme). This is because the Internet has failed to evolve architecturally. For instance the current TCP/IP protocol suite, the core protocols of the Internet, has changed little over the last two decades. The reasons for this are many and 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 (see for example, Laskowski and J. 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, since 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 just be delayed. Our approach is general, and the resources could include wired or wireless network capacity, buffering in a switch, or even (server) 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, and jitter (variation in delay) are important from an application's perspective, but there are good reasons to see within-network queueing delay decreasing as network's evolve, leaving latency solely determined by path length and physics. Indeed, Kelly (2000) provides scaling regimes where queuing delays decrease, and others (such as Enacheschu *et al.* 2006) have argued that buffers in the networks should be 'small', meaning $o(C)$ rather than $O(C)$ as at present, where C is the size of the link. This resonates with the introduction of optical switching, since optical buffering is costly.

We shall consider two types of traffic flows, fixed duration traffic, such as streaming or real-time traffic, and file transfers, which have a fixed volume. A flow is associated with some end-points in the networks and comprises a number of packet transfers. By fixed duration or fixed volume we mean that the duration/ or volume may be random, but is independent of the state of the network. In addition, we assume that flows of each type arrive as a stochastic process. We can then model the network as a graph with capacities, and ask how to manage the integration of these two types of traffic.

We first describe what can be done under the current architecture, in §2 , where we describe a utility maximization framework. We use a fluid limit to assess the performance under stochastic demands, and we are able to derive a characterization of the system equilibrium. The bandwidth sharing policy that results has the nice property of maximizing the schedulable region (we can carry as much traffic as possible), and can be achieved by selfish utility-maximizing users.

This assumes both data transfer, and real-time streaming share the same underlying bandwidth allocation policy, for example they both use TCP Reno, the predominant transport protocol in the Internet. However, we give a compact descriptions of the behaviour of the network under end-point admission control for real time streaming, and otherwise everything running over current TCP.

For the current architecture, the only feedback signals to a user are packet loss or delay†. In the event that there is a causal link between resource utilization and price signals which can be fed back to the users, we can have an optimization framework which seeks to maximize social welfare. We discuss the implications briefly in §2, and more fully in §6, where we discuss pricing and congestion pricing.

Multipath routing enables multiple routes between a source and destination, giving performance benefits over single path routing, and overcoming some of the architectural limitations of current networks. When combined with a rate or congestion control, further performance gains are possible (§3). Indeed, actively balancing

† The term 'user' may represent an end-user, or more realistically a protocol a user is running

across paths, using a so-called coordinated congestion controller (§3 a) maximizes the schedulable region and gives performance benefits over passive balancing using parallel paths. When choosing over which path to send data, users should be concerned with the underlying path cost rather than the actual rate they get. Indeed, if they rely on current, RTT-biased TCP, we show that they may receive degraded throughput.

If there are many possible routes between source and destination, we show that using a small number of routes (e.g 1 or 2) combined with a simple randomized selection scheme which allows routes choices to adapt to higher quality routes over time, can perform optimally. In other words, it is *as if* we were able to look at all routes simultaneously. We comment on some of the architectural requirements for multi-path routing, and on how it may be possible to develop a market if Internet Service Providers (ISPs) are able to attract traffic and charge for carrying it.

In § 4 we extend the flexibility given to end-users to allow them to choose both paths in series and in parallel, to optimize their performance. This enables them to fully exploit describe overlay networks, which are built on top of the current network and avoid limitations of the existing networks by using end-hosts to perform functions normally provided by network elements (such as routing, which is normally provided by routes). Indeed, they form the basis of so-call peer-to-peer(P2P) networks.

P2P networks are widely used for file-swarming, and have in addition to solve scheduling problems (deciding what to get from whom), which we mention briefly in § 5.

Our aim in this paper is to survey some of the fundamental ideas in resource allocation, applicable to controlling the current Internet. We show that it is possible to use smart schemes to avoid some of the limitations of the current Internet, without resort to pricing or congestion pricing. However, one can only go so far without some form of pricing, and throughput the paper and in the conclusions we discuss economic incentives and pricing.

2. Modelling and control of the current Internet

(*a*) *The utility maximization framework*

We now describe the framework for bandwidth allocations in a data network proposed by Kelly *et al.* (1998).

Data flows are classified 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 desired data rates $\{x_s\}_{s\in\mathcal{S}}$ should maximize the welfare function

$$W(\{x_s\}_{s\in\mathcal{S}}) := \sum_{s\in\mathcal{S}} n_s U_s(x_s) - \Gamma\left(\{n_s x_s\}_{s\in\mathcal{S}}\right), \tag{2.1}$$

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 use $n_s x_s$ for each flow type $s \in \mathcal{S}$.

For simplicity, we 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 ease of exposition, we also assume that the functions U_s, $s \in \mathcal{S}$ and Γ are differentiable.

In this context, 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 which has these properties is the continuous rate adaptation algorithm

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

where

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

is the so-called marginal cost for bandwidth of type s, and κ_s is some positive gain parameter†.

This framework can be related to flow control in the current Internet in the following manner. Let flow types be associated with a pair of IP source and destination addresses, which is appropriate for considering edge-gateways, or nodes within ISPs if we allow aggregation across source addresses. More specifically for end-hosts in the Internet, we associate flows with the so-called IP 5-tuple, comprising source and destination IP addresses, ports, and protocol type.

Let the network cost Γ be the sum of costs specific to links within the network, where the cost of link ℓ is some (convex, increasing) function Γ_ℓ of the aggregate rate y_ℓ through that link. Assume that routes for one particular flow type are fixed, and set $A_{s\ell}$ equal to 1 or 0 according to whether link ℓ is on the path of type s flow or not, so that

$$y_\ell = \sum_{s\in\mathcal{S}} A_{s\ell} n_s x_s,$$

and

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

Under these assumptions, the marginal cost p_s also reads

$$p_s = \sum_\ell A_{s\ell}\pi_\ell(y_\ell),$$

where $\pi_{ell}(y) = \Gamma_\ell'(y)$. One specific link cost function of interest consists in setting $\Gamma_\ell(y) = \int_0^y \pi_\ell(x)dx$, where

$$\pi_\ell(x) = \left(\frac{x}{C}\right)^B \frac{1 - x/C}{1 - (x/C)^{B+1}}.$$

The function $\pi_\ell(x)$ above is the packet loss probability in a single server queue with capacity C and finite buffer size B (the so-called M/M/1/B queueing system).

† This can be shown by invoking LaSalle's invariance principle. Such a scheme is referred to in the literature as a *primal* algorithm. Description of *dual*, as well as *primal-dual* algorithms, can be found in Srikant (2003) . An enlightening explanation of this terminology can be found in Voice (2006).

Several authors (Matthis *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 delay T_s along the same network path. Specifically, the *square-root formula* states that TCP sets x_s to equal $1/[T_s\sqrt{p_s}]$, expressed in suitable units.

As Kunnyur & Srikant (2000), observe, this can be interpreted as saying that TCP implicitly solves the utility maximization problem (2.1), with cost functions related to the path loss rates, and with utility functions $U_s(x) = -\frac{1}{T_s^2 x_s}$. In other words, the predominant flow control protocol in the current Internet can be thought of as implementing a utility maximization, as in the above framework. See also Kelly (2000) for a more complete discussion.

(*b*) *The case for the utility maximization framework*

We now provide arguments in favour of using the welfare maximization criterion given above as a suitable flow-control objective.

First, this objective can be practically achieved. Indeed, distributed schemes exist where data sources adapt their sending rates to the observed marginal cost of their network path, and network links adapt their marginal cost signal to the current flow they carry, leading to the desired allocations.

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 p_s reflects accurately the network costs, then the above framework is in line with microeconomic theory, with its corresponding emphasis on welfare maximization.

There are some subtleties here. Welfare maximization is a proper objective of a central planner. However, if the entities that determine rate regulation at the sources of data flows are selfish, their goal is to maximize their own individual utility, not the overall welfare. This raises the question of whether the congestion costs p_s are simply feedback signals, provided to help data sources adapt their rates, or whether data sources are actually charged for each unit of data sent, at the marginal cost of the corresponding network path.

In the former case, then there is no incentive for a selfish data-source to back off. In the latter case, the source's aim is to maximize its net benefit, $U_s(x_s) - x_s p_s$. If variations of x_s do not affect p_s, then indeed the selfish user will aim at equating its marginal utility $U_s'(x_s)$ with p_s: "price takers" have their selfish objective aligned with welfare maximization. Price makers, on the other hand, will typically aim to align $U_s'(x_s)$ with $p_s + x_s \frac{\partial p_s}{\partial x_s}$, a slight departure from the behaviour of price makers.

Nevertheless, Johari &Tsitsiklis (2004) established that, for data sources that are selfish price makers, the resulting welfare is no less than 3/4 times the optimal welfare†. To summarize, the motivation from microeconomics theory is predicated on data sources being effectively charged at the congestion cost.

The previous arguments rely implicitly on a scenario where data sources are long-lived, and the collection of active flows is essentially static. This is in sharp

‡ the Transmission Control Protocol

† In this discussion, we assume that the network resources set prices p_s exactly equal to marginal costs. The effect of selfish, revenue maximizing network resource providers is discussed in Ganesh *et al.* (2007)

contrast with the current usage of the Internet. There, data transfers have a limited duration. This duration typically depends on the achieved transmission rate for transfers of files that have an intrinsic volume. In that case, the utility of the transmission is related to the transfer duration. Real-time transmissions differ from file transfers: they have an intrinsic, fixed duration. Several utility models for real-time transfers are possible, depending on whether variable coding rates can be used or not.

We now elaborate on these aspects, and describe models of control and network performance in a dynamic setting. We shall establish desirable properties of utility maximizing allocations in such dynamic environments.

(*c*) *Flow dynamics under utility maximizing, "best effort" rate allocations*

We assume that for each type s, there are $n_s$file transfers and m_s real-time 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. Type s real-time connections are assumed to commence at instants of a rate κ_s Poisson process, and to last for an exponentially distributed duration, with parameter η_s.

We also assume that transfers, once initiated, take place in the network (i.e. there is no admission control, or rejection policy), and obtain a welfare maximizing rate allocation. That is, type s flows all receive a bandwidth share x_s such that the vector $\{x_s\}_{s\in\mathcal{S}}$ maximizes (2.1), where n_s should be replaced by (n_s+m_s), the total number of type s flows in our integrated traffic model. Where needed, we denote this rate allocation by $x_s(n+m)$ to stress its dependency on the current population of active flows.

The latter assumption can be motivated as follows. It 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. The appropriateness of such a best-effort service model can be questioned. For real-time transfers in particular, variable rate encoding should be implemented if the data consumer is to receive useful data in real time.

With the above statistical assumptions, the numbers n_s and m_s follow a stochastic evolution. However, to analyze the system's performance, we shall instead consider deterministic evolution equations, which correspond to the mean drift of the variables of interest in the original stochastic model†. Specifically, we consider the following differential equations

$$\frac{d}{dt}n_s = \nu_s - \mu_s n_s x_s(n+m), \quad s\in\mathcal{S} \tag{2.2}$$

$$\frac{d}{dt}m_s = \kappa_s - \eta_s m_s, \quad s\in\mathcal{S}. \tag{2.3}$$

The cost function Γ is again assumed to be convex. We let $\mathrm{dom}(\Gamma)$ denote the set of rate vectors z for which $\Gamma(z)$ is finite. We assume for convenience that utility

† 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.

functions U_s are differentiable, and that the cost function Γ is differentiable on the topological interior $\mathrm{int}(\mathrm{dom}(\Gamma))$ of its domain.

In this context, we introduce the following stability condition:

$$\rho \in \mathrm{int}(\mathrm{dom}(\Gamma)), \tag{2.4}$$
$$\exists \delta \in (0,\infty)^{\mathcal{S}} \text{ so that } U_s'(\delta_s) > \Gamma_s'(\rho+\delta), \quad s \in \mathcal{S}, \tag{2.5}$$

where Γ_s' denotes the s-th partial derivative of Γ, and ρ is the vector of loads ρ_s for each file transfer type s, defined as $\rho_s = \nu_s/\mu_s$.

The following result is established in Key & Massoulié 2006:

Theorem 2.1. *Under the conditions (2.4,2.5), the differential equations (2.2,2.3) have a unique invariant point,* $(\hat{\mathbf{n}}, \hat{\mathbf{m}})$. *It takes the form*

$$\hat{m}_s = \kappa_s/\eta_s, \quad \hat{n}_s = \rho_s/\hat{x}_s, \quad s \in \mathcal{S}, \tag{2.6}$$

where the equilibrium allocation $\hat{\mathbf{x}}$ *is the unique solution of*

$$\textit{maximize} \quad \sum_{s\in\mathcal{S}} \hat{m}_s U_s(x_s) - \Gamma\left(\hat{\mathbf{m}}\mathbf{x} + \rho\right) \tag{2.7}$$
$$\textit{over} \qquad x_s \geq 0, \quad s \in \mathcal{S}. \tag{2.8}$$

Moreover, it is possible to show (Key & Massoulié 2006) that all trajectories of the differential equations (2.2) and (2.3) converge to the equilibrium point. The main step of the proof consists in showing that the dynamics of file transfers in isolation (given by (2.2) with $\mathbf{m}$ set to 0) are asymptotically stable. This property follows by showing that the function

$$L(\mathbf{n}) := \sum_{s\in\mathcal{S}} \frac{1}{\mu_s}\left[\int_0^{n_s} U_s'\left(\frac{\rho_s}{x}\right) dx - n_s \Gamma_s'(\rho)\right]$$

is a Lyapunov function for these dynamics.

The invariant point identified in (2.1) has the following interpretation. The file transfers of type s contribute an irreducible load ρ_s on each resource they are associated with. The streaming (real-time) traffic then shares out what remains after the load has been accounted for. This in turn determines the rate that flows of type s receive and hence under our sharing assumptions also determines the rate that file transfers of type s receive.

Apart from the quantitative identification of limiting configurations, the previous result also implies the following qualitative property. When Conditions (2.4,2.5) are in force, response times will remain finite for all file transfers. Note that when ρ lies outside $\mathrm{dom}(\Gamma)$, this property must fail. Indeed, there must exist non-negative weights v_s, $s \in \mathcal{S}$ and a positive constant w such that the hyperplane

$$H := \{\mathbf{n} : \sum_{s\in\mathcal{S}} v_s n_s = w\}$$

separates ρ from $\mathrm{dom}(\Gamma)$, by convexity of Γ and the fact that Γ is non-decreasing. This implies that the weighted sum

$$\sum_{s\in\mathcal{S}} v_s n_s(t)/\mu_s$$

must diverge to infinity as $t \to \infty$. Therefore finite transfer times require ρ to belong to dom(Γ). In this sense, Condition (2.4) is "close to necessary" for finiteness of response times. Condition (2.5) is also "close to necessary": if $U'_s(0) \leq \Gamma'_s(\rho)$, then the convex Lyapunov function $L(\mathbf{n})$ has no finite local minimum, and hence trajectories $\mathbf{n}(t)$ must diverge to infinity.

The suitability of utility maximizing allocations for real-time flows is less obvious. As mentioned earlier, in the absence of variable-rate encoders, such real-time flows must proceed at a fixed rate. However in this case, denoting by π_s the desired rate of type s real-time flows, we may adapt the above best effort scheme as follows. Upon arrival, a type s real-time flow enters the system with probability $\min(1, x_s/\pi_s)$, where x_s is the utility maximizing allocation for type s flows. If it enters, it sends at a rate π_s, irrespective of the value of x_s. More generally, we may allow a real-time flow to choose one of a number of fixed rates, a limited form of rate adaptation in use today.

A similar scheme has been proposed by Karlsson *et al.* (2004). It can be shown (Key *et al.* 2004) that, when the equilibrium points $\hat{x}_s$ identified in the above Theorem are such that $\pi_s \geq \hat{x}_s$, then the equilibrium characteristics under the latter admission control scheme are as follows. The number of file transfers of each type s is still given by $\hat{n}_s = \rho_s/\hat{x}_s$, whereas the number of real-time transfers now reads $\hat{m}_s = (\kappa_s \hat{x}_s)/(\eta_s \pi_s)$. However the total load contributed by type s real-time transfers is unchanged, and equals $\hat{x}_s \kappa_s/\eta_s$.

Therefore the proposed admission control scheme, which may be better suited for real time traffic in the absence of adaptive encoding rate, achieves a performance closely related to that of the previous best effort scheme.

3. Multipath Routing and Congestion Control

We now show how combining multipath routing and congestion control fits naturally within a utility maximization framework.

Multipath routing, where a flow can choose or use several paths rather than one, provides performance and reliability benefits over single path routing. It is also a potential way of avoiding limitations of current architectures, where routing choices are dictated by network providers rather than end users. Additional benefits are obtained when congestion control and multipath routing are combined - offering the opportunity of cross-layer optimization. Benefits also accrue to service providers, since multipath routing enables resource pooling, where a set of resource can be considered as one.

Performance benefits are gained at different timescales. Using multipath routing gives reliability benefits, with reliability interpreted as the probability a path to between source and destination exists, a measure of connectivity. Path failures occur when resources (nodes or links) fail or are unavailable, which are relatively rare events, with mean time between failures expressed in days. The greatest benefits are obtained when the paths are disjoint, thus avoiding common mode failures.

Performance benefits expressed in terms of throughput operate on timescale of changes in congestion epochs and changes in traffic patterns, which operate on timescales ranging from sub-second (for congestion) to minutes, for traffic. To gain benefit here, multipath routing needs to be combined with congestion control. This enables routing on 'least-cost' routes, and in aggregate enables resource pooling.

Informally, such combined routing and congestion control is able to optimally pack traffic onto the network, in a way we shall make precise below.

Resource pooling gives significant performance benefits for networks, both for users and network operators, which are greater for queuing networks than for loss networks, see for example Kelly (1995). For instance, replacing two single queues which serve separate load by a combined queue which the combined capacity of the individual queues can halve the mean response time; a benefit to the user. Fast packet networks may behave as loss or queuing networks - performance depends on the timescale considered (for example, packet or flow) and the appropriate performance measure seen at a buffered resource, which also depends on the relative size of the buffer with respect to the service rate.

However, in either case *cut-set dimensioning* simplifies dimensioning: if the network is represented as a capacitated graph, then cut-set dimensioning allocates sufficient capacity to certain cutsets of the graph, rather worrying about the capacity on isolated resources. In other words, providing performance guarantees implies provisioning sets of resources adequately rather than individual resources. The assumption being that dynamic routing will find spare resource if it there, the exact location being less important. It follows from the max-flow min-cut theorem, that cut-set dimensioning typically gives sufficient conditions to provide satisfactory performance, however in certain cases we can derive simple necessary conditions. Provisioning becomes simpler and more robust. A simple example is given in Key *et al.* 2006a. References to generalized cut-sets for queueing networks can be found in Kelly (1995).

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

For multipath routing we assume users s can use path from the set $\mathcal{R}(s)$. With joint multipath routing and congestion control, for fixed numbers of users of type s, N_s the optimal rates for a class s users solve the welfare maximization problem

$$\text{Maximize } \sum_s N_s U_s \left(\sum_{r \in \mathcal{R}(s)} x_r \right) - \Gamma\left(N\mathbf{x}\right) \tag{3.1}$$

$$\text{over } x_r \geq 0, \quad r \in \mathcal{R}. \tag{3.2}$$

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

$$U_s' \left(\sum_{r \in \mathcal{R}(s)} x_r \right) \leq \Gamma_r'(N\mathbf{x}), \tag{3.3}$$

$$U_s' \left(\sum_{r \in \mathcal{R}(s)} x_r \right) < \Gamma_r'(N\mathbf{x}) \Rightarrow x_r = 0, \tag{3.4}$$

and, as a consequence, the allocation for user s only puts a non-zero allocation on paths whose price p_s 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 doesn't (i.e. allocates zero to these paths) the prices must be higher than ps.

When there are stochastic arrivals, by taking the fluid limit, we have convergence to the equilibrium point, $(\hat{\mathbf{n}}, \hat{\mathbf{m}})$, as in §2, that takes the form

$$\hat{m}_s = \kappa_s/\eta_x, \quad \hat{n}_s = \rho_s/\hat{x}_s, \quad r \in \mathcal{S}. \tag{3.5}$$

Moreover, under the stability conditions (2.4,2.5), the system of differential equations is asymptotically stable, in the sense of Lypaunov, and converges to the unique equilibrium point (see Key and Massoulié, 2006).

Consider the case that the penalty functions for path r decomposes into a sum of associated link penalty functions, $p_r = \sum_{j\in r} \Gamma'_j(\hat{y}_j)$, where

$$\hat{y}_j = \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}(s)} A_{jr}\left(\hat{n}_s + \hat{m}_s\right)\hat{x}_r \tag{3.6}$$

In equilibrium, let ρ_{rs} denotes the fraction $\hat{n}_s x_{rs}$ of load ρ_s that is carried on route r. Then this coordinated control chooses the load fractions to minimize the overall cost

$$\sum_{j\in J} \Gamma_j \left(\sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}(s)} A_{jr}\left(\rho_{rs} + \hat{m}_s \hat{x_r}\right) \right). \tag{3.7}$$

Moreover, when no streaming traffic is present, multipath routing splits load optimally across paths, to minimise $\sum_{j\in J} \Gamma_j \left(\sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}(s)} A_{jr}\rho_{rs}; C_j \right)$ which is *independent* of the specific form of utility functions U_s chosen.

Note that distributed rate control algorithms exist for all of the above optimization problems, e.g. Kelly & Voice (2005), Han *et al.* (2006). Kelly and Voice 2005 describe per-route controllers which are stable in the presence of feedback delays and where the gain parameters associated with each route are related to the corresponding round trip time.

(*b*) *Uncoordinated parallel routing*

A simpler form of multipath routing just uses parallel routing, or uncoordinated control, which solves

$$\text{Maximise} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}(s)} N_r U_r(x_r) - \Gamma(N\mathbf{x}). \tag{3.8}$$

over $x_r \geq 0$. This is simpler in the sense that it is potentially easier to implement in practice, since the underlying transport protocol does not need to be altered. In contrast to the coordinated controller, there is a utility per path, which may differ among the paths available to user s, where $N_s = \sum_{r\in\mathcal{R}(s)} N_r$. As an example, the utility on each path may represent a TCP controller, where the utility function on each path depends on the path and where the resulting throughput is inversely related to the round-trip time on each path. In the event that $U_r = U_s$ for all

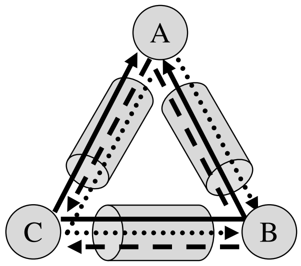

Figure 1. Example network where the use of parallel uncoordinated connections is inefficient

$r \in \mathcal{R}(s)$, Jensen's inequality shows that coordinated control gives at least as large welfare as the uncoordinated control.

When there are stochastic arrivals, by taking the fluid limit, we have convergence to the equilibrium point. However, the stability results for the system do not have such a simple representation as for uncoordinated routing. Indeed, we now give an example where parallel routing gives strictly smaller schedulable region than with a coordinated controller.

A coordinated router actively balances flows across paths, whereas uncoordinated parallel routing has a more limited form of load balancing - sending amounts on each path inversely proportional to the price on each path. For example, in the case that there are but two paths, with different prices, the coordinated controller will send all the traffic on the least cost path, whereas the uncoordinated controller will use both paths.

(*c*) *Uncoordinated vs coordinated control: the benefits of balancing across paths*

We now demonstrate that actively balancing across paths has advantages. We assume that only file transfers are present, which involves no loss of generality, since the effect of streaming traffic can captured through a redefined cost function (see Key and Massoulié 2006). For convenience, we look at the cost function which represents sharp capacity constraints, where $\Gamma = \sum_j \Gamma_j$ and

$$\Gamma_j(y; C_j) \begin{cases} 0 & \text{if } y \le C_j \\ \infty & \text{otherwise.} \end{cases} \tag{3.9}$$

where C_j is the capacity of resource j.

Consider the triangle network of Figure 1, with symmetric routing and capacities. There are three links, each has unit capacity, three source-destination pairs (such as A-B), and flows between pairs can either use the direct single-hop link (e.g. A-B) or the alternative two-hop path (eg A-C-B). Label the source destination pairs by the opposite vertex, so that ρ_C denotes the offered load from A-B. It is straightforward to show that the capacity region is

$$\rho_i + \rho_j \le 2, \quad \forall i, j \in \{A, B, C\}, i \ne j. \tag{3.10}$$

When all the offered loads are equal, the stability condition becomes $\rho \leq 1$, and stability is achieved without using alternate routes.

We now look at what happens under parallel routing, when each file transfer uses both paths (single hop and two-hop) in parallel. We assume that the allocation is so-called α-fair, with equal weights per connection, meaning that the utility functions are given by $U(x) = x^{(1-\alpha)}/(1-\alpha)$, for some $\alpha > 0$. Let n_i be the number of type i connections. Then type i connections receive an allocation of x_i, where

$$x_i = \left(\frac{1}{p_i}\right)^{1/\alpha} + \left(\frac{1}{p_j + p_k}\right)^{1/\alpha}, \quad i \neq j \neq k$$

where the p_i's, the Lagrange multipliers associated with capacity constraints. They are uniquely determined from

$$n_i \left(\frac{1}{p_i}\right)^{1/\alpha} + n_j \left(\frac{1}{p_i + p_k}\right)^{1/\alpha} + n_k \left(\frac{1}{p_i + p_j}\right)^{1/\alpha} = 1, \quad i \neq j \neq k$$

For symmetric offered load, the fluid limit equations are

$$\frac{d}{dt} n_i(t) = \nu - \mu x_i(\mathbf{n}(t)), \quad i = A, B, C. \tag{3.11}$$

We then have the following result,

Theorem 3.1. *Define ρ as λ/μ, and*

$$\rho^* := \frac{1 + 2^{-1/\alpha}}{1 + 2^{1-1/\alpha}}. \tag{3.12}$$

The solution $\mathbf{n}(t)$ to the system of differential equations (3.11) 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 $\mathbf{n}(t)$ decreases to zero in time at most $\theta[n_A(0)+n_B(0)+n_C(0)]$ for a suitable constant $\theta > 0$ whenever $\rho < \rho^*$.

The proof can be found in Key and Massoulié 2006. It is possible to show that ρ^* is indeed the exact capacity of the triangle network under symmetric loads. In other words, the schedulable region is smaller than for coordinated routing, and indeed approximately 30% less traffic can be offered when parallel routing is used instead of a coordinated controller. This is reminiscent of results in loss networks, where dynamic routing can lead to instabilities in symmetric networks at high load, and perform worse than single path fixed-routing.

Even when the schedulable regions are the same, parallel routing can incur higher network cost. For example, for the network of Figure 2, where each file transfer type r can use several paths $p \in \mathcal{P}_r$, and each such path consists of a single link, the stability region is has the form

$$\sum_{s \in S^*} \rho_s < \sum_{j \in \mathcal{L}(S)} C_j, \quad \forall S^* \subset \mathcal{S}, \tag{3.13}$$

where $\mathcal{L}(S^*)$ is defined to be the union of the sets $\mathcal{R}(s)$ for all $s \in S^*$. In this case, the schedulable regions are the same, however the network cost can larger — as shown in Key and Massoulié 2006.

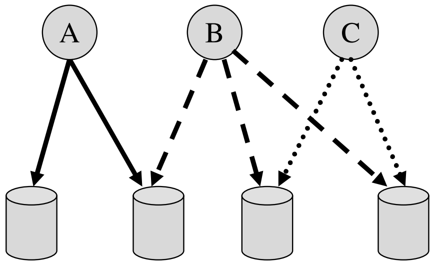

Figure 2. Example network with 1-hop routes: parallel uncoordinated connections achieve maximal stability region

(*d*) *Path selection and underlying path costs*

When choosing what path to send on, it is important that users chose react the correct underlying path cost, rather than merely looking at the rate they receive. For example, when using TCP Reno, the throughput is inversely related to the path RTT, whereas the network 'cost' is normally independent of an RTT component, and determined by the aggregate load on a resource. This distorts the incentive for end-systems, who are trying to trade-off throughput against RTT on different routes, and can lead to sub-optimal behaviour, as we now demonstrate. Recall §2 where we model TCP's behaviour by using utility functions of the form

$$U_r(x) = -\frac{1}{T_r^2 x}$$

(using the approach of Kelly 2001, for example). We consider the hexagonal network

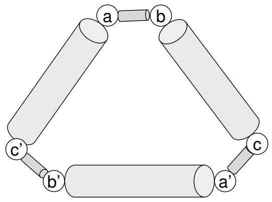

Figure 3. Network with alternation of fat and long with short and thin links.

of Figure 3, 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 time make. Now let each users independently and selfishly seek to chose the set of paths that maximize their rate. In other words, they are able to chose the path sets $\mathcal{R}(s)$ and 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 $\sum_r N_r x_r$. By allowing $\mathcal{R}(s)$ to include all possible route sets, it is sufficient to consider the allocation $\{N_r\}$. It is possible to show that the equilibria that results from users selfishly seeking to optimize their throughput is a Nash equilibrium, provided that an equilibrium exists.

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 short link. To fully utilize each thin link, each connection will receive a rate of $c/(2N)$; for this to happen, the Lagrange multipliers p must satisfy

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

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 that

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

But in this case the throughput is *half* what is would be if all connections were $l-s-l$.

If the RTT bias is removed, so that utility functions U_r are independent of path RTT, then in both the uncoordinated and the uncoordinated case, the Nash equilibrium is efficient and solves the global social welfare problem (3.1),(3.2).

But this assumes that end users have full access to all possible routes, and their states — for example the path quality or price on each route. In practice, when the set of possible routes is large, it may be infeasible to try all possible routes simultaneously, or assess their quality. What can we do in such circumstances? It turns out that provided we can sample routes dynamically in time, and move flows to higher quality routes, that we can do as well as if we had full knowledge of all routes. In other words, we attain the optimum. There are caveats when there can is a dependence on round trip time in our controllers/ utility functions. We now make these statements more precise.

(*e*) *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 only uses a small number of paths at any moment in time, while choosing to alter these paths over time as better ones become available. This is particularly appropriate when the end-systems are edge devices (gateway nodes), and have a large number of paths available. While if the end-systems are end-hosts, there are a practical limitations on opening 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

$$\sum_{c \in \mathcal{C}(s)} N_c = N_s, \quad s \in \mathcal{S}. \tag{3.14}$$

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

$$\text{Maximize} \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}(s)} N_c U_s \left(\sum_{r \in c} x_{c,r} \right) - \Gamma(N\mathbf{x}) \tag{3.15}$$

over $x_{c,r} \geq 0$ where $N\mathbf{x}$ 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 timescale compared to the timescale of arrivals and departures. The new route set is accepted provided that the *net benefit* that the user receives from the new route set is higher than that of the current route set. Users of type s can explore the entire space provided that for all s, for any $r \in cR(s)$ 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 the data rate obtained by users streaming along routes $r \in c$. This is the sum of the rates $x_{c,r}$ over $r \in c$, where $x_{c,r}$ is the sending rate along route r:

$$x_c = \sum_{r \in c} x_{c,r}.$$

We assume that uses use the following form of rate adaptation (see Kelly *et al.* 1998):

$$\frac{d}{dt} x_{c,r} = \kappa_{c,r} \left[U'_{s(c)}(x_c) - \partial_r \Gamma(X) \right] + \mu_{c,r}, , \tag{3.16}$$

where the term $\mu_{c,r}$ is non-negative and 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

$$B_c = U_s\left(x_c\right) - \sum_{r \in c} x_{c,r} p_r,$$

where

$$p_r = \partial_r \Gamma(X_r).$$

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

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

$$\dot{N}_c = \sum_{c'} N_{c'}\lambda_{c'c}\phi(B_c - B_{c'}) - \sum_{c'} N_c\lambda_{cc'}\phi(B_{c'} - B_c). \tag{3.17}$$

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

We can then show that this rate adaptation scheme and route selection policy does indeed maximize the welfare, in that it maximizes (3.15), and moreover solves the global optimization problem, (3.1,3.2). We shall assume the utility functions U_s and the penalty function Γ are continuously differentiable on their domain, that the former are strictly concave increasing, and the latter convex increasing, and that $U_s'(x) \to 0$ as $x \to \infty$. We the have

Proposition 3.1. *For a large population of users N_s, where users change from path c to c' at instants of a Poisson process with rate $\lambda_{cc'}$ provided they receive a higher rate on path c', then any absolutely continuous solution $(N_c, x_{c,r})$ to the system of ODE's defined by (3.16,3.17) converges to the set of maximizers of the welfare function*

$$\mathcal{W}(X,N) := \sum_{s\in\mathcal{S}}\sum_{c\subset\mathcal{R}(s)} N_c U_s(X_c) - \Gamma(X) \tag{3.18}$$

under the constraints (3.14). The corresponding equilibrium rates (x_r) are solutions of the coordinated welfare maximization problem (3.1–3.2).

This is a powerful result, which says that path re-selection 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 timescale dictated by the 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 round-trip time 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 gave equivalence with the coordinated controller we need to replace the utility function $U_s(x)$ for the coordinated controller by $x \to bU_s(x/b)$.

If we do not use resampling, how many paths. b, should we use to achieve good performance? For the example of Figure 2, where each resource node has capacity one, and we have fixed traffic demand (non-stochastic), with offered load of N, then in a perfect allocation all source nodes receive a rate of one and all resource nodes are fully utilized. If the resource allocation is less than perfect, some nodes will receive less than one unit: if we look at the minimum rate across nodes, this gives a performance measure related to unfairness or lack of load balance. Theory suggests $b = O(\log N)$ for a perfect allocation (with high probability). Experimental results for $N = 100$ suggest that with $b = 3$, for a coordinated allocation the minimum rate is over 0.95, whereas for an uncoordinated control is it only 0.5, and for $b = 10$ it is 0.8 (see Key *et al.* 2006a).

Peer-to-peer (P2P) applications such as Skype or BitTorrent effectively implement receiver driven-multipath congestion controllers. In BitTorrent, it answers the the question of who is a good peer to download from. A node is given a list of peers (say 50) and chooses a number of these to actively download from a fixed number of peers (say 4). It chooses the best set of peers by uploading to those from which it receives the best download rate, which both deters free-riders, and acts as an example of greedy-path selection. A node periodically examples a random node from its peer set to upload to and assess download quality: the randomization aspect allows a full search of the peer set in time. Transfers take place using TCP, where throughput depends on the round trip time, hence this is an example of a multipath uncoordinated congestion controller, which uses a greedy path selection algorithm to find best paths.

(*f*) *Architectures and charging*

To implement sender-side 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 AS's (autonomous systems), perhaps provided for resilience. For a home or "eye-ball" subscriber, currently there is 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.

Provided more that one ISP serves a neighbourhood, wireless can be used to access more than one ISP. Loose source-routing is currently under discussion again within the community, where a node is provided with a set of paths, and makes a choice between them. This is all we require for multipath to work: we do not need full source routing, where the source specifies the full path to the destination. The advent of IPv6 allows different addresses to be used for different interfaces for the same host, and both multipath and addressing is currently under discussion in the 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 eye-ball customers) or on the basis of 95%-ile charging for large customers connected by large fibre-optic pipes, with little relation to quality. In the UK, there is often a download limit associated with domestic tariffs, giving a volume component to the price. There is an incentive for ISPs to carry enough traffic to earn higher revenue, if it is more profitable. With respect to home users' traffic, once the higher tariff has been paid for, there is only an indirect incentive to carry the traffic. Additionally, ISPs have to pay transit ISPs for transit traffic, hence they have incentive to limit such traffic, and offload to peering points.

At the routing level, these incentives translate into policy-based routing decisions, used by ISPs to decide how route between themselves (using BGP, the Border Gateway Protocol). It also encourages "hot-potato" style routing, where packets are passed off to other ISPs as soon as possible, so that others bear the cost of carrying them. ISPs also keep information relatively private, this coupled with mixed incentives makes it unclear as to whether the current end-to-end routing, seen from a global or system perspective, is in any sense optimal.

But 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.* (2005), which could be advertised by a DNS-type service, the user could choose which stepping stone router to user. If advertised by an ISP then the ISP would then forward the traffic toward the destination, and the ISP might choose to advertise a set of such routers and vary the choice in time. The question of how to recover costs or charge end-users is vexed, and doesn't directly fit into current charging models for the Internet. But 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. With VOIP, alternate routes are being used, though not under necessarily under the control of the end-host.

4. Combinations of paths in series and parallel

As we have just discussed, suitable congestion controllers could combine paths in parallel to serve one data transmission, and still achieve the goal of aggregate utility maximization. Other possibilities exist. For instance, a source node S could send data to a destination node D, via a relay node R (R could be either an overlay node or a *stepping stone router* as previously discussed) 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 when R is a middle-box or proxy.

More generally, one could try to exploit arbitrary network paths made available to a particular transmission, in order to achieve welfare maximization for the overall source-destination flows. We discuss flow controllers which achieve this objective, by generalizing the notion of coordinated controllers.

Assume that for each flow f, with source $s(f)$ and destination $d(f)$, a collection of paths $\pi \in \Pi(f)$ can be used, 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^f_{ij} the rate sent for flow f along path (ij), and by W^f_i the amount of data buffered at i for flow f, with the convention that

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

In this context, the utility maximization problem that we would like rates x^f_{ij}

to solve is the following:

$$\text{maximize } \sum_f U_f\left(x^f\right) - \Gamma(y) \tag{4.1}$$

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

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

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

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

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

We consider the following scheme: for each path $(ij) \in \Pi(f)$, we set

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

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 is no data to forward from node i to node j at that particular time. This can be addressed in two different manners: 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

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

Thus, the vector y of path rates used in the definition of marginal costs in (4.7) reads

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

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 ODE's (4.6–4.11). This can be addressed by adding to the right-hand side of (4.6) a time-dependent, non-negative term u_{ij}^f such that $u_{ij}^f = 0$ if $x_{ij}^f > 0$.

Let us 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

$$\begin{aligned} &\textstyle\sum_j y_{ji}^f = \sum_j y_{ij}^f, \\ &W_i^f > 0 \Rightarrow y_{ij}^f = x_{ij}^f. \end{aligned} \tag{4.12}$$

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

$$\begin{aligned} &\mathbf{1}_{i=s(f)} U_f'(x^f) + \phi(W_i^f) - p_{ij} - \phi(W_j^f) \le 0, \\ &x_{ij}^f > 0 \Rightarrow \mathbf{1}_{i=s(f)} U_f'(x^f) + \phi(W_i^f) - p_{ij} - \phi(W_j^f) = 0. \end{aligned} \tag{4.13}$$

By (4.12), nodes i at which $\sum_j y_{ij}^f < \sum_j x_{ij}^f$ are such that $W_i^f = 0$. In view of (4.13), at such nodes i, it holds that either $x_{ij}^f = 0$, or $W_j^f = p_{ij} = 0$ for all outgoing paths $(ij) \in \Pi(f)$.

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

$$U_f'(x^f) = p_{s(f)i_1} + p_{i_1 i_2} + \ldots + p_{i_m d(f)}.$$

Furthermore, by the second part of (4.13), any concatenation of paths $(s(f)i_1)$, $\ldots$,$(i_m d(f)$ al ong which some rate x_{ij}^f equals zero is such that

$$U_f'(x^f) \le p_{s(f)i_1} + p_{i_1 i_2} + \ldots + p_{i_m d(f)}.$$

These last two properties coincide with the Kuhn-Tucker optimality condition for the welfare maximization problem (4.1–4.2). Thus, 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), and in the work of Srikant and 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, $\phi(W_i^f)$ 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 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 up-stream nodes.

5. More general traffic models: Multicast

The examples we have considered so far have tacitly assumed that there is a source node (or user, s), and a path to a "destination", a unicast connection. Other types of communication occur in the Internet: for example, multicast, where data is sent to several receivers simultaneously. It is distinguished from broadcast, in that only specific receivers are used. The particular case of a single source, can be handled within our utility maximization framework by having a route comprise a set of resources that provide paths to all the possible receivers.

An efficient way to deliver multicast is via distribution trees, connecting source(s) and receivers, where information is replicated at parent nodes and sent to child nodes; this contrasts with the most inefficient way of delivering multicast, which is to create N unicast sessions for N receivers. Consider the example of a network represented by a capacitated graph, where edge capacities represent hard capacity constraints. The max-flow min-cut theorem says that maximum flow between the source and a given receiver is given by the minimum cut in the graph. For unicast flows to all receivers, we have to take the union of the min-cuts across receivers to provide the required capacity for multicast , however the broadcast theorem of Edmonds 1972 shows that it possible to service all receivers at a rate determined by taking the minimum of the min-cuts (which is equivalent to saying that we can take the maximum for the min-cuts across receivers rather than the sum if we want to provide the maximum flows in the original graph to receivers).

However, IP multicast has failed to be widely deployed, partly because of the problem that partial deployment is difficult - it needs to enable at all nodes on a path, and partly because there is no incentive for downstream ISPS to deploy, since they do not see the efficiency gains. Instead, under the current charging model, they have an incentive to carry n unicast flow if they receive transit fees, rather than create a multicast flow which looks to them like a packet amplification.

Instead, peer-to-peer (P2P) multicast solutions have sprung up, initially used for file sharing, (such as Kazza, Napster etc). BitTorrent, which is a cooperative file swarming protocol, is a particularly interesting example for two reasons. First, in its original form it used a tit-for-tat mechanism to encourages users to swap pieces of a file with each other. Secondly, as we have already mentioned it uses a form of receiver-driven multipath routing with congestion control. In BitTorrent, a large file is broken into a number of smaller pieces, or chunks. Users who want this file register with a tracker, receive a part of the file, and are given the identity of a number of peers who also want the file and whom they can swap chunks with.

Such file-swarming applications need to solve a scheduling problem, deciding what pieces of a file to retrieve from whom, in addition to finding peers that can transfer at a high rate. For real-time live streaming applications, such P2P delivery also needs to ensure timely delivery of content. The scheduling problems can also be posed as optimization problems, but fall outside the scope of this paper .

6. Pricing and control

Congestion pricing is attractive economically and theoretically, and for fast-packet networks can create a version of the 'smart-market', first proposed by Varian & Mackie-Mason (1995). 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.

Moreover, if pricing is used as a means of control, different forms of user-behaviour may follow. For example, a more sophisticated real application than we have considered to date would require some minimum rate $xmin$ to function at all, but above this rate would employ an adaptive codec to vary the rate at which data is sent in the light of network conditions. This is equivalent to a utility function $U_k(x)$, for type k, which is zero for $x < xmin$ and concave for $x \geq xmin$. If the duration is T_k, then the overall utility of allocating a time-varying rate $x(t)$ equals $\int_0^{T_k} U_k(x(t))dt$. Conversely, file-transfers are usually more interested in minimizing the time to download, rather than the instantaneous rate they received. In other words, for a file transfer of type i, the utility of transferring in T seconds is $\tilde{U}_i(T)$ for some decreasing function $\tilde{U}_i$, and suppose they are limited to send at some maximum rate x_i. If now we apply a utility maximization framework, where users are charged a path-dependent price p_j per packet, then the following strategy is optimal: there is a critical price p_i^* such that file-transfer users connect if and only if the price is less that p_i^*, in which case they will send at the *maximum* rate x_i, otherwise they wait. In other words they perform self-determined admission control. While real-time users also have a critical price, above which they will not enter, p_k^*, corresponding to them receiving less than $xmin_k$. However if the price is below this they will *rate adapt* to this rate. This is contrary to current policies in the Internet, where, as we have described, file transfers are seen as 'elastic', and hence adapt their rate, while admission control is seen as only appropriate for real-time flows.

But this assumes price signals can be fed back to the users, perhaps via an intelligent packing marking scheme. Indeed, Intelligent packet marking has advocated as a means of more intelligently signalling congestion than the current practice of discarding packets. The simplest scheme would convey such information via single bit, such as the ECN bit† (Flloyd and Fall 1999), yet even this has not been adopted in the Internet, because of issues of backward compatibility and needing both hosts and end-systems to behave correctly when the bit has set. Briscoe *et al.* 2005 have proposed using an edge-based deployment of congestion notification, which may enable partial deployment in managed networks. But 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 based, quality based of 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.

† the IETF RFC actually uses 2 bits in the IP header to allow for negotiation and to provide a form of security via a nonce.

7. Concluding Remarks

We have shown that a utility maximization framework is a natural one for addressing questions of resource allocation, 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 standard's 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 such smart solutions produce versatile utility-maximizing allocations, which have nice properties of maximizing the schedulable region, and minimizing network cost. Indeed, some current P2P applications already implement variants of these controllers. Such controllers can work with current TCP, however the round-trip time 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 the this framework is real-time streaming or real-time applications in general, since it is not possible to give hard quality of service guarantees over a best-effort network. It is also not easy to implement rate-adaptive real-time streaming, since there is little incentive for an application to limit its rate. At the time of writing, the volume of real-time traffic is small, when compared to 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 is one where some form of pricing is used to relate path prices to real prices, via congestion prices. This brings nice advantages in terms of controllability and incentive compatibility, and allows for quality of service 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

Ben Fredj,S., Bonald, T., Proutière, A., Régnié, G. & Roberts, J. 2001 Statistical bandwidth sharing: a study of congestion at flow level. In *Proceedings of SIGCOMM 2001.*

Bonald T. & Massoulié, L. 2001 Impact of fairness on Internet performance. In *Proceedings of ACM SIGMETRICS 2001.*

Boyd, S. & Vandenberghe, L. 2004*Convex Optimization*, Cambridge University Press.

Briscoe B., Jacquet A., Cairano-Gilfedder, C.D., Salvatori, A. , Andrea Soppera & Koyabe, M. 2005 Policing Congestion Response in an Internetwork using Re-feedback. In *Proc. ACM SIGCOMM'05, Computer Communications Review 35(4), Sep 2005*

Edmond, J. 1972 Edge-disjoint Branchings. In *Combinatorial Algorithms*, R. Rustin, Ed., pp. 2131. Algorithmics Press.

Enachescu, M., Ganjali, Y., Goel, A., McKeown, N. & Roughgarden, T. 2006 Routers with very small buffers, *Proc IEEE Infocom, Barcelona*

Floyd, S. & Fall, K. 1999 Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Transactions on Networking* **7**, 458–472.

Ganesh, A, Laevens, K., and Steinberg, R. 2007 Congestion pricing and noncooperative games in communication networks. *Operations Research*, **55**(3), 430–438

Georgiadis, L., Neely, M. J. & Tassiulas, L. 2006 *Resource Allocation and Cross-Layer Control in Wireless Networks*. Now Publishers.

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. Networking, December 2006.*

Johari, R. & Tsitsiklis, J.N. 2004 Efficiency loss in a network resource allocation game. *Math. of Oper. Res*, **29**(3), 407 –435.

Karlsson, G., Lundqvist, H. & Ivars, I.M. 2004 Single-service quality differentiation. In *Proc. IEEE IWQoS 2004, June 2004, pp. 265-272.*

Kelly, F.P., Dynamic routing in stochastic networks 1995. In *Stochastic Networks* (Editors F.P. Kelly and R.J. Williams), Springer-Verlag, New York, pp 169 –186.

Kelly, F.P., Maulloo, A. K. & D. K. H Tan, 1998 Rate control in communication networks: shadow prices, proportional fairness and stability, *Journal of the Operational Research Society*, **49**, 237–252.

Kelly, F.P., Key, P.B. & Zachary,S. 2000 Distributed admission control. *IEEE Journal on Selected Areas in Communications*, **18**, 2617–2628.

Kelly, F. P. & Voice, T. 2005 Stability of end-to-end algorithms for joint routing and rate control, Computer Communication Review **35:2** 5–12.

Key, P., & Massoulié, L.2006 Fluid models of integrated traffic and multipath routing. *Queueing Systems*, 53(1):85–98, June 2006.

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

Key, P., Massoulié, L. & Towsley, D. 2006a Combining multipath routing and congestion control for robustness. In *CISS 2006, 40th Conference on Information Sciences and Systems*, Princeton, March 2006. IEEE Information Society.

Kumar, S. & Massoulié, L. 2005 Fluid and diffusion approximations of an integrated traffic model, Microsoft Research Technical Report MSR-TR-2005-160. Available at http://research.microsoft.com/users/lmassoul/MSR-TR-2005-160.ps .

Kunnyur, S. & Srikant, R. 2000 End-to-end congestion control schemes: utility functions, random losses and ECN marks. in *Proc. IEEE INFOCOM.*

Laskowski, P. & Chuang J. 2006 Network Monitors and Contracting Systems: Competition and Innovation. In *Proc. ACM Sigcomm 2006, Pisa, Italy, September 2006.*

MacKie-Mason, J.K, & Varian, H. 1995 Pricing Congestible Resources. *IEEE J. of Selected Areas in Communications, vol. 13, no. 7, Sept. 1995: 1141-49*

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

Massoulié, L. & Roberts, J. 2000 Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems* **15**, 185–201.

Mo J. & Walrand, J. 2000 Fair end-to-end window-based congestion control. In *IEEE/ACM Transactions on Networking* **8**, 556–567.

Padhye, J. , Firoiu,V., Towsley, D. & Kurose,J. 2000 Modeling TCP Reno performance: a simple model and its empirical validation. *IEEE/ACM Transactions on Networking* **8**, 133–145.

Raina, G. & and Wischik, D.J. 2005 Buffer sizes for large multiplexers: TCP queueing theory and instability analysis, In *Proceedings of EuroNGI conference, Rome, April 2005.*

Rockafellar, T. 1970 *Convex Analysis.* Princeton University Press.

Srikant, R. 2003, *The Mathematics of Internet Congestion Control*, Birkhauser, 2003

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

Zhang-Shen, R. and McKeown, N. 2004, Designing a predictable Internet backbone network. In *Proc. HotNets 2004.*