A Contract and Balancing Mechanism for Sharing Capacity in a Communication Network

Edward Anderson, Frank Kelly and Richard Steinberg
Australian Graduate School of Management, University of New South Wales
Statistical Laboratory, University of Cambridge
Judge Institute of Management, University of Cambridge[*]

August 2002
Revised August 2003
Second Revision October 2004

We propose a method for determining how much to charge users of a communication network when they share bandwidth. Our approach can be employed either when a network owner wishes to sell bandwidth for a specified period of time to a number of different users, or when users cooperate to build a network to be shared among themselves. We show how a Contract and Balancing Mechanism can be defined to mediate between rapidly fluctuating prices and the longer time scales over which bandwidth contracts might be traded. An important property of the process is that it avoids introducing perverse incentives for a capacity provider to increase congestion.

(Capacity Contracts; Congestion Pricing; Nash Equilibrium)

1 Introduction

In this paper we propose a Contract and Balancing mechanism, involving long term contracts and a short term balancing process, as a method for sharing capacity in a communication network. Such capacity is usually given in terms of *bandwidth*, which is the amount of data that can be transmitted per unit time—usually via fiber optic cables—specified in bits per second, or bps. The approach we propose allows volatile prices to be appropriately averaged, and may facilitate the creation of liquid markets in bandwidth.

[*] E-mail: eddiea@agsm.edu.au, f.p.kelly@statslab.cam.ac.uk, r.steinberg@jims.cam.ac.uk. Part of the research leading to this paper was performed while Frank Kelly was visiting the Graduate School of Business, Stanford University, supported by the Operations, Information and Technology Program and the Center for Electronic Business and Commerce, and while Richard Steinberg was visiting the Department of Operational Research, London School of Economics.

Dagstuhl Seminar Proceedings 05011
Computing and Markets
http://drops.dagstuhl.de/opus/volltexte/2005/204

At the present time large carriers trade capacity through long-term contracts known as *Indefeasible Right of Use* (IRU). An IRU is essentially a long-term lease of a portion of the capacity of a cable by the company, or companies, that built the cable, with fiber physically provided via switches along a single network, or segments of several networks. However, the procedure is unavoidably complex. *Fortune* magazine explains it this way (Kirkpatrick 2000):

> To deliver a broadband signal—say, a digitized video stream—over long distances generally requires that the data travel entirely on one dedicated fiber-optic network. Though these independent networks connect to pass data around, getting one network to reserve a pathway for another's traffic is complicated. It can require months of negotiation for deals that often last years at a locked-in price.
>
> So customers (mostly small telecoms and businesses that host Internet applications) face a dilemma. If they reserve enough bandwidth to handle only their regular needs, they'll have problems when usage spikes. But if they buy enough pipeline space to serve their maximum demand, they waste money on access they're not using. Customers would benefit enormously if they could mix and match networks to get the most flexibility and best price.

The difficulty is that existing long-term contracts were developed in an age of fairly predictable traffic carried over *circuit-switched* networks. In a circuit-switched network, e.g., the public telephone network, an accepted call reserves a circuit (which may be physical or virtual) and hence a fixed amount of capacity which it holds for the duration of the call.

However, communication networks have evolved towards a more flexible *packet-switched* technology, and there is a need for a more appropriate form of long-term contract. In a packet-switched network, the message to be transmitted is first broken down into small units, called packets, each containing a data portion and a header. Then—and this is the essential distinction—the packets flow through the network using an amount of capacity that depends on how many other flows are simultaneously in progress; at the destination the headers are stripped off and the data re-assembled into the original message.

In summary, a user of a packet-switched network will in general not know the exact capacity he will require; correspondingly, the network owner will not know the actual capacity requirements of any of the users, only the contracted amounts. Each user would like to pay only for the capacity he uses; the network owner will not invest in capacity for peak periods if he is only paid for contracted capacity. And yet, even if every user were able to estimate his average period needs precisely, requirements do change over time, and in a packet-switched network, they can fluctuate rapidly by the second. It might seem at first that a market for options could develop here; however, this would be unlikely, since the short intervals involved would imply high accounting costs compared with the levels of payment.

The flexibility of packet-switched technology has led to huge growth in the demand for communications capacity, at the same time as advances in router and optical technology have allowed a huge growth in the supply. But the rate of growth of both demand and supply has become much more difficult to predict than the steady growth experienced over earlier decades in circuit-switched networks. We show that the mechanism described in this paper can mediate between rapidly fluctuating demands and the long time scales over which bandwidth contracts may be traded, and may thus facilitate the creation of more liquid markets in bandwidth and will hence more ably support an industry facing considerable uncertainty of demand and supply.

1.1 Using price signals to manage congestion

One of the major advantages of the technology of the Internet over older circuit-switched technologies is that the Internet's congestion control mechanisms, in effect, share capacity among users so as to absorb random fluctuations in their various demands. The rate at which a source sends packets is controlled by TCP, the Transmission Control Protocol of the Internet, which is implemented as software on the computers that are the source and destination of the data (Clark 1996). TCP works as follows. When a resource within the network becomes overloaded, one or more packets are lost. Loss of a packet is taken as an indication of congestion; the destination informs the source, and the source immediately reduces its sending rate. The source then gradually increases its sending rate until it again detects the loss of packets. This cycle of increase and decrease thus serves to discover and utilize the available bandwidth and share it among flows.

Using dropped packets to signal congestion has obvious drawbacks. First, it is wasteful of system resources since a dropped packet may have already consumed resources at earlier stages of its route and may need to be resent. Second, the congestion signal is late; until packets begin to be dropped, users are unaware that congestion is becoming a problem.

These considerations have led naturally to proposals for the introduction of congestion marking, whereby packets that encounter congestion at a resource have certain bits in their headers set by the resource to indicate congestion. The procedure is called *Explicit Congestion Notification*, or ECN (Floyd 1994). Users detecting ECN marks should respond by reducing their transmission rates. The end result will be a system that can share resources without recourse to dropped packets, except in periods of exceptionally heavy use. ECN has now been made a "Proposed Standard" by the Internet Engineering Task Force (IETF), the body concerned with the evolution of the Internet architecture (Ramakrishnan, Floyd and Black 2001).

In addition to its designated use as an indicator of congestion, an ECN mark also has an intuitively appealing interpretation as a hypothetical congestion charge. When a link is well below capacity, there will be few if any marks generated and the appropriate charge should be quite low. Correspondingly, when a link is near capacity, many marks are generated and the charge should be higher. Indeed, theoreticians have developed an interpretation of TCP as a utility-maximizing algorithm, balancing the benefit to the user of its achieved flow rate against the impact on other users as signaled by lost packets or marks (Gibbens and Kelly 1999, Low, Paganini and Doyle 2002). Each link indicates its congestion by a scalar variable (termed *price*), and sources have access to the aggregate price of links on a route. The price at a link may be physically realized as, for example, a packet marking probability at the link, and can also be viewed as an implicitly constructed dual variable within an optimization framework. The packet-level interactions of sources and resources may then be viewed as a tatonnement process (Varian 1992), by which competing demands reach equilibrium. Under this interpretation, part of the remarkable success of TCP can be attributed to its ability to balance demand exceedingly rapidly, the speed of the process being limited only by network propagation delays.

Many choices are possible with regard to the level of aggregation at which marks might be reflected as costs or prices to users. For example, the marks could allow the dispersal of charging operations such as metering, accounting and billing to customer systems, as argued by Briscoe et al. (2000). Alternatively, as suggested by Key (1999), an Internet Service Provider might manage the risk associated with congestion pricing to provide end-users with a service defined in more traditional terms. Briscoe et al. (2003) provide an overview of the resulting network architecture, and discuss the potential engineering and commercial advantages over earlier Internet architectures such as Diffserv and Intserv, as a consequence of the improved flow of information *from* the network provided by congestion marking.

While there has been a considerable research effort on the connection between congestion marking and user demand in communication networks, there has been little work on the ways in which congestion marking might impact the supply of capacity. (For a review of the basic economic theory of congestion pricing see MacKie-Mason and Varian 1995.) One natural approach involves the owner of the link being paid based on the number of marks his link generates. However, this solution produces a perverse incentive for the owner to increase congestion, e.g., to make side payments to some users to generate sufficient traffic so as to maintain a state of high congestion. This would lead to high revenues for the owner and low utility to the users. Although it is possible that severe abuse by the owner would be punished by an eventual loss of business, it is not easy to see how systematic overcharging could be prevented.

What features should a charging scheme for bandwidth possess? We believe that there are four desirable characteristics. First, the scheme needs to allow a single network resource to be shared among different users when the particular profile of their requirements cannot be predicted in advance. In other words, it is not sufficient just to divide the resource between the different users according to some profile of predicted usage. Second, the charging scheme needs to promote the efficient use of the network resource, so that different users compete for bandwidth at times of high congestion in a way that will select the traffic with the highest utility (or user's willingness to pay). Third, the scheme needs to fix a payment to the network provider that depends on the total bandwidth made available to all the users (i.e., the size of the network resource) and not on their actual pattern of use. This makes the payment to the network provider match the cost of provision, in addition to removing the perverse incentive mentioned above. Finally, the users need to be able to control the amount that they pay; and, in particular, they should be able to protect themselves from high costs brought about through actions of other users. We will show that the Contract and Balancing Mechanism introduced in this paper satisfies all these requirements.

1.2 A contract and balancing mechanism

We propose a method in which the network provider sells contracts for usage. The novel part of our proposal is that users participate in a balancing process in which they make or receive payments based on a comparison of the marks they generated with the capacity they had contracted. More precisely, a contract for a particular link will entitle the purchaser to a certain proportion of the congestion marks generated on that link over a specified period of time. At the end of the period, users will make or receive payments according to whether they generated more or fewer congestion marks than their contracted amounts. We will call this the *Contract and Balancing Mechanism* (CBM).

As an example, suppose that two users have each contracted for part of a link with a capacity of 30 megabits per second (i.e., 30 Mbps), with user A contracting for 10 Mbps and user B for 20 Mbps. We suppose that the contractual rate is c dollars per Mbps for a period of a month, and the balancing charge is γ dollars per mark. Actual usage varies over time, and so do the congestion marks generated. Suppose that we fix on a "settling up" period of one month. At the beginning of the month, users A and B pay the network owner, respectively, $10c$ dollars and $20c$ dollars. Over the course of the month, if the number of marks generated by user A is exactly one-third the total marks generated, then at the end of the month no further payments are made. If, however, A generated greater than a third the total marks, then in the balancing process A will pay B the sum of

$$\gamma\,[z_A - (1/3)(z_A + z_B)],$$

where z_A and z_B denote the number of marks generated by A and B, respectively. If this expression is negative, i.e., if A generated *less* than a third of the marks, then A will *receive* this sum from B. Observe that the balancing payments are made among the users and do not involve the network owner; the only payments received by the network owner are the contractual payments at the beginning of the period.

Our proposal differs from that of Reichl et al. (2001), who suggest a long-term contract mechanism they call the *Cumulus Pricing Scheme*. In their scheme there are two time scales of interest, the contractual period (e.g., one year), and the measurement period (e.g., monthly). When actual usage over a measurement period exceeds a specified initial threshold, the user receives a red *cumulus point*; a user will receive a larger number of red cumulus points if his usage exceeds a further threshold, and so forth. Correspondingly, when measured usage over a month falls below a certain threshold, the user will receive a green cumulus point; if usage falls below a lower threshold, the user will receive a larger number of green cumulus points, and so forth. At the end of the contractual period, the total number of green points are subtracted from the total number of red points; the final tally dictates the renegotiation of the original contract. The Cumulus Pricing Scheme is designed to allow a user to overuse his allocated capacity in one period if it is balanced by a period of under use at another time.

There is a substantial literature on network and computing service pricing. Masuda and Whang (1999) provides a recent discussion of dynamic pricing schemes where the network owner charges in a way that induces optimal arrival rates to maximize the net value of the network, and Rump and Stidham (1998) consider the dynamic behavior of an input-pricing mechanism for a service facility in which self-optimizing customers base their future join/balk decisions on their previous experience of congestion. These papers model waiting times within queues explicitly, with demand decreasing as waiting times increase. In our model, congestion prices are used to induce traffic levels that remain below the capacity of the network: hence packet waiting times are small, and instead of a waiting cost there is a loss of utility as traffic is priced out of the market. (The delay incurred by TCP in downloading a large file is primarily a consequence of the limited rate that can be achieved across a congested network, rather than the times taken by individual packets to pass through router queues.) The simpler nature of our model, with its omission of externalities caused by queueing delays, allows us to concentrate on the game-theoretic issues that arise from our proposed Contract and Balancing mechanism. In our final section we remark on some of the other issues that would need to be addressed before our Contract and Balancing mechanism could be employed in models with significant externalities due to queueing delays.

In this paper we will show that a mechanism, involving long term contracts and a short term balancing process, will be an effective method for sharing resources in a communication network. A number of different issues need to be dealt with. We begin by providing a more detailed description of the balancing process and introducing the principle of price complementarity (Section 2). We next examine the short term choices on traffic volumes as each user attempts to maximize his utility given the contracts he holds (Section 3). In particular, we are interested in the case where users respond to congestion signals in the same way that they respond to congestion signals under TCP. This corresponds to price taking with a specific choice of utility function. Our main results are presented in Section 4, where we look at the interaction between short and long term user decisions regarding the amount of network capacity to purchase in the contract market. In Section 5, we consider a number of issues relating to implementation of this scheme; for example, we ask what information need be collected in order for the balancing process to be implemented in a network. Finally, in Section 6, we provide a brief summary of the results of the paper. Proofs of the propositions appear in the Appendix.

2 Fundamentals

2.1 The balancing process

We start by giving a more detailed description of the balancing process. Our setting is dynamic in which traffic levels vary over time. There exist complex issues concerned with the speed at which a system can adapt to changes in traffic through user responses to price signals. However, this will not be our focus here. We will work with time intervals that are assumed to be short enough that we can ignore changes in traffic characteristics during the interval, but at the same time sufficiently long compared with round trip time that users can easily adjust traffic levels so as to achieve their desired overall traffic rates. In dealing with behavior over a longer period of time we will just write traffic and price as functions of a continuous time parameter and assume that adjustments take place instantaneously.

We assume that there are n (≥ 2) users of a single link. Each user receives some price signal from the link (in the case of Internet traffic this will be a congestion mark or lost packet), and we write $p(t)$ for the price per unit of traffic in time interval t. This price is the same for each user of the link; it may depend on the overall traffic on a link, but not otherwise upon the traffic of any individual user. We assume that $p(t)$ is generated by the network on the basis of congestion in the link in interval t. The balancing process does not depend on any particular price-setting mechanism.

Furthermore, we assume that it has been agreed that balancing payments will be based on the price $p(t)$ in the following way. We suppose that the capacity of the link is Y and that user i has contracted for a total capacity of y_i during this interval. We write $\rho_i = y_i/Y$ for the proportion of the link contracted to user i. Then, for time period t at which traffic from user i is $x_i(t)$ and total traffic is $D(t) = \sum_{k=1}^{n} x_k(t)$, user i is required to make a payment of $p(t)(x_i(t) - \rho_i D(t))$.

This expression can be thought of as user i making a balancing payment at price $p(t)$ for his own traffic, $x_i(t)$, and then receiving back a proportion ρ_i of all the balancing payments made, $p(t)\,D(t)$. The actual payments are made on the basis of an integral of this expression over some convenient balancing period. We express the balancing period as the unit interval $[0, 1]$ to obtain

$$E_i = \int_0^1 [x_i(t) - \rho_i D(t)]\, p(t)\, dt \tag{1}$$

as the payment to be made by user i for the period $[0, 1]$.

In the TCP case that is our main focus, congestion marks are the price signals and we will assume that at any time t the price $p(t)$ is defined as a constant γ multiplied by the probability of a packet being marked in the time interval t. So γ is a price per mark. Then $\int_0^1 x_i(t)\, p(t)\, dt$ is the expected value of the marks generated by user i, and E_i is the expectation of the net payment into the balancing process by user i. We shall discuss the choice of the parameter γ, the charge per marked packet, in Section 5: until then it is convenient to standardize units so that $\gamma = 1$, so that the symbol $p(t)$ may be used for both the packet marking probability and the price at time t.

An alternative approach might be to apportion a fixed capacity charge among different users simply on the basis of the proportion of congestion marks each user generates. Then the payment made by user i over a period is given by

$$E_i \;=\; \frac{\int_0^1 x_i(t)p(t)\, dt}{\int_0^1 D(t)\, p(t)\, dt}\, cY$$

and there is no initial contract payment. This guarantees that the network owner receives a total payment of cY from the users. However this approach will leave individual users carrying a large risk. This follows because if other users do not send traffic at all, then a single user will end up paying the entire cost of the link. For example, suppose that user i expects to use about one tenth of the total link capacity, so that $\int_0^1 x_i(t)p(t)\, dt = Y/10$, then user i when sharing the resource with 9 similar users would expect to pay $cY/10$. However if the total traffic from other users is very low then very few congestion marks will be generated, but the proportion due to user i will be much larger. Thus the total cost to user i will be much higher than $cY/10$. In an extreme scenario there might only be one mark generated and if this falls onto the user in question then user i will pay cY. Thus we see that this approach does not have the last of the four desirable features we mentioned at the end of Section 1.1. In the same circumstance under CBM, the contract is likely to be for an amount $Y/10$ and the total cost to user i will have an upper limit of $cY/10 + \gamma(9/10)N$ where N is the total number of marks generated during the period. Since in these circumstances N is small, therefore the risk to user i is limited.

The Contract and Balancing Mechanism can be defined for a network. We suppose that a route through the network is identified with a nonempty subset of a set of links J. We suppose that there are price signals $p_j(t)$ for each link $j \in J$. We can define the balancing process in the network in various ways, but our starting point is just to carry out a link by link analysis. Suppose that a user with route r contracts for a capacity y_r over the balancing period $[0,1]$, and has actual traffic $x_r(t)$. If the total traffic in link j is $D^{(j)}(t)$ and its capacity is Y_j, then the net balancing payments made by this user for a single link $j \in r$ is

$$\int_0^1 [x_r(t) - \rho_j D^{(j)}(t)]\, p_j(t)\, dt$$

where $\rho_j = y_r/Y_j$, the proportion of the link capacity contracted to this route. Let the total price for route r be given by $p_r(t) = \sum_{j \in r} p_j(t)$. Then the total net payments made for route r are given by

$$E_r = \int_0^1 \left(x_r(t) p_r(t) - y_r \sum_{j \in r} \frac{D^{(j)}(t) p_j(t)}{Y_j} \right) dt. \tag{2}$$

As an example, suppose that the network consists of two links: Link 1 from point 1 to point 2 with capacity 20 Mbps, and Link 2 from point 2 to point 3 with capacity 30 Mbps. There are three users, with user A contracting for 5 Mbps from point 1 to point 2, user B contracting for 10 Mbps from point 2 to point 3, and user C contracting for 15 Mbps from point 1 to point 3. This can be summarized in the following table:

	Link 1	Link 2
User A	5 Mbps	·
User B	·	10 Mbps
User C	15 Mbps	15 Mbps

We suppose that for each link the contractual rate is c dollars per megabit per month, and the balancing charge is γ dollars per mark, with a contractual period of one month. At the beginning of the month, users A, B, and C pay the network owner, respectively, $5c$, $10c$, and $30c$ dollars. Let z_X^i denote the number of marks generated by user X on link i. Since user A contracted for one-fourth the capacity of Link 1, then his payment in the balancing process will be:

$$\gamma\, [z_A^1 - (1/4) z^1],$$

where $z^1 = z_A^1 + z_C^1$. Similarly, user B contracted for two-fifths the capacity of Link 2, and thus his payment in the balancing process will be:

$$\gamma\, [z_B^2 - (2/5) z^2],$$

where $z^2 = z_B^2 + z_C^2$. Finally, user C contracted for three-quarters the capacity of Link 1 and three-fifths the capacity of Link 2. His payment in the balancing process will be:

$$\gamma\, [z_C^1 + z_C^2 - (3/4) z^1 - (3/5) z^2].$$

Of course, the three balancing payments sum to zero.

2.2 Price complementarity

The price signal in a link is naturally thought of as a function of the traffic in the link. But sometimes it is convenient to model a link as having a fixed capacity that can be fully utilized. In this case the price signal is not determined from the overall traffic level when the link is full. We can have a link that is full, and hence has a given traffic level, but in different circumstances there can be different price levels (or congestion marks generated).

We can instead think of the determining factor as the level of traffic that is desired by the users. If the total desired traffic is less than the capacity of the link, then the link is not congested, and no congestion marks are generated, corresponding to a price of zero. If the desired traffic level is greater than the capacity of the link then the price is set to a level which brings actual total demand back to the capacity of the link.

This leads to a disjunction: either the link is full or the price is zero. We capture this in the following complementarity assumption:

Assumption 1: Price complementarity. *For each link j and time t, $p_j(t)(Y_j - D^{(j)}(t)) = 0$.*

We may regard the price complementarity assumption as an approximation when prices rise sharply from zero as the traffic levels approach the capacity of the link.

In the Internet, the mechanisms which place marks on packets at routers are generically termed *active queue management*. These mechanisms function by marking packets at a resource when measurements, such as arrival rate or queue size, indicate that congestion is near. Many of the recent suggestions for active queue management adapt marking rates so as to achieve a preset target for utilization or average queue size. Such adaptation can be interpreted in terms of a design goal to implement price complementarity. For details and discussion, see Low et al. (2002).

Assumption 1 enables us to simplify the equations for balancing payments. Specifically, equation (1) takes the form

$$E_i = \int_0^1 [x_i(t) - y_i] \; p(t) \, dt, \tag{3}$$

while the network version (2) simplifies to

$$E_r = \int_0^1 [x_r(t) - y_r] \; p_r(t) \, dt. \tag{4}$$

It is important to realize that, although the Contract and Balancing Mechanism has been designed to benefit users having highly variable bandwidth requirements, there are also substantial benefits to users whose bandwidth requirements are constant. Thus, in comparison with a conventional method that simply divides up available capacity among the users, under CBM a user who contracts for capacity on a link and only ever delivers that amount or less into the network (say, through a pipe with exactly this capacity) will never be required to make payments—and, in fact, might receive payments—in the balancing process. The simplest way to see this is to invoke the price complementarity assumption and then observe that the statement is a direct consequence of expression (3) or (4).

3 Choice of Traffic Volume

3.1 Price taking

We begin our modelling by looking at the short term choices to be made by players (where we use the term “player” rather than “user,” to include the possibility that an Internet Service Provider has purchased a contract on behalf of a collection of end-users). We start by considering a single link and a single time interval, so we drop t from our notation. We suppose that player i has already fixed a contract position y_i and we turn to the question of deciding on the demand x_i .

There are several alternative models we could consider. However, we will start by considering the simplest, which is to assume that all the players act as price takers, assuming that they can have no effect on price or total demand, where each player has quasi-linear utility $V_i(x_i)$. We ask what value of traffic is best for player i if the price, p, and the total demand, D, are given and independent of player i ’s choices.

This has the effect of setting x_i to be a price-dependent demand function, which we write as $D_i(p)$ for player i. In this case we must choose $x_i = D_i(p)$ to maximize

$$V_i(x_i) \;=\; u_i(x_i) - \,[x_i - \rho_i D]\, p \tag{5}$$

where $u_i(x_i)$ denotes the utility to player i if he generates a traffic volume x_i in the time interval. Under the normal assumption that the function $u_i(x_i)$ is strictly concave and differentiable, the maximum utility is achieved by choosing x_i to solve $u_i'(x_i) = p$, with $x_i = 0$ if $p > u_i'(0)$. So we end up with a demand function for player i of

$$D_i(p) = (u_i')^{-1}(p)$$

which, under our assumptions, is a well-defined decreasing function of p.

Though this discussion has been quite general, an important special case occurs if the Transmission Control Protocol (TCP) is used. The steady state behavior of TCP has been analyzed extensively. In equilibrium, the throughput x achieved by a connection is approximately $k\,/(T\sqrt{p})$, where T is the round-trip time of the connection, p is the packet loss or marking probability, and k is a constant (Floyd and Fall 1999). This motivates consideration of the demand function

$$D_i(p) = \frac{\alpha_i}{\sqrt{p}}, \tag{6}$$

where the parameter α_i is determined by the number of TCP connections of player i, and their various round-trip times.[1] This will correspond to price-taking behavior for a player with utility function of the form:

$$u_i(x) = K_i - \alpha_i^2/x, \tag{7}$$

for some arbitrary constant K_i. Gibbens and Kelly (1999) and Kunniyur and Srikant (2003) have made this observation when congestion marks are straightforward charges rather than being part of a balancing mechanism.

[1] Recall that we have standardized units so that the price per packet mark is 1.

The TCP protocol was designed for applications such as bulk data transfer, and other congestion control algorithms have been developed for applications such as streaming multimedia. Many of these algorithms are explicitly designed to have the same bandwidth usage as TCP when faced with the same marking probability (Floyd et al. 2000), and hence share the same approximate demand function (6). A more general class of demand functions

$$D_i(p) = \frac{\alpha_i}{p^{1/\beta}}, \tag{8}$$

where $\beta \in (0, \infty)$ has been studied in connection with more general congestion control algorithms. The cases $\beta = 1$, $\beta = 2$ and $\beta \to \infty$ correspond respectively to notions of *proportional fairness*, *TCP fairness* and *max-min fairness* (Mo and Walrand 2000), and there are currently proposals to alter the TCP algorithm in a manner which would vary the parameter appearing in its implicit demand function (8) from $\beta = 2$ downwards, perhaps as far as $\beta = 1$ (Floyd 2003, Kelly 2003). Our later development will use a time-varying version of this demand function,

$$D_i(t, p(t)) = \frac{\alpha_i(t)}{p(t)^{1/\beta}}. \tag{9}$$

corresponding to a time-varying utility function for player i of

$$u_i(t, x_i(t)) = \alpha_i(t)^\beta \, \frac{x_i(t)^{1-\beta}}{1-\beta}, \qquad \beta \neq 1$$

$$u_i(t, x_i(t)) = \alpha_i(t) \log x_i(t), \qquad \beta = 1.$$

We assume that all players share the same choice of β, but we allow $\alpha_i(t)$ to fluctuate stochastically, as connections come and go. Note that the demand function (9) is unbounded as $p \to 0$, and thus if the price complementarity condition holds, then for a link with capacity Y, we have $\sum D_i = Y$, and so the price on the link is given by

$$p(t) = \left(\frac{\sum_1^n \alpha_i(t)}{Y} \right)^\beta. \tag{10}$$

3.2 More sophisticated player behavior

Next, we consider how a player might be motivated to deviate from price-taking behavior if he takes into account the impact of his choices on the price p and the total demand D.

We will suppose that price complementarity holds. This enables us to carry out an analysis of the optimal choice of traffic when a player knows the demand functions for the other players (or at least the aggregate of these). As before, we consider a single link and a single time interval. Thus, from (3), player i chooses x_i to maximize

$$u_i(x_i) - p(x_i - y_i).$$

Under price complementarity, player i views the price p as a function $p(x_i)$ of its choice of traffic x_i, where $p(x_i)$ is determined by

$$D_{-i}(p(x_i)) + x_i \;=\; Y,$$

where D_{-i} is the aggregate demand function of the other players. The viewpoint here is that a player, rather than responding to a price signal, is controlling the price. We have

$$\frac{\partial p}{\partial x_i} \;=\; \frac{1}{-D'_{-i}(p)}.$$

Hence player i chooses x_i so that

$$u'_i(x_i) - p - (x_i - y_i)\frac{\partial p}{\partial x_i} \;=\; 0,$$

which can be rewritten as

$$(x_i - y_i) \;=\; D'_{-i}(p)[p - u'_i(x_i)]. \tag{11}$$

When we consider individual decisions on the demand x_i, it is not very satisfactory to assume that a player will ignore actual price feedback in preference for a calculation based on an estimate of the aggregate demand function D_{-i}. However, we can use (11) to suggest an adjustment to the demand function that would arise from the simpler price-taking approach. Under price taking, player i would have selected x_i so that $u'_i(x_i) = p$. Now $D'_{-i}(p) < 0$ and we assume that utility is concave, so u'_i is a decreasing function. Hence from relation (11) we see that, in the short term, player i departs from price taking by understating/overstating his demand, according as his demand is greater/less than his contracted capacity. He thus moves his demand in the direction of his contracted capacity, a relatively benign deviation from price-taking behavior. Note also that it is not necessary that player i be small for his behavior to be well approximated by price taking: it is enough that the mismatch between his optimized demand and his contract capacity be small.

A more extended version of the above argument would assume that each player j chooses a demand function $D_j(p)$, and would look for a Nash equilibrium among the demand functions. This could give some indication of the equilibrium that might result when players repeatedly adjust their demand behavior over time, as they learn of the other players' choices. This is essentially the same approach as has been suggested by Klemperer and Meyer (1989) within a supply function—rather than demand function—context, and has also been used for electricity (Green and Newbery 1992, Newbery 1998). It is perhaps less reasonable for us to suppose that a player has so much knowledge of other players: as pointed out in Ganesh, Laevens and Steinberg (2000), it is impractical for players in networks to be aware of even the *number* of other players sharing resources with them. They also point out that in their setting, which applies equally here, there can be multiple Nash equilibria and that sufficient conditions for uniqueness of the Nash equilibrium are apparently neither simple nor intuitive. Further, there seems to be no mechanism available that would lead players towards an equilibrium solution in the time scales we are considering.

In our later development we shall assume price-taking behavior. We are motivated to do this by the discussion of this section, which shows that more sophisticated behavior can be viewed as a perturbation of price-taking, and by the observation that the steady state behavior of TCP can be interpreted as price-taking.

Assumption 2: Price-taking. *For his short-term choice of traffic volume, player i acts as a price-taker with demand function $D_i(p)$ related to his utility via the equation $u'_i(D_i(p)) = p$.*

3.3 Inappropriate incentives

As mentioned in the introduction, one important feature of the proposed method is that it provides no incentive to the network owner to increase congestion in the network by adding additional traffic.

We might ask whether there is ever an incentive for a player to boost his traffic artificially, thereby gaining from the balancing process more than his potential losses from marked packets as the link reaches capacity. This will happen only in some very specific circumstances. For example consider a player, A, who has contracted for half the capacity of a link and knows that the other players are using 60% of the link without any significant peakiness in their traffic characteristics. This would enable A to benefit by pushing artificial traffic into the link so that it is essentially at capacity, thereby pushing the price up. The other players will, in this scenario, make a net payment into the balancing process of around 60% of all payments and end up contributing to player A a total of 1/6 of all their payments in the balancing process. Notice, however, that the only occasion when there is an incentive for a player to add artificial traffic occurs when some other player is sending in a consistent way more than their contract amount of traffic and is also operating with inadequate methods for reducing traffic as prices increase. This is precisely the kind of behavior that we might wish to discourage, and the balancing process will have the desirable deterrent effect.

We can interpret this same hypothetical situation in terms of non-price-taking behavior as expressed in (11). Sending artificial traffic corresponds to a choice of x_i for which $u_i'(x_i) = 0$. This can be a solution of (11) only when $x_i < y_i$, so that player i's traffic volume remains less than his contracted amount. Moreover the sending of artificial traffic with the aim of creating a significant benefit in the balancing process can only occur with high prices p, which in turn will imply from (11) a small value of $D'_{-i}(p)$, i.e. low price sensitivity by the other players.

4 Choice of Capacity

In this section we consider the two stage process where in the first stage players decide on the size of contract they wish to purchase over the balancing period $[0, 1]$, and these decisions determine the size of the link (or network) which is built or purchased. The second stage of the process occurs as the players make short term choices with respect to the quantity of traffic they wish to send at each time point $t \in [0, 1]$.

Since the CBM scheme does not require a match between contractual amounts and actual traffic level, it is of interest to investigate how close these will be. There is nothing to prevent a player from contracting for more than his expected usage so as to benefit from payments in the balancing market or, correspondingly, from contracting for only a small (or zero) amount and paying more directly for his actual usage via the balancing process. We shall identify circumstances where a player has an incentive to contract for a capacity closely related to his anticipated usage.

4.1 Choice of capacity for a link

Consider the CBM scheme for a link with $n \geq 2$ players. Here we will assume player i ($i = 1, 2, \ldots, n$) contracts for a capacity of y_i over the balancing period $[0, 1]$, at an immediate cost to him of $C_i(y_i)$. A link capacity $Y = \sum_i y_i$ is then available for use by all n players over the period $[0, 1]$. We will assume that C_i is a linear function: $C_i(y_i) = c_i y_i$. We have in mind a situation in which decisions on the capacities y_i are taken before the construction of the link and it is reasonable to take the cost for one user as independent of the costs for others.

The case that is of primary interest to us is that where a player—e.g., an ISP—is contracting on behalf of end-users who are each operating under TCP. In this case the end-users are constrained to operate as though their utility functions were of the form given by equation (7). We model a situation in which this is indeed their utility function and the player who contracts on their behalf uses this utility function when making tradeoffs between the cost of contracting for a greater amount and the benefit to end-users as the size of the link is increased.

At time $t \in [0, 1]$ we assume that the demand from player i is a function $D_i(t, p(t))$ of a price $p(t)$, with

$$D(t, p(t)) = \sum_1^n D_i(t, p(t)) \; \leq Y. \tag{12}$$

The total expected cost to player i of the contract for capacity y_i is

$$W_i = C_i(y_i) + \int_0^1 \mathbb{E}\left[\left(D_i(t, p(t)) - \rho_i D(t, p(t))\right) p(t)\right] \, dt \tag{13}$$

where $\rho_i = y_i/Y$ is the proportion of the link contracted to player i. The utility to player i at time t is $u_i(t, D_i(t, p(t)))$. Thus, the expected utility to player i over the period $[0, 1]$ is

$$U_i = \int_0^1 \mathbb{E}[u_i(t, D_i(t, p(t)))] \, dt. \tag{14}$$

Hence player i, whom we assume to be risk neutral, will choose capacity y_i so as to maximize

$$V_i \; = U_i - W_i = \int_0^1 \mathbb{E}[u_i(t, D_i(t, p(t))) - (D_i(t, p(t)) - \rho_i D(t, p(t)))\, p(t)] \, dt \; - \; C_i(y_i). \tag{15}$$

Equation (15) indicates that the utility to player i has three components: (i) a positive benefit corresponding to his actual demand over the period, (ii) a dis-benefit from having to pay for his actual demand, net the payment reduction from the balancing process; and (iii) a dis-benefit from the payment for the contracted capacity.

Next we consider the optimal choice of contract amount for player i. In addition to the price complementarity assumption, we take each D_i as given by (9), where the probability distribution for $(\alpha_i(t), t \in [0, 1], i = 1, 2, ...n)$ is common knowledge. We suppose player i's utility $u_i(t, x_i(t))$ is consistent with his demand function, so that $u_i'(t, D_i(t, p(t))) = p(t)$, using u_i', and later D_i', to denote the partial derivative with respect to the second argument. Let $\alpha(t) = \sum_j \alpha_j(t)$, and write y_{-i} for $\sum_{j \neq i} y_j$.

Proposition 1 *For given values of y_j, $j \neq i$, player i has an optimal choice of contract quantity y_i which is unique. The choice is zero if*

$$\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{y_{-i}}\right)^\beta \left(1+\beta\frac{\alpha_i(t)}{\alpha(t)}\right)\right]\, dt < c_i, \tag{16}$$

and is otherwise given by the solution of the following equation:

$$\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta \left(1+\beta\frac{\alpha_i(t)}{\alpha(t)} - \beta\frac{y_i}{Y}\right)\right]\, dt = c_i \tag{17}$$

The condition (16) gives a bound on the contract price, beyond which the cost of participating in a contract is sufficiently expensive that a player will choose not to contract for any amount, but pay for usage entirely through the balancing process. Note that $(\alpha(t)/y_{-i})^\beta$ is just the anticipated price if $y_i = 0$, so this condition can also be seen as giving the factor by which c_i must exceed this anticipated price, if contracts are to be uneconomic.

Now we turn to the case where each of the players are individually seeking to maximize their overall utility—thus we consider a Nash equilibrium between the players with respect to their choice of y_i. First we consider a simple case in which one player is distinguished from all the others.

Example 1: A network constructor. Suppose that one player, a network constructor, is about to build a link at cost c_1 per unit of capacity. This player has no demand of its own but will profit by selling capacity to other users at a rate of c_2 per unit of capacity using the Contract and Balancing Mechanism. We assume symmetry of the demand structure among the other players. An option for the network constructor is to build more capacity, by an amount y_1, than the other players require in total, and then to benefit from payments in the balancing process.

In looking for a Nash equilibrium in the quantities y_i we will optimize over y_1 for given $y_2, ...y_n$ and hence the payment $(c_2 - c_1)\sum_{i=2}^n y_i$ does not change the choice of optimal y_1. So we need to consider the case where $c_1 < c_2 = c_3 = ... = c_n$, $D_1(t, p(t)) = 0$, $t \in [0,1]$, the joint distribution of $(\alpha_j(t),\ j = 2, 3, ..., n)$ is invariant under permutation of players $j = 2, 3, ..., n$, and $\beta \geq 1$. To find a Nash equilibrium we look for a solution $y_1 > 0$, $y_2 = y_3 = ... = y_n > 0$ to relations (17). The relation for player 1 gives

$$\left(1 - \beta\frac{y_1}{Y}\right)\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta\right] dt \;=\; c_1, \tag{18}$$

while adding relation (17) over all n players gives

$$\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta\right] dt \;=\; \frac{c_1 + (n-1)c_2}{n}. \tag{19}$$

The left hand side of equation (19) is decreasing in the total capacity Y, and admits a unique solution for Y. Since $c_1 < c_2$, relation (18) then admits a unique solution for y_1, which lies in the range $0 < y_1 < Y/\beta$. Equation (17) can then be solved by $y_2 = y_3 = ... = y_n = (Y - y_1)/(n-1)$. Thus we have a Nash equilibrium where player 1 chooses a positive value of y_1, in order to benefit later from the balancing process.

Next, we show that this Nash equilibrium is unique. First any equilibrium must have $y_1 > 0$, since if $y_1 = 0$ then from Proposition 1

$$\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta\right]\, dt \le c_1,$$

while adding relation (17) over i such that $y_i > 0$ gives that the same expression is not less than c_2, a contradiction since $c_1 < c_2$. Hence, at an equilibrium, $y_1 > 0$, and so relation (18) holds. Hence $y_1 < Y/\beta \le Y$, and so at least some of $y_2, y_3, ..., y_n$ are positive. That all of these variables are positive at an equilibrium now follows, since if relation (17) holds for a value $i \in \{2, 3, ..., n\}$ then inequality (16) cannot simultaneously hold for a different value i in this range, by the symmetry assumption on the joint distribution of $(\alpha_j(t),\ j = 2, 3, ..., n)$. Finally $y_2 = y_3 = ... = y_n$, from relation (17) and the assumed symmetry of the demand structure.

Finally, we consider whether player 1 might have an incentive to artificially induce congestion later, in the second stage of the game (violating our assumption that he acts as a price-taker at this stage). Suppose, then, that at a later time t, when the aggregate demand function from players $2, 3, ..., n$ is $D(t, p(t)) = \alpha(t)/p(t)^{1/\beta}$, player 1 chooses to send an amount of (valueless) traffic $d(t)$ in an attempt to increase his benefit from the balancing process. Then the net payment to player 1 at time t will be

$$(y_1 - d(t))p(t) = (y_1 - d(t))\left(\frac{\alpha(t)}{Y - d(t)}\right)^\beta.$$

This expression will be maximized by the choice $d(t) = 0$ provided $y_1 < Y/\beta$. Thus players $2, 3, ..., n$ can be assured that, provided player 1's share of capacity is bounded by $1/\beta$, he will have no incentive to send spurious traffic in the second stage of the game. □

Example 2: Cournot competition. Suppose $c_1 = c_2 = ... = c_L < c_{L+M} = c_{L+2} = ... = c_{L+M}$, $D_i(t, p(t)) = 0$, $i = 1, 2, ..., L, t \in [0, 1]$, and the joint distribution of $(\alpha_j(t),\ j = L{+}1, L{+}2, ..., L{+}M)$ is invariant under permutation of players $j = L{+}1, L{+}2, ..., L{+}M$. We look for the existence of a Nash equilibrium where $y_1, y_2, ..., y_L > 0$ and $y_{L+1} = y_{L+2} = ... = y_{L+M} = 0$, so that players $1, 2, ..., L$ supply all the capacity and players $L{+}1, L{+}2, ..., L{+}M$ act only as customers, with all their costs arise in the balancing market, in which they simply pay the price $p(t)$ for any traffic they generate.

The relation (17) for players $1, 2, ..., L$ becomes

$$\left(1 - \beta\frac{y_i}{Y}\right)\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta\right] dt \;=\; c_1, \quad i = 1, 2, ..., L \tag{20}$$

while $y_{L+1} = y_{L+2} = ... = y_{L+M} = 0$ implies that

$$\left(1 + \frac{\beta}{M}\right)\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^\beta\right] dt \;\le\; c_{L+1}. \tag{21}$$

We deduce that, provided

$$\left(1+\frac{\beta}{M}\right)c_1 \le \left(1-\frac{\beta}{L}\right)c_{L+1}, \tag{22}$$

there exists a Nash equilibrium with $y_{L+1} = y_{L+2} = ... = y_{L+M} = 0$ and $y_1 = y_2 = ... = y_L$, where y_1 is the unique solution to the equation

$$\left(1-\frac{\beta}{L}\right)\int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Ly_1}\right)^{\beta}\right] dt \;=\; c_1. \tag{23}$$

This equilibrium is unique, by a variant of the argument used to show uniqueness in Example 1: observe that if any one of $y_{L+1}, y_{L+2}, ..., y_{L+M}$ is positive then they must all be equal, by the symmetry assumption on the joint distribution of $(\alpha_j(t),\ j = 2, 3, ..., n)$.

Next we consider the relationship of the above model with the Cournot model of an oligopoly. Suppose that player $i, i = 1, 2, ..., L$, chooses y_i to maximize

$$y_i\left(P\left(\sum_1^L y_j\right) - c_1\right)$$

where

$$P(Y) = \int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^{\beta}\right] dt,$$

the time-averaged expected price if a total capacity Y is constructed. Then there is a unique Nash equilibrium; at this equilibrium $y_1 = y_2 = ... = y_L$, where y_1 is the unique solution to equation (23).

Hence the condition (22) is necessary and sufficient within our model for players $L+1, L+2, ..., L+M$ to act only as customers, and for players $1, 2, ..., L$ to act as if playing within a Cournot game. Observe that the condition becomes easier to satisfy the larger the number of suppliers L, the number of customers M, or the ratio of costs c_{L+1}/c_1.

If condition (22) is not satisfied then players $L+1, L+2, ..., L+M$ will contract for positive capacities, and by adding relation (17) over all players we obtain that the time-averaged expected price is

$$\int_0^1 \mathbb{E}[p(t)]\ dt \;=\; \int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^{\beta}\right] dt \;=\; \frac{Lc_1 + Mc_{L+1}}{L+M},$$

generalizing equation (19). □

The symmetry assumptions on demand in Examples 1 and 2 were important in establishing the uniqueness and the form of the Nash equilibrium. Now we consider the case where each player has the same unit cost $c_i = c$, $i = 1, 2, ..., n$, but are not otherwise identical, and we show the existence of a unique Nash equilibrium.

Proposition 2 *Under price complementarity, and assuming that all players follow a price-taking policy, there is a unique Nash equilibrium for the contract quantities y_i, $i = 1, 2, ..., n$. At* *the Nash equilibrium, the time-averaged expected price is the cost per unit of capacity:*

$$\int_0^1 \mathbb{E}[p(t)]\ dt\ = c, \tag{24}$$

and player i's optimal choice of contract quantity y_i is given by

$$y_i = c^{-1/\beta} \left(\int_0^1 \mathbb{E}[\alpha(t)^\beta]\, dt \right)^{\frac{1}{\beta}-1} \int_0^1 \mathbb{E}\left[\alpha(t)^{\beta-1} \alpha_i(t) \right]\ dt\ . \tag{25}$$

Moreover y_i satisfies the following equation:

$$y_i = \frac{\int_0^1 \mathbb{E}\left[p(t) D_i(t, p(t))\right]\ dt}{\int_0^1 \mathbb{E}[p(t)]\, dt}. \tag{26}$$

Observe that the form (26) exhibits player i's choice of contract quantity y_i as a price-weighted integral of $D_i(t, p(t))$, the anticipated usage by player i of the link. A similar form, but with a more general weight function, will occur in the next section.

Efficient investment in capacity occurs when the price on a link (what the users are prepared to pay for more capacity) is the same as the cost of additional capacity. However the situation here is complicated by the fact that prices fluctuate over time. The appropriate measure becomes the time average of the expected price. Hence equation (24) establishes that the Nash equilibrium induces players to contract for quantities which result in efficient investment in the capacity of the link.

Example 3: The impact of sharing on capacity built. Using Proposition 2 we see that the total capacity constructed by n players will be

$$Y\ = \sum_{i=1}^n y_i = c^{-1/\beta} \left(\int_0^1 \mathbb{E}[\alpha(t)^\beta]\, dt \right)^{\frac{1}{\beta}}. \tag{27}$$

If each player had constructed his own dedicated capacity, the total capacity built would be

$$\sum_{i=1}^n c^{-1/\beta} \left(\int_0^1 \mathbb{E}[\alpha_i(t)^\beta]\, dt \right)^{\frac{1}{\beta}}. \tag{28}$$

If, for $i = 1, 2, ..., n$, the parameter $\alpha_i(t)$ is neither time-varying nor stochastic, so that $\alpha_i(t) = \alpha_i$ for $t \in [0, 1]$ where α_i is a constant, then the total capacity constructed, Y, will be the same in both cases. Further, from equation (10), the price on either the shared link, or on each of the n dedicated links, will be $p(t) = c$, for $t \in [0, 1]$. For example, suppose that $\beta = 2$, and that player i carries a constant number, α_i, of connections over the interval $[0, 1]$. Then each connection will receive a throughput $c^{-1/2}$. Moreover, if the utility of each connection is $u(x) = -1/x^2$, then these values, for capacity initially constructed and throughputs later achieved, are exactly the values which maximize the sum of utilities over all connections minus the cost of the capacity.

Suppose that $\beta > 1$. Then the total dedicated capacity (28) is larger than the shared capacity (27), by Jensen's inequality, unless $\alpha_i(t) = \alpha_i$ with probability one for almost every $t \in [0, 1]$, $i = 1, 2, ..., n$.

Example 4: Effect of variability in demand. It is interesting to ask how variability in anticipated demand affects the size of the contract that a player will take. We use the Nash equilibrium result to explore this question. Consider a situation in which two players have the same average level of traffic so that $\int \mathbb{E}\alpha_1(t)\, dt = \int \mathbb{E}\alpha_2(t)\, dt$ but player 1's demand has higher variability, in the sense that $\int \mathbb{E}\alpha_1(t)^2\, dt > \int \mathbb{E}\alpha_2(t)^2\, dt$. We ask which of the two players will, in a Nash equilibrium, take the higher contract amount? The analysis here makes no distinction between variation over time and variance of the values of α_i at any fixed time. Thus we might be considering a case where player 1 knows that his traffic volume will fluctuate significantly according to the time of day while player 2 has a constant amount of traffic, but equally we might consider a situation in which both players have the same expected demand profile over the day, but for player 1 this is a forecast with considerable uncertainty, while player 2's usage can be predicted with near certainty.

Consider the case $\beta = 2$. Observe that y_1 and y_2 as given by (25) differ only in the term $\int_0^1 \mathbb{E}[\alpha(t)\alpha_i(t)]\, dt$. Now

$$\begin{aligned}
&\int_0^1 \mathbb{E}\left[(\alpha_1(t)+\alpha_2(t))\alpha_1(t)\right]\, dt - \int_0^1 \mathbb{E}\left[(\alpha_1(t)+\alpha_2(t))\alpha_2(t)\right]\, dt \\
&= \int_0^1 \mathbb{E}\alpha_1(t)^2\, dt - \int_0^1 \mathbb{E}\alpha_2(t)^2\, dt \ > \ 0.
\end{aligned}$$

Thus player 1, the player with higher variability of demand, will take the larger contract in this case. The result depends on the parameter β: if $\beta = 1$ the optimal choice of contract quantity (25) depends only on the expectation $\int \mathbb{E}\alpha_i(t)\, dt$ and not on higher moments. □

4.2 A network model

We next discuss a stylized network model, rather than a single link. Although we cannot give general sufficient conditions for a Nash equilibrium to exist, we shall see that, under certain conditions, any interior Nash equilibrium must take a form generalizing expression (26).

Associate each player with a single route, which is just some subset r of the set of links, J. Thus our stylized network model does not allow a player to control more than one route. It would certainly be desirable to remove this restriction, and to allow a player to distribute his traffic over several routes, but this is not allowed in the model considered here.

Suppose that each link j of the network is associated with a cost c_j per unit capacity. Let R be the set of players. For each $r \in R$, player r contracts for a capacity y_r over the balancing period $[0, 1]$ at an immediate cost to him of $y_r \sum_{j\in r} c_j$. If this is a consortium of players building a network, then a capacity of

$$Y_j = \sum_{r:j\in r} y_r \tag{29}$$

is built on link j at a cost of $c_j Y_j$. Thus the sum of the immediate costs to the players $r \in R$ exactly match the sum of the build costs of the links $j \in J$. In any case we will assume that each link is fully contracted so that (29) always holds.

At time $t \in [0,1]$ the demand from player r is a function $D_r(t, p_r(t))$ of a price $p_r(t)$, where $p_r(t)$ satisfies

$$\sum_{r:j\in r} D_r(t, p_r(t)) \leq \ Y_j \qquad j \in J \tag{30}$$

$$p_r(t) = \sum_{j\in r} p_j(t) \qquad r \in R. \tag{31}$$

The total cost to player r of the contract is

$$C_r \ \ = \ \ y_r \sum_{j\in r} c_j \ \ + \int_0^1 [D_r(t, p_r(t)) - y_r]\, p_r(t)\, dt. \tag{32}$$

Note that equations (30, 31) do not involve the costs $(c_j,\, j \in J)$, although we should expect these costs to influence the choice of $(y_r,\, r \in R)$ and hence of the capacities Y_j, $j \in J$.

We shall assume that the set of routes R includes $\{j\}$ for each $j \in J$, so that for each link of the network there is a player able to provide capacity on just that link. This ensures that the link-route incidence matrix has rank J. Observe that for simplicity of notation we are using J to indicate the total number of links, as well as set of links itself. Similarly, we shall write y_k for $y_{\{k\}}$.

We now describe a simple example, to illustrate the notation and one of the new features present in a network model.

Example 5: A hub and spokes network. Consider a network where the set of routes is $R = \{\{j\}, j \in J, \{i,j\}, i \neq j, i,j \in J\}$, and $D_r(p_r) = \alpha/p_r, r \in R$. We may view the J links of the network as forming a hub and J spokes, with each route comprising either a single link, or two links. From (30, 31) we have that

$$\sum_{i\neq j} \frac{\alpha}{p_i + p_j} + \frac{\alpha}{p_j} = Y_j.$$

Differentiating these equations with respect to y_k we obtain,

$$-\sum_{i\neq j} \frac{\alpha}{(p_i+p_j)^2}\left(\frac{\partial p_i}{\partial y_k} + \frac{\partial p_j}{\partial y_k}\right) - \frac{\alpha}{p_j^2}\frac{\partial p_j}{\partial y_k} = 0 \quad k \neq j$$

$$-\sum_{i\neq k} \frac{\alpha}{(p_i+p_k)^2}\left(\frac{\partial p_i}{\partial y_k} + \frac{\partial p_k}{\partial y_k}\right) - \frac{\alpha}{p_k^2}\frac{\partial p_k}{\partial y_k} = 1 \quad k = j.$$

Now suppose that $Y_2 = Y_3 \cdots = Y_J$, so that $p_2 = p_3 = \cdots = p_J$, and consider the effect of varying y_1. We can solve for the partial derivatives, and it follows that

$$\frac{\partial p_1}{\partial y_1} = -\frac{(p_1+p_2)^2}{J\alpha} + o(J^{-1})$$

while

$$\frac{\partial p_2}{\partial y_1} = \frac{2p_2^2}{J^2\alpha} + o(J^{-2})$$

as $J \to \infty$. Thus both partial derivatives decay with an increase in the number of links J in the network. But note especially that the second of the above partial derivatives decays much more quickly than the first. This is an intuitively plausible result: varying the capacity of link 1 should be expected to have a more significant effect on prices at link 1 than on prices at other links. □

In general

$$\frac{\partial p_r(t)}{\partial y_r} = \sum_{j \in r} \sum_{k \in r} \frac{\partial p_j(t)}{\partial y_k}.$$

We shall make the approximation that

$$\frac{\partial p_r(t)}{\partial y_r} = \sum_{j \in r} \frac{\partial p_j(t)}{\partial y_j}. \tag{33}$$

This approximation ignores the cross-derivatives $\partial p_j(t)/\partial y_k$ for $j \neq k$; we have seen that they are of smaller order than the diagonal terms $\partial p_j(t)/\partial y_j$, at least for the simple network described in Example 5. As well as this approximation, we assume price complementarity and that players act as price takers in the short term. However we make no assumption on the forms for the demand functions D_r, other than that they are decreasing and continuously differentiable.

Proposition 3 *If y_r, $r \in R$, is a Nash equilibrium at which $y_r > 0$, $r \in R$, then y_r satisfies the equation*

$$y_r \;=\; \frac{\int_0^1 \mathbb{E}\left[w_r(t) D_r(t, p_r(t))\right]\, dt}{\int_0^1 \mathbb{E}\left[w_r(t)\right] dt}, \tag{34}$$

where $w_r(t) = \partial p_r(t)/\partial y_r$. Further, the time-averaged expected price on link j is the cost per unit of capacity on link j:

$$\int_0^1 \mathbb{E}[p_j(t)]\; dt \;=\; c_j. \tag{35}$$

As in Proposition 2, the optimal choice of contract quantity (34) is just a weighted integral of the anticipated usage. Whether there exists a useful sufficient condition for equations (34, 35) to identify a unique Nash equilibrium in the network case remains an open question. Even in the single link case the assumption (9) on the form of the demand functions is critical; without this assumption it is possible to construct a single-link example where equations (34, 35) identify a point which is not a Nash equilibrium.

5 Implementation Issues

There are a number of issues which need to be dealt with in considering the implementation of the Contracting and Balancing Mechanism. Recall that with ECN marking, we suppose that $p_j(t)$ is given by a multiple γ of the proportion of packets passing through link j at time t that are marked. (For the purpose of this discussion we assume dropped packets are exceptional.) One critical issue is the information requirements. With a single link this is quite straightforward. The balancing process needs to know for each player the total number of packets that have been marked during the balancing period. If player i accounts for a total of z_i packets marked during the balancing period, then player i pays into the balancing process a net amount of

$$E_i \;=\; \int_0^1 [x_i(t) - \rho_i D(t)]\, p(t)\, dt \;=\; \gamma \left[z_i - \rho_i \sum_{k=1}^{n} z_k \right].$$

In implementing this scheme we might wish to have some check on the accuracy of reported numbers of marked packets (cf. Briscoe et al. 2000).

When we come to consider the information requirements for a network there are a number of options. First, consider the case where for each link j in the network we record Q_j, the total number of marks generated on that link during the balancing period (we do not need to assign these marks to individual routes). If we also know the total number of packets marked by each player, then from (2) the player associated with route r will make a net payment into the balancing process of

$$E_r \;=\; \gamma \left[z_r - y_r \sum_{j \in r} \frac{Q_j}{Y_j} \right]. \tag{36}$$

This calculation assumes that Q_j is not incremented when a packet that is already marked is marked again (otherwise, the sum of all balancing payments might not be zero).

In order for the Contract and Balancing Mechanism to work well in a network we need the price for a route r to be given by the sum of the prices on the links of that route (i.e. for relation (31) to hold). If the marking probabilities on any link are small, then it is unlikely that the same packet will be marked twice and this assumption will be sufficiently accurate.

In the case where we do not have access to marked packet counts for individual links in the network, then it can still be possible to make estimates for the amounts to be paid in the balancing process. Suppose that price complementarity holds. Then E_r is given by (4) and the estimation problem becomes that of estimating $\int_0^1 p_r(t)\, dt$. It is possible that players would agree to this being estimated for each route by the network owner sending uniformly distributed "probe packets" on that route. Then, if P_r is the proportion of probe packets marked or dropped during the balancing period,

$$E_r = \gamma(z_r - y_r P_r).$$

Another possibility is that the network owner might identify a subset of a player's packets, uniformly distributed in time, as probe packets. Notice that we cannot sensibly use the overall proportion of a player's packets marked or dropped (i.e. $Q_r / \int_0^1 x_r(t)\, dt$) as an estimate of $\int_0^1 p_r(t)\, dt$. Doing so would just encourage players to send artificially high amounts of traffic at quiet times in order to decrease this ratio

A remaining issue is the choice of γ, the charge per mark. This figure has to be agreed to by the users at the outset as part of the contract arrangement - it is simply a scale factor applied to the price. It is interesting that if price complementarity holds then the choice of γ will make very little difference to the outcome in terms of the contract amounts y_i or the transmission rates that actually occur. It turns out that changing the value of γ does not lead to a change in the price, which is the product of γ and the marking probability. Instead, increasing γ just leads to a scaling down of the marking probability with the average price in the link staying the same (as is shown by (23)). As we discussed earlier, changes in the marking probability are not necessarily related to changes in traffic volume. In this case both y_i, given by (25), and the traffic volume, $D_i(t, p(t))$, remain unchanged. Even if price complementarity does not hold exactly, this will still be approximately true. Hence we can set an appropriate value for γ, the charge per mark, by deciding on a desired rate of marked packets, and then using (23) to determine γ. In this way we can minimize the risk that the marking probability becomes high.

6 Conclusion

In this paper we have proposed a Contract and Balancing Mechanism, involving long term contracts and a short term balancing process, as a method for sharing resources in a communication network. The approach allows volatile prices to be appropriately averaged, so as to mediate between rapidly fluctuating congestion prices and the longer time scales over which bandwidth contracts might be traded, and eliminates the incentive for a capacity owner to increase congestion.

In Section 1.1 we set out four characteristics that a bandwidth charging mechanism should possess: sharing of resources when demand is unpredictable; allocation of resources in a way that reflects the users' underlying utilities; payment to the network provider based on bandwidth provided rather than demand; and protection of users from high costs when they have low usage These are features of the mechanism we propose, but it is not the only mechanism that has these properties. For example, we could use a balancing mechanism based on the square of the number of marks received over the balancing period. More generally, for any increasing function $f(\cdot)$, a user who generates z_i marks could pay into the balancing mechanism an amount $f(z_i)$ and receive back a proportion ρ_i of the total payments made by all users. This gives a balancing payment of

$$E_i = \int_0^1 [f(x_i(t)\, p(t)) - \rho_i \sum_{k=1}^{n} f(x_k(t)\, p(t))]dt.$$

The CBM approach is the simplest mechanism of this form and has the benefit of inducing appropriate investments in link capacity (as shown by Proposition 2).

We have studied the existence and form of Nash equilibria for players' choices of capacity using a framework in which the players each begin by buying some capacity at the first stage, then the traffic eventuates and finally payments are made to other players as a result of the balancing process. We have been able to show that in many cases the choice of capacity at the equilibrium will be close to the predicted traffic demand at the anticipated price.

Specifically, we have three main results. The first states that each player will have a unique optimal choice of contract quantity for a link given any set of contract quantities by the other players. The second states that if the players have the same marginal link cost and if they all follow a price-taking policy, then there is a unique Nash equilibrium for the contract quantities; further, the time-averaged expected price is the cost per unit of capacity. Finally, the third result generalizes the second result, under certain conditions, by finding the form of an interior Nash equilibrium.

The Contract and Balancing mechanism we propose can also be employed in other situations in which users compete for a service resource with significant negative externalities, so that costs are imposed on other users when one of the users increases his use of the resource. For example this occurs when users queue for a limited capacity resource. In order to use the mechanism it would be necessary to record for each user, as a congestion charge, an estimate of the externality (Dewan and Mendelson 1990) imposed on other users. However, rather than pay the aggregate congestion charges to the service provider after the event, the user contracts with the service provider in advance for a proportion of the total congestion charges. This payment would be made in addition to any more conventional usage charges. Then, at the end of each balancing period, a balancing process occurs: the proportion of congestion charges contracted for is compared with the proportion of congestion charges incurred. Users who turn out to be responsible for more than their contracted proportion of congestion charges need to make a further payment, while users who end up with a smaller than contracted proportion of congestion charges will receive a payment. The advantage of this sort of charging scheme is that it provides appropriate price signals for the user while avoiding uncertainty in income for the service provider.

An important further issue relates to competition among network providers. Though we do not deal with this issue explicitly in this paper, we do in Example 2 consider a form of competition between suppliers of capacity on the same link. It would be of considerable interest to consider competition among providers of capacity on different links. Where the links are direct substitutes, it is natural to consider the case where routing is sufficiently flexible to enable an equalization of congestion prices on the various links. In this case, CBM applied to each link separately will have the same net result as if it were applied to the combined link. An analysis similar to that of Example 2 could then be carried out, with Cournot outcomes from the capacity-setting game played by different providers. This analysis can still be carried through when different links have different marking procedures, or different values of γ associated with them. (See Kreps and Scheinkman (1983) for another example of Cournot outcomes when there are pre-committed capacities followed by price-setting.) Further research could consider more general network contexts in which routing flexibility complements the Contract and Balancing Mechanism.

Acknowledgement

We are grateful to Ou Jihong for pointing out an error in an earlier version of Section 4.2 and for a simplification in the proof of Proposition 1, and to the referees for their helpful suggestions and comments.

References

Briscoe, B., V. Darlagiannis, O. Heckman, H. Oliver, V. Siris, D. Songhurst, and B. Stiller. 2003. A market managed multi-service Internet. *Computer Communications* **26**, 404-414.

Briscoe, B., M. Rizzo, J. Tassel and K. Damianakis. 2000. Lightweight policing and charging for packet networks. *Third IEEE Conference on Open Architectures and Network Programming.*

Clark, D.D. 1996. Adding service discrimination to the Internet. *Telecommunications Policy*, **20**, 169-181.

Dewan, S. and H. Mendelson. 1990. user delay costs and internal pricing for a service facility. *Management Science* **36**, 1502-1517.

Floyd, S. 1994. TCP and Explicit Congestion Notification. *ACM Computer Communication Review* **24**, 10-23.

Floyd, S 2003. HighSpeed TCP for large congestion windows. Network Working Group, Internet Engineering Task Force, December.

Floyd, S., M. Handley, J. Padhye, and J. Widmer. 2000. Equation-based congestion control for unicast applications. In *Proc. ACM SIGCOMM 2000*, pages 43-54, Stockholm.

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

Ganesh, A., K. Laevens and R. Steinberg. 2000. Dynamics of congestion pricing. Microsoft Research Technical Report MSR-TR-2000-70, August. Revised August 2004.

Gibbens, R.J. and F.P. Kelly. 1999. Resource pricing and the evolution of congestion control. *Automatica* **35**, 1969-1985.

Green. R.J. and D.M. Newbery. 1992. Competition in the British electricity spot market. *J. Political Econ.* **100**, 929–953.

Kelly, T. 2003. Scalable TCP: improving performance in highspeed wide area networks. *Computer Communication Review* **33**, 83-91.

Key, P. 1999. Service differentiation: congestion pricing, brokers and bandwidth futures. *9th International Workshop on Network and Operating Systems Support for Digital Audio and Video.*

Kirkpatrick, D. 2000. Enron takes its pipeline to the net. *Fortune*, January 24, pp. 127-130.

Klemperer, P.D. and M.A. Meyer. 1989. Supply function equilibria in oligopoly under uncertainty. *Econometrica* **57**, 1243-1277.

Kreps, D.M. and J.A. Scheinkman. 1983. Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell Journal of Economics*, **14**, 326-337.

Kunniyur, S. and R. Srikant. 2003. End-to-end congestion control: utility functions, random losses and ECN marks. *IEEE/ACM Transactions on Networking* **11**, 689-702. Low, S.H., F. Paganini, and J.C. Doyle. 2002. Internet congestion control. *IEEE Control Systems Magazine* **22**, 28-43.

MacKie-Mason, J.K. and H.R. Varian. 1995. Pricing congestible network resources. *IEEE Journal on Selected Areas in Communications* **13**, 1141-1149.

Masuda, Y. and S. Whang. 1999. Dynamic pricing for network service: Equilibrium and stability. *Management Science* **45**, 857-569.

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

Newbery, D.M. 1998. Competition, contracts and entry in the electricity spot market. *RAND Journal of Economics*, **29**, 726-749.

Ramakrishnan, K.K., S. Floyd, and D. Black. 2001. The addition of Explicit Congestion Notification (ECN) to IP. Transport Area Working Group, Internet Engineering Task Force, September.

Reichl, P., P. Flury, J. Gerke, and B. Stiller. 2001. How to overcome the feasibility problem for tariffing Internet services. *IEEE International Conference on Communications (ICC' 2001)*, Helsinki.

Rump, C.M. and S. Stidham, Jr. 1998. Stability and chaos in input pricing for a service facility with adaptive customer response to congestion. *Management Science* **44**, 246-261.

Varian, H.R. 1992. *Microeconomic Analysis.* Third Edition. Norton: New York.

Appendix

Proof of Proposition 1

The first order conditions for an optimal choice of contract quantity require

$$\frac{\partial V_i}{\partial y_i} \;=\; \frac{\partial U_i}{\partial y_i} - \frac{\partial W_i}{\partial y_i} \;=\; 0. \tag{37}$$

From (14),

$$\begin{aligned}\frac{\partial U_i}{\partial y_i} &= \int_0^1 \mathbb{E}\left[u_i'(D_i(t,p(t)))D_i'(t,p(t))\frac{\partial p(t)}{\partial y_i}\right] dt \\ &= \int_0^1 \mathbb{E}\left[p(t)D_i'(t,p(t))\frac{\partial p(t)}{\partial y_i}\right] dt,\end{aligned}$$

using the assumption that player i acts as a price taker for the choice of $x_i(t) = D_i(t,p(t))$. Under the assumption of price complementarity,

$$W_i = C_i(y_i) + \int_0^1 \mathbb{E}[(D_i(t,p(t)) - y_i)\,p(t)]\,dt$$

and

$$\frac{\partial W_i}{\partial y_i} \;=\; c_i + \int_0^1 \mathbb{E}\left[\big(p(t)D_i'(t,p(t)) + D_i(t,p(t)) - y_i\big)\,\frac{\partial p(t)}{\partial y_i} - p(t)\right] dt.$$

Thus at a stationary point

$$\frac{\partial V_i}{\partial y_i} \;=\; \int_0^1 \mathbb{E}[p(t)]\;dt - \int_0^1 \mathbb{E}\left[(D_i(t,p(t)) - y_i)\,\frac{\partial p(t)}{\partial y_i}\right] dt - c_i = 0 \tag{38}$$

Since $D_i(t,p(t))$ is given by (9), $p(t)$ is given by (10). Now consider $\partial p(t)/\partial y_i$, which measures how price varies with changes in capacity at time t, and is equal to $\partial p(t)/\partial Y$ for each i. We have that:

$$\frac{\partial p(t)}{\partial y_i} = -\beta\left(\frac{\alpha(t)}{Y}\right)^{\beta}\frac{1}{Y} \tag{39}$$

Substituting (9), (10), and (39) into (38) yields:

$$\begin{aligned}\frac{\partial V_i}{\partial y_i} \;&=\; \int_0^1 \mathbb{E}\left[\left(\frac{\alpha(t)}{Y}\right)^{\beta}\right]\;dt + \int_0^1 \mathbb{E}\left[\left(\frac{\alpha_i(t)}{p(t)^{1/\beta}} - y_i\right)\beta\left(\frac{\alpha(t)}{Y}\right)^{\beta}\frac{1}{Y}\right] dt - c_i \\ &= \frac{1}{Y^{\beta}}\int_0^1 \mathbb{E}\left[\alpha(t)^{\beta}\left(1 + \beta\frac{\alpha_i(t)}{\alpha(t)} - \beta\frac{y_i}{Y}\right)\right]\;dt - c_i,\end{aligned} \tag{40}$$

which gives the formula (17) at a stationary point.

Observe that for y_i very large, Y becomes large and $\partial V_i/\partial y_i$ is negative. We will show below the following property of V_i: for every value of y_i at which the second derivative of V_i is zero or positive, the derivative of V_i is negative. This property is enough to show that either there is exactly one stationary point which is a maximum, or the maximum is achieved at $y_i = 0$. (This in turn will establish the result.) First observe that this property implies that at any stationary point, V_i must have strictly negative second derivative and hence be a local maximum. Thus if $\partial V_i/\partial y_i < 0$ when $y_i = 0$ it can never have a turning point at positive y_i, and V_i achieves its maximum at 0. If $\partial V_i/\partial y_i \geq 0$ at 0 then there will be at least one stationary point in $[0, \infty)$. Clearly a stationary point of V_i implies (17). But now note that if there are two stationary points then the second derivative of V_i cannot be negative throughout the interval between them, and we obtain a contradiction by considering the largest value of y_i between them at which the second derivative is not negative.

It only remains to prove the property we have referred to. Now

$$\frac{\partial^2 V_i}{\partial {y_i}^2} = \int_0^1 \mathbb{E}\left[2\frac{\partial p(t)}{\partial y_i} - (D_i(t,p(t)) - y_i)\frac{\partial^2 p(t)}{\partial y_i^2} - D_i'(t,p(t))\left(\frac{\partial p(t)}{\partial y_i}\right)^2\right] dt. \tag{41}$$

Since

$$\frac{\partial^2 p(t)}{\partial y_i^2} = -\frac{D''(t,p(t))}{(D'(t,p(t)))^2}\frac{\partial p(t)}{\partial y_i}.$$

we have

$$\frac{\partial^2 V_i}{\partial {y_i}^2} = \int_0^1 \mathbb{E}\left[\left(2 + (D_i(t,p(t)) - y_i)\frac{D''(t,p(t))}{D'(t,p(t))^2} - \frac{D_i'(t,p(t))}{D'(t,p(t))}\right)\frac{\partial p(t)}{\partial y_i}\right] dt.$$

For this form of demand we have

$$D''(t,p(t)) = \frac{(\beta+1)\alpha(t)}{\beta^2 p(t)^{(2\beta+1)/\beta}}$$

and so

$$\frac{D''(t,p(t))}{(D'(t,p(t)))^2} = \frac{(\beta+1)}{Y}.$$

Hence

$$\begin{aligned}\frac{\partial^2 V_i}{\partial {y_i}^2} &= -\int_0^1 \mathbb{E}\left[\left(2 + \left(\frac{\alpha_i(t)}{p(t)^{1/\beta}} - y_i\right)\frac{(\beta+1)}{Y} - \frac{\alpha_i(t)}{\alpha(t)}\right)\beta\left(\frac{\alpha(t)}{Y}\right)^\beta\frac{1}{Y}\right] dt \\ &= -\frac{\beta}{Y^{\beta+1}}\int_0^1 \mathbb{E}\left[\left(2 + \left(\frac{\alpha_i(t)}{\alpha(t)} - \frac{y_i}{Y}\right)(\beta+1) - \frac{\alpha_i(t)}{\alpha(t)}\right)\alpha(t)^\beta\right] dt \\ &= -\frac{\beta}{Y^{\beta+1}}\int_0^1 \mathbb{E}\left[\left(2 + \beta\frac{\alpha_i(t)}{\alpha(t)} - (\beta+1)\frac{y_i}{Y}\right)\alpha(t)^\beta\right] dt.\end{aligned}$$

Thus if $\partial^2 V_i/\partial {y_i}^2$ is non-negative we must have

$$\int_0^1 \mathbb{E}\left[\left(2 + \beta\frac{\alpha_i(t)}{\alpha(t)} - (\beta+1)\frac{y_i}{Y}\right)\alpha(t)^\beta\right] dt \leq 0. \tag{42}$$

But note that for every t, and every realization of the $\alpha_i(t)$,

$$1 + \beta\frac{\alpha_i(t)}{\alpha(t)} - \beta\frac{y_i}{Y} \leq 2 + \beta\frac{\alpha_i(t)}{\alpha(t)} - (\beta+1)\frac{y_i}{Y}.$$

Thus if inequality (42) holds, then the integral term in the equation (40) is non-positive, and hence $\partial V_i/\partial y_i < 0$ whenever the second derivative is non-negative. □

Proof of Proposition 2

Since each of the players chooses short term traffic volumes on a price-taking basis and chooses capacity optimally, we can use the development we gave for the proof of Proposition 1. We look first for a Nash equilibrium at which each y_i is non-zero, and so the set of y_i will satisfy (38). Because of the form of the demand functions we know that, except when $\alpha(t) = 0$, we have $D(t, p(t)) = Y$. On the other hand if $D(t, p(t)) < Y$ then from price complementarity $p(t) = 0$ and $\partial p(t)/\partial Y = 0$. Hence under any realization of demands

$$\int_0^1 (D(t, p(t)) - Y) \frac{\partial p(t)}{\partial Y} \, dt = 0.$$

Thus if we sum (38) over i, the terms involving $\partial p/\partial y_i$ sum to zero, giving

$$\int_0^1 \mathbb{E}[p(t)] \, dt \;=\; c. \tag{43}$$

Thus, again from (38),

$$\int_0^1 \mathbb{E}\left[(D_i(t, p(t)) - y_i) \frac{\partial p(t)}{\partial y_i}\right] \, dt = 0$$

at the optimizing choice of y_i. Since

$$\frac{\partial p(t)}{\partial y_i} = -\left(\frac{\alpha(t)}{Y}\right)^{\beta} \frac{\beta}{Y},$$

the equation (26) for y_i follows from (10).

This can be simplified by substituting for p and D_i to obtain

$$\begin{aligned} y_i &= \frac{\int_0^1 \mathbb{E}\left[\alpha(t)^{\beta} \alpha_i(t)/p(t)^{1/\beta}\right] \, dt}{\int_0^1 \mathbb{E}[\alpha(t)^{\beta}] \, dt} \\ &= \frac{Y \int_0^1 \mathbb{E}\left[\alpha(t)^{\beta-1} \alpha_i(t)\right] \, dt}{\int_0^1 \mathbb{E}[\alpha(t)^{\beta}] \, dt}. \end{aligned}$$

But using the expression (10) for $p(t)$ we can rewrite (43) as

$$\int_0^1 \mathbb{E}\left[\alpha(t)^{\beta}\right] \, dt = Y^{\beta} c,$$

and so

$$y_i = c^{-1/\beta} \left(\int_0^1 \mathbb{E}[\alpha(t)^{\beta}] \, dt\right)^{\frac{1}{\beta}-1} \int_0^1 \mathbb{E}\left[\alpha(t)^{\beta-1} \alpha_i(t)\right] \, dt$$

as required. This is the only Nash equilibrium with all $y_i > 0$, but we need to rule out the possibility of some y_i being zero.

At a solution with $y_i = 0$, we have

$$\int_0^1 \mathbb{E}\left[\alpha(t)^{\beta} \left(1 + \beta \frac{\alpha_i(t)}{\alpha(t)}\right)\right] \, dt < Y^{\beta} c$$

and hence

$$\int_0^1 \mathbb{E}\left[\alpha(t)^\beta\right]\ dt < Y^\beta c. \tag{44}$$

Now each of the non-zero y_j satisfies

$$\int_0^1 \mathbb{E}\left[\alpha(t)^\beta \left(1+\beta\frac{\alpha_j(t)}{\alpha(t)}-\beta\frac{y_j}{Y}\right)\right]\ dt = Y^\beta c.$$

Let $J \subset \{1,2,...,n\}$ be the set of indices for which $y_j \neq 0$. Summing this over $j \in J$ gives

$$\int_0^1 \mathbb{E}\left[\alpha(t)^\beta \left(1+\beta\frac{\sum_{j\in J}\alpha_j(t)}{\alpha(t)}-\beta\right)\right]\ dt = Y^\beta c.$$

But $\sum_{j\in J}\alpha_j(t) < \alpha(t)$ and so this implies $\int_0^1 \mathbb{E}\left[\alpha(t)^\beta\right]\ dt > Y^\beta c$, giving a contradiction from (44). □

Proof of Proposition 3

Consider first the case without time-dependence or stochastic effects. Then the total cost to player r of the contract is

$$W_r \ = \ y_r \sum_{j\in r} c_j \ + [D_r(p_r) - y_r]\, p_r. \tag{45}$$

Thus

$$\frac{\partial W_r}{\partial y_r} \ = \ \sum_{j\in r} c_j \ + \ \frac{\partial}{\partial y_r}\left\{[D_r(p_r)-y_r]\,p_r\right\}.$$

But

$$\frac{\partial\, U_r(D_r(p_r))}{\partial y_r} = p_r D_r'(p_r)\frac{\partial p_r}{\partial y_r}.$$

The first order conditions for a Nash equilibrium, relation (14) with i replaced by r, require that we set these derivatives equal, which gives

$$\sum_{j\in r}(c_j - p_j) = [y_r - D_r(p_r)]\frac{\partial p_r}{\partial y_r},$$

where we have made use of (31). Similarly the first order conditions in the time-varying stochastic case are

$$\sum_{j\in r}(c_j - \int_0^1 \mathbb{E}\left[p_j(t)\right] dt) = \int_0^1 \mathbb{E}\left[(y_r - D_r(t,p_r(t)))\frac{\partial p_r(t)}{\partial y_r}\right] dt$$

From (33) we deduce that for each $r \in R$

$$\sum_{j\in r}(c_j - \int_0^1 \mathbb{E}\left[p_j(t)\right] dt) = \sum_{j\in r}\int_0^1 \mathbb{E}\left[[y_r - D_r(t,p_r(t))]\frac{\partial p_j(t)}{\partial y_j}\right] dt.$$

The incidence matrix $A = (A_{jr})$ of links on routes has rank J, and hence $A^T z = 0 \Rightarrow z = 0$. We can thus deduce that

$$c_j - \int_0^1 \mathbb{E}\left[p_j(t)\right] dt = \int_0^1 \mathbb{E}\left[[y_r - D_r(t,p_r(t))]\frac{\partial p_j(t)}{\partial y_j}\right] dt.$$

But at each time t and every realization of the $\alpha_i(t)$, either relation (30) holds with equality, or $\partial p_j(t)/\partial y_j = 0$. Hence, summing the above equality over routes r such that $j \in r$, we obtain

$$c_j = \int_0^1 \mathbb{E}\left[p_j(t)\right] dt,$$

and so

$$y_r \int_0^1 \mathbb{E}\left[\frac{\partial p_j(t)}{\partial y_j}\right] dt = \int_0^1 \mathbb{E}\left[D_r(t, p_r(t)) \frac{\partial p_j(t)}{\partial y_j}\right] dt$$

giving the formula we require. □