Research Papers
in Management Studies

UNIVERSITY OF
CAMBRIDGE
Judge Institute of Management
cambridge's business school

PRICING INTERNET SERVICE

R Steinberg

WP 18/2002

The Judge Institute of Management
University of Cambridge
Trumpington Street
Cambridge CB2 1AG, UK
www.jims.cam.ac.uk

PRICING INTERNET SERVICE

R Steinberg

WP 18/2002

Dr Richard Steinberg
Judge Institute of Management
University of Cambridge
Tel: +44 (0) 1223 339638
Fax: +44 (0) 1223 339701
Email: r.steinberg@jims.cam.ac.uk

Please address enquiries about the series to:

Research Support Manager
Judge Institute of Management
Trumpington Street
Cambridge CB2 1AG, UK
Tel: 01223 339700 Fax: 01223 339701
E-mail: research-support@jims.cam.ac.uk

Pricing Internet Service

Richard Steinberg
University of Cambridge
Judge Institute of Management
Trumpington Street
Cambridge CB2 1AG
England

March 2003

To appear in: A.K. Chakravarty and J. Eliashberg, eds., *Managing Business Interfaces: Marketing, Engineering, and Manufacturing Perspectives*, Dordrecht: Kluwer Academic Publishers/ISQM, 2003.

Abstract

Internet service needs to be appropriately priced, because Internet Service Providers want to remain competitive, and because pricing is increasingly seen by engineers as an essential tool for controlling network congestion. Thus, the pricing of Internet service lies at the interface of marketing and engineering. From this perspective we present a review of the major proposals for pricing Internet service. Topics include: Smart Market, Paris Metro Pricing, Max-Min Fairness, Proportional Fairness, CPE Pricing, and the Contract and Balancing Process. Background information on the history and technology of the Internet is also provided.

1 Introduction

How should Internet service be priced? The answer depends on who is pricing to whom. In this chapter, we will for the most part be concerned with how an Internet Service Provider (ISP) should price to its customers, the "end users." The methods we consider here can, for example, help AOL determine how it should price Internet service to your home, but also how much it should charge a business customer for a direct connection from the business's own networks to the Internet. We will also discuss Internet pricing from another—but closely related—point of view, viz., how the owner of a dedicated communication network should price bandwidth to businesses that host Internet applications. Internet service needs to be appropriately priced, both because Internet Service Providers want to remain competitive, and because pricing is increasingly seen by engineers as an essential tool for controlling network congestion. Thus, the pricing of Internet service lies at the interface of marketing and engineering. It is from this perspective that we present a review of the major proposals for pricing Internet service.

Our naïve opening question is quite general, so let us consider some specifics: Can Internet pricing be determined by an auction mechanism? Can that elusive concept, Quality of Service, be endogenously determined, leaving the ISP, so to speak, off the hook? Can Internet pricing be based on the philosophical concept of fairness? Can the individual packets that comprise Internet traffic be priced via a consumer demand function? And finally, can the pricing of bandwidth be based on a contractual agreement? In this chapter, we consider all these questions and the often ingenious methods that have been suggested for addressing them.

In the following section, we discuss what the Internet is, how it developed, and how it works. In Sections 3 through 7, we examine, respectively, Smart Market, Paris Metro Pricing, Max-Min Fairness, Proportional Fairness, and CPE Pricing. In Section 8 we look at a proposal for bandwidth pricing, the Contract and Balancing Process. Section 9 concludes with related issues, as well as a brief mention of several other pricing proposals.

A review such as this is necessarily subjective, and other authors would have different lists of the "major" proposals. However, any credible review of Internet pricing must include two: Smart Market, because it was the first such proposal and has been enormously influential, and Proportional Fairness, because it is grounded in a deep and elegant theory that generalizes several earlier approaches. However, all the proposals discussed here are significant, and it is impossible to say which one, if any, will become the standard.

Finally, note that we do not consider the pricing of Internet *content*, i.e., digital information goods; for this topic, especially with regard to bundling, see Bakos and Brynjolfsson (2000). For a comprehensive overview of how the Internet is affecting marketing generally, see Barwise, Elberse and Hammond (2002).

2 The Internet

The Internet is the publicly accessible network of computers that exchange data worldwide. However, this simple definition omits the Internet's most distinguishing trait. As Papadimitriou (2002) lucidly puts it:

> The most novel and defining characteristic of the Internet is its nature as an artifact that was not designed by a single entity, but emerged from the complex interaction of many economic agents (network operators, service providers, users, etc.), in various and varying degrees of collaboration and competition.

The truth of Papadimitriou's remark is evident from a recapitulation of the Internet's ontogeny.

The Internet appeared in embryo as the U.S. Defense Department's *Advanced Research Projects Agency Network*. ARPANET was designed to be a robust communication network that could survive a nuclear attack, even if one or more sites were destroyed. When ARPANET first went into operation in 1969, it was comprised of four host sites: UCLA, Stanford Research Institute, UC Santa Barbara, and the University of Utah. (Here *host* refers to any computer that has full access to other computers on the network.) By 1971, other educational and research institutions had joined ARPANET, so that it was now comprised of fifteen hosts. At this time, it was used primarily by a small number of computer scientists interested in developing networking technologies.

The number of hosts grew modestly over the next few years. In 1974, the word "internet" appeared in the title of a document by Vinton Cerf, a DARPA[1] scientist at Stanford Research Institute, and two Stanford University graduate students, Yogen Dalal and Carl Sunshine (Cerf, Dalal, and Sunshine 1974). They used the term as a short form for "internetwork," in the phrase "internetwork transmission control program." This refers to the method of transmitting files by breaking them down into small units of information called "packets" which are then sent over the network. The Transmission Control Program, they explained, "acts in many ways like a postal service since it provides a way for processes to exchange letters with each other" (Cerf, Dalal, and Sunshine, 1974, p. 3).

In 1978, the Transmission Control Program was split into the Transmission Control Protocol (TCP) and the Internet Protocol (IP), known jointly as TCP/IP, which remains to this day the standard communication language of the Internet. In 1980 and 1981 another—completely independent—network was founded, called CSNET (Computer Science NETwork), through grant funds from the National Science Foundation (NSF). By the end of 1981, ARPANET had two hundred hosts.

[1] In 1973, ARPA had been renamed DARPA (D for "Defense").

By 1983, all research networks were converted to TCP/IP and were inter-connected through ARPANET or CSNET. It was around this time that the capitalized term "Internet" came to mean a connected set of networks, specifically those using TCP/IP. By 1984, there were more than a thousand hosts.

In 1986, the National Science Foundation created a research network called NSFNET, which was significant for its high-speed backbone. (A *backbone* is a large transmission line for long-distance connection that carries data gathered from smaller lines that interconnect with it.) At the same time, the NSF supported the development of regional networks that could carry traffic from individual organizations, such as government agencies and universities, to the national backbone service. This led to the number of hosts multiplying from about two thousand early in the year, to more than five thousand by the end of the year.

In 1987, NSF commissioned Merit Network, Inc., MCI, IBM, and the State of Michigan to manage the NSFNET backbone project, and by year-end the number of hosts exceeded ten thousand. As more networks joined up, the Internet became truly international, and by 1989, there were over a hundred thousand hosts worldwide. In 1990, the ARPANET was retired, but NSFNET continued to grow. By now, the phrase "the Internet" had appeared in print (Wall and Schwarz 1990, Chapter 6, p. 260), and the number of hosts had passed three hundred thousand.

Also in 1990, a computer scientist working at CERN (European Organization for Nuclear Research), Tim Berners-Lee, developed the World Wide Web as a tool for collaboration for the high-energy physics community. "The Web" was publicly released by CERN the following year. Also in 1991, the NSF removed commercial restrictions on the NSFNET. By 1992, the number of hosts had reached the million mark.

In 1995, NSFNET reverted back to a research network, and the main US backbone traffic was re-routed through various interconnected network providers. Thus, the Internet—in what is essentially its current form—was born. At the time, it had five million hosts. By one estimate, in July 2002 there were in excess of 165 million Internet hosts (Zakon 2003).

2.1 The Web

Most people erroneously consider the Internet to be synonymous with the *World Wide Web*. However, unlike the Internet, the Web has no physical existence and is simply a method for accessing information via the Internet.

Every host computer on the Internet has associated with it a unique number known as an *IP address*. Through the use of browsers such as Internet Explorer and Netscape, the Web allows users to automatically search the Internet for the IP address of the computer that contains the requested file—e.g., a text document, graphic image, or a sound, video or multimedia file—and call it up for "viewing." Web pages are formatted in a system of symbols called HTML (HyperText Markup Language) that supports links to other files, i.e., access via a simple click of the mouse. The Web is just one of many ways that information can be disseminated over the Internet; among the others, email is probably most familiar.

2.2 Network Externalities and Metcalfe's Law

The extraordinary rate of growth of the Internet can be explained by the commonplace that its value rises with the number of computers connected to it—this effect is called "network externalities"—which leads to a snowballing effect. By how much does the value rise? The most popular view is provided by *Metcalfe's Law*, which was proposed by George Gilder in a discursive column in Forbes ASAP (Gilder 1993). Gilder named the principle in honor of Robert Metcalfe who, as a member of the research staff of Xerox PARC (Palo Alto Research Center) in 1973, predicted the potential of a system for connecting hundreds of computers within a building using hardware running from machine to machine.[2] Gilder later provided a succinct formulation (Gilder 1999):

> Metcalfe's Law: The observation by Robert Metcalfe that the performance and value of a network rises by the square of the total power of the computers compatibly attached to it.

The idea behind this square law is very straightforward. Assume for simplicity that the power of each of the computers is roughly the same, which we set to be 1. Then if there are n computers attached to the network, the value of the network to each one of them is the number of computers, or n. Thus, the total value of the network is the number of computers times the value to each computer, or n^2.

Although Metcalfe's Law is frequently and enthusiastically cited, it also has its detractors. For example, Odlyzko (2000) dryly observes: "If this 'law' were indeed valid, there would be no need to worry about general connectivity. Two equal size networks would double the total value, and so the two carriers would have an irresistible economic incentive to link up, either through merger or some standard interconnection." Although he expands further on this idea, Odlyzko eventually concedes: "However, this law does reflect a fundamental truth, that the value of a point-to-point communication network increases faster than just in proportion to the number of users." And this is the only point we wish to observe here. But how exactly does the Internet work?

[2] Metcalfe went on to design one of the more popular systems for doing this, called *Ethernet*.

2.3 Packets, Bytes, and Bits

Although much of Internet traffic is carried over the same lines as telephone calls, the technology of the Internet—called *packet switching*—is very different from the technology of the public telephone network—called *circuit switching*. In a circuit switched network, a specific path to the destination is obtained for the message, and a fixed amount of capacity is reserved for the duration of the call, during which time no other information is sent along the same path.

By contrast, in a packet-switched network the file to be transmitted is first broken down into small units, called packets, which are numbered and affixed with a header containing the packet number, the destination computer's IP address, and other information. (The header also contains unassigned fields which can be used for future applications, e.g., congestion notification.) The packets are then sent through the network along individual paths, where the choice of path depends on the other flows simultaneously in progress. At the destination, the headers are stripped off and the packets re-assembled into the original file.

The data in a packet is stored in units of information called *bytes*. A byte in text might correspond to a single character, such as a letter, number, or typographic symbol (e.g., 't', '9', or '@'), but in a visual image might correspond to a color or a pixel location. A byte is typically eight bits long, where a *bit* (i.e., *bi*nary digi*t*) is the smallest unit of data and has a value of 0 or 1.

The *data transfer rate*, or simply *rate*, at which data is sent through the network is measured in bits per second (bps)—or, more usually—megabits per second (Mbps). The rate depends on the load on the *resources* involved, i.e., any elements of the network required for communication that might be potential bottlenecks. An example of a resource would be a fiber optic cable between New York and London.

2.4 TCP/IP and Explicit Congestion Notification

TCP/IP (Transmission Control Protocol/Internet Protocol) is the basic communication language, or protocol, of the Internet. TCP enables two host computers to establish a connection, and manages the breaking down of the file into packets at the source (the sending computer) and the re-assembling of the message at the destination (the receiving computer). IP specifies the format of packets, as well as the addressing scheme, so that the packets get to the right destination. Although all packets from the same file will have the same destination IP address, in general each packet will have its own routing through the network. When the packets arrive at the destination computer, it is TCP on that machine that strips off the headers and reassembles the data into the original file. TCP on the receiving machine also sends "acknowledgement packets" back to the source computer to verify the integrity of the data received; if no acknowledgement packet comes back after a specified time interval, the packets that were not acknowledged are retransmitted by the sending machine.

TCP also has the responsibility for adapting the sending rate of the source computer to the capacity of the network. 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. Consequently, 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 congestion via the loss of packets, and the cycle begins anew.

Formally, *congestion* can be defined as the loss of network performance that occurs when the users of the network collectively demand more resources than the network can provide. Users of course need to know when the network is congested, but dropping packets is a crude way of notifying them: it is both wasteful (a dropped packet might have already consumed resources; it might also need to be resent) and untimely (until their packets are dropped, users are unaware of the existence of congestion).

These drawbacks have led to proposals for the introduction of congestion marking, whereby a packet encountering a long queue will have a specific bit in its header set to indicate "congestion experienced." The procedure is called *Explicit Congestion Notification*, or ECN. Users detecting ECN marks should respond by reducing their transmission rates. This will result in the network being able to share resources without having to drop packets, except in periods of exceptionally heavy use. Explicit Congestion Notification has now been made a "Proposed Standard" by the Internet Engineering Task Force (IETF), the body concerned with the evolution of the Internet architecture.

2.5 Bandwidth and Elastic Traffic

Capacity in the Internet is usually defined in terms of *bandwidth*, the amount of data that can be transmitted per unit time, i.e., the maximum rate, usually over fiber optic cable. A related concept is *throughput*, which refers to the amount of data transferred successfully from one place to another in a given amount of time.

Fiber optic cable, or simply *fiber*, refers to both the medium and the technology associated with the transmission of packets as light pulses along a glass or plastic fiber. Fiber has significantly more bandwidth than conventional copper wire and is in general not subject to electromagnetic interference. Most telephone company long-distance lines are now of optical fiber, and increasing share of Internet traffic is carried over fiber.

Resources of the network are managed by network *servers*, which are computers and other devices that manage network traffic, and which may be *dedicated*, i.e., performing no other task other than traffic management. One often talks about a *client/server* relationship between two computer programs in which one program, the client, makes a service request from another program, the server, which fulfills the request. Thus, for example, your computer uses TCP/IP to make client requests for files in other computers on the Internet. Another example is your Web browser, which is a client program that requests services, i.e., the sending of Web pages or files, from a Web server in another computer somewhere on the Internet.

In a network, a *node* is a connection point, either an end point—called simply an *end node*—or a redistribution point, for packet transmission. The management of traffic between nodes in a network so that the receiving device can handle all the incoming data is called *flow control*. This is particularly important where the sending device is capable of sending data much faster than the receiving device can receive it, a not uncommon situation. In cases were traffic can adjust to changes in delay and bandwidth while still meeting the needs of its applications, the traffic is called "elastic" (Shenker 1995; Kelly, Maulloo, and Tan 1998). A mathematical definition of elastic traffic is given in Section 5.

2.6 Quality of Service

Internet Quality of Service (QoS) refers to the extent that transmission rates and other service characteristics can be objectively measured and improved. The Internet currently operates on a single service quality, viz., what is usually described as "best effort" service. This means that the network attempts to serve all users, transmitting packets on a first-come, first-served basis without making any explicit commitment to any user with regard to rate, guarantee of packet arrival, or any other measure of service quality. It is the users, not the network, who are expected to detect the existence of congestion and reduce their transmission speeds. Discussions of Internet QoS can be found in Odlyzko (1998) and Park, Sitharam and Chen (2000).

2.7 Internet Service Providers

An *Internet Service Provider*, or ISP, is a company that provides access to the Internet. The ISP owns a host computer that is permanently connected to the Internet. Typically, for a monthly fee, the ISP provides individuals with a software package, a username and password, and an access phone number. The user can then log on to the Internet and make unlimited use of a number of services such as access to the World Wide Web and email. Such a pricing system is called a *flat rate scheme*. In addition to providing Internet access, some ISPs have their own online independent content. ISPs also serve large companies, providing a direct connection from the company's own networks to the Internet.

In March 2002, the Internet Service Provider reported to have the largest share of global Internet usage (i.e., percentage of Internet users worldwide using a particular ISP) was America Online (AOL) at 13.58%, with its nearest competitor, Road Runner, having a share of only 2.76%. Road Runner is owned by America Online's parent company, AOL

Time Warner. The third largest share was held by UUNET at 2.18% (WebSideStory 2002). Although these figures have undoubtedly changed since this study was conducted, the dominance of AOL in the ISP market at this point in time appears undisputed.

As of February 2003, AOL has four main pricing plans. Under the standard plan, customers pay a flat rate of $23.90 per month for unlimited usage. There is also an annual flat-rate plan payable in advance that works out to be $19.95 per month. Under the "limited usage" plan, customers pay $9.95 per month for five hours of usage, plus $2.95 for each additional hour. Finally, under the "light usage" plan customers pay $4.95 per month for three hours of usage, plus $2.50 for each additional hour. Each of these plans provides customers with AOL-specific content in addition to access to the Internet.[3]

3 Smart Market

Mackie-Mason and Varian (1995, 1996) contend that users should face prices that reflect the resource costs they generate. Specifically, the authors propose that, in any efficient pricing mechanism, the price a user pays should have three components: (1) a fixed connection fee; (2) a close-to-zero charge per packet when the network is not congested; and (3) a positive charge per packet when the network is congested, where the price varies so as to reflect the existing degree of network congestion.

In particular, they present their *Smart Market* scheme in which a per-packet charge is levied against the user application whenever the network is congested. By "smart," they are referring to the feature that packets are priced in real time to reflect the existing degree of network congestion. Under Smart Market, each packet has in its header a bid field that indicates how much its sender is willing to pay for transmission. The bids would be determined by input from three parties: the local administrator controlling access to the net, the actual user of the sending computer, and by the computer software itself.

Operating as a multi-unit Vickrey auction,[4] the network admits all packets with bid prices that exceed the current threshold value, which is determined by the marginal congestion cost imposed by the next additional packet. The users thus pay not the price they bid, but the market-clearing price, which is lower than the price of all admitted packets.

What is clever about this scheme is that users can be expected to bid their true values, the dominant strategy in the Vickrey auction (see Vickrey 1961, 1962).[5] Mackie-Mason

[3] These prices are for the U.S. market. AOL has similar pricing plans in other countries.

[4] A Vickrey (1961) auction is a second-price sealed-bid auction for a single item. That is, the bidders simultaneously submit their bids for the item without knowledge of the bids of the other players, and the winner of the auction is the one with the highest bid, who is required to pay the second-highest bid price. This is generalized to auctions with m identical objects in Vickrey (1962).

[5] Mackie-Mason and Varian mention in both of their papers cited above that the idea of using a second-price auction to allocate network resources goes back to Waldspurger et al. (1992), although in a different

and Varian provide some simple analytic models based on Smart Market pricing. They show that capacity should be expanded when the revenues from congestion fees exceed the cost of providing the capacity and, further, that this result holds in a competitive setting.[6] Further, they show that the competitive price will result in the optimal degree of congestion.

Ganesh, Laevens and Steinberg (2000) observe that the Smart Market proposal corresponds to the Bertrand model of competition (Tirole 1988), where each producer sets a price and is willing to supply any demand at that price, and consumers choose the quantity demanded based on the market price. In Smart Market, this corresponds to each user specifying a price he is willing to pay per packet, and the network accepting packets from a collection of highest bidders. However, Ganesh et al. claim that Smart Market, while efficient from a theoretical point of view, is impractical to implement. They provide two reasons for this. First, there is the considerable difficulty of conducting repeated auctions at the speeds at which the Internet operates. Second, there is a fairness issue connected with comparing a bid on a packet that has recently entered the network with a bid on a packet that has been waiting for an extended period of time. Smart Market does not make an allowance for this.

What is notable in the Smart Market scheme is that it does not promise the customer a specific level of QoS—which might, in fact, be difficult to measure and maintain—but simply guarantees a priority level for packet transmission. A packet with a high bid will gain access sooner than one with a low bid, but the technology of the Internet means that delivery time cannot be guaranteed. This concept of abandoning the chimera of a QoS guarantee, and instead employing a more tangible and objective measure as a proxy for service level, is itself a significant contribution of the Smart Market proposal. We will see this approach used in other proposals we discuss; Paris Metro Pricing takes this idea the furthest.

4 Paris Metro Pricing

Once upon a time the Paris Metro employed an unusual pricing scheme. Users were offered a choice of travelling First or Second class, where the only difference between classes was the price charged. First class carriages, being more expensive, typically had fewer passengers and thus were less congested. Hence, users with a strong aversion to congestion were willing to pay the higher First-class fare. Whenever First-class carriages became too popular, some users decided it was not worth the extra cost and switched to travelling Second, thus reducing congestion in First class and restoring the quality

context.

[6] The importance of evaluating the viability of a pricing proposal by considering it in a competitive setting cannot be overstated. See, for example, Section 5.

differential. In a paper entitled “A modest proposal for preventing Internet congestion,” Andrew Odlyzko (1997) made the inspired suggestion to apply this idea to the Internet:

> Following in the footsteps of Jonathan Swift,[7] I propose to turn a perceived burden into a solution, and rely on usage-sensitive pricing to control congestion, bypassing most of the complexity of other solutions. This should allow for simpler networks that are easier to design and deploy and operate faster.

Odlyzko calls his proposal *Paris Metro Pricing*, or PMP. [8]

Under PMP, the Internet Service Provider first partitions its network into several logically separate subnetworks, each having a fixed fraction of the capacity of the network, and then applies different charges to each subnetwork. The ISP offers no guarantees of service quality; however, on average, higher-priced networks will be less congested. Users will sort themselves according to their aversion to congestion and the prices charged on the subnetworks. Odlyzko’s attractively simple proposal raises two issues, discussed below.

4.1 The Acceptability of PMP

Was the Paris Metro a special case, or would a pricing scheme with endogenously-determined quality be acceptable to consumers in other contexts? There are in fact a number of other examples of PMP in practice, many from India. For example, De Palma and Leruth (1989) observed a two-price system in effect for the busses of New Delhi. The busses were all identical except for the price charged. Displayed on the front of each bus was a sign that read either “1R*e*” or “2R*s*,” where—as might be expected—the 1 rupee busses were considerably more crowded.

In his original paper, Odlyzko (1997) warns that the analogy between the Paris Metro and the Internet should not be overdrawn. He points out that, on the Paris Metro, all passengers arrived at the destination at the same time; the different prices only paid for the expected differential in discomfort caused by congestion (e.g., probability of getting a seat). With PMP applied to the Internet, packets on a lower-priced network would have a higher probability of being dropped and thus a higher probability of requiring re-sending; therefore, the file could be expected to arrive later than on a higher-priced network. The pricing for New Delhi buses clearly has much in common with that for the Paris Metro. If we assume that the arrival rates for the two classes of busses was about the same, then bus choice would not have made a significant difference in arrival time.

[7] Here Odlyzko cites Swift’s 1729 satire, “A Modest Proposal for Preventing the Children of Poor People in Ireland from Being A Burden to their Parents or Country, and for Making them Beneficial to the Public.”

[8] He has further discussion on PMP in his follow-up papers, Odlyzko (1998, 1999a).

But there exist other examples of real-world PMP that involve a time element via queuing, and thus arguably form closer analogies to Odlyzko's proposal for Internet pricing. Chander and Leruth (1989) cite the case of a government hospital in New Delhi, where two different options were available for treatment of routine ailments, fee and free. Users could choose to pay the fee and expect a shorter queue, although the same set of doctors treat the non-paying patients. Gibbens, Mason and Steinberg (2000) report a similar arrangement for medical care in the United Kingdom where, as an alternative to receiving free (at the point of delivery) medical care via the National Health Service, people can choose to pay for private treatment at what they will expect will be a greatly reduced waiting time.

In December 1995, a four-lane toll highway called the *91 Express Lanes* opened in the median of a 10 mile section of the Riverside Freeway (State Route 91) in California (Transportation Research Board 2002). This was the first privately-financed toll road in the U.S. in over fifty years. According to the California Private Transportation Company, "the private sector would take the risk and the State would get congestion relief at no cost to taxpayers."[9]

Perhaps the most unusual instance of PMP in practice is at Tirupati, a pilgrimage center in Andhra Pradesh, India where devotees go for *darshan* ("viewing of the deity"). Devotees can have free darshan, called *Sarvadarsanam* ("darshan for all"). However, they may also choose to pay a fee, which is called *special darshan*. With regard to the latter, the official Tirupati website[10] explains: "Pilgrims who use this queue will have a shorter waiting time." Special darshan is available in two categories, at 40.00 rupees per head and at 50.00 rupees per head. The timings are the same for Sarvadarsanam and the special darshan, and the queues merge at the inner sanctum. Remarkably, the system handles more than fifty thousand pilgrims per day.

Odlyzko's motivating idea can be summarized as follows: A simple pricing scheme can induce users to separate themselves into classes that provide differing qualities of service, and the division can be self-stabilizing. The above examples tend to support that belief for a wide variety of services. However, there is another issue connected with PMP that deserves close scrutiny.

4.2 The Viability of PMP Pricing under Competition

The Paris Metro is run by a government monopoly, the RATP - Régie Autonome des Transports Parisiens. Therefore, for purposes of applicability to Internet pricing, we might

[9] http://www.91expresslanes.com/. The congestion toll in effect as of September 2002 ranges from as high as $3.60, for eastbound traffic, 7:00 am to 8.00 am, Monday through Thursday, and $4.75, for westbound traffic, 5.00 pm to 6.00 pm, Monday through Friday; and as low as $1.00, 11.00 pm to 3.00 am, every day in both directions.

[10] http://www.tirupati.org/.

ask what occurs when competition is introduced into Odlyzko's PMP model. Gibbens, Mason, and Steinberg (2000) address this question through the use of game theory.

Their model is specified as follows. There are two competing, profit-maximizing Internet Service Providers, each of which may offer either one or two service classes. In the case where an ISP chooses to offer two service classes, it forms them by logically dividing its network in two and charging separate prices on each subnetwork. Congestion on a subnetwork is determined in equilibrium by two factors: the fraction of the ISP's total network capacity allocated to a subnetwork, and the number of users on the subnetwork. The immediate consequences are that a network with low capacity and many users will have high congestion. Further, quality is demand-dependent, determined (in part) by the equilibrium choices of the prices. Gibbens et al. ask: In equilibrium, how many service classes will the two ISPs choose to offer?

They begin with the following assumptions. Each user joins one and only one network, resulting in Q_i users on network i. Upon joining a network, a user receives utility with three components: (i) a positive benefit, V, independent of the network joined, (ii) a dis-benefit, dependent on the degree of congestion, $K(Q_i)$, and preference for (i.e., aversion to) congestion, θ, and (iii) a dis-benefit from having to pay a price, p_i, to the ISP. Further, users are assumed heterogeneous in their preference for congestion. Those users with elastic traffic receive little dis-benefit from congestion, and will have low values of θ; those with inelastic traffic will be very sensitive to congestion, and will have high values of θ. A user's utility in joining network i is $U(\theta, i) = V - \theta\, K_i(Q_i) - p_i$, and the costs are set to zero so that the profit of the network is $\Pi = p_i\, Q_i$.

The authors make two simplifying assumptions, one about the form of the congestion function and the other concerning the distribution of consumer disutility for congestion. First, congestion on a network is assumed to be proportional to the number of users divided by the capacity of the network. Second, consumer disutility for congestion, θ, is uniformly distributed.

Gibbens et al. find that the unique equilibrium outcome is that neither ISP subdivides its network, i.e., neither firm will employ Paris Metro Pricing; further, the two firms charge the same price. In their analysis, the authors assume that the capacities of the subnetworks are fixed and symmetric. However, they also conducted numerical analysis which indicates that their results are robust to the assumption of fixed, equal, and symmetrically split capacities; in all cases considered, the ISPs each maximize their own profits by charging a single price and offering one network. The authors provide the economic interpretation of their results: As more products are introduced, the cost of increased competition outweighs the benefits from greater segmentation of the market.

The existence, uniqueness and symmetry of the equilibrium gives rise to the question of the extent to which these results depend on the form of the congestion function and the distribution of consumer disutility for congestion. Haimanko and Steinberg (2000) offer the following partial result. Consider any congestion function and any consumer disutility cumulative density function with the only assumptions being that these functions are each continuously differentiable and strictly increasing. Then with fixed, equal capacities in the case where neither ISP subdivides its network: (1) there does not exist an asymmetric equilibrium, and (2) if an equilibrium does exist, then it is unique. This tends to indicate that the results of Gibbens et al. are robust to the form of the congestion function and the distribution of consumer disutility for congestion. However, further research in this direction might yield new insights.

The above results tend to indicate that Paris Metro Pricing will not be viable in a competitive market. However, in order to simplify their model, Gibbens et al. assumed that there is a fixed number of firms in the market (i.e., two). The authors suggest that the process of free entry may be a mechanism by which a range of prices and qualities can arise in equilibrium. This would be a worthwhile area for further research on a proposal that has garnered a considerable amount of attention.

4.3 The True Origins of Paris Metro Pricing

One may wonder how the Régie Autonome des Transports Parisiens arrived at this innovative pricing scheme. The answer is that it was an accident of history.

When the Paris Metro opened in July 1900, intercity rail service in France had differential classes of service, and so it was decided that the Metro would have differential classes of service as well. (Although there were at the time three classes of service on the intercity rail, it was felt that two classes would suffice for the Metro.) The difference in pricing between Metro classes was justified by the differences in materials used in the carriages, and hence in the comfort levels for passengers. Due to this price differential, First Class was less crowded.

From the 1930s onwards, the basis of the Metro's rolling stock was the Sprague-Thomson carriage, which had upholstered seats in First class and wooden seats in Second. In the early 1980s, these carriages were replaced with modern equipment. With modern cars and materials, comfort became harmonized between the two classes. Thus, for a time, essentially the only difference between the two classes was the price and, consequently, the level of congestion. Voilà: PMP.

Given that there was no longer any distinction between the carriages, Claude Quin, the president of RATP and a member of the Communist Party, advocated equating the rates so as to eliminate the class differences; this was ultimately implemented in August 1991. Although the change was putatively motivated by egalitarian reasons, it was felt that this would lead to greater utilization of train capacity as well.[11]

[11] I am grateful to Henri Zubar of Régie Autonome des Transports Parisiens and Emile Quinet of ENPC, CERAS, Paris for providing historical information.

5 Max-Min Fairness

In the presence of congestion, not all users can be allocated their desired rate, and the network faces the task of having to decide what rate to allocate to each user. Understandably, it might want to do this in a manner that is considered to be "fair." But is fairness truly an appropriate concept for communication networks? Kelly, Maulloo, and Tan (1998) make the argument that it is. They point out that, although fairness has traditionally been considered an *economic* issue, involving as it does comparisons of utility, any discussion of the performance of a rate control scheme *must* address the issue of fairness, since a given scheme might maximize the rate of information passing through the network—an important criterion—but at the same time deny access to some users.

That being said, it is all too obvious that fairness can be interpreted in a variety of ways. In the setting of communication networks, the first criterion to be widely accepted is due to Jaffe (1980, 1981), and has come to be known as *Max-Min Fairness.* Jaffe begins by citing one interpretation that requires that all users obtain equal throughputs. He points out, however, that in a network with different users using lines of different capacities (it seems this would be a rather common situation), it is unlikely that such a policy would be desirable. He argues that the general interpretation of fairness as "all users are treated equally" may not be desirable in practical networks, as one user may be more important and thus more deserving of a higher rate. The criterion he arrives at is an equilibrium condition:

> A set of rates is *max-min fair* if no rate can be increased without simultaneously decreasing another rate that is equal or smaller.

If the network has a single bottleneck resource, then under Max-Min Fairness each user either receives his desired allocation or an equal share of the bottleneck resource. (For more details on Max-Min Fairness in communication networks, see Chapter 6 of Bertsekas and Gallagher (1992).)

5.1 Mathematical Formulation

It is not difficult to provide a precise mathematical statement of Max-Min Fairness. We begin with some definitions. Consider a network with a set J of resources, and let C_j be the capacity of resource $j \in J$. A *route* r is a nonempty subset of J, where R denotes the set of routes in the network. Each route r can be associated with a user of the network. When user r is allocated a rate x_r, let $U_r(x_r)$ denote the utility to user r.

We assume that traffic on the network is such that, for all r, utility function $U_r(\cdot)$ is increasing, strictly concave, and continuously differentiable on $[0, \infty)$. Under these assumptions, the network is said to experience *elastic traffic* (Shenker 1995). Further, we assume that utilities are additive. That is, the aggregate utility of the vector of rates $x = (x_r,\, r \in R)$ is the sum of the utilities, i.e., $\sum_{r\in R} U_r(x_r)$.

Define the matrix $A = (A_{jr},\, j \in J,\, r \in R)$ by:

$$A_{jr} = \begin{cases} 1 & \text{if} \quad j \in r \\ 0 & \text{if} \quad j \notin r \end{cases}$$

Thus, A describes which resources lie on which routes. Define the grand capacity vector $C = (C_i(\cdot),\ j \in J)$.

A vector of rates $(x_r, r \in R)$ is said to be *feasible* if $x \geq 0$ and $Ax \leq C$. A feasible vector of rates is *max-min fair* if for any other feasible vector y:

$$\exists\, r \ :\ y_r - x_r > 0 \quad \Rightarrow \quad \exists\, x_s \leq x_r \ :\ y_s - x_s < 0.$$

The compactness and convexity of the feasible region imply that such a vector x_r exists and is unique (see Kelly 1997a).

5.2 Nash Bargaining Solution

In 1991, Mazumdar, Mason and Douligeris (1991) proposed that fairness issues for communication networks be placed in a rigorous game-theoretic framework. They embark on the following sequence of reasoning. First, games can be broadly classified into cooperative and non-cooperative. However, in non-cooperative games each user acts individually to optimize his own performance without regard to the performance of the others. This leads to a Nash equilibrium solution, which might turn out to be Pareto inefficient. Since Pareto inefficiency is undesirable, a cooperative game is the preferable framework. Finally, they argue that the *Nash bargaining solution* is a suitable candidate for a fair, optimal operation point in the sense that it satisfies certain axioms of fairness and is Pareto optimal.

What the Nash bargaining solution does is predict an outcome based only on information about each bargainer's preferences, as modelled by an expected utility function over the set of feasible agreements and the outcome that would result in the case of disagreement. For more details, see Nash (1950) and Roth (1979). Although Mazumdar et al. make some connections with Jaffe's work, they do not compare the Nash bargaining solution with Max-Min Fairness, and so leave the relationship between the two solution approaches unresolved.

In a lecture before the London Mathematical Society, Frank Kelly (1997b) observed that the concept of Max-Min Fairness can be grounded in the moral philosophy of John Rawls. Rawls had developed essentially the same concept, which he called "maximin," as a component of his "Theory of Justice" (1971, 1999). As Rawls explains it, this rule for "choice under uncertainty" is one that

> a person would choose for the design of a society in which his enemy is to assign him his place. The maximin rule tell us to rank alternatives by their worst possible outcomes: we are to adopt the alternative the worst outcome of which is superior to the worst outcomes of the others.

Although Max-Min Fairness has attractive properties as a method for rate control, it cannot be considered as an option for pricing Internet service as it is lacking one essential feature: a market mechanism.

6 Proportional Fairness

In an influential series of papers, Frank Kelly and his co-authors describe a pricing and rate control model for elastic traffic in which each user chooses the charge per unit time he is willing to pay. Thereafter, the network determines each user's data transfer rate according to a fairness criterion applied to the rate per unit charge. Kelly shows that his criterion, which he calls *Proportional Fairness*, achieves a system optimum when the users' choices of charges and the network's choices of allocated data transfer rates are in equilibrium. We outline the main results here. For proofs and further development, see Kelly (1997a) and Kelly, Maulloo and Tan (1998).

There are three interrelated optimization problems to consider: the System Problem, the User Problem, and the Network Problem. We make use of the same notation as was used in Section 5.1. We also define the grand utility function vector $U = (U_r(\cdot),\ r \in R)$.

The System Problem

The *system optimal* rates are given by the solution to the following problem:

SYSTEM(U, A, C)

$$\begin{aligned} \text{maximize} \quad & \sum_{r \in R} U_r(x_r) \\ \text{subject to} \quad & Ax \leq C \\ \text{over} \quad & x \geq 0 \end{aligned}$$

In words, the system problem is to maximize aggregate utility, subject to the capacity constraints. This may appear to be a straight-forward problem, but Kelly, et al. (1998) point out that, despite the putative tractability of optimizing a strictly concave function over a convex set, the network operator is unlikely to know the grand utility function vector U. This brings us to the second of the three problem formulations:

The User Problem

Suppose that each user r can choose an amount to pay per unit time, w_r, and receive in return a flow proportional to w_r. Here w stands for *weight* or—in particular—*willingness-to-pay*. Thus, the rate allocated to user r will be:

$$x_r = \frac{w_r}{\lambda_r}$$

where λ_r is the charge per unit flow by the network to user r, which is presumed to be known to user r. Then r seeks to solve the following problem:

USER$_r(U_r;\, \lambda_r)$

$$\begin{array}{ll} \text{maximize} & U_r\left(\dfrac{w_r}{\lambda_r}\right) - w_r \\ \\ \text{over} & w_r \geq 0 \end{array}$$

That is, user r chooses an amount to pay per unit time, w_r, and receives in return a rate $x_r = w_r/\lambda_r$.

The Network Problem

As observed above, the network operator is unlikely to know the grand vector of user utility functions, $U = (U_r(\cdot),\ r \in R)$. However, it is not unreasonable to assume that he could be knowledgeable about the grand vector $W = (W_r,\ r \in R)$ of user willingness-to-pay. This brings us to the third optimization problem:

NETWORK$(A, C; w)$

$$\begin{array}{rl} \text{maximize} & \displaystyle\sum_{r \in R} w_r \log x_r \\ \\ \text{subject to} & Ax \leq C \\ \text{over} & x \geq 0 \end{array}$$

Thus, this problem is formulated as if the network maximizes a logarithmic utility function, but with the constants $(w_r, r \in R)$ chosen by the users. These three optimization problems are beautifully interrelated by the following theorem (Kelly 1997a):

Problem Decomposition Theorem. *There exist vectors* $\lambda = (\lambda_r,\ r \in R)$, $w = (w_r,\ r \in R)$ *and* $x = (x_r,\ r \in R)$ such that

i $w_r = \lambda_r x_r$ for $r \in R$

ii w_r *solves* $\mathrm{USER}_r(U_r;\ \lambda_r)$

iii x *solves* $\mathrm{NETWORK}(A, C; w)$

In addition, the vector x is the unique solution to the $\mathrm{SYSTEM}(U, A, C)$ *problem.*

Thus, the SYSTEM problem can be solved by simultaneously solving the NETWORK and USER problems.

6.1 Proportional Fairness Theorem

For a rate vector x, the *proportional change* with respect to a rate vector y_r is:

$$\frac{y_r - x_r}{x_r}$$

A feasible vector of rates $x = (x_r,\ r \in R)$ is called *proportionally fair* if there exists no other feasible rate vector y for which the sum of the proportional changes over all users is positive. In other words, a rate allocation is proportionally fair if no aggregate improvement is possible. In symbols, a rate vector x is proportionally fair if, for all feasible rate vectors y:

$$\sum_{r \in R} \frac{y_r - x_r}{x_r} \leq 0$$

The concept of proportional fairness can be generalized to *weighted* proportional fairness. In particular, we can use as weights each user's willingness-to-pay. A feasible vector of rates is *proportionally fair per unit charge* if, for all feasible rate vectors y:

$$\sum_{r \in R} w_r \frac{y_r - x_r}{x_r} \leq 0$$

Thus, a feasible set of rates $\{x\}$ in a network is proportionally fair per unit charge if, for any other set of feasible rates $\{y\}$, the sum of the weighted proportional changes is negative or zero. This leads to the following result (Kelly 1997a):

Proportional Fairness Theorem. *A rate vector x solves* $\mathrm{NETWORK}(A, C; w)$ *if and only if it is proportionally fair per unit charge.*

In a series of lectures presented to the Netherlands operations research community in January 2000, Richard Weber (2000) demonstrated that a proportionally fair allocation will also be the Nash bargaining solution, and that a weighted proportionally fair vector is the Nash bargaining solution where the players have unequal bargaining power, thus reconciling the work of Mazumdar, Mason, and Douligeris (1991) with that of Jaffe (1980, 1981). Details are provided in Chapter 10 of Courcoubetis and Weber (2003).

6.2 Proportionally Fair Pricing

Gibbens and Kelly (1999) make use of the theoretical model developed in Kelly, Maulloo, Tan (1998) to develop an implementable pricing scheme that makes use of *packet marking*. The motivating idea is that it might be easier for the network to achieve an efficient allocation by conveying information on congestion to the users, rather than asking the users to supply information to the network; in addition, the network is relieved of the burden of eliciting information from users about their preferences.

Called *proportionally fair pricing*, the scheme is very simple in its operation. Time is broken down into discrete slots. The network marks a packet if it arrives in a time slot in which aggregate arrivals exceed capacity. Here marks reflect the fact that such packets imposed a cost, in terms of delay or loss on another packet. End users are informed as to whether their packet was marked, and they are free to choose how to adapt their sending rates. Of course, a natural way to implement packet marking would be through the use of the ECN bit to carry the mark.

Gibbens and Kelly suggest that for a network with "potentially uncooperative" end users, it may be necessary for each mark to be associated with a small fixed charge to the user, ensuring that the end user will have both the appropriate incentive and the necessary information in order to use the network efficiently. Proportionally Fair Pricing aims to combine the flexibility of Smart Market with the simplicity of Paris Metro Pricing; further, it is a pricing system that can evolve naturally from existing standards and proposals (e.g., TCP and ECN).

7 CPE Pricing

Rather than using one bit to mark packets, why not make use of several bits and actually *price* packets? Ganesh, Leavens and Steinberg (2000, 2001) do exactly this, allowing the network to assign a price to each packet rather than to simply mark them.[12] The price assigned by the network reflects the degree of congestion encountered by the packet; users

[12] In their analysis the price is taken to be a real number, but they suggest that in practice a small number of bits of price feedback should provide sufficient accuracy.

are informed of how much they were charged and thus will be motivated to make use of this information to adjust their sending rates. Since the prices reflect the "social cost" imposed on others, users will have the correct incentives to adapt to congestion.

Like Gibbens and Kelly (1999), Ganesh et al. consider time to be broken down into discrete slots, in which a single network resource is shared by N users. In each time slot t, user i transmits a quantity of packets $x_i(t)$. The unit price of packet transmission, $p(t)$, is determined as a function $\phi(x)$ of $x = x_1 + x_2 + \ldots + x_n$, the traffic arriving at the resource in that time slot. Thus:

$$p(t) = \phi(x(t)), \qquad \text{where } x(t) = \sum_{i=1}^{N} x_i(t).$$

Each user i derives a utility, $u_i(x_i(t))$, which is a non-decreasing function of the number of packets he sends in time slot t. His total utility is assumed to be the sum of his utilities over all N time slots. In time slot t, user i chooses $x_i(t)$ so as to maximize:

$$u_i(x_i(t)) - x_i(t)\,\phi(x(t))$$

This expression denotes the single-stage payoff to user i as a function of the actions of all the users.

Now if the utility functions of all the users are common knowledge, and if the price function is also common knowledge, then this is the well-known Cournot competition model (Tirole 1988), and a Nash equilibrium will be sought. Ganesh et al. remark that this approach does not suffer from the problems associated with the "Bertrand" approach of Smart Market pricing (discussed in Section 3).

Nevertheless, they find this model to be unsatisfactory. Specifically, they point out that there can be multiple equilibria, and that sufficient conditions for uniqueness of the Nash equilibrium are neither simple nor intuitive. They find the Nash equilibrium approach to be unrealistic in this setting, as it assumes that the users would be aware of the utility functions of all the users of the network; however it is unlikely that a user will even know the *number* of users on the network.

7.1 An Expectation-Based Model

Ganesh et al. point out the only information that a user needs to know in order to choose his sending quantity is the price in that time slot; it is irrelevant as to how that price was arrived at from the auctions of all the users. Essentially, the only information that a user could be assumed to have is the history of the prices and of his own actions. This leads to the following alternative modelling framework.

Each user i forms his own estimate, $\hat{p}_i(t)$, of the price in time slot, t, by exponentially smoothing all the price information available to him, viz., all the prices up until time $t-1$.

He then chooses his transmission rate, $x_i(t)$, so as to maximize his payoff as a function of this expected price. Thus, he chooses $x_i(t)$ so as to maximize:

$$u_i(x_i(t)) - x_i(t)\,\hat{p}_i(t).$$

The authors now need to make three assumptions pertaining to, respectively, the form of the utility functions, the form of the price function, and the method of user price-estimate updating. First, the utility functions are assumed to be of the form:

$$u_i(x) = w_i(x^\beta - 1)/\beta, \qquad 0 \le \beta < \infty.$$

If $\beta = 0$, then the formula is interpreted to mean $u_i(x) = w \log x$.

Second, the price function is assumed to be iso-elastic:

$$\phi(x) = \left(\frac{x}{C}\right)^k,$$

where C is a scale parameter which is associated with the physical capacity of the resource, and $k \ge 1$ defines the steepness of the penalty for demand in excess of capacity. The use of the iso-elastic price function is to keep resource utilization close to C, but at the same time to prevent demand from exceeding the resource capacity. The form of the price function, which they name the *CPE price mechanism*, corresponds to a constant price elasticity demand function, which has empirical support in the context of consumer demand (Simon 1989).

Third, users are assumed to use exponential smoothing to update their respective estimates of the price:

$$\hat{p}_i(t) \;=\; \alpha_i\, p_i(t{-}1) + (1{-}\alpha_i)\,\hat{p}_i(t{-}1)$$

where α_i is the smoothing constant, $0 < \alpha_i < 1$.

There are two major results in Ganesh et al. (2000). One is that there is a price q^* which is “self-consistent” in the following sense. If all users had q^* as their price estimate and optimized their transmission rates accordingly, then the price would in fact be q^*, and this q^* is unique. The other result is that the price estimates of individual users converge to q^* if each smoothing constant α_i is sufficiently small.

In Ganesh et al. (2001), the authors generalize the earlier results to allow for delays in the price information being provided to the users. They also present simulation results that support their theoretical results.

8 Pricing Bandwidth

Consider now the owner of a *dedicated* (i.e. private) IP network who intends to lease capacity and would like to know how to price it. Potential customers would include

Content Service Providers, i.e., businesses that host network applications, such as Blockbuster Video, which delivers full length films to customers online as well as through its rental stores, and Streetmail, which delivers information traditionally available from local newspapers and community calendars. Other customers could include small telecommunications companies requiring additional capacity, as well as companies with special bandwidth needs, such as a financial services firms requiring, e.g., the capability to make video broadcasts to its analysts.[13] How should the network owner charge its customers for bandwidth? Let us examine how this problem is now handled.

Currently, large network owners trade capacity via a complex procedure involving long-term leases known as *Indefeasible Right of Use* (IRU). An IRU is effectively the temporary ownership of a portion of the capacity of a fiber optic cable, and is granted by the company, or consortium of companies, that built the cable. IRUs typically cover a period of twenty to twenty-five years. In some cases, the IRU forbids the customer to resell the capacity during the term of the lease. IRUs are usually paid for up front in a single cash payment.

The role of IRUs is clear. A telecommunications firm interested in extending its business into a new region could lease capacity in the region, allowing it to enter the market but bypass the costs associated with a buildout. A major player could thus essentially piece together a communication network without having to incur the enormous expense of actually building the network. In this way it would assure itself of having sufficient capacity for when it requires it.

IRUs originated in the days of the Bell System monopoly prior to the break-up of AT&T in 1984. IRUs allowed competitors to utilize the large undersea cables that had been laid by AT&T to build out their networks. The IRU concept had a resurgence in popularity around 1997, when a number of new backbone providers arrived and had a need to build their markets quickly.

However, IRUs have recently come under criticism, with the accusation that they can be used to inflate revenues. Legal problems aside, there are serious practical problems with IRUs. In the antediluvian days of the Bell System, traffic was circuit-switched and reasonably predictable. However, the flexibility of the newer packet-switched technology has lead to enormous growth in demand for capacity, which has become far more difficult to predict, and IRUs have diminishing utility.

8.1 Contract and Balancing Process

The ECN bit in the packet header has an appealing interpretation as a hypothetical congestion charge. When a transmission line in a network is well below capacity, there

[13] Cable & Wireless issued a press release announcing a three-year global communications contract with UBS Warburg (Salgado, Valder Couser 2002). Secure data networks was among the services mentioned.

will be few if any marks generated and the appropriate charge should be low; when a line is near capacity, many marks will be generated and the charge should be high. It seems plausible, then, that congestion marks could be used for controlling the supply of capacity.

A natural approach would involve the owner of a network transmission line being paid by each user based on the number of marks the user generates on the line. However, this simplistic method would produce a perverse incentive for the owner to increase congestion; for example, by making side payments to some users to generate traffic to maintain a state of high congestion. Further, given the likelihood of rapid variation in the rates at which congestion marks would be generated—a necessary consequence of their role in balancing demand over very short time scales—there is the additional issue of how prices can be appropriately averaged.

Anderson, Kelly, and Steinberg (2002) propose a *Contract and Balancing Process* (CBP) that addresses both issues. With CBP, bandwidth pricing is handled by a contractual process based around the use of ECN marks. It operates as follows. The network owner specifies a period of time for which potential users can contract a portion of the transmission line capacity. At the start of this period, the users pay the network owner according to their respective contractual amounts. During the period, the number of ECN marks received by each user is tabulated. At the end of the period, the users engage in a balancing process *with the other users*, based on each user's individual usage as compared with his respective contractual amount. The proportion of ECN marks generated by each user is employed as a proxy for usage, so that users who generated a larger-than-contracted proportion of ECN marks will make payments to the other users, while those who generated a smaller-than-contracted proportion of ECN marks will receive payments from the other users.

Observe that this approach allows volatile prices to be appropriately averaged, and thus mediates between rapidly fluctuating congestion prices and the longer time scales over which the bandwidth contracts would be written. Also note that CBP eliminates the incentive for the network owner to increase congestion. Finally, note that high usage by a user in one period—which may create considerable costs to other users—cannot be cancelled out by him through low usage in another period (a period in which aggregate demand might in fact be very low).

8.2 An Example

Consider the following illustrative example involving three parties, network operator, N, and two customers, User 1 and User 2. Network owner N has a fiber optic cable with a capacity of 100 megabits per second (Mbps), and is offering use of the cable over a six-month period. He has set the contractual rate at c dollars per megabit and the balancing charge at γ dollars per mark. User 1 chooses to contract for 40 Mbps of capacity, while

User 2 contracts for 60 Mbps.

At the beginning of the contractual period, Users 1 and 2 make upfront cash payments to the network owner of $40c$ dollars and $60c$ dollars, respectively. Suppose over the course of the six-month period the number of marks generated by User 1 is exactly 40% of the total marks generated. Then, at the end of the contractual period no further payments are made. If, however, User 1 generated more than 40% of the total marks, then in the balancing process, User 1 will pay User 2 an amount equal to the balancing charge, γ, times the excess number of marks he generated. In symbols, User 1 will pay User 2:

$$\Delta_{1,2} = \gamma[z_1 - 0.40(z_1 + z_2)]$$

where z_i = number of marks generated by User i. If, however, User 1 generated less than 40% of the marks, then the expression above will be negative, and User 1 will *receive* this amount from User 2. Observe that the balancing payments are made only among the users and do not involve the network owner; the only payments received by the network owner are those made to him at the beginning of the period.

8.3 Results

Anderson et al. (2002) presents three main results. Consider the CBP scheme for a line with $n \geq 2$ users, where each user i contracts for a specified capacity y_i over the balancing period at an immediate cost to him of $C_i(y_i)$. A line capacity $Y = \sum_i y_i$ is then available for use by all n users over the period.

The first result is that each user will have a unique optimal choice of contract quantity, given any set of contract quantities by the others. Further, a bound on the contract price is provided, beyond which the cost of participating in a contract is sufficiently expensive that a user will choose not to contract for any amount, but instead will pay for usage entirely through the balancing process.

The second result is that if the users have the same marginal cost and 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 unit cost of capacity.

Finally, under certain conditions, the second result can be generalized from a line to a network. Note that this leaves open the question as to whether there exists a useful sufficient condition to identify a unique Nash equilibrium in the network. This would be a good starting place for future research in this area.

9 Conclusion

Throughout this chapter, we have seen that, in order to properly address the topic of pricing Internet service, one needs to understand the current operation of the Internet and anticipate future technological developments, as well as be cognizant of the underlying marketing issues, including competition, real-time pricing, market segmentation, quality of service, differentiated services, and consumer perceptions of fairness. Fairness appears to have special significance; in addition to Proportional Fairness most of the Internet pricing proposals can probably be re-stated in terms of some type of fairness criterion.

In recent work, Bolton, Warlop and Alba (2003) study "consumer perceptions of price (un)fairness," and recommend that future work in that area consider whether consumers perceive fairness not only through comparisons with past prices, competitor prices, and perceived costs—their study's reference points—but also with respect to prices paid by other consumers. The authors suggest that consumers may be particularly sensitive to prices paid by others in the presence of price discrimination. Clearly, research in this marketing area may have a significant influence on how Internet service pricing is modelled in the future.

In order to implement any of these proposals, there must be in place a workable billing and accounting system so that users can be properly charged. Mackie-Mason and Varian (1995, 1996) propose that the accounting system for Smart Market be handled by the routes that compare the packet bid price with the threshold price. They recommend that accounting information on every nth packet be sent (they suggest $n = 1000$) to a dedicated accounting machine that would determine the equilibrium access price; however, Smart Market would require a new field in the IP header to carry the requisite price information. An ISP currently using flat-rate pricing that wishes to convert to PMP will require a way to keep track of each user's choice of network (which of course could change over time). Proportional Fairness requires little by way of billing and accounting, since each user is charged by his willingness-to-pay. The major implementation issue for CPE Pricing is that, like Smart Market, it would require a new field in the IP header to carry the price information. For the Contract and Balancing Process, Anderson et al. (2002) cite Briscoe et al. (2000), who argues that ECN marks could allow the dispersal of charging operations, including accounting and billing, to customer systems. Implementation is clearly an understudied issue in the pricing of Internet service and is a fertile area for future research.

There are a number of other interesting pricing proposals; we have space to mention only a few. Clark (1997) develops a pricing method based on the observation that marginal costs are nonzero only during congestion. In Clark's scheme, called *Expected Capacity Pricing*, there is a contract between the users and owner of the network as to the expected capacity that the network will provide the user. Based on this contract, the incoming traffic is metered and each packet is tagged as being either an *in* packet, i.e., within the profile of expected capacity, or an *out* packet, not within the expected capacity profile. At any point of congestion, only "out" packets can be dropped (or be recipients of ECN marks). Thus charging is related not to actual traffic, only expected traffic, which should lead to operational simplicity, since measurements within the network are not required.

Gupta, Stahl, and Whinston (1996) model user preferences as depending on the service acquired as well as the expected waiting time. In their scheme, called *Priority Pricing*, an incoming user service request is associated with an instantaneous value for the service, and a delay cost for this value. The user request can be fulfilled through several alternatives, each associated with a price and an expected waiting time. The user seeks the lowest total cost for a service, where the cost includes the cost of delay. Thus, an optimal choice is reflected by a choice of an alternative in a particular priority class, dependent upon the user's instantaneous value and delay cost. A billing and accounting system would be especially easy to implement for Priority Pricing, as the Internet Protocol already provides for a priority field in the packet header.

Marbach (2002) presents a pricing scheme that employs priorities to provide differentiated quality of service, where the users are free to choose the priority of their own traffic, but are charged accordingly. Under the assumption that users choose an allocation of priorities to optimize their own net benefit, Marbach shows that there exists a unique equilibrium, and that the resulting allocation is max-min fair. Jarray and Wynter (2002) provide an extension of Smart Market, in which they consider the determination of static prices over a network. For a thorough discussion of other Internet pricing proposals, the reader is referred to Courcoubetis and Weber (2003).

In conclusion, Kelly, Maulloo and Tan (1998) have predicted that, in the future, we can expect networks to have intelligence embedded at their end points which will act on behalf of human users. The authors suggest that this development is likely to lessen the distinction between engineering and economic issues and increase the importance of an interdisciplinary view. Further discussion on this intriguing topic can be found in the introduction and conclusion sections of Kelly (1996), as well as the introduction to Kelly (2000).

References

Anderson, E., F. Kelly, and R. Steinberg (2002). A contract and balancing mechanism for sharing capacity in a communication network. Working Paper LSEOR 02.50, London School of Economics, Department of Operational Research, July.

Bakos, Y. and E. Brynjolfsson (2000). Bundling and competition on the Internet: Aggregation strategies for information goods. *Marketing Science* 19 (1), 63–82.

Barwise, P., A. Elberse, and K. Hammond (2002). Marketing and the Internet. In: B.A. Weitz and R. Wensley (eds.), *Handbook of Marketing.* London: Sage, pp. 527–557.

Bertsekas, D. and R. Gallagher (1992). *Data Networks.* Second Edition. Englewood Cliffs, New Jersey: Prentice-Hall. 1992.

Bolton, L.E., L. Warlop, and J.W. Alba (1993). Consumer perceptions of price (un)fairness. *Journal of Consumer Research.* March.

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 (OpenArch 2000)*, 77-87.

Cerf, C., Y. Dalal, and C. Sunshine (1974). Specification of internet transmission control program. Request for Comments 675, NIC [Network Information Center] 2, INWG [International Network Working Group] 72, Network Working Group, Stanford Research Institute, December.

Chander, P. and L. Leruth (1989). The optimal product mix for a monopolist in the presence of congestion effects: A model and some results. *Journal of Industrial Organization*, 7 (4), 437–449.

Clark, D. (1997). Internet cost allocation and pricing. In: L.W. McKnight and J.P. Bailey (eds.), *Internet Economics.* Cambridge, MA: MIT Press, pp. 215–252.

Courcoubetis, C. and R.R. Weber (2003). *Pricing Communication Networks: Economics, Technology and Modelling.* Chichester: John Wiley and Sons.

Ganesh, A., K. Laevens, and R. Steinberg (2001). Congestion Pricing and User Adaptation. *IEEE INFOCOM 2001*, 959-965.

Gibbens, R.. (2000). Control and pricing for communication networks. *Philosophical Transactions of the Royal Society of London*, Series A 358, 331–341.

Gibbens, R., R. Mason, and R. Steinberg (2000). Internet service classes under competition. *IEEE Journal on Selected Areas in Communications* 18 (12), 2490–2498.

Gilder, G. (1993). Metcalfe's law and legacy. *Forbes ASAP*, 13 September. [http://www.gildertech.com/public/telecosm_series/metcalf.html]

Gilder, G. (1999). *The Telecosm Glossary: An Opinionated Lexicon.* Second Edition. Gilder Technology Group.

Gupta, A., D.O. Stahl, and A.B. Whinston (1996). Priority pricing of integrated services networks. In: L.W. McKnight and J.P. Bailey (eds.), *Internet Economics.* Cambridge, MA: MIT Press, pp. 323–352.

Haimanko, O. and R. Steinberg (2000). Price symmetry in a duopoly with congestion, CORE Discussion Paper, No. 2000/56, December.

Jaffe, J.M. (1980). A decentralized, "optimal", multiple user, flow control algorithm. *Proceedings of the Fifth International Conference on Computer Communications*, pp. 839–844.

Jaffe, J.M. (1981). Bottleneck flow control. *IEEE Transactions on Communications.* COM-29 (7), 954–962.

Jarray, F. and L. Wynter (2002). An optimal smart market for network pricing and resource allocation. INRIA Research Report RR-4310, Institut national de recherche en informatique et en automatique.

Kelly, F. (1996). Modelling communication networks, present and future. *Philosophical Transactions of the Royal Society* A358 (2000) A354, 437–463.

Kelly, F. (1997a). Charging and rate control for elastic traffic. European Transactions on Telecommunications 8, 33–37. [Revised version available at http://www.statslab.cam.ac.uk/∼frank/elastic.html]

Kelly, F.P. (1997b). Pricing and rate control for communication networks. 1997 Naylor Prize Lecture, London Mathematical Society. November 21. [Available at http://www.statslab.cam.ac.uk/∼frank/TALKS/LMS/]

Kelly, F. (2000). Models for a self-managed Internet. *Philosophical Transactions of the Royal Society* A358 (2000) 2335–2348.

Kelly, F. (2001). Mathematical modelling of the Internet. In B. Engquist and W. Schmid (eds.), *Mathematics Unlimited - 2001 and Beyond.* Berlin: Springer-Verlag, 685–702.

Kelly, F. P., A.K. Maulloo, and 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.

Lovelock, C. (2001). *Service Marketing: People, Technology, Strategy*, Fourth Edition. Upper Saddle River, New Jersey: Prentice Hall.

MacKie-Mason, J.K. and H. R. Varian (1995). Pricing the Internet. In: B. Kahin and J. Keller (eds.), *Public Access to the Internet*. Cambridge, MA: MIT Press, pp. 269–314.

Mackie-Mason, J.K. and H.R. Varian (1996). Some economics of the Internet. In: W. Sichel and D.L. Alexander L. (eds.), *Networks, Infrastructure, and the New Task for Regulation*. Ann Arbor: The University of Michigan Press, pp. 107–136.

Marbach, P. (2002). Priority service and max-min fairness. *IEEE INFOCOM 2002*, 266-275.

Mark, R. (2002). AOL maintains strong lead among ISPs Worldwide. Press Release, WebSideStory, March 7, http://www.websidestory.com/

Mazumdar, R., L.G. Mason, and C. Douligeris (1991). Fairness in network optimal flow control: optimality of product forms. *IEEE Transactions on Communications* 39 (5), 775–782.

Nash, J. (1950). The bargaining problem. *Econometrica* 18, 155–162.

Odlyzko, A.M. (1997). A modest proposal for preventing Internet congestion. Unpublished manuscript, AT&T Labs, September.

Odlyzko, A.M. (1998) The economics of the Internet: Utility, utilization, pricing, and Quality of Service. AT&T Labs Research, working paper, July [Available at http://www.dtc.umn.edu/∼odlyzko/doc/networks.html].

Odlyzko, A.M. (1999a). Paris Metro pricing for the Internet. *Proceedings ACM Conference on Electronic Commerce* (*EC'99*), ACM, 1999, 140–147.

Odlyzko, A.M. (1999b). Paris Metro pricing: The minimalist differentiated services solution. *Proceedings 1999 Seventh International Workshop on Quality of Service* (*IWQoS '99*), IEEE, 159–161.

Odlyzko, A.M. (2000). The history of communications and its implications for the Internet. AT&T Labs Research, working paper, June [Available at http://www.dtc.umn.edu/∼odlyzko/doc/networks.html].

Owen, D. (1999). Paris rail abolishes first class travel. *Financial Times*, September 1, 1999, p. 2.

de Palma, A. and L. Leruth (1989). Congestion and game in capacity: A duopoly analysis in the presence of network externalities. *Annales d'Economic et de Statistique* 0 (15–16), 389–407.

Papadimitriou, C.H. (2002). Learning the Internet. In: J. Kivinen and R.H. Sloan (eds.), *Computational Learning Theory* (*Proceedings of the 15th Annual Conference on Computational Learning Theory, COLT. Lecture Notes in Computer Science* 2375, Springer, p. 396.

Park, K., M. Sitharam, and S. Chen (2000). Quality of service provision in noncooperative networks with diverse user requirements. *Decision Support Systems, Special Issue on Information and Computation Economies* 28, 101-122.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Rawls, J. (1999). *A Theory of Justice*, Revised Edition. Cambridge, MA: Harvard University Press.

Roth, A. (1979). *Axiomatic Models of Bargaining.* Berlin: Springer-Verlag.

Salgado, T., H. Valder, and C. Couser (2002). UBS Warburg selects Cable & Wireless as global telecoms supplier, Press Release, Cable & Wireless, September 5.

Shenker, S. (1995). Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communication* 13 (7), 1176–1188.

Simon, H. (1989). *Price Management.* New York: North Holland.

Tirole, J. (1988). *The Theory of Industrial Organization.* Cambridge, MA: MIT Press.

Transportation Research Board 2002). 91 Express Lanes Orange County, California. *Innovative Finance for Surface Transportation*, National Cooperative Highway Research Program, May [http://www.innovativefinance.org/projects/highways/91.asp].

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16 (1), 8–37.

Vickrey, W. (1962). Auction and bidding games. In: *Recent Advances in Game Theory.* The Princeton University Conference, 15–27.

Waldspurger, C.A., Hogg, T., Huberman, B.A., Kephart, J.O., and Stornetta, W.S.

(1992). Spawn: A distributed computational economy. *IEEE Transactions on Software Engineering* 18 (2), 103–117.

Wall, L. and R.L. Schwartz (1990). *Programming Perl*. O'Reilly & Associates, Inc. 1990.

Weber, R. (2000). Pricing communications services. Lectures presented to the Netherlands OR community, Lunteren, 11–13 January.

WebSideStory (2002). AOL maintains strong lead among ISPs worldwide. Press Release, March 7.

Zakon, R.H. (2003). Hobbes' Internet timeline. v6.0, February [http://www.zakon.org/robert/internet/timeline/].