1999

Internet Resource Management and Pricing

Christopher J. Clark
Edith Cowan University

Edith Cowan University

Copyright Warning

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

Internet Resource Management and Pricing

by

Christopher J Clark

A Thesis Submitted in Partial Fulfilment of the
Requirements for the Award of
Bachelor of Science Honours (Computer Science).

At the Faculty of Communications, Health and Science, Edith Cowan University, MtLawley.

Date of submission: 26th November 1999

ABSTRACT

Originally conceived and funded as a research project, the Internet has grown into a commercial, global and integrated service network. This has changed the nature of traffic on the Internet with the increasing use of things like video conferencing and time critical transactions. These forms of Internet usage place high demands on bandwidth. Added to this is the fact that the number of users is increasing at a dramatic rate and shows no signs of slowing. This is leading to a 'tragedy of the commons' where endemic congestion will reduce the value of the Internet to everyone. It also implies the introduction of some form of quality of service (QoS) to differentiate time critical traffic from less time critical traffic.

Pricing usage has been shown to be effective in controlling congestion by promoting more effective resource allocation. To provide the necessary QoS, there is an argument that simply increasing the available bandwidth will achieve this, while at the same time maintaining the simple model of the current Internet. However, there is also an argument that a more complex model may be needed that provides various levels of QoS with an associated pricing scheme to manage usage of these levels of QoS. A major part of the debate on this subject surrounds the trade-off between efficiency, economics and complexity that exists in introducing QoS and pricing to the Internet. This document discusses some of these issues, presents some of the current proposals for pricing Internet usage and finally compares the presented pricing proposals.

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

(ii) contain any material previously published or written by another person except where due reference is made in the text; or

(iii) contain any defamatory material.

Signature 

Date 26/11/1999

ACKNOWLEDGMENTS

Acknowledgements go to Dr. S. P. Maj for his encouragement, helpful guidance and informed comments, which were invaluable and given freely throughout the preparation of this document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background

The Internet is a network of networks that uses packet-switching communications technology based on the Transmission Control Protocol / Internet Protocol (TCP/IP). In this scheme, no open connection is maintained during a communication session as in the circuit-switched technology used in telephony. Instead, all users share a common communications medium that uses 'statistical multiplexing' to facilitate this sharing. The stream of data to be sent is broken up into pieces called 'packets' which are sent on to the network. When one computer is not sending a packet other computers can use the line to send packets. The breaking up of the data into packets and the reassembley on arrival is handled by the TCP. IP provides the addressing needed by computers on the Internet (routers) to forward the packets to the next link on their journey to the destination.

This 'packetisation' allows for efficient use of the communications medium. Take the case of an interactive terminal session to a remote computer. The majority of the time the user is thinking and the network is only needed when they strike a key. When this happens the data is encapsulated in a packet and sent across the network. Maintaining a connection the whole time would be wasteful and would stop other users using the communication line.

As there is no connection maintained during communication, the Internet is referred to as 'connectionless'. Each packet is routed independently to its destination and it is possible for packets belonging to one message to take different routes and arrive out of order. When a packet is received by a router, it examines the destination address in the IP header and passes it to the next router, which is chosen according to a routing algorithm. The routing is dynamically calculated to provide the best route to the next

hop, this means it is possible for packets to take different routes and arrive in a different order. This provides a very good failure recovery mechanism; if a node on the route fails, the routing information will be updated and the packets will be able to take a different route. TCP ensures that on arrival the packets are reassembled in the correct order and that none are missing, and IP ensures that packets reach the correct destination.

The network makes no commitments about how long a packet will take to be delivered, or even if it will be delivered at all. In other words it is a 'best effort' service. These delays and lost packets can occur during times of congestion. If the arrival rate of packets at a switch (router) is greater than the outgoing rate, queues of packets build up in the switch. These queues can cause packet delays, and if buffer space runs out, packets will be dropped. There is also a time-to-live value in the IP header that, when exceeded, will also cause the packet to be dropped.

In addition to the delay and dropping response of routers, TCP has mechanisms that are sensitive to congestion. These mechanisms are described by Clark and Fang (1998, p. 366), but simply put, when TCP detects that a packet has been lost, through not receiving an acknowledgement, it slows down its rate of transmission (or 'backs off') and then resends the lost packet. When packets are not being lost, TCP will increase its transfer rate to try to fill the network links fully.

This simple 'connectionless', 'best-effort' architecture has been very successful to date. However, there are current moves to extend this architecture in two ways. The first is the introduction of multicasting. At present, when a sender wants to send a packet to multiple receivers, the packet must be duplicated once for every receiver. This is not an efficient use of resources. Multicasting will allow the source to send the packet only once, and it will be replicated by the network only when necessary, that is when the

transmission paths diverge. The packet's destination address, then, becomes the multicast group address, which doesn't convey information about the receiver's location as it is a logical address. Receivers wishing to become part of a multicast send a 'join' message to the nearest router. The routing algorithm then creates distribution trees that connect every source to every receiver. An important characteristic to this discussion is that in multicasting the sender is not aware of who is receiving the packets.

The second proposed extension to the Internet architecture is the introduction of different Qualities of Service (QoS). This is because the current best-effort service may not be able to support some of the future video and voice applications. Also, providing the same level of service to all applications may not be an efficient use of bandwidth; providing differential QoS may allow scarce resources to be devoted to the applications that are the most performance sensitive. This has implications on pricing the Internet as Shenker, Clark, Estrin and Herzog (1996, p. 200) point out "Offering multiple qualities of service requires some form of incentives, such as pricing, to encourage the appropriate use of the service classes". However, the introduction of levels of QoS is not a foregone conclusion and there is strong debate on the subject.

Due to the changing nature and rapid growth in Internet use there is increasing congestion and a case can be made for introducing some form of resource management. Usage-based pricing has been proposed as an effective way to achieve more effective management. However, there is much debate on how usage-based pricing should be implemented, or whether it should be implemented at all. Included in this debate are the issues of how to handle differential QoS and multicasting, which both add complexity to any usage-based pricing solution. Many of these issues remain unresolved. This document discusses some of those issues and describes some of the possible usage-based pricing solutions put forward by various researchers.

1.2 Research Questions

- What are the contemporary issues surrounding resource usage management on the Internet?
- How does usage-based pricing relate to resource usage management on the Internet?
- What are the advantages and disadvantages of current pricing proposals for Internet usage?

2. THE CASE FOR USAGE-BASED PRICING

2.1 The Expanding Internet

The use of the Internet in general is increasing at a rapid rate as a 1998 study by Daley (cited in IRM, 1998, p. 1) shows:

1. Internet traffic is doubling every 100 days.
2. Between 1993 and 1997, the number of Internet users rose from 3 million to over 100 million.
3. Just 4 years after the Internet was opened to the public, 50 million people were connected. It took radio 38 years and TV 13 years to reach that mark

According to Coffman and Odlyzko (1998, p. 2), traffic and capacity for the public internet grew at 100% per year in the early 1990s, and during 1995 and 1996 this increased to1000%. In 1997 this settled back again to 100%. They also say that most reports on the growth rate of the Internet vary widely, however, all show a rapid rate of growth.

The 100% growth in traffic per year is also demonstrated by the Western Australian Regional Network Organisation (WARNO) that provides Internet access through the Perth Academic Research Network (PARNet) to universities and research organisations in Western Australia. Figure 1 shows the statistics for the traffic transmitted by WARNO from November 1997 to October 1999 (data sourced from PARNet, 1999). This shows a general upward trend in traffic transmitted for all institutions, with the overall total for Feb-99 being almost double of that for Feb-98 and the total for Sep-99 more than double that for Sep-98. Statistics for traffic received by WARNO show a similar trend.

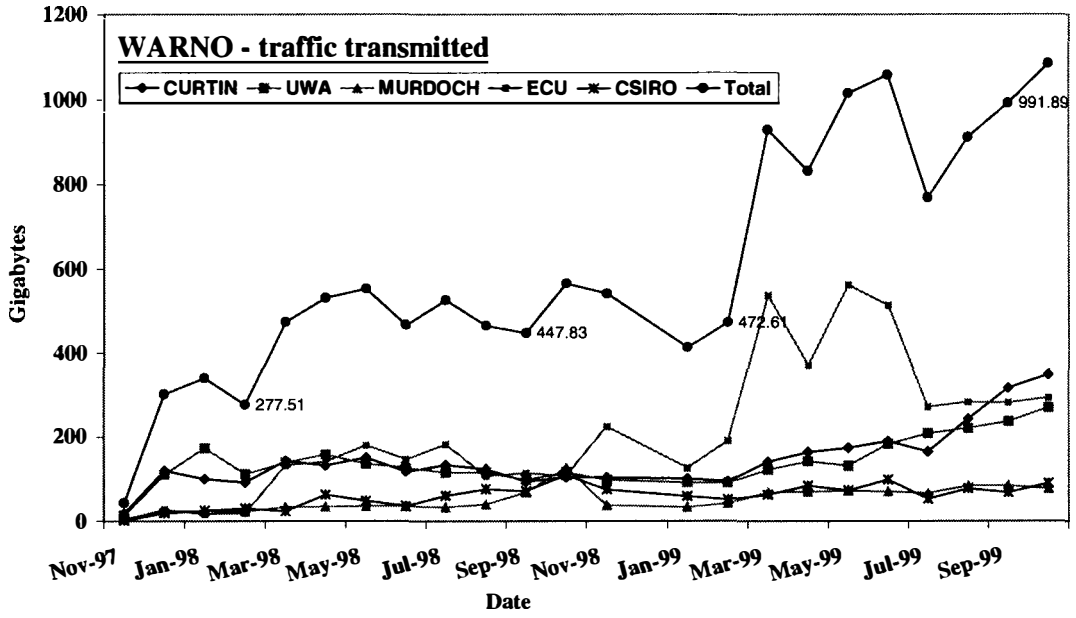

Figure 1: Traffic transmitted by WARNO (produced from statistics from PARNet, 1999).

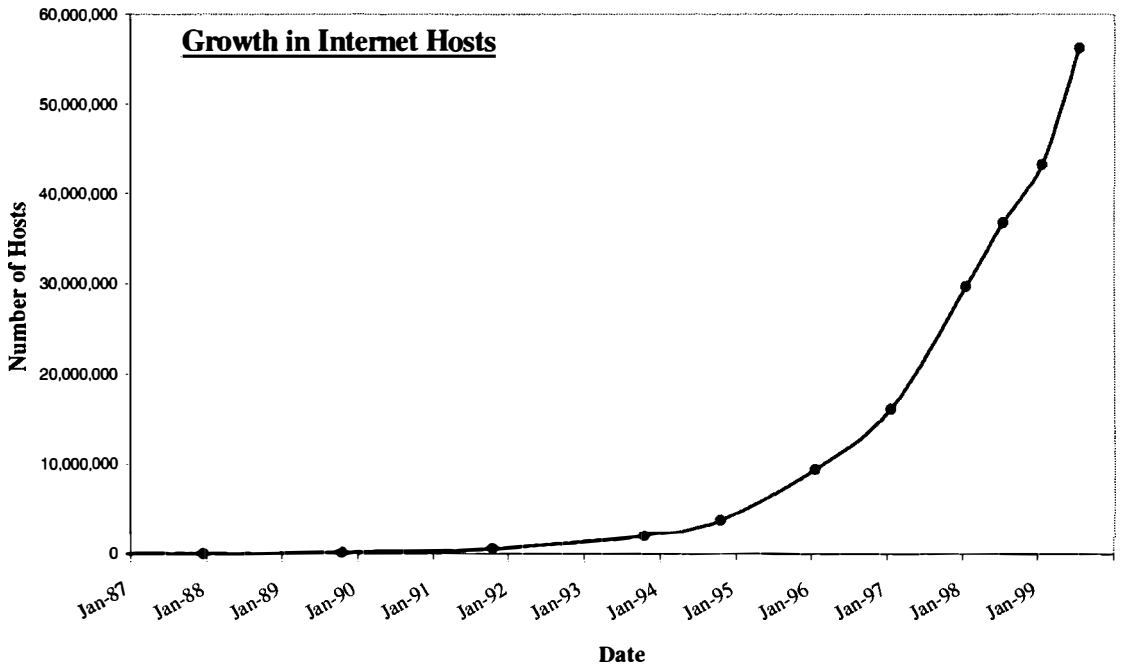

Figure 2: Growth in number of Internet hosts (produced from statistics from ICS, 1999).

The number of hosts on the Internet is also increasing rapidly. Data collected by the

Internet Software Consortium was used to produce the graph in Figure 2. It shows an exponential growth in the number of hosts, reaching 56,218,000 in 1999, and as the curve indicates the expected total for the year 2000 to be much higher.

The increasing number of people using the Internet is obviously contributing to the increase in traffic. However, the content of network traffic is also becoming more sophisticated and the messages are getting larger. This is due to the use of such things as streaming video and audio, and web pages with high graphical content. These two forces are causing a rapid increase in the demands on bandwidth and contribute to congestion, which at times can cause unacceptable delays in transmission. Some applications, such as real-time video, are sensitive to delay, and so are rendered useless at times of congestion. The other aspect is that these bandwidth hungry applications can also cause delays to other users, and this has already happened. For example, "during the weeks of November 9 and 16, 1992, some packet audio/visual broadcasts caused severe delay problems, especially at heavily used gateways to the National Science Foundation Network (NSFNET) backbone and in several mid-level networks" (MacKie-Mason & Varian, 1993, p. 9). In an earlier example "the Internet effectively collapsed in 1986 before TCP was redesigned to avoid congestion" (Schnizlein, 1998, p. 52).

The reason the nature of Internet traffic is changing is that the Internet, which originally started as primarily a research tool is now being used for things such as e-commerce and entertainment. This has led to the increasing use of things like video conferencing and time critical transactions. These forms of Internet usage, as well as placing demands on bandwidth, imply that the introduction of some form of QoS to differentiate them from less time critical traffic may be of value. There are arguments that simply increasing the available bandwidth will provide this QoS, while at the same time maintaining the simple model of the current Internet. However, there are also arguments that a more complex model may be needed that provides various levels of QoS with an associated

pricing scheme to manage usage of these levels of QoS. A major part of the debate on this subject surrounds the trade-off between efficiency, economics and complexity that exists in introducing QoS and pricing to the Internet. This document will be discussing some of the issues and also presenting some of the current proposals for pricing Internet usage.

2.2 Charging for Internet Access - the Present Situation

A common misconception is that the Internet is free. However, this is not the case. Pricing is on the basis of a flat-rate monthly access fee (with no charge for incremental usage). Many people are insulated from this cost by being members of an educational institution or corporate intranet that absorbs the cost of access. At Edith Cowan University (ECU) for example, Internet access is provided free to students and staff. This is becoming a problem as the cost is rising significantly each year. The budget allocation to the communications department at ECU for Internet usage is fixed at $106,000 annually. The bill from the university's service provider, Australian Academic Research Network (AARnet) for 1998 was $423,000. The budgetary gap is expected to be larger this year with the 1999 second quarter bill at the time of writing at $227,000 and the projected total for 1999 to be $740,000.

The reason the Internet has flat-rate access charges with no explicit restrictions on the incremental use of bandwidth is historical. The Internet originated as the Advanced Research Project Agency Network (ARPANET), a defence research project funded by the Advanced Research Project Agency (ARPA). The primary design goal of the Internet was "survivability in the face of failure" (Clark cited by Fang, 1996, p. 105). Owing to this and its military application, there was little or no emphasis on an accounting infrastructure. As a result most pricing and billing on the Internet is access-based, where users are charged a fixed fee per month for access with no incremental usage fees. The great advantage of this is that it is simple and gives predictable costs.

However, the nature of the Internet has changed and is changing rapidly, what was originally conceived and funded as a research project has grown into a commercial, global and integrated service network.

The types of flat-rate access in use today, according to Fang (1996, p. 105) can be divided into two groups:

- **By access pipe** - dedicated lines are provided to enterprises by Internet service providers (ISPs) and charged according to the bandwidth supported by the line. This usually includes a fixed start-up fee and a flat monthly connection fee.
- **By access time** - The most common for individual users. Users pay a flat monthly fee for unlimited access or a fixed amount for a given number of hours with an extra charge for each hour over the quota.

Another charging scheme that is an extension of the access pipe is used by the Australian Academic Research Network 2 (AARNet2) which provides the Internet to the universities in Australia. This scheme provides differential charging. Three bands of charging exist, on-net (domestic) traffic, off-net (international) traffic and local traffic between each Regional Network Organisation (RNO). This scheme takes into account that more distant destinations may require using more expensive links and so it regains some of the costs for providing those links. However, neither this scheme nor the ones above take into account the nature or volume of that traffic, effectively allowing unrestricted use of the line capacity once connected and providing undifferentiated service to both time-critical and non-time-critical applications alike. This can result in an effect called the 'tragedy of the commons', which can cause a general degradation in service.

2.3 'Tragedy of the Commons'

Unrestricted use of a shared resource, according to MacKie-Mason and Varian (1995, p. 1141), can lead to what is referred to as the "tragedy of the commons" where public goods are accessed without regulation, leading to a possible total loss of access. An example of this is where a public area or "common" is made available for people to graze their animals, free of charge or restriction. The end result is that people will take advantage of this and use the common without thought of the consequences to the general good. In the end the land is overgrazed and useless to everyone. This could be equated to the present day Internet, with the "common" being the available bandwidth. The lack of regulation is that users can indiscriminately use the line capacity without concern for the effect it is having on other users. In light of the increasing use of the Internet combined with more demanding applications this could lead to a congested network with reduced value for everyone.

Once the infrastructure of a network is in place, "the short-term incremental cost of providing passage through fixed capacity computer networks is essentially zero". (Gupta, Stahl, & Whinston, 1999, p. 58). Given, then that physically sending a packet costs nothing, the true cost may lie in adding to the "tragedy of the commons". That is, the cost caused by the congestion and resulting loss in overall value of the network caused to other users by sending a packet. For this reason, many researchers have focussed on externality (or congestion) pricing (for example, (Clark & Fang, 1998; Fang, 1996; Gupta et al., 1999; MacKie-Mason & Varian, 1993; Stahl, Whinston, & Zhang, 1998)) . In this scheme the amount of congestion a request for service will cause on a network is assessed and used to place a value on that service. If the request is made at a time when the network is not congested, then the cost is essentially zero. This can be seen as a method of balancing network resource usage, as users will tend to spread their network use over more economical periods of lower congestion.

There have been many simulations (e.g. Cocchi, Shenker, Estrin, & Zhang, 1993; Gupta et al., 1999; Stahl et al., 1998) that demonstrate the effectiveness of externality pricing in providing a more optimal balance in network resource usage than flat-rate pricing. An example is provided by Gupta el al. (1999) who carried-out a simulation of their computational approach that produces what they term 'optimal pricing'. 'Optimal prices' are arrived at by dynamically adjusting prices according to the current level of congestion. They found this form of usage-based pricing gave significant gains over fixed charges and time-based charges. The benefits shown include performance enhancements as well as monetary benefits. Figure 3 shows the benefits gained plotted against varying levels of traffic volume (exogenous load) placed on the network for the different pricing schemes with the 'optimal pricing' of their computational approach coming out well ahead.

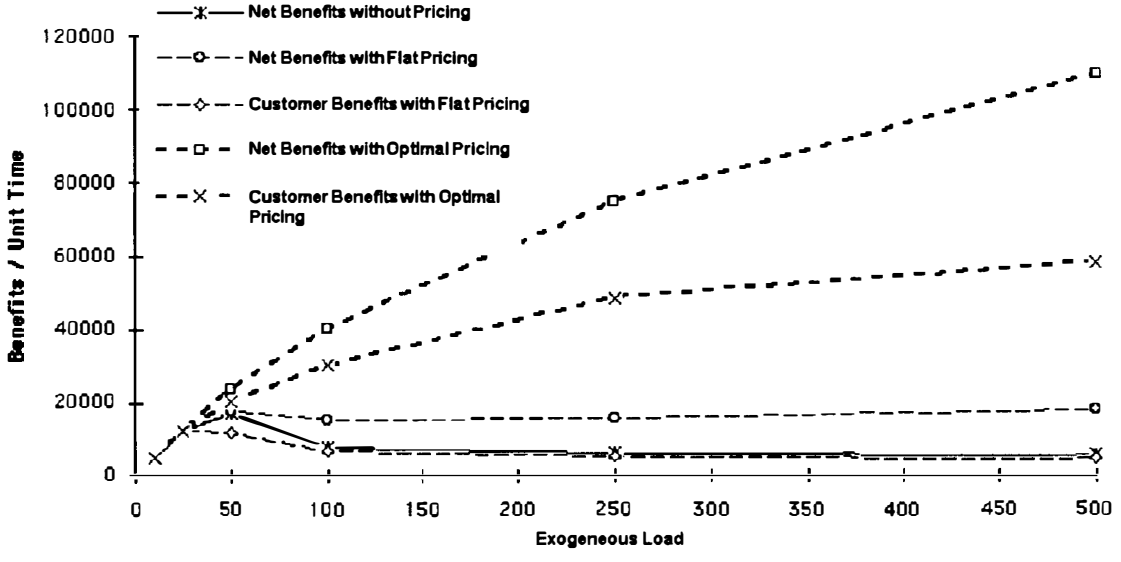

Figure 3: Benefits of Optimal Pricing (Gupta, Stahl, & Whinston, 1995b, p. 11)

These results are supported by a simulation study performed by Stahl, Whinston and Zhang (1998) that compared Gupta et al.'s optimal pricing with flat-rate charging as used by America Online (AOL). The findings showed that both users and service

providers would benefit under an optimal usage-based pricing scheme. They also found that people are more willing to pay usage based fees than flat-rate fees for Internet use in return for reduced congestion and will voluntarily choose a subnetwork with optimal usage-based pricing over another with fixed access fees and no usage charges. The optimal usage-based charging scheme showed seven times the benefits over the flat-rate scheme. These benefits being the reduction in congestion and waiting time. The service provider's benefits were increased profits with the optimally priced usage-based subnetwork generating five times the profits as flat-rate charging.

Apart from balancing resource usage by spreading network use, the benefits of usage-based pricing stem from the avoidance of what can be referred to as the 'perverse effects' of flat-rate pricing, where a congested network can actually provide incentives to create more congestion.

2.4 Perverse effects of Flat-rate Pricing

Odlyzko (1997, p. 11) describes the case where flat-rate pricing and a congested, slow network can actually encourage the increase of data transfers. Jokingly referring to the 'World Wide Web' as the 'World Wide Wait' he says tools have been developed to get better performance while surfing the Web. One such tool is 'PeakJet' that uses the time a user spends looking at a Web page to download all the pages linked to that page, so that if the user decides to look at a page later it is quickly available from the local hard disk. Another similar tool is 'WebWhacker' that can be left to download web pages all night in case the user wants to view them the next day. The problem is that the more congested and slower the network is, the more incentive there is to use these tools. Most computers usually only use a fraction of the capacity of their link to the Internet, these tools, however, exploit the full available bandwidth. Odlyzko (1997, p. 11) estimates that it would take less than 200,000 Personal Computers (a very small fraction of those currently connected) with a connection of 28.8Kbps to completely saturate the Internet

if they were all downloading web pages at the full rate. Modems can now run at double this speed, suggesting that it would take far fewer computers to achieve saturation.

Another effect of flat-rate pricing in an undifferentiated QoS network is that it leads to Internet Service Providers (ISPs) offering tiered quality service. That is, the demands of users who require a high QoS are catered for by offering separate tiers of service with different flat-rate charges. A high-speed link will be offered at a higher price (usually much higher) than a lower speed link. Once a user subscribes to a certain tier, they are committed to stay with it, no matter how their need for QoS changes. Edell and Varaiya (1999) point out that in this situation, consumers and suppliers lose out on two counts:

> Since tiered service is flat-rated, and demand is quality-sensitive, the waste at higher speeds is greater, and the flat rate charge is correspondingly higher. Also, many light users will not subscribe to flat-rated higher quality service, even though they would occasionally subscribe if the charge were usage-based. So those users are denied the benefits of higher quality service, and producers are denied the revenues from those subscribers. (p. 11)

This model using tiered service results in the numbers of subscribers to the higher-speed links being limited, which reduces the revenues generated by those links. This will "lower the pace of equipment cost reduction" (Edell & Varaiya, 1999, p. 11) and provide a disincentive to ISPs to invest in expanding and improving those links. In this way, flat-rate pricing can be seen to be inhibiting the improvement of QoS on the Internet.

These kinds of symptoms are the result of a system that is not efficient both in terms of economics and resource allocation. Usage-based pricing has been demonstrated to increase efficiency, control resource usage and provide economic benefits to both users and providers. This was shown in above in Figure 3. Other examples of simulations that clearly demonstrate these benefits can be found in Cocchi, Shenker, Estrin, & Zhang (1993), Gupta, Stahl, & Whinston (1997), MacKie-Mason & Varian (1995), Parenteau

& Rishe (1997), Rupp, Edell, Chand, & Varaiya (1998) and Stahl, Whinston and Zang (1998).

Apart from these 'perverse effects', there is also the question of fairness, as different user applications generate different amounts of traffic and have different bandwidth requirements, but pay the same access fee as each other under a flat-rate scheme.

2.5 Traffic Volumes and Bandwidth Requirements

Edell, McKeown and Varaiya (1995, p. 1162) point out that different users generate widely different amounts of traffic. A pricing scheme dividing costs equally among users would therefore not seem appropriate. Why should a user with low bandwidth requirements, who is generating small amounts of traffic, pay the same as a user with more demanding requirements? For example, during a telnet session a user requires high speed, low volume data transfer for real-time interaction with a server. Whereas a video-conference needs clarity and co-ordination of picture and sound which requires high-volume, high-speed transfers with low variability in transfer rates. The telnet user, generating little traffic, may be experiencing unacceptable delays when sharing the network with users involved in a video conference, and yet pay the same access fee. It also may be the case that the users involved in the video-conference would be willing to pay more for higher bandwidth during the video-conference to guarantee the QoS.

The difference in bandwidth requirements can be illustrated by considering the dramatic difference between ASCII text and multimedia. According to Lucky (cited in MacKie-Mason & Varian, 1994):

> ASCII text uses about 44 bits per word. Telephone-quality voice uses 21,000 bits per word, and stereo CD uses 466,000 bits per word. Network quality video without compression is about 100 *megabits* per second. With compression, it's about 45 Megabits per second! - which is the entire capacity of the NSFNET backbone. (p. 3).

Presently in the Internet all users are charged the same, regardless of the demands made on the capacity of the link they have access to. We have the case of light users subsidising heavy users. ISPs may even find heavy users unprofitable as they place large demands on link capacity, but pay the same flat rate as light users.

Not only do users generate different amounts of traffic, but also they place different values on their usage. Sound economic theory dictates that this valuation of service should be linked to a price.

2.6 User's Valuation of Service

With the present 'best effort' service of the Internet there is no way for a user to indicate the value they place on a service. All users are treated equally. As Parenteau and Rishe (1997) point out:

> There is no deterrent to the user downloading many megabytes in which he has only limited interest. Whereas the user who is performing time critical communications, is placed in the same queues as the casual user. Clearly the latter's preferences are not being served. Nor is the interest of the service provider since the time-critical user would likely be willing to pay for improved performance. (p. 93)

This is not an economically efficient state of affairs. As the Internet is moving away from its research oriented beginnings into a more commercial realm it may have to adopt policies in line with economic theory. As Schnizlein (1998, p. 52) notes, "an economically efficient system generates revenue for expansion by matching prices to user's valuation of the services". There should be incentives for users to use only the bandwidth they really want, in other words what they are willing to pay for. Linking the prices to a user's valuation of a service will do this and also give incentives to carriers to carry out expansion as providing higher levels of offered service means higher income for the carriers. The willingness of users to pay for a certain level of service would also

give indicators to service providers of whether expansion is viable or necessary.

To date there has been little or no empirical evidence of how users value different levels of service. The Internet Demand Experiment (INDEX) currently being run at Berkeley aims to rectify this situation and has already produced some preliminary results.

2.7 The Internet Demand Experiment (INDEX)

In a model of the Internet that has various levels of QoS, efficiency can be obtained if the combinations of price and QoS match user needs. (Rupp et al., 1998, p. 85) says that to do this service providers must understand "the structure of user demand". A group at the University of California in Berkeley set out to reach this understanding through the INDEX experiment, which is a:

> Real-world market trial seeking to provide this information [the structure of user demand] and measure how individuals value Internet usage when they are offered different Quality of Service choices. (Rupp et al., 1998, p. 85)

This trial started in April 1998 and is scheduled to run for two years. It provides Internet access to about 70 users from the Berkeley campus over Integrated Services Digital Network lines. Users select network services from a menu of QoS-price offerings and have to pay for their usage. This involves real monetary costs, which provides the necessary incentives to users to choose services based on their true valuation of network resources. Users are offered a sequence of service plans, which make up various experiments that last about 6-10 weeks each. Users can instantaneously change their QoS-price choice, even during a session. They also have instantaneous feedback in the form of a price meter that shows how much they are spending. The user choices are monitored and the value a customer places on a service is measured by the time and money they spend on that service. The experiment only monitors TCP traffic, the authors citing the fact that TCP makes up the majority of traffic at Berkeley to justify

this.

The main goals of INDEX are:

- Measurement of user demand for Internet access as a function of QoS, pricing structure, and application.
- Demonstration of an end-to-end system that provides access to a diverse group of users at attractive price-quality combinations.

The preliminary findings, as published in Edell and Varaiya (1999) show:

- Charging subscribers a small amount per megabyte caused a 35% drop in network traffic. (This translated into about $3 per month).
- When usage is measured by connect time the waste induced by flat-rate pricing is large.
- If users are offered various levels of service quality, the demand for more than one service quality increases. (shows the loss to consumers and providers of tiered service where consumers are locked into one service quality)
- The demand for use of high quality services increases if there are a greater number of high levels of QoS on offer. This is because quality sensitive applications behave better when the QoS is higher, so there is more attraction to use those applications. Also, the intangible cost of waiting time of the user is reduced with a higher QoS, encouraging more use.
- Under flat-rate tariffs, light users are subsidising heavy users.

The results provide empirical evidence demonstrating that flat-rate pricing is actually detrimental to the growth of the Internet as it fosters waste and the proliferation of tiered levels of service, which as discussed earlier, are a negative influence on Internet expansion. The results also support the introduction of different levels of QoS as it was

shown that users would choose different levels of QoS if they were available. Additionally, the availability of higher levels of QoS would encourage the use of more sophisticated applications that take advantage of the higher QoS and so provide more incentive for capacity expansion and the provision of the higher levels of QoS. This would have the effect of increasing the level of user satisfaction while using the Internet, and may encourage even greater numbers of users to subscribe to the Internet.

The results of INDEX support placing the choice of QoS in the hands of the user and show that benefits will be obtained by doing so. This view is also held by MacKie-Mason, Murphy & Murphy (1995) and Gupta et al. (1999). The reason users should be able to select the QoS is that the value a user places on a service is subjective, and varies across time and application. The QoS, then, should not be tied directly to the application. Prices can be the controlling link between the users subjective valuation and the QoS an application receives.

2.8 Quality of Service and Pricing

QoS from a user's perspective can be defined by things like window size, download time or audio quality. These can be translated into technical terms by considering probability of packet loss, data transfer rate and consistency of delays of packets. Different applications require different levels of QoS. For example, email can be delivered without loss of quality no matter the delay or order of arriving packets. An interactive game or audio conversation, however, requires a minimum data transfer rate and an appropriate ordering of packets. However, defining QoS solely on an application basis has its drawbacks as Gupta et al. (1999, p. 58) point out. Sometimes a user might want urgent, high priority email and so would require a higher QoS to that mentioned above for email. A user may want to download a video to view at a later date and so a lower QoS than mentioned for video would be sufficient. Furthermore, users could develop 'masking' where an application requiring low QoS can be made to look like an

application requiring a higher QoS and so gain preferential treatment. A pricing mechanism can be used as a tool to overcome these problems. Pricing can be used to differentiate between levels of QoS required by users, if they need urgent email, they specify this by paying for it. Users would not be forced into paying a high fixed price for video if they are going to view it later. Pricing would discourage 'masking' by providing an incentive to maintain an appropriate level of QoS. Gupta et al. (1999, p. 58) contend " that to sustain an e-commerce environment in which each application and user will require a different QoS, pricing network traffic based on usage will be a necessity".

For these reasons MacKie-Mason and Varian (1995, p. 3) advocate bringing users back into the loop by using a form of feedback called responsive pricing. In this scheme, the QoS (and thus the price) is not set at an application level, but moved back into the control of the user. The response to this feedback need not be human with its limited response times. Very sophisticated user behaviour could be automated using pre-programmed network interfaces. This form of automatic feedback control is already demonstrated by the Transmission Control Protocol (TCP) congestion control algorithms, and so has a pretext for its implementation. Mackie-Mason et al. envisage that this feedback and response would be dynamic and occurring on a very finely-grained time scale. Discouraging adaptive users from transmitting when the network is congested and encouraging users to send traffic during lulls, thus increasing network efficiency. Gupta, Stahl and Whinston (1996) have developed a computational model to achieve this dynamic adjustment of prices and this is described below in 'Priority Classes'.

There are several schemes for implementing QoS and, as the main focus of this document is pricing the Internet, only a brief description of some of these are given in the section that follows along with their relationship to pricing.

2.9 Ways of Providing QoS

2.9.1 Asynchronous Transfer Mode (ATM)

ATM is designed to handle the communication requirements of various network services including data, images, voice, and video. ATM provides the necessary QoS for time sensitive data such as voice and video by providing a path through the network with a guaranteed delay time associated with that path. It achieves this by using 'virtual connections' which are maintained by ATM switches. An ATM network has a similar operation to the switched telephone network. In ATM, before sending information relating to a call, a communications path is established. All cells (similar to packets, but of a fixed size) relating to a call are then made to follow this path and are delivered in the same order as they were sent. ATM therefore provides QoS using a connection-oriented method. It is because of this connection-oriented nature that Odlyzko (cited in Tebbutt, 1998, p. 67) has doubts as to whether ATM is suited to the Internet.

Odlyzko's (cited in Tebbutt, 1998, p. 67) doubts stem from the fact that most Internet traffic uses the Hypertext Transfer Protocol (HTTP) (see Figure 4) and that a characteristic of HTTP is that it produces 'bursty' traffic. That is, there are long periods of inactivity followed by short bursts of traffic. For example, a user will not be generating any traffic while reading a Web page, but will cause a sudden burst of activity when a link is clicked on. Added to this is the possibility that every link on a web page could generate traffic from any one of the almost 60,000,000 hosts on the Internet. It would be a waste of resources, then, to maintain a connection during the whole time a user is viewing a web page and also the overheads of establishing new connections every time a different link is clicked would be very large. The bursts of traffic generated by using HTTP also usually consist of a small number of packets and so this traffic does not have the well-defined flows that ATM was designed for. It can be concluded then that ATM doesn't resolve the problem of minimising delay for time critical customers with 'bursty' traffic who are willing to pay for a reduction in traffic.

According to Odlyzko these customers are important and should not be ignored by relying solely on ATM to provide QoS:

> Web browsing encompasses a variety of important applications, many of them mission-critical, such as when your customers might be ordering from you or your doctor might be looking up information relevant to your sickness. It's all transacted through the HTTP protocol (Odlyzko cited in Tebbutt, 1998, p. 68)

According to Odlyzko, cited in the same article, this importance will continue into the future and so should be a concern in any QoS proposals.

Statistics on the WARNO from July 1997 to November 1999 (taken from PARNet, 1999) support the view that HTTP traffic is the most predominant on the Internet with HTTP making up 61% of the traffic (see Figure 4).

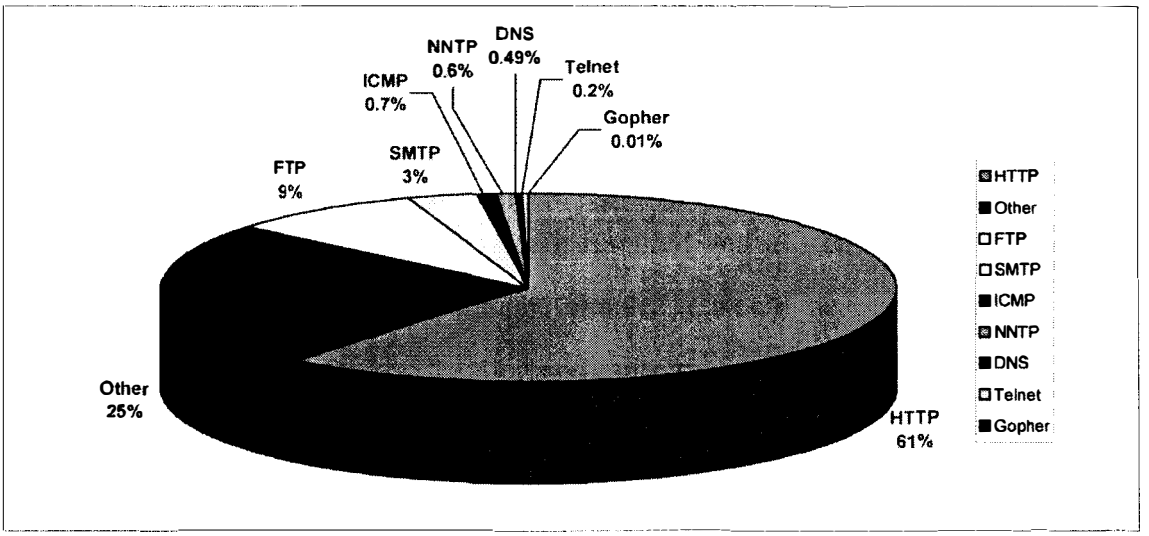

Figure 4: Traffic received by WARNO - Jul'97 to Nov'99, by protocol (produced from statistics from PARNet, 1999)

However, regardless of its lack of value in providing QoS for HTTP traffic, ATM is effective for things like video and audio due to their need for a guaranteed QoS and their characteristics of consisting of flows of traffic that can't tolerate out of order

packets or long delays.

2.9.2 Resource Reservation Protocol (RSVP)

RSVP is a method of providing QoS in Ethernet/IP networks. RSVP:

> reserves bandwidth between a sender and a receiver by examining each link along the route. The sender interrogates each network device [router] along the route. If each device has spare bandwidth available the path is established, if not the sender is informed (Engel & Maj, 1999, p. 2).

In RSVP a reservation commits resources to a user request, in which case Schnizlein (1998, p. 53) notes "admission control for reservations is the logical place to handle commitment to pay for those resources". Policies that define the admission control would be held on a separate 'policy server'. When a user makes a request for resources, routers would request a policy decision from the policy server before allocating the requested resources. The policy server would interact with the billing system to produce the necessary pricing information for the request.

Since the resources for a transmission are reserved and all the characteristics of the reservation are known, billing would be simple. It would be much like billing for telephone calls, that is, based on duration and capacity.

A disadvantage to RSVP is pointed out by Schnizlein (1998, p. 53). All the routers along the transmission path would have to maintain the state for each reservation. This can be expensive, especially when you consider that "reservations can be specified for individual flows from any application on any computer to another" (Schnizlein, 1998, p. 53). This means that the potential number of reservations is larger than the number of pairs of communicating computers. Schnizlein goes on to say "The feasibility of supporting the potentially huge number reservations aggregated near the center of the Internet is questioned."

RSVP is suitable for consistent flows such as real-time audio or video because it reserves the necessary resources along the transmission path to guarantee the necessary QoS. However, as regards 'bursty' HTTP traffic, it suffers the same problems as ATM Its main advantages, though, are that it provides ATM-like QoS over an Ethernet/IP network and a simple model for billing.

2.9.3 Internet Protocol Version 6 (IPv6)

IPv6 is the next generation Internet protocol that, among other things, will introduce a way of specifying the QoS of packets. It will do this by introducing traffic class and flow label fields that allow nodes in the network "to distinguish certain packets for possible special treatment by a router" (Lee, Lough, Midkiff, IV, & Benchoff, 1998, p. 30). The class field which was formerly known as the priority field in Internet Protocol Version 4 (IPv4), can be used to specify the type of application the data belongs to, for example a real-time application. (This would be set to all zeros for normal applications). The flow label field's function is to allow packets that have the same flow label to be treated by routers with the same specialised processing. For example, packets belonging to a video transmission would be tagged with the class field for video, and so that all packets in that transmission can be identified as belonging to the same stream, they would all be marked with the same flow label.

The possibility of being able to specify different levels of quality of service (QoS) with IPv6, as with other methods, means that it may be necessary to introduce some form of control to prevent everyone choosing the highest QoS all the time. Without this control the point of having differential QoS would be lost. Placing a higher economic value on a higher level of service has been shown to provide this control, as shown by simulations by groups such as Gupta et al. (1999) (see Figure 3). There is currently a movement to introduce IPv6 which underlines the need to implement some form of Internet pricing.

2.10 Do we need differential QoS?

Another aspect to QoS is that, in general, the lower the utilisation rate on a network, the higher the QoS the network provides. This is noted by Odlyzko (1998b):

> Even the notoriously congested trans-Atlantic links do appear to provide good performance for applications as demanding as packet telephony in the early hours of Sunday morning. What this says is that even without any new QoS technologies, one can provide excellent quality by lowering utilization. (p. 9)

Lower utilisation can be achieved by increasing the amount of available bandwidth. This leads to the idea of over-engineering the networks and providing a 'fat pipe' that will give a high QoS to everyone. Another way would be to provide different 'pipes' or 'channels' with varying levels of utilisation and so provide a range of QoS. This idea is outlined by Odlyzko (1997) in the Paris Metro Pricing (PMP) scheme (see below).

Taking the approach of providing QoS through reducing utilisation by increasing bandwidth (as proposed by Odlyzko (1998a; and 1998b)) has the great advantage of being simple. This approach maintains the current best effort service of the Internet, which has proved very successful to date. It avoids the likely high cost and complications of implementing a differential QoS scheme. It also avoids a serious defect of many proposed QoS schemes, which is that they require deployment throughout the Internet to be effective. The problem with this is that the Internet is a heterogeneous environment that lacks any central control, making it difficult to install end-to-end schemes.

Apart from adding complexity, Odlyzko provides economic reasons why implementing differential QoS may not be effective. In the U.S., data communications cost about $80 billion in 1997, which was 13% of total information technology (IT) spending. Actual data transmission accounted for:

> only 20% of total for data communications, and 2.6% of total for all of IT. Thus data lines are a small part of the entire IT picture, and any scheme that attempts to improve their performance has to be weighed against costs that it might impose on the rest of the system. It is better to double the spending on transmission than to increase the average cost of all other IT systems by 3%. (Odlyzko, 1998b, p. 7)

When you consider the complexity and cost of implementing a QoS scheme on the Internet and the pressure this will place on developers and network managers, along with the declining cost of bandwidth, it may be more cost effective to bypass the QoS issue by improving transmission capacities.

Further to the above, Odlyzko in a 1998 study found that, contrary to popular belief, most of the Internet is not congested. He found that long distance data links, backbone links and corporate private line links are lightly used:

> While the long distance circuit switched voice network has average utilization of about 33%, the Internet backbone links appear to have average utilizations closer to 10% to 15%, and corporate long-haul links (which is where the bulk of data transport capacity is) have utilizations in the 3% to 5% range. (Odlyzko, 1998a, p. 3)

He contends that the congestion occurs at key choke points, for example public exchange points and network access points. The key to solving the problem of congestion of bandwidth then, could be to address the key choke points, which Odlyzko (1998b, p. 5) says "should not be too expensive to eliminate".

However, Odlyzko admits that there is some contention about the significance of these points to congestion, and "there is still no consensus as to what causes the poor observed performance" (Odlyzko, 1998b, p. 4) of the Internet. A possible answer is put forward by Mackie-Mason and Varian (1994, p. 3), who say that low average utilisation rates on backbones are misleading because: "IP traffic is very bursty and

peak usage can be 10 times the average". An example of this is described by Smarr and Catlett (cited in MacKie-Mason & Varian, 1993):

> If a single remote visualization process were to produce 100Mps bursts, it would take only a handful of users on the national network to generate over 1Gbps load. As the remote visualization services move from three dimensions to [animation] the single-user bursts will increase to several hundred Mbps...Only for periods of tens of minutes to several hours over a 24-hour period are the high-end requirements seen on the network. With these applications, however, network load can jump from average to peak instantaneously. (p. 9)

This highlights the problem with all averages, there may be peaks in usage that could occur at a critical time which causes delays to other users. These peaks, however, will not be visible when an average is taken over a period of time.

Although there is contention about some of the causes of delays, it is not disputed that there are a large number of problems which cause delays that are not related to deficiencies in data transmission. For example, the main sources of consumer complaints are delays in email which are often caused by mail-server problems, and Web surfing delays which are mainly caused by server overloads. There is doubt whether a QoS scheme would solve these problems. According to Odlyzko (1998b, p. 5), it is possible that "making the entire system more complicated, increasing the computational burden on the routers, and increasing the numbers and lengths of queues" by introducing a QoS system, would aggravate the situation.

There are differing opinions on whether sufficient bandwidth will be available to provide 'fat pipes', these arguments are discussed in the next section.

2.11 Free Bandwidth?

There is an argument that bandwidth will soon be virtually free as "semiconductor densities are doubling every18 months, photonic bandwidths are doubling every 12 and wireless bandwidths are doubling every 9" (Metcalfe, 1998, p. 130). (These advances even outstrip the gains in microprocessor power, which has been doubling every 12 months). Advocates of this argument maintain that it would be better to keep the simple flat rate charging scheme and provide a single high quality of service to all users. However, in the same article Metcalfe also points out that Internet traffic doubles every 4 months, the number of Internet users is rising, the amount of time each user spends on the Internet is going up and the bandwidth usage of new applications is increasing. Regardless of bandwidth getting cheaper, these increasing demands will make bandwidth scarce, this scarcity naturally affecting its price. Metcalfe (1998, p. 130) says that these pressures mean "the Internet can no longer be an economics free zone". This is backed up by Fishburn and Odlyzko (1998, p. 129) who say "more than two decades of experience have shown that any bandwidth gets saturated quickly".

The argument that bandwidth is scarce and that any advances in bandwidth supply will soon be saturated is not clear-cut, however. Odlyzko, in a paper on the economics of the Internet goes against previous arguments for providing differential QoS, giving evidence that "providing enough bandwidth for uniformly high quality transmission may be practical" (Odlyzko, 1998b, p. 1). He gives the example of the University of Waterloo where a "12 fold jump in network bandwidth from 128Kbps to 1.5Mbps in July 1994 did not cause traffic to jump suddenly by a factor of 12. Instead, it continued to grow at its usual pace." (Odlyzko, 1998b, p. 3). This steady rate of growth is also displayed in the Internet (apart from the anomalous period in 1995 and 1996) with traffic doubling every year. The argument is that although the growth rate is rapid, it is predictable, and expansion to networks can be planned to cope with it. Added to this is that unit prices for transmission capacity will fall and total spending on high bandwidth

connections will rise, making expansion economically feasible.

If the technical problems of Wave Division Multiplexing (WDM) can be overcome, the argument that bandwidth will remain scarce may not be valid. WDM is a technique that "allows separate communications channels to be sent on different colours of light" (Tebbutt & Taylor, 1998, p. 95). This would provide huge increases to bandwidth, even in existing fibre-optic lines, so much so that "the whole world's Internet traffic could be pumped down one fibre" (Payne cited in Tebbutt & Taylor, 1998, p. 96). However, the technical problems are large and Tebbutt goes on to say: "WDM transmitters remain tremendously expensive and sensitive devices with fine-grained laser controls and mirror adjustments" that make them impractical at the moment. There is also the fact that this would only benefit areas connected to fibre-optic lines.

The fast growth and rapidly changing nature of every aspect of computing makes it difficult to predict whether sufficient gains will be made in bandwidth provision or if any gains will be quickly negated by higher usage. One thing is certain though, and that is computing is becoming more and more complex and difficult to manage. The tendency, then, may be to adopt the simpler approach to QoS. This view is upheld by Odlyzko (1998b, p. 6), saying: "optimality is unattainable, and we should seek the simplest scheme that works and provides necessary transmission quality".

3. THE CASE AGAINST USAGE-BASED PRICING

3.1 Accounting and Transaction Cost

The major objection to usage-based pricing is the accounting and transaction cost. The comparison can be made to telephony where the administrative costs can be 50% or more of the cost of making a call (McKnight & Bailey, 1995, p. 11). MacKie-Mason and Varian (1994, p. 2), however, believe that it depends on how well the prices are designed as to whether the benefits exceed the transaction and accounting costs. Certainly is has been shown in simulations (as mentioned above) that usage-based pricing can be an effective means of improving network efficiency and providing benefits to both users and network access providers. So, in simulations at least, the benefits would seem to outweigh the costs.

3.2 Overheads

Apart from the economic concerns, Shenker et al. (1996, p. 188) doubt that pricing based on congestion is even implementable due to technical difficulties (these are discussed further in the description of the 'Smart Market', below). One of these difficulties is the overhead involved in accounting for traffic on a packet level considering the number of packets traversing the Internet. In one month in 1997 the estimated traffic on Internet backbones was 3000 terabytes (Coffman & Odlyzko, 1998, p. 24). This already represents a huge number of packets and is currently doubling every year. Researchers such as Shenker et al. (1996) and Odlyzko (1998b) are concerned about this overhead.

Usage-based pricing may also introduce overwhelming complexity. This can be in the form of calculating the degradation to network performance and loss of utility to other users of performing a transmission in order to set a price. (Shenker et al. (1996, p. 192)

maintains that this may not even be possible). It could also be the complexity of new protocols that indicate whether the sender or receiver is going to pay. These issues are discussed more fully below in connection with each proposed pricing scheme.

Network managers may also be concerned about extra complexity and added monetary costs. The development and installation of a usage-based pricing may be expensive in terms of hardware and software. Additionally, monitoring and maintaining a usage-based pricing mechanism may add complexities and extra work to the daily tasks of an already busy network manager.

The complexities that could be introduced by usage-based pricing may destroy the simplicity of the connectionless, best-effort service that has so far made the Internet so successful. Part of this simplicity is also provided by the flat-rate pricing scheme currently in use. Flat-rate pricing is generally preferred by users and as a result there is resistance to introducing usage-based pricing.

3.3 Preference for Flat-rate Pricing

The preference for flat-rate pricing has been demonstrated in telephony where, in a study of the Bell System in 1970 (Cosgrove & Linhart cited in Odlyzko, 1997, p. 12), it was found that consumers are willing to pay more for a flat-rate plan than they would under a usage-based scheme. In data networking large organisations also show a preference for flat-rate pricing. For example, branches of the United States armed forces built their own networks when the United States (US) Defence Data Network introduced usage-based pricing (Bailey cited in Odlyzko, 1997, p. 12).

Odlyzko (1997, p. 12) cites Cosgrove & Linhart giving three reasons why consumers prefer flat-rates:

1. Predictability - Users know in advance how much the service will cost. This avoids

the worry of receiving a sudden large bill that hasn't been budgeted for.

2. Overestimate of usage - Users typically overestimate how much they use a service, and so calculate that they will be better off paying a flat-rate.
3. When users are charged on how much they use they tend to worry whether their use of the service is worth the cost. The result is that usage is reduced. This was observed in the US where charges for local calls had the effect of shortening the length of calls.

Service providers also prefer flat-rates. With flat-rates there will be no need for the potentially costly development of a traffic measurement and charging mechanism. Accepting the fact that consumers are willing to pay more for a flat-rate plan, there may be more profits to be gained by service providers. (Although simulations of usage-based pricing have actually shown increased profits for service providers, see Stahl et al. (1998)). Also, flat-rates allow for more profitable marketing strategies, such as bundling strategies (selling combinations of goods at a single price) and bulk-buy offers. An example of this would be the current offer of Telstra Australia for international phone calls in blocks of half an hour at a reduced price.

Parallels with telephony are always made in the argument between flat-rate and usage-based pricing. However, data network pricing is different. To make a call on a telephone takes time, so even if the telephone was only priced with a flat monthly fee, a user would still not spend all day making calls. Odlyzko (1997, p. 13), on discussing flat-rate charging for local calls in the US, says that average households make only about five local calls a day of about four minutes each, even though making more calls would cost no more. However, a computer could be connected to the Internet all day and be using bandwidth without human intervention. Without usage-based pricing there is no incentive not to do this, even if the content being downloaded is of little interest.

4. WHO PAYS?

4.1 Sender or Receiver?

A major issue in implementing any usage-based pricing scheme is who pays, the sender or the receiver? This has two aspects, one is the mechanics of indicating who pays, which can introduce further complications in an already complicated scenario. The various methods will be discussed with each alternative billing solution.

The second aspect is who should pay? In the case of surfing the Internet, a user visiting different sites will cause content to be downloaded to their computer simply by viewing a page. This could include things like graphics, audio or video clips. Parenteau and Rishe (1997, p. 98) suggest that the user should naturally be billed for this traffic as they are the one who initiated the transfer However, in these days of commercialism on the Internet, why should a user pay for the transfer of advertising material, simply because an organisation chooses to put it on their Web page? Surely the organisation that placed that material on the Internet should pay, as they are the ones who benefit from it. On the other hand, making everyone who places content on the Internet pay for its transfer will severely restrict amount of information available on the Internet. It would transform the Internet into a purely advertising domain. However, a case where it would be appropriate to make a user pay for downloading files from the Internet is where a link to a file is placed on a page that a user has the option of downloading, for example a demonstration version of a game. It may prove a difficult issue to decide who is benefiting from a data transfer, then, if that can be resolved, there still remains the complex issue of how to indicate the willingness to pay.

Deciding who will pay may involve some form of negotiation with a request for payment being made and a corresponding acceptance or refusal. There may also be cases where the sender and receiver should share the cost of transfer, which may also need some form of negotiation. This implies the introduction of new protocols and

increased complexity, which may prove difficult to implement. The complexities increase further when multicasting is considered.

4.2 What About Multicasting?

Multicasting brings up the issue of how to share the cost of a single data flow that is shared among many receivers. In general it would be appropriate to bill the receivers because they initiate the joining of a multicast group. (However, there could be cases where a sender-receiver, cost splitting would be appropriate). The question is how to share the cost amongst a group of receivers that could be potentially very large and also be changing in size dynamically as members join and leave. Herzog, Shenker and Estrin (1997) put forward accounting mechanisms and policies that can be used to determine these cost shares on a finely grained scale. However, what is not addressed in this proposal is how to reduce the costs of the first few members who join, as the cost will be high when it is divided amongst a small number of receivers. Shenker et al. (1996, p. 199) suggest a potential member could put a cap on the amount they are willing to pay to limit their exposure to cost to avoid this. A problem with this though, as Shenker et al. also point out, is that it could lead to receivers getting a 'free-ride' on other receivers who have joined a group using a low price.

The proposals of Herzog et al. (1997) use cost sharing approaches that depend on the number of receivers downstream of each link involved in a multicast session. Service providers who have a large number of receivers in their network would therefore be charged a larger proportion of the total cost of the multicast than service providers having less receivers. For the cost sharing to work, service providers would have to reveal the true number of receivers on their network. As Shenker et al. (1996, p. 199) note, the problem here is that there is no incentive for providers to reveal this number as it would expose them to greater costs.

Many of the issues of multicasting as well as those mentioned in previous sections remain unresolved to date. There are numerous proposals for usage-based pricing mechanisms that attempt to resolve some of these problems. A sample of these are discussed in the following sections along with further elaboration of the issues and the way they may be resolved.

5. ALTERNATIVE USAGE-BASED PRICING SCHEMES

5.1 Volume Pricing

One of the easiest and simplest methods of usage-based pricing is charging simply on the total number of bytes or packets that pass through a subscriber's interface, known as volume pricing. This could be seen as fairer than flat-rate pricing, as a heavy user would be paying costs which are more in proportion to their use. This was shown in an INDEX experiment using volume pricing:

> The three heaviest users account for about 3GB of data and 35 percent of total expenditures in the actual experiment. Under a flat-rate tariff, however, these three users would only account for ... 5 percent of the total expenditures. (Edell & Varaiya, 1999, p. 13)

Furthermore, the experiment found that the heaviest 30 percent of users would pay less under a flat-rate scheme and the remaining 70 percent of light users would pay more. This means that volume pricing would avoid the case of light users subsidising heavy users.

Volume pricing has already been successfully implemented in New Zealand (NZ) as described in Brownlee (1994). This provided a charging scheme for the shared use of a common Internet link by NZ universities. The goals of the scheme were to:

- Measure traffic in both directions through NZGate (NZ's link to the Internet) for each participating site and charge for it by volume: i.e. for the number of Megabytes moved in and out each month.
- Charge enough to cover actual costs, plus a percentage for development
- Use the resulting development funds to buy more capacity as demand grows (Brownlee, 1994, p. 1)

According to Brownlee (1994, p. 6) the success of this scheme was demonstrated by subsequent substantial upgrades that were funded by the generated revenues. He also adds that "users have always perceived that their payments were closely related to the benefits they derived". This perception being enhanced by the improvement in service given by the network upgrades.

To provide predictability of costs, they introduced 'committed traffic volume' per month. The price per Megabyte would decrease as the volume purchased increased. A site would commit to paying for a certain volume per month, then if actual traffic volumes fell outside the committed amount for more than a month, the committed volume would be changed to the actual rate. This cost predictability was another factor in the scheme's success.

Congestion was taken into account on a coarse scale by discounting on a time of day basis to encourage off-peak usage. This scheme introduces overheads, as traffic peaks have to be monitored consistently to keep up with shifts in peak usage. Shifts in usage patterns can actually be caused by introducing time of day discounts, as users move to a more economical off-peak time, creating a new peak. (The measurement of the shifting of traffic peaks with time-of-day pricing will be a subject of a future INDEX experiment (Edell & Varaiya, 1999, p. 4)). According to Gupta et al. (1999) this traffic pattern monitoring with the subsequent updating of prices to match, along with the accounting of individual packets introduces similar overheads to more optimal pricing proposals that relate the cost of transmission directly to the cost of sending a packet. This reduces the attractiveness of volume pricing. However, although Brownlee (1994, p. 6) says that "The overheads of charging are significant" he goes on to say "The benefits provided by charging are, however, well worth their cost to us".

Schnizlein (1998) is not convinced that volume pricing is an answer and in addition to the concerns of high overheads he describes a 'perverse effect' of volume pricing that could encourage network use in times of congestion:

> TCP will increase its rate until congestion signals the limit of capacity. Total traffic moves more quickly when the network is unloaded - at no effective cost - which increases charges per unit of time. Users may prefer to use the network during peak periods when traffic is slower and it is easier to limit the volume of their traffic (p. 54).

The result is that users may not be encouraged to use the network at off-peak times even though this would improve network efficiency and improve response times.

Volume pricing places a value on network traffic. Additionally, different types of traffic can be given different values. For example, in NZ, low priority traffic such as email is discounted at 30% and the full rate charged for high priority traffic such as TELNET. This can give more efficiency by reducing waste as well as providing a fairer system that charges less for traffic that is less demanding. However, it doesn't take into account the value a user places on their traffic. The high priority traffic will still be charged the full rate at any time (within the time of day block). Email will still be treated the same no matter how urgent it is. The only choice for the user is to send or not send. Volume pricing, then, does provide some management of resource usage, mainly by reducing waste, and by spreading usage on a coarse scale using time of day discounts. In summary there are:

Advantages:

- It is fairer than flat-rate pricing. Additionally, users feel that payments are directly related to the benefits derived.
- A value is placed on the traffic generated, and even though this is not linked to the social costs of congestion, it does provide some benefits in controlling wasteful usage.

- It allows for bulk selling of volume, with its associated benefits of cost predictability and discounts.
- It is simple to implement, with the software to perform the accounting already available and in use.
- It has been tried before and proven successful.
- It provides indicators and revenue for expansion.
- It is not built into the network architecture.
- No changes are required to protocols or routers.
- It does not assume the Internet is a homogeneous environment.

Disadvantages:

- There is no provision for different levels of QoS.
- It does not take into account fluctuations in congestion levels and so is unaware of the social costs of a demand for service.
- Apart from choosing a more economical time of day, the only choice for the user to economise is not send a packet.
- Time of day usage patterns have to be continually monitored to provide time of day discounts.
- There are high overheads in traffic accounting and charging.

Although there are some very attractive benefits to Volume Pricing, that at least in NZ outweigh the disadvantages, it does not provide the finely-grained control promised by 'optimal pricing' schemes. 'Optimal' schemes link prices directly to the social cost of congestion and allow users to decide the value they will place on a service. This is becoming of increasing importance with the growing need to introduce some form of differential service to the Internet. Two of these 'optimal pricing' schemes of note are the 'Smart Market' proposed by Mackie-Mason and Varian (1995), and 'Priority Classes' proposed by Gupta et al. (1996).

5.2 Smart Market

According to MacKie-Mason and Varian (1995, p. 1) "A communication network is as good or as bad as its users perceive it to be. Network performance should therefore be measured in terms of overall user satisfaction." This philosophy led them to propose the 'Smart Market'. This scheme is based on the idea that a user's demands on the network affect other users and their satisfaction. That is, the traffic generated by one user will consume a certain amount of resources on the network, and so degrade the performance of the network for other users. In times of heavy traffic the network will become congested and display a noticeable degradation of performance for all users. The cost to a user's performance, then, can be directly related to the congestion that another user's demands cause on the network. MacKie-Mason and Varian contend that to achieve optimal efficiency, usage-based charges must equal this cost of congestion caused by a user's actions.

The Smart Market is based on a generalised Vickrey auction, which is:

> a well known scheme for assigning a good to the agent who places the highest value on it, when individual valuations are private information. The idea is to solicit bids and award the good to the highest bidder, but charge the second highest bid as the price. Bidding ones true valuation is a dominant strategy for each agent (MacKie-Mason, 1997, p. 12).

Bidding the true valuation is a well-known and important part of the Smart Market.

Using the Smart Market (see Figure 5) a user would send a packet with a bid for the price they are willing to pay in the current interval. The network gateway sorts the bids and admits, in descending order, only those it can accommodate without degrading the network performance below a certain bound. Users are charged only the maximum bid of packets not allowed in. Thus users only pay the congestion cost and get to keep all the excess value above the cut-off bid. It also means users only get service if it costs

them less than their valuation of the service. If the network is not congested, then all packets will be admitted and the cost will reduced to a minimum. During times of congestion the low bid packets would be delayed until the burst of congestion eases, with the high bid packets being given priority. The reduced delay for high bid packets would give them a higher QoS.

As mentioned, this scheme provides incentives for bidding the true value, so users won't be tempted to select an inappropriate bid to obtain a higher level of service than needed. This is so because the price a user pays is not set by the priority they set but by the bid of the first packet rejected from the network. (a proof of this is provided in MacKie-Mason (1997, p. 12)).

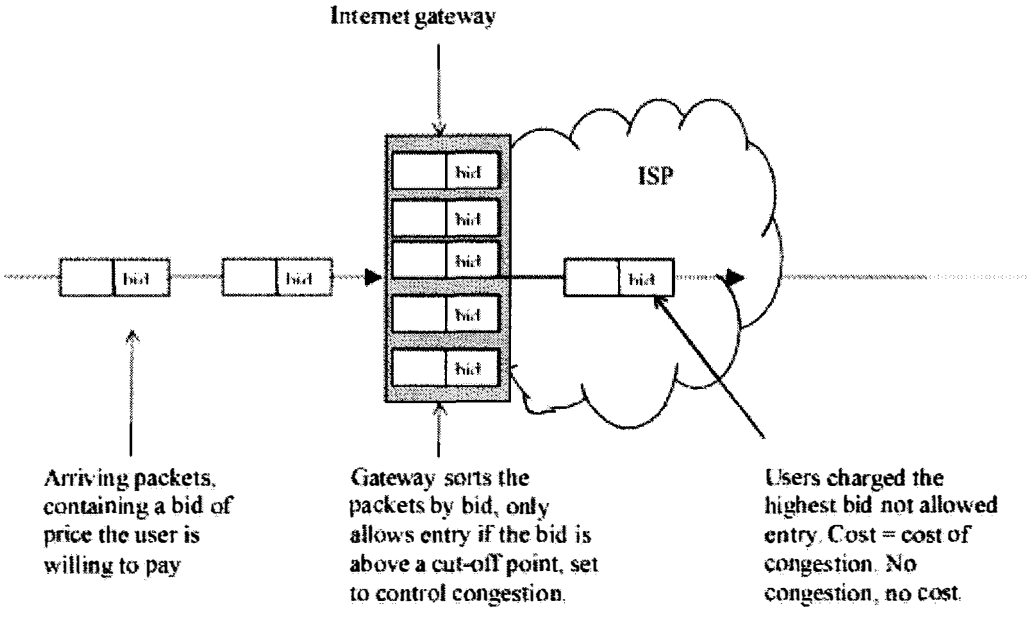

Figure 5: Smart Market

Advantages

It can be concluded that:

- Those with the highest cost of delay get served first.
- It relates a user's valuation of a service to the price as well as the QoS the user application receives. This puts the user in control by enabling the level of QoS to be selected based on how much the user is willing to pay, avoiding the problem of linking QoS directly to the application type.
- Prices are directly linked to the 'true' cost of sending a packet. This is a positive influence on resource management and economics by spreading demand to less congested times and leaving the network to those willing to pay the price of sending packets during congestion.
- It provides indicators for capacity expansion of the networks because the price for network usage is set to the value of the packets not admitted to the network. If the cumulative value of those rejected packets is greater than the cost of expanding the network, then it is appropriate to go ahead with that expansion.
- It provides a certain predictability in charges, because the actual price paid will always be equal to or lower than the bid.
- It adjusts prices dynamically, with the price automatically reduced to zero at times of low congestion. When there is congestion the price will increase, but never more than the bid price.

Economically, the Smart Market is a reasonable proposal, but there are concerns over technical implementation issues. The Smart Market, as well as other schemes such as that proposed by Gupta et al. (1999) (described below), involve computing the cost of congestion and using this to create an optimal balance in network usage that contributes to the common welfare of users. Shenker et al. (1996, p. 188) contend that, apart from possibly not being cost effective, it is questionable whether such a mechanism is actually implementable because the costs of congestion are extremely difficult to

calculate and fundamentally unknowable. They give several reasons, which are included here in the disadvantages of the Smart Market:

Disadvantages:

- Accurate bids cannot be submitted - Users have to submit a bid based on their valuation of the service. A losing bid on a packet will cause some unknown amount of delay rather than a complete loss of the service. This is because the rejected packet will be retransmitted at a later time. This means the bid must reflect the utility loss caused by the delay, rather than the valuation of the service itself. So the delay associated with each bid level needs to be known to make an accurate bid, and this delay is unknown.
- There is a complex relationship between the fate of packets caused by congestion and the resulting change in a user's utility. Some applications are very sensitive to delay or dropping of packets and some are not. An optimal pricing scheme would have to take these different delay and drop sensitivities into account
- Advances in technology can quickly change the relationship between packet loss or delay and application utility. For example, advances in congestion control could decrease an application's sensitivity to packet loss.
- It also assumes the Internet to be monolithic, which it is not. When a user bids a price for a message, how is that apportioned over multiple networks that forward the message? Parenteau and Rishe (1997, p. 94) point out that even in a single network packets may traverse many nodes and it's not clear what happens when a packet's bid price enables it to be forwarded from one or more nodes, but then delayed indefinitely in a higher priced node. To avoid this, the bidding method would have to be extended to evaluate the entire path and "this entails a distributed multiple good auction of daunting complexity" (Shenker et al., 1996, p. 189).
- The queuing and sorting of bids at the gateways would introduce delays in transmission.

- Most applications involve a sequence of packets and the effect on utility of delaying or dropping one packet depends on the treatment given to the other packets. For example, almost all the packets in a transmission may have been transmitted without incurring significant charges, but the last crucial few may be dropped, leaving the user wishing that the bids had been concentrated on the last packets. Alternatively, a few crucial packets may get dropped first, but after other packets had started their journey, effectively wasting the bids on the later, now worthless packets.
- It will be unsatisfactory to users because, considering the above reasons, it will be impossible to predict how much it will cost to send any single packet.

In light of these concerns, Shenker et al. (1996, p. 191) suggest that "It is important to allow prices to be based on some approximation of congestion costs, but it is important to not force them to be equal to those congestion costs." They propose that no pricing policy should be embedded in the network architecture, to allow different providers a choice in how they charge. This gives rise to their scheme of Edge Pricing.

5.3 Edge Pricing

Edge Pricing describes the location at which pricing occurs, which is on the edge, or outer border of the network. This means traffic accounting and charging is done only at the points of entry and exit to a network. Each service provider then becomes responsible for allocating the cost of packets crossing their network, and no more. There is no need to calculate costs on an end-to-end basis in the case of packets having to transit several intermediate networks to reach their destination. Edge Pricing also avoids having to impose a homogeneous pricing system on the heterogeneous make-up of the Internet.

Shenker et al. (1996) put forward a model for Edge Pricing that is an alternative to responsive pricing schemes such as MacKie-Mason and Varian's (1995) Smart Market

and Gupta et al's. (1996) Priority Classes. Due to Shenker et al's. (1996, p. 192) claims that the true congestion costs, which responsive schemes rely on, are inaccessible (see the disadvantages of the Smart Market above), Shenker et al. propose two approximations which can be used instead to place a price on various offered service classes:

1. Use the expected congestion conditions at a particular time of day. The time of day is related to expected congestion conditions within the realm of a particular network or service provider. Different time-zones with the different congestion patterns are not an issue in Edge Pricing as each network is only responsible for packets entering and leaving its sphere of influence.
2. Replace the cost of the actual path with the cost of the expected path through the network, with the charge only dependent on the source and destination of the flow and not on the route. In the case of a packet crossing an intermediate network the path from source to destination in that network would trace the route from the entry point to the exit point.

Using these two approximations the price can be worked out before a transfer occurs because you know the expected congestion along an expected path. This means the price can be calculated locally at the access point, or edge, of the network where the user's packets enter. Where networks are interconnected, the network providers purchase service from each other, the same way that users purchase service. If a packet has to cross two networks to travel from user to destination, the first network provider, using its own pricing policy, can calculate the total cost the packet will incur from the entry point to the exit point of its network and bill the user. When the packet enters the second network it can then bill the first based on its own pricing policy. Each provider takes full responsibility for each packet entering its network (see Figure 6). Bilateral agreements between connected providers provide the necessary cost shifting, these

agreements being based on aggregate usage and not on individual flows. This simplifies costing and enables different service providers to develop their own costing strategies and offer things like special deals, such as bulk discounts, as found among the various telephone service providers.

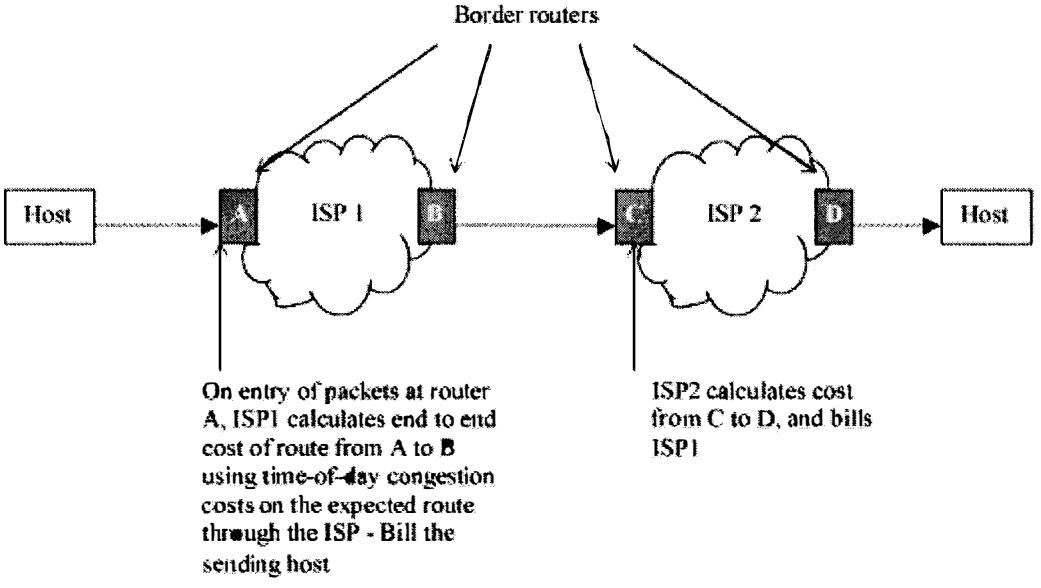

Figure 6: Edge Pricing

The above describes the case where the sender of the traffic is billed, however, there are many cases where the receiver should be billed. A lot of Internet traffic involves users accessing servers and downloading material. If the servers had to pay for this traffic, they would soon disappear. To be able to bill the receivers of the traffic, the billing must occur in the opposite direction. For this to happen there must be some way for the receiver to indicate a willingness to pay, for example, a new control message generated by the receiver indicating this willingness. Then the packets have to be charged according to the receiver's contract with its provider, rather than the sender's contract. This means the packet, on entry to a network, must be charged to the next hop, instead of the previous one.

Using the first approximation of expected congestion conditions at a time of day is

similar to telephone networks where time-of-day pricing encourages users to shift their usage to more economical times. Unlike responsive pricing schemes, this doesn't take into account any instantaneous fluctuations in congestion. Packets sent with a high QoS would still be charged the high price during a lull in network traffic. However, Shenker et al. (1996, p. 192) claim that this is not an issue because the onus can be moved back to the user to monitor the fluctuations in congestion. A lower service class will give equal performance to a higher service class in times of low congestion. So instead of changing the prices, the requested QoS can be changed by the user. This means that the network doesn't reduce costs in times of lower congestion, but the user selects a lower service class and is charged less for that service. Shenker et al. (1996, p. 192) suggest the user in this case, can be automated in the form of adaptation routines embedded in the application, allowing rapid and complex modifications of service class. Moving the onus to the user keeps the network simple and moves the responsibility to adapt to current network conditions outside the network, which is inline with the present Internet design philosophy. Shenker et al. (1996, p. 193) make the point that it would also seem sensible to make the applications themselves, with their varying sensitivities to network conditions, responsible for their own behaviour, giving the reason that "it seems preferable to place the bulk of the variability where it can be done in the most informed way."

The second approximation of using the expected path, rather than the actual path is proposed by Shenker et al. to take the uncertainty out of setting prices and also to provide a fairer pricing system:

> Having the price of the service depend on the network's decision about routing seems an unnecessary source of price variation that makes it harder for the user to make informed plans about network use. Moreover, when alternate paths are taken by the network in response to congestion, the extra cost due to the congestion should not necessarily fall only on those flows that have been redirected. (Shenker et al., 1996, p. 193)

Shenker et al's model, using the two approximations, is a specific example of a general concept. Edge pricing in general only describes the point where the charges are calculated, not how they are calculated. Providers are free to choose a pricing scheme, or mixture of schemes within their network. This could be usage-based, flat-rate or capacity-based charging. This gives rise to the main advantage of Edge Pricing, described in the first point below.

Advantages:

- The pricing policy is not built into the network architecture. This gives freedom to individual providers to be more innovative in designing their own policies, allowing a natural evolution or a 'natural selection' to occur with pricing policies, leading to the development of the best scheme for a particular network.
- Prices are more predictable than in responsive pricing schemes. This would be a positive factor to subscribers as it takes some of the uncertainty out of budgeting network costs (this applies to Shenker et al's model described above).
- Complexity is moved to the edges of the network away from the performance sensitive core.

Some of the disadvantages result mainly from unresolved issues in Shenker et al's proposal.

Disadvantages:

- It does not adjust prices to instantaneous fluctuations in congestion and so does not have the ability to make adjustments to bursts of traffic.
- There would be overheads in monitoring congestion to keep the approximation of expected congestion conditions at the time of day correct.
- Similarly, the expected path would have to be monitored to update it in accordance with network changes.

- In the case of the receiver paying, they mention the receiver can indicate a willingness to pay but not how a maximum limit can be put on how much they are willing to pay.
- It is undefined what will happen if a user has reached the maximum they are willing to pay and refuses to accept more packets. In this case, how do you treat packets already transported, possibly across several networks?
- There is no explanation of how the receiver and sender could share the cost of transfer.
- There is also the issue of charging in multicasting. (described above in 4.2 What About Multicasting?)
- Trying to resolve the issues mentioned in the previous points could lead to overwhelming complexity.
- Although individual network providers are free to choose their own pricing policy, Edge Pricing does require all networks to participate in the general concept to allow the necessary cost shifting between networks.

The proposal for edge pricing is a preliminary one and Shenker et al. acknowledge these unresolved issues. However, Edge Pricing is mainly given by Shenker et al. as an alternative to externality pricing schemes with the idea to "initiate a dialog about such pricing schemes and hopefully stimulate the creation of other pricing paradigms" (Shenker et al., 1996, p. 200)

Another scheme, similar to Edge Pricing, is the Metro Card scheme which is proposed by Fang (1996) and described in the following section.

5.4 Metro Card

As described in Shenker et al's (1996) version of Edge Pricing, the Metro Card performs its accounting at the borders of a network. The difference here though, is that each

network has to participate in the same scheme to facilitate the accounting. The basic idea is that a packet will accrue tolls as it crosses border routers of a network. These tolls will accumulate on the packet's journey to its destination. On arrival a notification is then sent back to the originator, in the form of a bill packet that contains the value of tolls accumulated on the total journey. The tolls are dynamically adjusted to reflect the congestion and delay on a particular route. Various routes will be available for a packet to take, the route chosen will depend on the toll willing to be paid for the packet, the service level required and the congestion currently existing on those routes.

In the case of packets crossing two different networks, represented by ISP1 and ISP2, before reaching its destination, ISP2 would charge the tolls to ISP1. ISP2 does not have to know who originated the packet. ISP1 would then recoup this cost by charging the originator of the packet. The similarity to Edge Pricing is that costs are resolved only between bordering networks.

There are three components to the Metro Card system

1. An accounting field in the IP header: This contains four elements - Service_Level, Accu_Tolls, Budget and Max_Charge.
   - Service_Level indicates the level of service the packet desires.
   - Accu_Tolls is the accumulated tolls incurred within a particular network. This field is subtracted from the budget and then set to zero whenever the packet enters a new network.
   - Budget shows how much value remains. As the packet travels this is decremented by the tolls accumulated.
   - Max_Charge is the maximum the packet is willing to pay for its travel. This is not changed during travel and the Budget is initially set to Max_Charge.
2. The tolls: Tolls are calculated and stored by the routers. There are different tolls for different routes and a packet will have various routes it can take to a destination.

The tolls are dynamically adjusted to reflect the congestion existing on the routes. Routing tables will have to be enlarged to accommodate multiple routes for a particular destination and the delay existing on that route. The Service_Level is used here to match the best possible route to the level of service desired.

3. Accounting: Accounting tables are maintained in the border routers. These tables are used to record the tolls incurred by a packet within the network. This information can then be used to bill bordering networks for traffic. Included here are the bill packets, which complete the accounting cycle by sending the total cost of sending the packet back to the originator. These bill packets are treated by routers as a special case and do not incur tolls.

Advantages:

- This is a responsive pricing scheme that adjusts dynamically to congestion and so would provide benefits to resource usage balancing.
- The issue of who pays (receiver, sender or cost sharing) is not built into the scheme and so is left in the hands of the users or user applications. Fang (1996, p. 107) citing the advantage as being that the participants in a transmission have better knowledge of who is the beneficiary.
- It allows an explicit statement of the desired QoS to be indicated.

This proposal suffers from many of the problems of other schemes mentioned in this document.

Disadvantages:

- There would be the same difficulties as the Smart Market with calculating the costs of congestion. Therefore, there would be problems in setting optimal tolls and service levels.
- It is uncertain what would happen if a packet does not have enough budget to reach

its destination. This is similar to the problems of the Smart Market mentioned by Parenteau and Rishe (1997, p. 94).

- There is unpredictability in the charges. Setting the budget high does not provide a guarantee that the packets will arrive, or receive a certain level of QoS. This is because the accumulated charges are dependent on the congestion encountered and the level of congestion is unknown at the point of sending a packet.
- It has the drawback of all schemes in that the accounting of individual packets my cause large overheads and be difficult and expensive to implement.
- The implementation of this scheme requires changes to the IP protocol and IP header, which is a barrier to its introduction.
- Similarly significant changes are required in routing software.
- The Metro Card assumes a homogeneous network. For example, bill packets have to be treated in the same way by all intervening networks. What would happen if an intervening network had a policy of charging for all packets, regardless of type?
- It remains undefined as to what would happen if bill packets were lost. This introduces a point of weakness to the whole system.

The Metro Card, as do the other schemes mentioned, suffer because of the complexities they may introduce as well as the fact that they also may destroy the simplicity of the current service of the Internet that has proven so successful to date. A scheme that side-steps this complexity and also maintains a form similar to the current Internet is the Paris Metro Pricing scheme.

5.5 Paris Metro Pricing

Odlyzko (1997) presents an idea inspired by the Paris Metro rail system that is still used successfully today in some regions. 1st and 2nd class rail cars are provided that are identical in number and quality of seats. The only difference being that 1st class tickets cost twice as much. This leads to the obvious result that 1st class cars are less crowded

than 2nd class cars. People choosing 1st class do so on the basis that they can sit down, and only the people willing to pay for this privilege do so. This is a self regulating system in that if 1st class become too crowded, people will be less willing to pay the extra and move to 2nd class, so reducing the congestion in 1st class and restoring its quality of service.

Paris Metro Pricing (PMP) is a simple pricing scheme to control congestion on networks. The main network would be partitioned into several logically separated channels. Each channel would have a fixed portion of the capacity of the entire network. The channels would operate in the same best effort manner of the present Internet with all packets on a particular channel treated equally. The only difference between the channels would be price. The idea being that the higher priced channel would be less congested and so offer a higher QoS (see Figure 7). Odlyzko (1997, p. 5) argues that the QoS offered by a lightly loaded best effort service would be satisfactory for most needs. This being demonstrated by the acceptable performance of the present day Internet at uncongested times, such as in the early morning. Schnizlein (1998, p. 52) also holds this veiw, saying that the service on the Internet is quite good, only being degraded when it is under load.

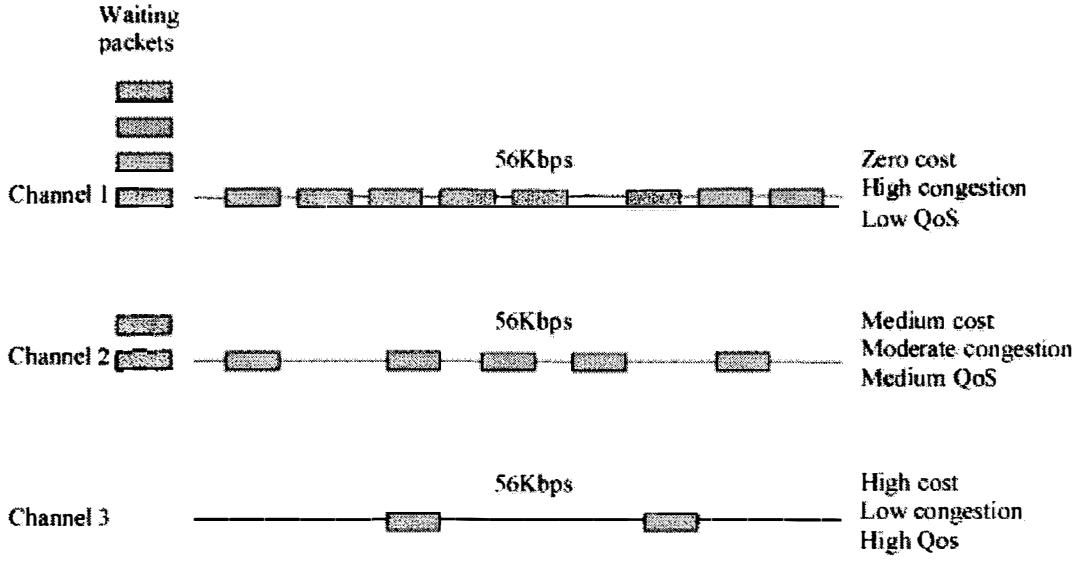

Figure 7: Paris Metro Pricing

This scheme could be implemented using IPv4. The currently unused 3-bit priority field could be used to indicate the channel. Changes would have to be made to router software to maintain logically separate queues or to give appropriate priority to packets belonging to different channels. The major change would be the necessary introduction of hardware or software to count the packets for each user. This, as in all usage-sensitive pricing schemes, is a complicating factor. A mitigating aspect of PMP though, is that this accounting could be done at the edge of the network, rather than in the performance sensitive core. The accounting could also be simplified using sampling to reduce the overhead of counting every single packet.

Advantages:

- It is simple, there is none of the complexity of things like the bidding procedures of the Smart Market, and the need to dynamically calculate congestion costs. The pricing is constant and easily understood. (Optimally, the charges for the channels should remain fairly constant to maintain some predictability and allow the self-regulation of congestion to take effect as described in the example of the Paris train system).
- It can be implemented using the current Internet protocols. This would reduce the impact of introducing the new system and it could be implemented without having to wait for the implementation of something like IPv6.
- The general consumer preference for flat-pricing can be accommodated by selling large blocks of transmission capacity. For example, selling 100Mb a week on a low quality channel, or 60Mb on a higher quality channel. This is also preferable for service providers as it allows for a more predictable income.
- Reduces traffic management tasks by inducing users to separate themselves into classes. In this way congestion control is achieved virtually for free as far as the network is concerned, the users manage themselves.
- The lowest QoS channel can be left as a free service, to users this will appear the

same as the current Internet and will ease the transition to usage-based pricing.

Disadvantages:

- With best effort service there would be occasional service degradations, even on the higher QoS channels. "For PMP to work, the performance of the different networks has to be predictable, at least on average. Unfortunately the fractal nature of data traffic means that we have to expect that all PMP channels will experience sporadic congestion." (Odlyzko, 1997, p. 8)
- The service degradation could upset the whole system. For example, if the lowest channel is congested for a long time, delays would cause users to move to a higher priced channel. The higher priced channel would then become congested, degrading its QoS. Odlyzko (1997, p. 8) suggests that providing a high price barrier between channels would discourage users switching too readily to a higher QoS channel. He suggests this could be done by only selling the capacity to send large blocks of packets, rather than single packets.
- To maintain a high level of QoS on the premium channels the load would have to be low. It is difficult to ascertain whether the capacity utilisation would be so low as to make it unprofitable (Odlyzko, 1997, p. 6)
- It is difficult to set the prices and capacities of the separate networks. Odlyzko (1997, p. 8) says that these could be set by taking note of customer surveys and customer complaints. Also, time of day variations in traffic patterns could be used to adjust prices and capacities to reach an optimal balance.
- PMP has the same disadvantage of other schemes, that is the accounting overhead. The problems involved in tracking usage on a packet level and resolving ownership of those packets down to a user level could be overwhelming.

PMP is not designed to be an optimal solution to pricing on the Internet. It is put forward as a simple alternative that would provide some form of congestion control as

well as a form of differential service quality, while at the same time maintaining some of the characteristics of the current Internet. There are researchers, however, who are investigating 'optimal' pricing schemes that provide a more accurate and dynamic pricing mechanism that is very sensitive to changes in congestion and provides truly differential QoS. One such scheme is Priority Classes, proposed by (Gupta et al., 1996).

5.6 Priority Classes

Gupta et al. (1996) propose a pricing scheme they call Priority Classes. They define a network as a series of interconnected servers and clients. The servers provide various services and the clients make demands on those services in the form of service requests. The demands on each of the servers are made through a priority queue system that is attached to each of the servers. These queues provide the various priority classes. A price and expected waiting time are associated with each priority class at each server. A client can use the expected waiting time and price to select a priority class and server that will minimise costs and provide a desired level of service.

The process of selection of a server and priority class is:

> Upon the arrival of a service request, the type of service required is identified (a service is characterized by the amount of computational cycles required at a server). Then, the current estimates of prices and predicted waiting times are obtained for all the servers offering the particular service. The user then evaluates the total expected cost of this service in terms of her delay cost and the service cost against her value of the service. If the total cost of the service is higher than her value for the service, the user quits the system; otherwise, she submits the request for obtaining the service. (Gupta et al., 1995b, p. 8)

The user (or a software tool such as a 'Smart Agent' working on behalf of the user) can select the QoS based on how much they value the service. This is of benefit to the users as they can request services based on their valuation of that service. Additionally, it discourages misuse by providing incentives for the appropriate selection of QoS. Load

is also distributed to less congested nodes through the process of selecting the server that offers the best service for the price. The benefits to the network service providers are that they can monitor the loads at different servers and set prices according to the load imposed by those servers on the backbone. Furthermore, the demand experienced by the different service classes at a server can give indicators as to whether a particular service is really needed, or if expansion is necessary.

The prices are calculated and adjusted dynamically at each node for each priority class. The dynamic computation of the prices is carried out using "approximations of performance parameters estimated based on short-term historical data collected at individual network nodes" (Gupta et al., 1999, p. 59). This provides a pricing policy that is sensitive to instantaneous fluctuations in demand, and so is an attempt at an 'optimal' pricing scheme that directly relates a users impact on the network to prices.

They derive formula for 'optimal prices' by priority class that "maximize total welfare of all users" (Gupta et al., 1996, p. 73). (Evidence of the optimality and benefits was shown in a simulation, the results of which are shown in Figure 3). The formula adjust prices so as "to optimise the trade-off between greater throughput volume and longer waiting times" and so that "aggregated user demands don't exceed optimal levels and waiting-time expectations are correct" (Gupta et al., 1996, p. 73).

The optimal prices depend on the traffic flow at the site, the size of the packets, the priority class, and the social cost of time. Shenker et al. (1996) express doubts about calculating the latter, the social cost of time, pointing out that knowing the utility loss as a result of service degradation due to delay is fundamentally unknowable. In addition:

> the problem of denial of service leading to some delay, rather than an eternal denial of service, makes the valuations of the flows not directly related to congestion costs. Consequently, determining optimality in the presence of fluctuating demand is extremely difficult. (Shenker et al., 1996, p. 7).

However, Gupta et al. (1999, p. 61) claim to have shown that "reasonable estimates of users' delay based utility loss can be computed by a Baysian computational approach based on current prices and observable user actions". In this way Gupta et al. reduce the difficulties of determining optimality by using "reasonable estimates" to arrive at "approximations of 'optimal' prices" (Gupta et al., 1999, p. 60)

(A Baysian approach is a form of inference mechanism. It includes the fact that we have some knowledge about a process being investigated before obtaining the data in the inferencing process. This prior knowledge has some influence on the conclusions drawn from that data. Inferencing should therefore be based on the prior knowledge as well as the data. Baysian inference is the mechanism for drawing inference from this combined knowledge (Coles, 1999).)

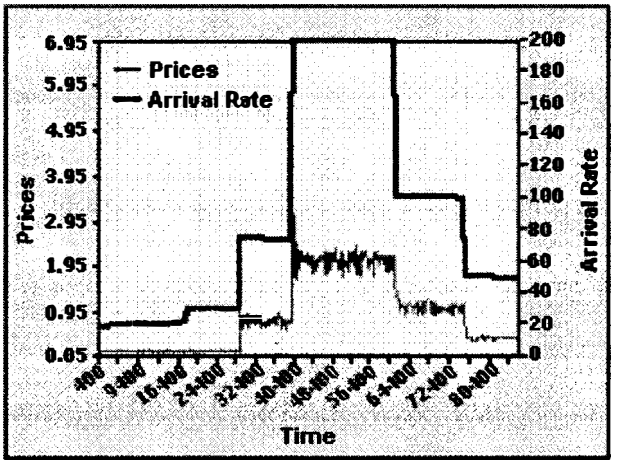

Figure 8: Prices with changing exogenous demand (Gupta et al., 1999, p. 60)

Gupta et al. (1999) demonstrated, by means of simulation, that their dynamic pricing mechanism responds well to fluctuations in demand. Figure 8 displays the results. It shows the demand pattern for a day covering times when the network is both under-utilised and over-utilised. The price fluctuations are superimposed on this. The left y-axis represents the prices, the right y-axis represents arrival rates of demands and the x-axis the time in seconds.

It can be seen that prices adjust quickly as the demand changes, with the price being almost zero when the network is uncongested and significantly higher when the network is very congested. This adaptation is automatic and does not require any prior knowledge of the demand characteristics of users. This is because the adaptation is based on "observable system performance" (Gupta et al., 1999, p. 60). The same behaviour was also displayed during periods of what they call 'fractal demand', which represents bursts of traffic, with the prices keeping pace with the rapidly changing demand caused by the bursts.

Accounting and Billing

Gupta, Stahl, & Whinston (1995a, p. 12) suggest that accounting and billing could be carried out in two ways. First, each server could meter charges at its location and periodically send a bill to the client machine which processes the charges to produce a monthly bill. Second, a service request could include a bill portion so that when a service request is made the server records the charges and when the request returns to the client, it contains a complete bill.

The user would not receive a bill from each node and link of the network, but would rather receive one bill from their access provider for their use of a server. The access provider would be charged for its connection to the next link in an identical manner to Edge Pricing, with cost shifting occurring between intervening networks as described above. It would be the responsibility of the first access provider in the chain to recover the total cost to itself from the user who originated the service request.

Advantages:

- Prices are directly related to congestion and so encourage spreading network usage over time and to less congested nodes.
- Dynamic price adjustment takes into account fluctuating demand and so prices

reflect the 'true' cost to the network.

- Because price adjustment is dynamic and based on demand, there is no need to continuously monitor changing demand patterns to update prices.
- Approximations are used that simplify the calculation of 'optimal' prices.
- It offers various priorities in the form of 'Priority Classes'. This allows users to select different levels of QoS based on the application and the subjective valuation of the service.
- Computation of prices is decentralised and requires no network-wide information to compute prices at a particular node.
- It takes advantage of the processing power of individual nodes, spreading the work and thus the overheads of computation.
- The profitability of each server can be used to guide investment decisions for expansion.

Disadvantages:

- It could lead to starvation for low priority classes. This is expressed by Odlyzko (1997, p. 15): "low priority classes could fail to get any bandwidth at all if enough traffic from higher priority classes show up".
- Although computation of prices is distributed the scheme does have substantial overhead. It requires "collecting and processing extensive information about the network" (Odlyzko, 1997, p. 15).
- The issues of resolving who pays and how this is indicated would introduce further overheads.
- As in other schemes, traffic accounting would introduce significant overheads.
- Individual nodes on the networks would have to 'know' about the price adjusting computations and the priority classes, so this scheme requires a uniform deployment across the whole network.
- The process of searching for a server with acceptable prices and waiting times for a

particular service class, then presenting the user with a choice, introduces delays and overheads. These delays could be significant if the number of servers offering the desired service is large. (However, this could be alleviated using 'Smart Agents'.)

- The development and deployment of the software to implement the scheme, (including the software to perform the tasks mentioned in the previous point) would be costly and have a large impact on the current Internet.

These disadvantages have to be balanced against the evidence of Gupta et al's. simulations (see Figure 3) which show large benefits of adopting their usage-based scheme. For example, if it is accepted that this scheme contributes to the general welfare then it could be assumed that low priority classes are included in these benefits. The low priority classes, then, may not experience starvation. Additionally, as expressed by Brownlee (1994, p. 6) in the NZ experience, the overheads of the system may be well worth it, considering the benefits gained.

The following section explores the relationship between the overheads involved in the various pricing schemes and the 'optimality' of the pricing schemes.

5.7 Comparison of Pricing Schemes

This section provides a comparison between the pricing schemes based on two veiwpoints:

1. **Optimality:** the concept of an optimal pricing scheme as proposed by Gupta et al. (1999).
2. **Implementation issues:** such as the impact made on the network, the overheads involved and the impact on users.

5.7.1 Optimality

Gupta et al. (1999, p. 59) propose a set of characteristics that a desirable (optimal) pricing mechanism should possess:

1. Prices should encourage users to use the network when it is less congested by shifting their demands across time.
2. Prices should take into account the impact of current load on future demand.
3. Pricing should preferably be coarser than packet level pricing so that is easier and less costly to implement.
4. Prices should reflect the load status of the network nodes (routers, gateways).
5. Prices should yield effective load management by redistributing the load from highly loaded nodes to lightly loaded nodes.
6. The pricing scheme should be implemented in a completely decentralised manner, for example, by requiring performance information at an individual node to set prices at that node but not requiring any system-wide information. Otherwise, the overhead costs involved in computing the prices may negate any potential benefits of the pricing method.
7. There should be multiple priorities in order to take into account the different QoS required by different applications and users.
8. Prices should encourage the appropriate use of different levels of QoS.
9. The pricing scheme should be implemented in such a way that service providers have incentives to provide the required QoS based on the profits they derive from pricing methods.

Pricing mechanisms with these characteristics would promote efficient resource usage by spreading load on the network across different nodes as well as time. They would also provide different levels of QoS that are directly linked to prices and the user's valuation of a service. The prices would encourage appropriate use of QoS and give

incentives to service providers to invest in expansion. Points 3 and 6 include the characteristics that a pricing mechanism should be computationally viable.

Table 1 (see Appendix) provides a comparison of the various pricing proposals discussed based on Gupta et al's. proposed characteristics of a desirable pricing mechanism. Gupta et al's. 'Priority Classes' scheme is revealed as the most optimal using these characteristics. The 'Smart Market' and the 'Metro Card' schemes are also shown as being highly optimal, only lacking in that they require accounting on a packet level. Next lowest on optimality are 'Edge Pricing' and the 'Paris Metro Pricing' schemes. This is also not surprising as these were designed to achieve approximations to optimality in order to overcome the difficulties in designing a truly optimal scheme. Volume pricing and flat-rate pricing come last, with volume pricing providing more benefits that flat-rate pricing.

5.7.2 Implementation issues

Table 2 (see Appendix) shows a comparison of the pricing schemes based on other issues. It was chosen to call these 'implementation issues' as they may affect the implementation of a pricing scheme both from the user's point of view and from aspect that they may introduce complexities into the Internet. Categories 1 to 3 reflect the impact on users and categories 4 to 11 the technical implications to the network of implementing a pricing scheme. These comparisons are included to balance the conclusions that may be drawn from assessing a pricing scheme purely on whether it displays the desirable characteristics of an optimal scheme. As will be shown, the most optimal pricing schemes actually have the greatest disadvantages from an implementation point of view.

The categories used to describe 'implementation issues' are:

1. **Perverse effects:** effects such as caused by flat-rate pricing where a congested network can actually encourage network use.
2. **Fairness of prices:** whether prices are related to the social cost of network use, that is, whether prices reflect the degradation to service of other users caused by a user's service request being carried out.
3. **Predictable cost:** can users predict how much their network usage will cost them. It was shown that part of the reason users prefer flat-rates is that the costs are predictable, so this may be an important factor.
4. **Correct prices easily arrived at:** whether the parameters used to calculate the prices are actually knowable or accessible, and whether the calculation of those prices involve large overheads.
5. **Built into the architecture:** building the pricing mechanism into the architecture may be difficult due to the lack of central control. It also would not be as flexible as a mechanism that was independent of the network.
6. **Can indicate who pays:** as discussed earlier, this would introduce many complexities, and so a pricing scheme that has solved these issues would be at a great advantage.
7. **Simple:** many of the pricing proposals would introduce further complexity to the Internet and destroy many of the advantages of the present system. A complex pricing scheme would also be difficult and expensive to implement.
8. **Implementation overheads:** this is related to the point above and includes things like development of new software and protocols.
9. **Management overheads:** this could include monitoring of shifts in time of day usage patterns, adjusting prices and setting priority levels for different service classes.
10. **Assumes Internet homogeneous:** the Internet is heterogeneous and any scheme that relies on a uniform deployment over the Internet to function may have barriers to its introduction.

11. **Changes to protocols/routers needed:** a pricing scheme that requires changes to existing Internet infrastructure or protocols may experience barriers to its introduction. This would also be a further overhead in deployment of a pricing scheme.

5.7.3 Optimality vs Implementation

In order to provide a comparison between the results shown in Table 1 and those shown in Table 2 a numerical value was associated to a result based on whether it represented an advantage or a disadvantage. A graph (Figure 9) was then produced from the totals for each pricing scheme using these values. Values were attributed in the following way:

Table 1

- All categories - No = 0, Yes = 1, partial advantages = 0.5.

Table 2

- Categories 1, 5, 10, 11 - No = 1, Yes = 0, partial advantages = 0.5.
- Categories 2, 3, 4, 7 - No = 0, Yes = 1, partial advantages = 0.5.
- Categories 8, 9 - Low = 1.5, Moderate = 1, High = 0.
- Category 6 - Not included as none of the schemes really addresses this issue.

This is a crude estimate of the advantages. However, it serves the purpose of illustrating the trade-off between implementablity and achieving optimality.

The results in Figure 9 show that the more optimal pricing schemes, Smart Market, Metro Card and Priority Classes, have the least advantages concerning implementation. This mainly stems from the complexity, implementation overheads and management overheads they may introduce to the Internet.

The least optimal schemes, Flat-rate and Volume pricing show the greatest advantages concerning implementation. However, this must be balanced against the negative affects already described of having a pricing scheme that is far from optimal from the perspective of resource management.

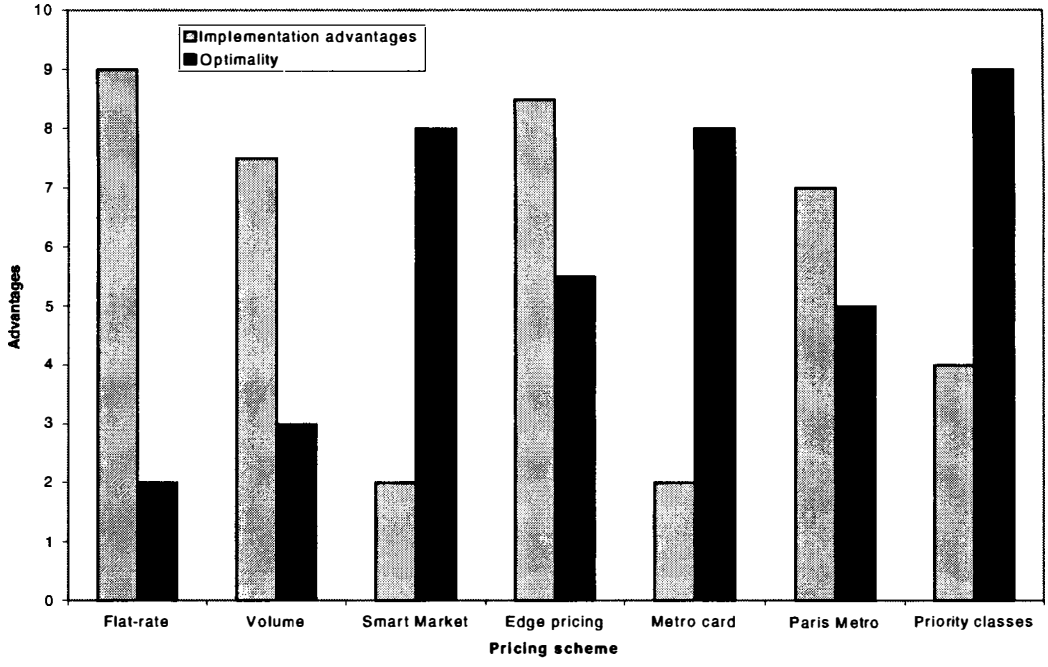

Figure 9: Comparison of Pricing Schemes

The pricing schemes that show more of a balance between optimality and implementability are Edge Pricing and Paris Metro Pricing. This result is a reflection of the acknowledgement by the designers of the problems of achieving true optimality in a pricing scheme. In Edge Pricing, Shenker et al. (1996) uses the approximations of expected congestion conditions at a time of day along an expected path to eliminate some of the difficulties. In Paris Metro Pricing, Odlyzko (1997) retains some of the benefits of the current Internet by the use of differently priced channels, avoiding the need for complex price calculations.

These results underline the fact that these pricing schemes are proposals (except for

Flat-rate and Volume pricing) and require more research and experimentation before being implemented. The research so far on optimal pricing has involved the use of simulations. This has shown many benefits of adopting an optimal pricing scheme such as Priority Classes. However, further research could be done in the form of a real-life trial, similar to the INDEX experiments, using an optimal pricing scheme. This may test the implementability of the pricing scheme and also provide some empirical evidence of the benefits.

6. CONCLUSION / SUMMARY

The Internet is growing rapidly and the demands being made on transmission capacity are increasing in volume and sophistication. This is leading to a situation where the current best-effort service and flat-rate pricing on the Internet may no longer provide the service desired by users.

Flat-rate pricing can give rise to 'perverse' effects that can cause increased use of the network in times of congestion and disincentives to invest in capacity expansion. This, combined with the fact that incremental usage of bandwidth is not charged for may lead to a 'tragedy of the commons' where the value of the Internet for both service providers and users is reduced.

The best-effort service of the Internet may not provide sufficient guarantees of QoS for delay sensitive applications such as real-time video and audio. However, it may be less than optimal to link QoS directly to the application, as different users value their network traffic differently at different times. Pricing can be used as a tool in the provision of differential QoS. The price a user is willing to pay can indicate the value the user places on a service and so the level of QoS they desire for that service. This can encourage the appropriate use of QoS and also provide indicators to service providers for capacity expansion.

There are various ways of providing differential QoS, such as ATM and RSVP. However, they have limited value in the case of HTTP traffic, which makes up the majority of traffic on the Internet. The possibly imminent introduction of IPv6 with its ability to specify QoS may mean that some form of usage-based pricing be implemented to control appropriate use of QoS.

The arguments for the introduction of usage-based pricing and differential QoS are not

clear cut, however. There are arguments that differential QoS may not be needed due to advances and falling costs in the provision of bandwidth. Additionally, the introduction of differential QoS and a usage-based pricing scheme would probably be a source of complexity and overheads. It may be better to avoid this and invest in increasing available bandwidth to keep the Internet simple. On the other hand, there are arguments that any gains in bandwidth may be quickly negated by new, more demanding applications and that the 'tragedy of the commons will make differential QoS and usage-based pricing a necessity. Adding to the uncertainty there is also contention about the presence of congestion on the Internet and the real cause of delays. Another factor is that users have a preference for flat-rate pricing and so may resist the introduction of a usage-based scheme.

There are various proposals for usage-based pricing schemes, some of which were discussed in this document. Other examples include Token Bucket (Schnizlein, 1998), Multiple Priority Queues (Parenteau & Rishe, 1997), Allocated Capacity Framework (Clark & Fang, 1998) and the use of the IP precedence field (Braun, Claffy, & Polyzos, 1993). Of the proposed schemes discussed, there seems to be a trade-off between optimality and implementability, with the most optimal scheme (Priority Classes) being the least implementable. The two schemes that take the difficulties of achieving optimality into account (Edge Pricing and PMP) seem to display more of a balance between optimality and implementability.

The problem of whether the receiver, sender or a combination of both pay and how to indicate this is unresolved in many of the proposed schemes. It can be seen as a negative aspect and potential source for added complexity in the implementation of usage-based pricing. This problem is exacerbated in the context of multicasting.

Further research could be carried out in the area of indicating who pays (receiver,

sender or a combination). This could involve some guidelines for deciding who should be responsible for paying as well as the mechanics of indicating who pays. An area for particular attention could be in resolving the issues surrounding who pays in multicasting. Another area for research could be in providing empirical evidence of the benefits of usage-based pricing using one of the proposed schemes in a real-life experiment similar to INDEX.

7. REFERENCES

Braun, H.-W., Claffy, K., & Polyzos, G. (1993). A Framework for Flow-Based Accounting on the Internet. Paper presented at the Proceedings of IEEE Singapore International Conference on Information Engineering 93: Communications and Networks for the Year 2000.

Brownlee, N. (1994). New Zealand Experiences with Network Traffic Charging, [WWW]. Available: http://www.press.umich.edu/jep/works/sourcefiles/brownlee.html [1999, 15 Oct].

Clark, D., & Fang, W. (1998). Explicit Allocation of Best-Effort Packet Delivery Service. IEEE/ACM Transactions on Networking, 6(4), 362-373.

Cocchi, R., Shenker, S., Estrin, D., & Zhang, L. (1993). Pricing in Computer Networks: Motivation, Formulation, and Example. IEEE/ACM Transactions on Networking, 1(6), 614-627.

Coffman, K., & Odlyzko, A. (1998, 2 Oct 1998). The Size and Growth Rate of the Internet, [WWW]. AT&T Labs - Research. Available: http://www.firstmonday.dk/ [1999, 13 Oct].

Coles, S. (1999). Statistical Inference: (Bayesian Inference), [WWW]. Available: http://www.maths.lancs.ac.uk/~coless/btch/btch.html [1999, 24 Oct].

Edell, R., McKeown, N., & Varaiya, P. (1995). Billing Users and Pricing for TCP. IEEE Journal on Selected Areas in Communications, 13(7), 1162-1175.

Edell, R., & Varaiya, P. (1999). Providing Internet Access: What we learn from the INDEX Trial (16 April 1999), [WWW]. Dept. of Electrical Engineering & Computer Science, University of California, Berkely. Available: http://www.INDEX.Berkeley.EDU/reports/99-010W [1999, 6 Nov].

Engel, B., & Maj, P. (1999, 3-5 Sept 1999). Towards a Quality of Service on the Internet - an Educational Case Study. Paper presented at the 3rd Baltic Regional Seminar on Engineering Education, Goteborg, Sweden.

Fang, W. (1996). Building an Accounting Infrastructure for the Internet. Paper presented at the IEEE Global Telecommunications Conference - Communications: The Key to Global Prosperity.

Fishburn, P., & Odlyzko, A. (1998). Dynamic behaviour of differential pricing and quality of service options for the Internet. Paper presented at the Proceedings of the first international conference on Information and computation ecomomies.

Gupta, A., Stahl, D., & Whinston, A. (1995a). Pricing of Services on the Internet, [WWW]. Available: http://cism.bus.utexas.edu/alok/pricing.html [1999, 16 Nov].

Gupta, A., Stahl, D., & Whinston, A. (1995b). A Priority Pricing Approach to Manage Multi-service Class Networks in Real-time, [WWW]. Available: http://www.press.umich.edu:80/jep/econTOC.html [1999, 16 Nov].

Gupta, A., Stahl, D., & Whinston, A. (1996). An Economic Approach to Network Computing with Priority Classes. Journal of Organizational Computing and Electronic Commerce, 6(1), 71-95.

Gupta, A., Stahl, D., & Whinston, A. (1997). A Stochastic Equillibrium Model of Internet Pricing. Journal of Economic Dynamics and Control, 21, 697-722.

Gupta, A., Stahl, D., & Whinston, A. (1999). The Economics of Network Management. Communications of the ACM, 42(9), 57-63.

Herzog, S., Shenker, S., & Estrin, D. (1997). Sharing the "Cost" of Multicast Trees: An Axiomatic Analysis. IEEE/ACM Transactions on Networking, 5(6), 847-860.

ICS. (1999). Internet Software Consortium, [WWW]. Available: http://www.isc.org/ds/ [1999, 27 Oct].

IRM. (1998). Internet Resource Management - An Enterprise-wide Concern, [WWW]. Sequel Technology Corporation. Available: http://www.sequeltech.com/products/download.asp [1999, 5 September].

Lee, D., Lough, D., Midkiff, S., IV, N. D., & Benchoff, P. (1998). The Next Generation of the Internet: Aspects of the Internet Protocol Version 6. IEEE

Network(Jan/Feb), 28-33.

MacKie-Mason, J. (1997). A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network, [WWW]. Ann Arbor, Dept. of Economics, University of Michigan. Available: http://www-personal.umich.edu/~jmm/papers/reserve3.html [1999, 5 Oct].

MacKie-Mason, J., Murphy, L., & Murphy, J. (1995). The Role of Responsive Pricing in the Internet, [WWW]. Presented at MIT workshop on Internet Economics March 1995. Available: http://www.press.umich.edu/jep/works/MackieResp.html [1999, 5 Oct].

MacKie-Mason, J., & Varian, H. (1993, April). Pricing the Internet. Paper presented at the Public Access to the Internet, JFK School of Government.

MacKie-Mason, J., & Varian, H. (1994). Some FAQs about Usage-Based Pricing, [WWW]. University of Michigan. Available: http://www.press.umich.edu/jep/works/mackiemason.usage.html [1999, 5 Oct].

MacKie-Mason, J., & Varian, H. (1995). Pricing Congestible Network Resources. IEEE Journal on Selected Areas in Communications, 13(7), 1141-1149.

McKnight, L., & Bailey, J. (1995). An Introduction to Internet Economics, [WWW]. Available: http://www.press.umich.edu/jep/works/McKniIntro.html [1999, 5 Oct].

Metcalfe, B. (1998). Pay-as-we-go Internet Puts Your Money Where Your Consumption Is. InfoWorld, 20(38), 129-130.

Odlyzko, A. (1997). A Modest Proposal for Preventing Internet Congestion. AT&T Labs Internal Report. Available: http://www.research.att.com/~amo [1999, 15 Oct].

Odlyzko, A. (1998a). Data networks are lightly used and will stay that way, [WWW]. AT&T Labs - research. Available: http://www.research.att.com/~amo [1999, 20 Oct].

Odlyzko, A. (1998b). The Economics of the Internet: Utility, Utilization, Pricing and

Quality of Service, [WWW]. AT&T Labs - research. Available: http://www.research.att.com/~amo [1999, 20 Oct].

Parenteau, B., & Rishe, N. (1997). Internet Pricing and Prioritization. Paper presented at the IEEE 4th International Workshop on Community Network Proceedings.

PARNet. (1999). Perth Academic Research Network, [WWW]. Available: http://www.parnet.edu.au [1999, 3 Nov].

Rupp, B., Edell, R., Chand, H., & Varaiya, P. (1998). INDEX: A Platform for Determining how People Value the Quality of their Internet Access. Paper presented at the Proceedings of the 6th IEEE/IFIP International Workshop on Quality of Service.

Schnizlein, J. (1998). How can routers help Internet economics? Paper presented at the ACM Proceedings of the first international conference on Information and computation economics.

Shenker, S., Clark, D., Estrin, D., & Herzog, S. (1996). Pricing in Computer Networks: Reshaping the Research Agenda. J Telecommunications Policy, 20(3), 183-201.

Stahl, D., Whinston, A., & Zhang, K. (1998). A simulation study of competitive Internet pricing: AOL flat rates versus GSW usage prices. Paper presented at the Proceedings of the first international conference on Information and computation economies.

Tebbutt, D. (1998). Quality and Equality. Australian Personal Computer(November), 67-68.

Tebbutt, D., & Taylor, N. (1998). Seven Key Internet Technologies. Australian Personal Computer(November), 86-96.

8. APPENDIX

Table 1: Comparison of pricing schemes based on characteristics of 'optimal' pricing policies.

Pricing Scheme	Flat-rate	Volume	Smart Market	Edge Pricing	Metro Card	Paris Metro	Priority Classes
1. Encourage use at less congested times	No	On a coarse basis if time of day discounts offered	Yes	Only on a coarse time or day basis	Yes	No	Yes
2. Include impact of current load on future demand	No	No	Yes	No	Yes	No	Yes
3. Coarser than packet level	Yes	No	No	Yes	No	Yes	Yes
4. Reflect load status	No	No, not dynamically	Yes	No (not dynamically)	Yes	No	Yes
5. Provide load management	No	No	Yes	User's responsibility	Yes	User's responsibility	Yes
6. Decentralised	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7. Multiple priorities	No	No	Yes	Yes	Yes	Yes, limited to number of channels.	Yes
8. Encourage appropriate use of QoS	No	No differential QoS	Yes	Yes	Yes	Yes	Yes
9. Incentives to service providers	No	Yes	Yes	Yes	Yes	Yes	Yes

Table 2: Comparison of pricing schemes based on implementation issues.

Pricing Scheme	Flat-rate	Volume	Smart Market	Edge Pricing	Metro Card	Paris Metro	Priority Classes
1. Perverse Effects	Yes	Yes, probably not marked	No	No	No	Some instability caused by self-regulation	No
2. Fairness of prices	No, Unfair to light users	Yes, fairer than Flat-rates	Yes, directly based on congestion caused	Yes, however they are based on approximations	Yes, directly related to congestion	Yes, users pay for the channel they need at the time	Yes
3. Predictable cost	Yes	Yes, can use bulk selling of volume	No, however, the upper limit is known.	Yes	No, however, the maximum budget is known	Yes	Yes, Can accept or reject cost of current session
4. Correct prices easily arrived at	Yes, simply an access fee	Yes	No	Yes, moderately	No	No	Automatically calculates correct prices
5. Built into architecture	No	No	Yes	No	Yes	Yes	Yes
6. Can indicate who pays	Subscriber pays for access	Traffic identified on whether sent or received.	Unresolved	Unresolved	No	User pays for access	No, the user pays for services received
7. Simple	Yes	Yes	No	Yes	No	Yes	No
8. Implementation overheads	Low	Moderate	High	Moderate	High	Moderate	High
9. Management overheads	Low	High	High	Moderate	High	Moderate	High
10. Assumes Internet homogeneous	No	No	Yes	No	Yes	No	Yes
11. Changes to protocols/routers needed	No	No	Yes	Yes	Yes	Can use IPv4 , but changes to routers needed	Yes