

INTERNATIONAL TELECOMMUNICATION UNION



Telecommunication Development Bureau

Final report

IP-based networks:

Pricing of telecommunication services

Programme 4.1, Activity 4335 of the Operational Plan VAP 2002

Market, Economics & Finance Unit

January 2003

Table of contents

List of Figures	iv
List of Tables	iv
Executive Summary.....	1
1 Introduction	3
2 Overview of the internet and IP	5
2.1 Network Layers.....	5
2.2 The structure of Connectivity in the Internet.....	6
2.3 Internet addressing	8
3 Quality of service: Technological aspects	10
3.1 Service quality on IP networks.....	10
3.2 Categories of service quality	12
3.2.1 QoS	12
3.2.2 ToS and CoS.....	13
3.2.3 GoS.....	13
3.3 Categories of service quality	14
3.3.1 IntServ	14
3.3.2 DiffServ	16
3.3.3 QoS and ATM.....	17
3.4 Technical constraints on the development of QoS and CoS services	19
3.4.1 QoS problems at the Internet's edges and within networks	19
3.4.2 QoS problems at borders	20
4 Quality of service: pricing and congestion	22
4.1 The Internet is not an economic 'public good'	22
4.2 Prices and cost structures	23
4.3 Class of service pricing	28
4.4 Pricing and QoS between ISPs and their transit provider	30
4.4.1 The structure of settlement prices	30
4.4.2 QoS guarantees for transit	33
4.5 QoS and the Next Generation Internet.....	34
4.6 Conclusion regarding QoS and CoSes	36
5 Present and future of "real-time" IP service.....	38
5.1 Existing VoIP	38
5.1.1 Public IP telephony on Proprietary IP networks	39
5.1.2 International Accounting Rate bypass.....	40
5.1.3 Private IP telephony on corporate IP networks	43
5.1.4 Proprietary IP routing technology	43
5.2 Future "real-time" IP networks	44
5.2.1 Technological aspects	44
5.2.2 Pricing and settlements.....	45

6	Analysis of research on Internet service pricing	47
7	IP regulatory issues.....	50
7.1	Regulation and ISPs	50
7.2	Regulation and incumbent PSTN operators	52
7.3	Regulation and next generation mobile operators	54
	References	55
	Glossary.....	57
	Annex I:	I
	QoS and the limitations of cheap bandwidth	I
	Annex II:	IV
	Research papers on pricing and CoS	IV
	Annex III	IX
	QoS attributes of ATM networks.....	IX

List of Figures

Figure 2-1:	OSI and Internet protocol stack.....	5
Figure 2-2:	Vertical and hierarchical interconnection in the Internet	7
Figure 2-3:	The IPv4 address expressed as a dotted decimal notation.....	8
Figure 3-1:	Application specific loss and delay variation QoS requirements	11
Figure 3-2:	IP Precedence ToS field in an IPv4 header	13
Figure 3-3:	The Three Levels of End-to-End QoS Are Best-Effort Service, Differentiated Service, and Guaranteed Service.....	16
Figure 3-4:	Coordination and end-to-end superior service quality	21
Figure 4-1:	Demand for Internet service deconstructed	29
Figure 4-2:	Fitting CoSes within service QoS requirements.....	35
Figure 5-1:	International VoIP using a Wide Area Ethernet Network.....	40
Figure 5-2:	VoIP international Bypass	41
Figure 6-1:	Depicting the cross-fertilisation of economic ideas to computer network design	48

List of Tables

Table 4-1:	Traffic hierarchies in next generation networks	35
Table 0-1:	Suitability of ATM Forum service categories to applications	IX

Executive Summary

This study is concerned with pricing and different categories of quality of service over IP networks, with a view to the provision of real-time IP services. In providing information about this topic the report also analyses several academic studies (supplied by the ITU) in terms of their contribution to the topic, and explains the business case for international bypass offered by IP networks.

The main point to be emphasised is that class of service technologies are not well developed, and do not operate on the public Internet to any great degree, but are mainly restricted to a small number of campus networks. Moreover, no system of service category pricing (or class-of-service pricing) currently exists. Neither accounting, billing, or user interface software have been developed that would enable this service. As such, discussion of class-of-service options is typically reserved for academic papers.

The **research papers** discussed in this report have no practical application for real IP networks at this time. They are exercises in mathematical model building. This should be puzzling to readers who are not highly Internet literate as there is a great deal of literature that discusses class and grade of service on IP networks including the Internet. However, most of this literature is technical or comes from firms that are interested in selling in-house solutions for the communication needs of corporations. In such cases quality of service (QoS) is more manageable and Intranets can now provide a solution that includes voice, data, and video. However, no pricing to end-users operates on such networks.

Where VoIP does operate outside of corporate intranets, it appears to be mainly an international **Accounting Rate bypass** service of low quality, and largely limited to single transit ISP networks. Typically phone-to-phone calls pass through gateway devices that packages the data received from PSTN phone lines and sends it into an IP network. At the far end the packaging is reversed. This type of service is at an early stage of development.

On IP networks no dedicated circuit is held open for the duration of a communication, as occurs with the PSTN. Rather information is digitised and placed into packets, and sent with other packets from different sources in a randomised fashion, ultimately to reach their destinations. This randomising of packets means that all packets are treated with equal priority be it a packet from a voice conversation, or an email. Where congestion occurs packets that are earlier in the queue will be forwarded first, i.e. packets containing voice will have to wait for any packets that are not time-critical and that are earlier in the queue. As a rule, the reliability and quality of 'virtual' connections on the Internet falls well short of what can be provided over the PSTN.

Network management designed to control congestion is therefore the key to real-time

quality of service on IP networks including the Internet. A great deal has been done technologically to address congestions problems but ultimately it will fall to demand management if congestion problems are to be largely overcome.

Demand management is mainly a matter of pricing, where often the **structure of pricing** is more important than the price level. To be economically efficient the structure of the prices offered to users should match the structure of the costs users cause; that is, the way costs are caused should be reflecting in the way liability is incurred by the customer. Pricing on the Internet does not work like this at present. Internet users presently pay a periodic subscription fee, and except for dialup users who pay per minute charges, face no additional price for sending extra packets even at times of peak demand.

Arguably the main pricing tool required to bring about a reliable real-time IP service is a **congestion price**; that is a price that varies so that all those demanding real-time service at that price and during a period of peak demand, would receive it. Such a pricing system would also work to communicate the optimal level of investment in network capacity. In conjunction with a system that enabled users to select the quality of service (e.g. the class-of-service {CoS}) they required, it would suggest the convergence of IP networks and the Internet, with other platforms such as the PSTN and CATV (cable TV) networks.

It appears to be many years off before specified service qualities that are combined with more sophisticated pricing than occurs with Internet service presently, will become generally available for Internet users. The main reasons for the lack of sophisticated pricing and QoS options are technical and can be summarised as software and hardware problems that exist between ISPs that result in less than seamless interoperability between them.

In addition to QoS problems at network borders, there are other problems that will need to be overcome before sophisticated pricing and QoS options are developed: These are: congestion management problems *on* IP networks – mainly the pricing of services (as outlined above); a lack of accounting information systems able to provide the necessary measurement and billing between networks, such as would be required to support several levels of service quality that subscribers may select from depending on the type of communications service they are engaging in at that time, and the absence of an interface with end-users that enables different QoS to be chosen in a way that provides value to users.

1 Introduction

According to the terms of reference, this study is concerned with the roll of pricing in regard to different categories of quality of service over IP networks, with a view to the commercial viability of real-time IP services. In providing information about this topic the report provides an analysis of several academic studies (supplied by the ITU) in terms of their contribution to the topic, and explains the business case for international bypass offered by the Internet and possibly other cross-boarder IP networks.

When looking at real-time service over IP networks, we are compelled to address issues of service quality. Real-time service over IP requires certain service quality characteristics which are problematic for IP networks - especially a public network like the internet. This report is fairly ambitious in its approach, setting out structural and technical reasons explaining why service quality is problematic on IP networks, and in doing so it explains the various categories, grades and classes of service quality which are possible on IP networks in controlled environments.

In practice quality of service options are not widely available on the Internet at present although some are making an appearance on private IP networks and campus networks. However, there are no accounting or payment systems that would enable users to pay for a higher QoS on the basis of usage or packet throughput. There is thus much to be done to marry future quality of service developments (the supply-side) with the introduction of pricing options (the demand management side).

The main quality of service problem that must be addresses in order for real-time services to be viable, is congestion. The report discusses congestion problems and explains that when packets that require real-time service quality are randomised with other packets, as occurs with the Internet, either all packets must receive real-time service quality, or packets requiring priority treatment must be able to be targeted for special treatment.

In addition to a prioritised system of admission to the network, there are at present chiefly two methods that can be used to improved service quality on the Internet:

1. Through reserving capacity on connections between which the higher QoS is required, or
2. To include a facility that enable packets that are marked (tag) accordingly to receive priority treatment.

Neither of these two methods presently functions on a sufficient scale on the Internet for VoIP or any other real-time service to be a serious competitive challenge to PSTN service. The reasons for this are technically complex but appear to be sufficiently serious to rule out a broadly based takeoff of real-time services provided over the

Internet occurring in the next few years. However, in many countries that retain high Accounting Rate charges for international PSTN calls, International VoIP voice services are growing rapidly. Such services are likely to be built around single transit ISP networks and in this regard it could be argued that they fall slightly short a voice over the public Internet service. The quality of this service will be relatively low. With the ongoing development of Wide Area Network technologies which are international in scale, this type of service is likely to continue growing especially in regions of the World that try to hold onto the old Accounting Rate system.

In order for the any-to-any concept to operate with the Internet, pricing services according to the quality characteristics of the service will be required in order to provide a reward to service providers for developing what is undoubtedly a higher cost service. However, software and hardware that would enable this to occur appear to be some way off, not least because a solution to the technological problems of providing differential service quality options also appears to be some way off; i.e. class and grade of service pricing. As these service quality options have not yet developed into a commercially viable technology, it is thus not surprising that more sophisticated pricing models that allow users to pay for service quality, have not yet developed.

The report is structured as follows: Chapter 2 provides an overview of IP networks and the Internet in particular. It explains the layers of software that enable the Internet to function, and describes how connectivity among ISPs provides a structure to the Internet. Chapter 3 discusses the topic of categories of service quality and the technical problems that prevent multiple service quality categories developing on the internet. Chapter 4 addresses economic aspects of congestion management – principally the use of pricing to cost-effectively manage peak demand in a way that improves the economic welfare of society. Chapter 5 looks at “real-time” IP services including voice over IP (VoIP) Accounting Rate by-pass. Chapter 6 discussed the academic research papers provided by the ITU, with Chapter 7 reserved for regulatory issues relating to IP, especially real-time services.

2 Overview of the internet and IP ¹

2.1 Network Layers

The Internet is comprised of well over 100,000 networks which operate different software and hardware solutions on top of TCP/IP protocols. This diversity is both a strength and a weakness. It enables networks using non standardised equipment and operating with non standardised and diverse software, to connect and communicate with each other over the Internet. On the negative side this diversity makes it difficult to overcome technical (hardware and software) obstacles that stand in the way of seamless interoperability between networks.

Figure 2-1: OSI and Internet protocol stack

Applications and Services	Layer 7 – Application Layer 6 – Presentation Layer 5 – Session
TCP or UDP	Layer 4 – Transport
IP	Layer 3 – Network
Layer 2 - Data Link	Layer 2 - Data Link
Layer 1 - Physical	Layer 1 – Physical

Source: Smith and Collins (2002.)

One of the principle reasons for this lack of seamlessness is hinted at in Figure 2-1, which shows the seven-layer protocol stack that makes up the Internet, known as Open Systems Interconnection (OSI).² IP operates at level 3, with applications and services protocols operating above that at layers 5 to 7.^{3,4}

-
- ¹ At various stages in this report I have drawn heavily on a report I and D. Elixmann authored for the European Commission. That report is included in the references as WIK (2002). The technical advice I received from Alberto E. García and Klaus Hackbarth while doing that study has proven useful in the present study, although clearly any technical or other errors are my own.
 - ² The OSI describes how information from a software application in one computer moves through a network medium to a software application in another computer. It is the primary architectural model for inter-computer communications.
 - ³ In their combined form as written TCP/IP signifies a suite of over 100 protocols that perform lower level functions. IP (Internet protocol) and TCP (transmission control protocol) do, however, bear the largest share of the workload in layer 3.

The technical obstacles that stand in the way of seamless interoperability will need to be overcome in order for the Internet to converge with other platforms such as the PSTN and CATV networks. In order for the converged Next Generation Internet to become reality solutions will need to be found that improve the quality and reliability of connectivity, and the bandwidth and quality of service (QoS) available to end users. At present interoperability occurs between different networks operating IP, but it is not completely seamless, and not all functionality is retained between networks.

2.2 The structure of Connectivity in the Internet

The Internet is arranged in a loose hierarchy of entities, with IP communications devices like PCs, workstations and servers (also called hosts) at the outer edges, connected to Local Area Networks (LANs)⁵, which are connected to one or more regionally focused ISPs (called local ISPs). Local ISPs are typically connected to national ISPs, which are themselves connected to international ISPs. At the top of this loose hierarchy are core ISPs, also known as IBPs (internet backbone providers), or Tier 1 ISPs. In some cases an ISP may traverse two or more of these loose hierarchical levels. A diagrammatic representation of this hierarchical structure can be seen in Figure 2-2.

Although prior to the mid 1990s it was only large ISPs that maintained interconnection with several other ISPs, it is now the case that large numbers of regional ISPs connect to several other regional ISPs (horizontal connectivity). Moreover, many ISPs connect to more than one transit provider (vertical connectivity).

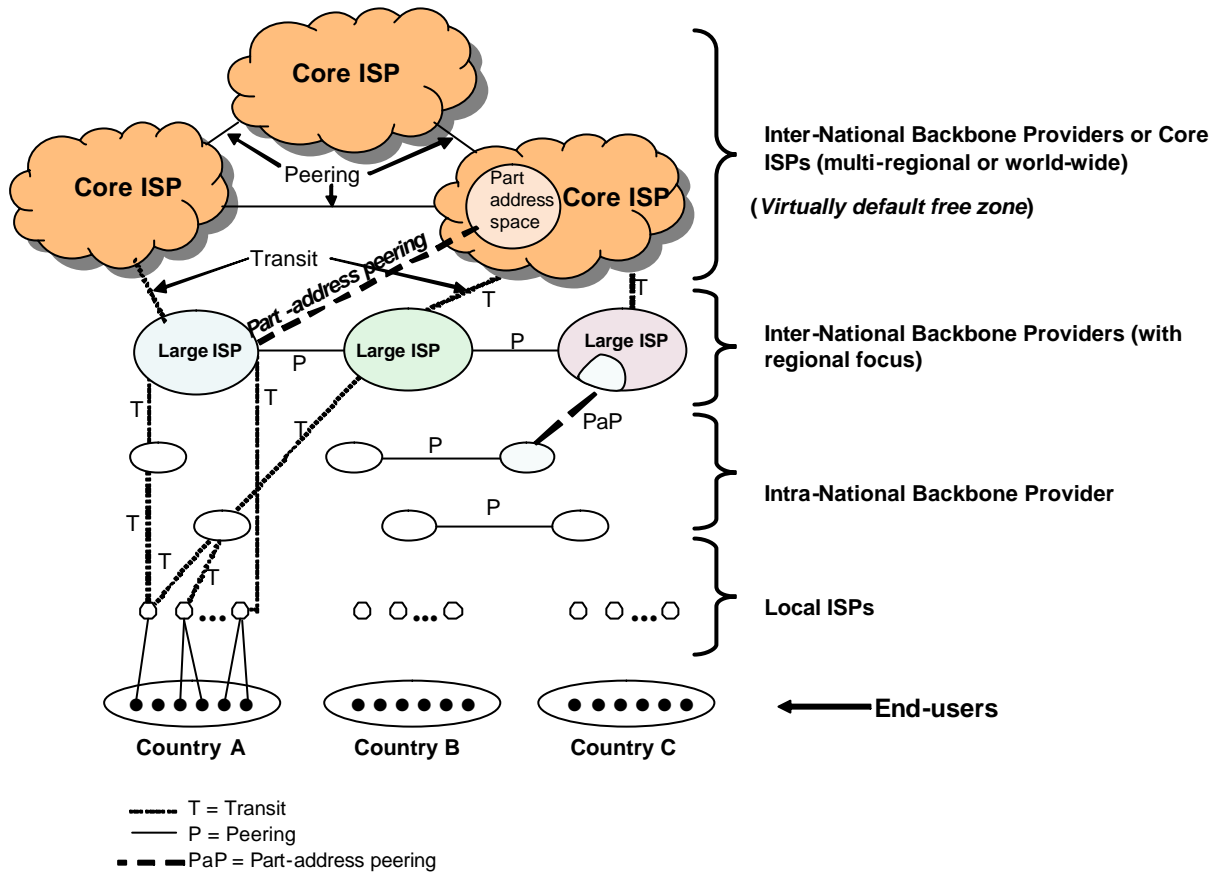
Horizontal interconnection is known as *peering*, although where it occurs between ISPs that are not major backbone providers it is commonly known as *secondary peering*. Peers only accept traffic from other peers that is for termination on their own network. Under a peering arrangement where packets are sent with addresses that are not recognised on the receiving ISP's network, the packets are dropped. With rare exceptions peering relationships do not involve payment between the peering partners. The price each charges to accept traffic from the other for termination is simply the cost of the reciprocal arrangement; it is a *sender-keeps-all* interconnection arrangement.

Although it is rare, ISPs that are dissimilar in terms of size (Km of network or customer address space) may agree to peer, but in this case the larger ISP will only agree to provide the smaller peering partner with a subset of its total address space, and typically this will comprise roughly the same number of addresses as the smaller ISP is able to make available to the larger ISP. This we have referred to as *part-address peering*.

4 At layer 1 and 2 there are a multitude of different fixed-link networks (e.g. ISDN, LANs, ATM-networks, SDH-networks, and (D)WDM) that can transport IP traffic. The Internet Protocol (IP) at layer 3 is completely independent of the lower levels.

5 LANs are collections of several IP communications devices connected to one another.

Figure 2-2: Vertical and hierarchical interconnection in the Internet



Source: Derived from WIK 2002

Vertical interconnection is what is covered by **transit** contracts. ISPs pay for transit. Unlike in a peering relationship, the ISP selling transit services will accept traffic that is not for termination on its network (i.e. datagrams with addresses that are not recognised by the larger ISP's routing tables), and will route this transit traffic to its peering partners, or will itself purchase transit where the termination address is not recognised. As such, a transit agreement offers connection to all end-users on the Internet, which is much more than is provided under a peering arrangement. Starting in the late 1990s, many smaller ISPs started to take transit contracts with more than one transit providing ISP.⁶ This is known as **multi-homing**. The main reasons explaining why ISPs may choose to multi-home appear to be:

- upstream service resilience, and

⁶ The main reasons for the increase in connectivity between ISPs has been the development of routing protocols that enable low cost hardware and network management options that have enabled smaller ISPs to choose between alternative routes and transit providers when sending traffic up the hierarchy, and in the declining cost of leased infrastructure, and the growth in underlying infrastructure. For smaller and medium sized ISPs, multi-homing was made economically viable by the development of BGP4 and subsequently by cheap and easily operated routing equipment that uses it. See for example, BoardWatch Magazine, July 1999

- to assist with the optimisation of traffic flows.⁷

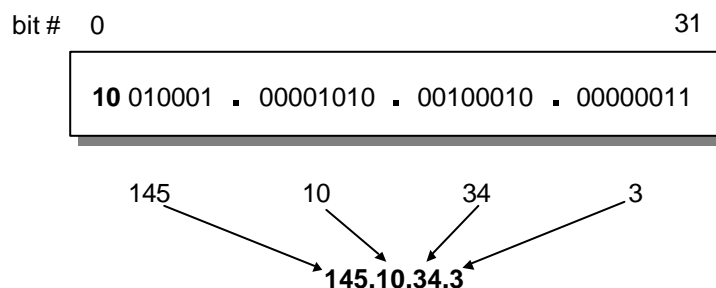
There are other forms of connectivity that substitute for peering and transit contracts. These are known as: *Hosting*, *Caching*, *Mirroring*, and *Content Delivery Networks* (CDNs). They are all rather similar in their purpose which is to hold content (e.g. web pages) closer to the edges of the Internet, and in so doing reduce the cost of transit for ISPs, and improve response times for information requests. All entail minor variations on this theme.

2.3 Internet addressing

Although the hierarchy of the Internet is loose it is an integral part of the Internet's character. It avoids the need for each ISP to interconnect with each and every other ISP. Such an arrangement would be entirely impractical. Among other things, it would require all ISPs to maintain complete address tables in their edge routers, and the Internet would need to be fully meshed, a structure which is clearly impossible given the number and geographic dispersion of ISPs today.

For these reasons the Internet uses a hierarchical addressing and routing system. IP datagrams are initiated at the outer edges. These datagrams have addresses in the form of an IP number in the header of each datagram. As a rule ISPs will only have the address (IP numbers) recorded on its routing tables for its own customers or those customers of ISPs it **peers** with. If the address is not found in this list, as is commonly the case among smaller ISPs, the datagram is sent up the hierarchy to a larger ISP with which the smaller ISP has a transit service contract. This procedure is followed until a network is found on which the addressed is recognised, in which case the packet will be sent for termination.

Figure 2-3: The IPv4 address expressed as a dotted decimal notation



Source: Semeria (1996)

IP addresses, or more correctly IP numbers, are attached to every IP packet. Once the digitised information is packed into the datagram's payload, up to the moment it is

⁷ See Huston (2001a).

delivered to the receiver, the header is the only part of the datagram that is inspected. Since the early 1980s the IP datagram in use has been IPv4 (Internet protocol version 4). It has a 32-bit IP address which is divided into 8-bit fields, each expressed as decimal number and separated by a dot. Figure 2-3 provides an example of the dotted decimal notation.

In practice, addressing is not as disorganised as the above paragraph implies as the routers of ISPs exchange information with other routers concerning the best route to send datagrams, and this information is stored in routing tables which are updated periodically. The choice of path that datagrams will follow in order to reach their termination address is also decided by routing protocols, and these can be manipulated by the network manager.⁸

⁸ A collection of routers that is under the administrative control of a single organisation forms an *Autonomous System (AS)* – also known as a *routing domain*.

3 Quality of service: Technological aspects

3.1 Service quality on IP networks

IP networks are based on packet switching technology. Information is digitised and placed into packets, and sent with other packets from different sources in a randomised fashion, ultimately to reach their destinations. This randomising of packets means that all packets are treated with equal priority be it a packet from voice conversation, or an email. Where congestion occurs packets that are earlier in the queue will be forwarded first, i.e. packets containing voice will have to wait for any packets that are not time-critical and that are earlier in the queue.

There is no dedicated circuit held open in an IP network for the duration of a communication, as occurs with the PSTN.⁹ As well as the ability to interoperate with diverse systems, one reason for the attractiveness of packet networks like IP compared with circuit switched networks is that they provide for a much greater level of flexibility in the bandwidth requirements of connections, and provide greater utilisation efficiency of available capacity. It also suggests the possibility of service integration, not something associated with the PSTN.

As a rule, the reliability and quality of 'virtual' connections on the Internet falls well short of what can be provided over the PSTN. The gap in quality between the two platforms has, however, closed considerably as processor power and the data carrying capacity of networks has leaped forward. But the gap in service quality is still very apparent, such that voice services on the Internet can be provided, but under quite specific technical arrangements and with widely variable service quality being experienced by end users.¹⁰ Generally, outside of the most modern of private intranets, the experience of users with VoIP is of a service of variable and poor quality, which mainly operates on international links out of countries where the price of PSTN calls is regulated at very high levels, i.e. it is an international bypass service.¹¹

In terms of QoS most things concerning Internet traffic are uncertain and have to be defined probabilistically. Strictly speaking packets do not receive equal treatment in the Internet, however, as they are randomised irrespective of whether they are voice, email, or for some other purpose, then in this sense no packet can be said to be targeted for superior treatment.

⁹ In the case of long distance calls technological convergence has resulted to all traffic being send on the same underlying infrastructure and optical circuits. This means that PSTN circuits are even now more virtual than real.

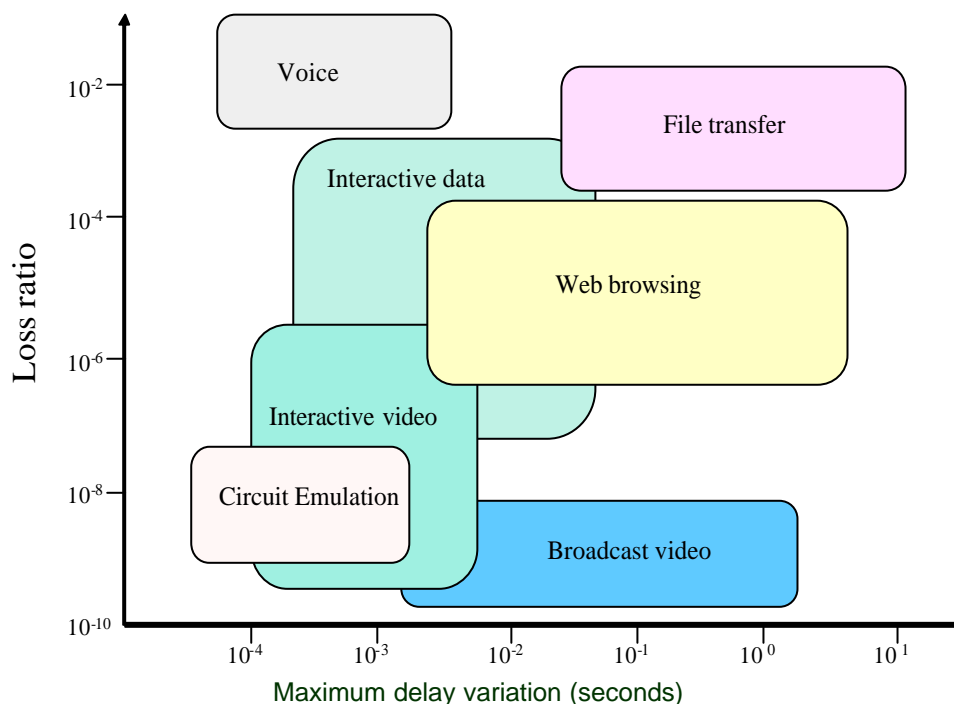
¹⁰ It is not uncommon for buffer times of 1 second to be required.

¹¹ We discuss the IP international bypass model in Chapter 6.

There are clearly many different types of service that can be provided over the Internet. These include WWW, streaming video, file transfer, email, and real-time voice services. All these products make different demands on QoS. VoIP is said to tolerate a certain level of latency (delay), jitter (delay variation) and bandwidth. Video streaming requires higher bandwidth, although can tolerate slightly more latency and jitter than does VoIP. To the extent that 'real time' applications, can adjust the time of playback¹² then latency and jitter statistics required of the network will be correspondingly lower. While adjusting play-back times is possible for streaming video, it is not for VoIP.

In Figure 3-1 can be seen loss and delay variation parameters applicable to various applications. It should be clear that if each of these services is to be provided over the Internet to a quality such that it competes head on with the traditional platform that has provided the service (e.g. PSTN, CATV, FTA broadcasting), then either all datagrams will need to receive premium quality QoS and without significant packet loss or queuing at nodes, or packets will need to be treated differently in line with the QoS required of the application, (i.e. if an end-user wants to use the internet for a real-time IP telephone conversation he or she will need to get a service with especially short delay times for datagram delivery).

Figure 3-1: Application specific loss and delay variation QoS requirements



Source: McDysan (2000)

The Internet was not envisaged with this ability in mind. It was designed to provide a low

¹² It does this by 'buffering' – i.e. by holding packets in memory for very short periods of time until the packets can be 'played back' in the correct order without noticeable delay.

cost data delivery service, where QoS and controls on queuing to be admitted to the service (called Grade of Service {GoS} discussed below) were not important attributes. As has been noted already, datagrams from many sources share the same transport pipe, channels (one direction) or circuits (bidirectional). This is part of the traffic management function in IP networks and is known as *statistical multiplexing* which involves the aggregation of data from many different sources so as to optimise the use of network resources.

3.2 Categories of service quality

As a popular term *quality of service* (QoS) is often used to include rather more than its technical definition. Often when Internet commentators use the term QoS they mean something akin to the service experienced by end-users. For the purposes of this report we shall identify 3 different aspects of service quality as understood by Internet technologists. These are:

- Quality of service (QoS)
- Type of service (ToS) field and within this, Class of service (CoS)
- Grade of service (GoS)

3.2.1 QoS

Because QoS conveys a lot of information about how the Internet works, this issues is discussed in rather more detailed than for the other 2 bulleted topics. QoS is defined by a set of parameters that describe a flow of packets or cells (call them datagrams) produced by an Internet session. This can involve point to point, or multipoint, multicast, and broadcast. The most important QoS parameters in packet/cell switched networks are:

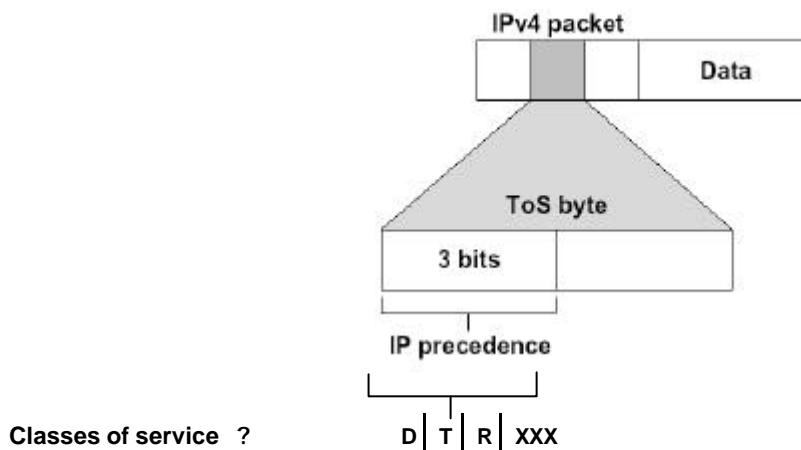
- Latency; the time it takes for packets or cells to go from sender to receiver;
- Jitter; the variation in latency;
- The rate of packet or cell loss or arrival which is too late to be useful, and
- Errors in labelling / addressing.

Altogether, these statistics describe the quality of service of a particular flow of datagrams. It is important that these QoS statistics are maintain from origination to termination, within the limits required in order to provide a VoIP service that is of sufficient quality to attract consumers in large enough numbers to make the service viable.

3.2.2 ToS and CoS

Type of Service (ToS) refers to the type of service field that exists in IP datagrams for a Class of Service (CoS) to be specified for each packet. IPv4 headers provide 3 bits which in practice today enable four different classes of service to be specified in the ToS field settings. Figure 3-2 shows the ToS field in an IPv4 packet header.

Figure 3-2: IP Precedence ToS field in an IPv4 header



Source: Black (1998), and Cisco Systems

In Figure 3-2 the D is for specifying whether the datagram can be delayed, the T is for throughput priority, R is for indicating whether a reliable subnetwork is required, and 'XXX' is reserved for future use.

This field can be used to support CoS but in order to do so routers have to be programmed to support it. At present this is not done. Rather, an architecture called Differentiated Services (*DiffServ*) uses this field although in a slightly altered form and renamed the Differentiated Services Field. We discuss *DiffServ* in Section 3.3.2.

3.2.3 GoS

While QoS relates to the statistical properties of a particular flow, Grade of Service (GoS) refers to the statistics that describe the probability of having your packets admitted in the first place. In a PSTN environment this property is referred to as *blocking*. Both GoS and QoS must be known before we can describe service quality between two points that are 'virtually' connected over an IP network(s). Thus, a GoS parameter needs to be added to the above QoS parameters:

- The probability that the service (as described by the QoS parameters) is availability.

3.3 Categories of service quality

Many technological changes have occurred in recent years which enable QoS to be managed more effectively within an IP network, although between networks there are still QoS problems. In recent years two approaches to providing a superior service category have been widely discussed. These are *IntServ* and *DiffServ* technologies. In the remainder of this section we discuss these two contribution and also identify certain architectural features of the Internet that have a significant bearing on QoS. Then in Section 3.4 we discuss the QoS problems that are preventing real-time services being widely available on the public Internet, and are preventing the development of different categories of service quality.

The Internet provides two main ways in which traffic can be managed selectively for QoS.

1. To mark the packets with different priorities (tagging), or
2. To periodically reserve capacities on connections where higher QoS is required.

The first one provides preferential treatment for packets that are marked accordingly. This approach to QoS must be implemented in all routers through which packets can pass. It provides for different queues for priority (tagged) and non-priority packets, where selection of the next packet to go out is determined through a weighted queuing algorithm. This approach does not guarantee fixed values of QoS. Rather, QoS still needs to be viewed probabilistically, i.e. in terms of QoS statistics that are superior in some way to what is standard. It involves the use of virtual circuits (VC) or virtual paths (VP), and not types of end-user services. It is implemented by the real-time transport protocol (RTP) and is supplemented by a corresponding control protocol known as real-time control protocol (RTCP) which controls the virtual connection used by this technology.¹³

In the second case a form of signalling is introduced which tries to guarantee a minimum value of capacity for the corresponding packet flows which require a higher QoS than standard. A brief discussion of these technologies is provided immediately below.

3.3.1 IntServ

In point 1 above the technology is known as *IntServ* (for integrated services). Along with RSVP (Resource ReSerVation Protocol) *IntServ* works through admission control. During periods of heavy usage each flow request would only be admitted if it did not

¹³ See RFC1889, and RFC1890

crowd out other previously admitted flows. Packets that are not labelled as requiring priority will form the group from which packets are dropped when network congestion begins to reach the point when stated QoS statistics are threatened.

In 2000 the *IntServ* model offered two service classes, with standards specified by the Integrated Services Working Group of the IETF:

- (i) the controlled load service class, and
- (ii) guaranteed QoS class.

The QoS offered by the former during periods when the network is in high demand, is similar to the QoS provided by an unloaded network not using *IntServ* technology, such as is provided today on a backbone during uncongested periods. For this option to work the network needs to be provided with estimates of the demands required by users' traffic so that resources can be made available.

The guaranteed QoS class focuses on minimum queuing delays and guaranteed bandwidth. This option has no set-up mechanism or means of identifying traffic flows, and so needs to be used along with RSVP.¹⁴ The receiver of packets needs to know the traffic specification that is being sent so the appropriate reservation can be made, and for this to occur, the path the packets will follow on their route between sender and receiver needs to be known. When the request arrives at the first routers along this path the path's availability is checked and if confirmed the request is passed to the next router. If capacity is not available on any router on this path an error message is returned. The receiver will then resend the reservation request after a small delay.

IntServ technology does not use the ToS field in IP packets, but rather works with emulated VCs.

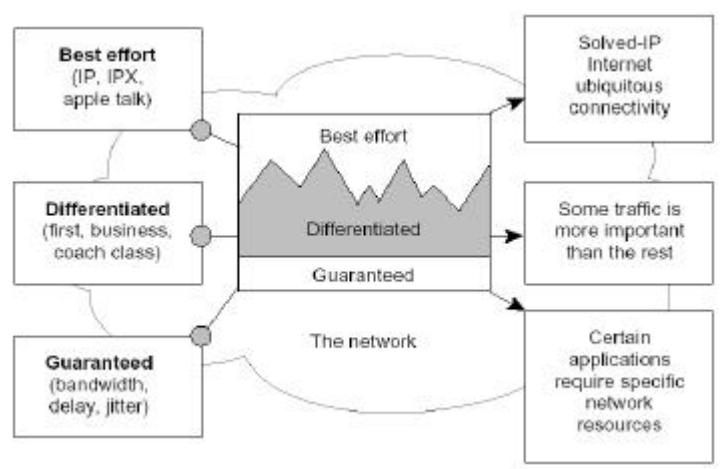
Although the *IntServ* / RSVP model can be used on the Internet, there are severe problems that in practice prevent this from occurring:

- There are severe scalability problems that prevent its use in large networks, and
- The *IntServ* model is a technical solution only. No attention was paid in its design to the need for the service to be priced in order to manage demand and supply.

At this time *IntServ* is limited to use on private IP networks.

¹⁴ RSVP is a control protocol which does not carry datagrams. Rather, these are transported after the reservation procedures have been performed through the use of RTP. RSVP requires in addition, signalling protocols to make these reservations, (discussed further below). However, reservation is only possible if all routers involved in the transmission support RSVP. RSVP uses a token bucket algorithm. Tokens are collected by a logical token bucket as a means of controlling transmission rate and burst duration. The simple token bucket algorithm relies on two parameters: the average transmission rate and logical bucket depth. Arriving packets are checked to see if their length is less than the tokens in the bucket.

Figure 3-3: The Three Levels of End-to-End QoS Are Best-Effort Service, Differentiated Service, and Guaranteed Service



Source: Cisco Systems

3.3.2 DiffServ

DiffServ (differentiated services) architecture is designed to operate at the edges of networks based on *expected congestion* rather than actual congestion along paths. As is implied by this description, there is no guaranteed QoS for any particular flow. As with the standard Internet, *DiffServ* technology is still based on statistical bandwidth provisioning. *DiffServ* technology is intended to lift QoS statistics for packets that are marked accordingly.

DiffServ will support several different QoS standards in the ToS field of the slightly modified IPv4 packet header. In its slightly altered form it is known as the Differentiated Services Code Point (DSCP).¹⁵ Marking of the DSCP will normally only need to occur once, at a DS network boundary or in the user network. All data shaping, policing and per flow information occurs at network edges. This means that *DiffServ* has considerable scaling advantages over *IntServ*.¹⁶

The flexibility of the system allows service providers to match expected of QoS to expected performance levels, such that numbers of different performance levels (and

¹⁵ Under IPv6 *DiffServ* can not apply the TOS field because the basic header does not contain it. However, it will be implemented through an extension to the basic header.

¹⁶ *DiffServ* requires that a service profile is defined for each user, the pricing of which will be determined between the ISP and end-user. The subscriber is then allocated a virtual token bucket which is filled at a set rate over time, and can accumulate tokens only until the bucket is full. As packets arrive for the user, tokens are removed. However, all packets whether tagged or not, arrive in no particular order (as occurs with the present Internet). Under congested conditions, while the user has tokens in credit, all her packets will be marked as "in profile", and packets not tagged as "in" form the group from which packets are dumped under congested conditions. Otherwise routers do not discriminate between packets. This is said to make the system much easier to implement than is the case with *IntServ*.

prices) can in principle be provided. There are, however, no specified standards for the detailing of expected capacity profiles. This function is left open for ISPs enabling them to design their own service offering. The down-side of this is that without agreement and performance transparency between networks, the service would only operate “on-net”, i.e. it will not operate to any scale on the Internet. *DiffServ* has not therefore developed as an Internet architecture. It too appears limited in its use, mainly to private IP networks, although some large regional ISPs say they are using it in parts of their networks.

Development of pricing, accounting and billing systems to be used with *DiffServ* is necessary for service providers to build a value chain. Accounting architectures are currently being developed that support management of Internet resources. These architectures will also manage pricing and accounting of different classes of services and service levels.¹⁷ Issues that remain to be addressed are technical as well as strategic.¹⁸ In the last couple of years the IETF has been looking into accounting and billing systems.

Network designers appear to be looking elsewhere than *DiffServ* for a long-term solution to the problem of providing a real-time QoS on the Internet. The research focus seems to have moved to facilitating convergence between optical and data network layers (layers 2 and 3) under the concept of Packet over SONET (PoS) (see Figure 2.1).

3.3.3 QoS and ATM

At the beginning of 2002, most significant ISPs were using ATM to transport IP datagrams.¹⁹ IP datagrams are loaded into ATM cells for transport.²⁰ IP over ATM is an overlay model involving two different protocol architectures that were not originally designed to work with each other. However, ATM routing of IP has much improved price / performance compared to IP routing, although with technological progress this advantage may not last into the medium term.^{21,22} Other reasons for using ATM

¹⁷ Middleware technologies such as enhanced IP-multicast facilitate a new range of communication applications. See Internet Engineering Task Force: <http://www.ietf.org/html.charters/diffserv-charter.html>. A market based bandwidth management model for *DiffServ* networks with the implementation of bandwidth brokers has been proposed recently by Hwang, et al (2000).

¹⁸ Examples of technical issues are: what kind of accounting architectures should be developed for the next generation of Internet, and what type of middleware components are necessary. Strategic issues include the evolution of the Internet service portfolio, the influence of technologies and architectures on the opportunities for existing players and new entrants, the strategic importance of technologies and the development of alliances.

¹⁹ Some have started using MPLS or a similar technology, but most of these are likely to still be using ATM as the main ‘transport’ technology.

²⁰ ATM relies on routing at the edges and switching in the core, consistent with the modern approach to network design – “route once and switch many”.

²¹ IP packet headers contain the information which enables them to be forwarded over the network. IP routing is based on the destination of the packet, with the actual route being decided on a hop-by-hop basis. At each router the packet is forwarded depending on network load, such that the next

include its QoS advantages.²³

ATM requires the ATM adaptation layer (AAL) in order to link with upper layer protocols (see the ISO scheme - Figure 2-1). AAL converts packets into ATM cells and at the delivery end it does the contrary. Data comes down the protocol stack and receives an AAL header which fits inside the ATM payload. It enables ATM to accommodate the QoS requirements specified by the end-system.²⁴

A feature of ATM is that QoS statistics are predictable and measurable allowing ISP transit providers to offer service level agreements for connections that deliver a specified quality of service. Classes supported by UNI 4.0 are: constant bit rate (CBR); variable bit rate, real-time (VBR-rt); variable bit rate, non-real-time (VBR-nrt); available bit rate (ABR), and unspecified bit rate (UBR), the latter being recommended for the Internet.

In practice many of the QoS attributes of ATM are not readily available to ISPs as ATM must be used with other embedded protocols, and because protocols that link IP and ATM layers are complex and do not readily provide for the QoS attributes of ATM to be usefully deployed by IP over ATM. The development of application programming interfaces would have the effect of making the QoS attributes of ATM more accessible to end-systems running IP over ATM. This would increase the prospect of IP over ATM providing QoS features that are useful to end-users such as where ATM runs from desktop to desktop.

While 4 or 5 years ago, ATM was thought by many to be the means by which the next generation Internet would become a reality, it appears to be at the mature stage of its product life-cycle and this being the case its popularity is predicted to decline among large ISPs. Some large ISPs have already converting to MPLS although they are probably still using ATM for IP packet transport.

hop is not known with certainty prior to each router making this decision. This can result in packets that encapsulate a particular communication going via different routes to the same destination. This design means that packets arrive in different order than they are sent in, requiring buffering.

22 While IP is a packet oriented soft state (connectionless) technology located at layer 3 on the ISO scheme, ATM is a cell oriented hard state (connection oriented) technology located at layer 2 of the ISO scheme (see Figure 2-1).

23 This section draws mainly on Black (1999), Marcus (1999) and McDysen (2000); Kercheval (1997).

24 There are four AAL protocols: AAL1, AAL2, AAL3/4 and AAL5:

- AAL1: constant bit rate (suitable for video and voice);
- AAL2: variable length, low bit rate, delay sensitive (suitable for voice telephony and the fixed network part of GSM networks);
- AAL3/4: intended for connectionless and assured data service (not thought to cope well with lost or corrupt cells), and
- AAL5: intended for non assured data services, which is **recommended for IP** (others may be contracted for although I am told this does not happen for IP).

3.4 Technical constraints on the development of QoS and CoS services

3.4.1 QoS problems at the Internet's edges and within networks

Many service quality problems occur at points of congestion within transit ISP networks, and at borders with other networks. However, service quality problems also occur around the outer edges of the Internet. Arguably the main ones are:

- The relatively slow speed provided by most residential access lines;
- The bottleneck in the access network part e.g. in xDSL access, or the virtual connection between the DSLAM and the backbone network point, and
- Bottlenecks that occur within LANs and end-user ISPs, and in WAINs and at respective interconnection points.

More generally, there are several factors presently holding back the development of the convergence of the Internet with other platforms (e.g. the PSTN). These can be grouped into several overlapping categories:

- Congestion management *on* IP networks is not yet especially well developed, and often results in inadequate quality of service for some types of service, e.g. VoIP;²⁵
- The superior QoS or the offer of several service categories with different qualities, can not be retained *between* ISP networks due to technical reasons, such as software and even hardware incompatibility (an ISP's software/hardware may not support the QoS features provided by another ISP);
- There is a lack of accounting information systems able to provide the necessary measurement and billing between networks, such as would be required to support several levels of service quality that subscribers may select from depending on the type of communications service they are engaging in at that time;
- There is no interface with end-users that enables different CoSes to be chosen in a way that provides value to users, and

²⁵ Limitations in QoS values are caused either by processing in the host, at the network access interface, or inside the network. Hence, network connections must provide limited values for loss ratio, insertion rate, and delay and delay variation, in order to fulfil specific QoS service parameters. This holds generally, even for "best effort" service (even if it has no special QoS requirement), and for a network which is correctly dimensioned in order to avoid congestion.

According to McDysan (2000), a number of resource types may be a bottleneck in a communications network, the main ones being the following: transmission link capacity; router packet forwarding rate; specialised resource (e.g. tone receiver) availability; call processor rate, and buffer capacity.

- The quality of access networks is presently insufficient for QoS problems between backbones to be noticed by end-users under most circumstances.

These issues remain largely unresolved, although considerable effort is being undertaken to overcome them.

3.4.2 QoS problems at borders

When traffic is exchanged between ISP networks it becomes what is termed "off-net" traffic.²⁶ A range of QoS problems arise with off-net traffic.

The main off-net QoS problems appear to be explained by the following:

- (i) Where interconnecting networks use different vendor equipment and this equipment does not involve wholly standardised industry offerings, a number of problems tend to arise that impact on QoS. Network design structures that enhance QoS are lost, and management systems are not standardised.
- (ii) Service level agreements (SLAs) offered by transit providers are all different. The statistical properties of ISP networks are different and moreover are not readily comparable for reasons that include differences in the way this data is collected.
- (iii) The specifications of ATM's VBR service (used for Internet traffic) sometimes differ between networks, with the consequence that when traffic crosses borders QoS is not maintained.
- (iv) Equipment that is two or more years old is less likely to provide the QoS capabilities that are being offered by newer equipment.

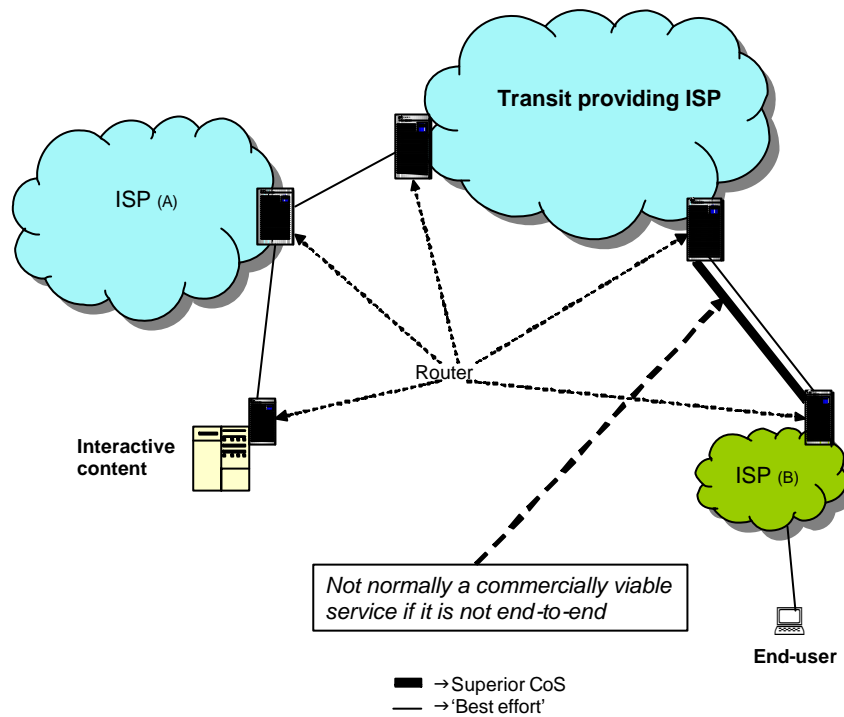
The upshot of these four types of problems is that degradation of QoS at borders is very common.

There is also reason to suspect that possible solutions to these QoS problems may be delayed due to a coordination problem. All networks that handle datagrams being sent between communicating hosts ('terminal equipment' in PSTN language), need to be able to retain the QoS parameters that are provided by the originating network if those QoS parameters are going to be actualised between the hosts or communicating parties. In other words, if one of the ISP networks involved in providing the communication imparts a lower QoS in its part of the network than is provided by the others, the QoS of the flow will be correspondingly reduced. The situation is shown in

²⁶ On-Net traffic means traffic that is interchanged between hosts connected to the same AS and hence routed with an interior gateway routing protocol (IGP), in contrast to Off-Net traffic which is routed between different ASes by an exterior gateway routing protocol (EGP).

Figure 3-4. In this regard individual networks may be reluctant to invest in higher QoS without there being some way of coordinating such an upgrade with others in the chain.

Figure 3-4: Coordination and end-to-end superior service quality



4 Quality of service: pricing and congestion

In this section we look at the relevant economic factors relating to QoS. This includes a discussion of possible roles for pricing and demand management in improving QoS as well a discussion about claims that technological developments will avoid the need for demand management; in particular, that cheap bandwidth and faster processing power will overcome congestion (i.e. scarcity).

The basis of the arguments outlined in this section will be useful in understanding the research papers that are analysed in Annex II below.

4.1 The Internet is not an economic 'public good'

The discussion in this section is based on the condition that congestion management can not be satisfactorily accomplished with technology alone, such as by finding more cost effective ways to utilising existing capacity, or by "throwing bandwidth" at the problem. A general defence of this condition, especially in regard to over-engineering the Internet, can be found in Annex I. However, many of the specific problems entailed in this approach are discussed immediately below.

Given the condition above, technological development of the Internet should have as one of its goals, to allow for mechanisms through which demand management can function. Very generally, the way this would occur is by increasing the price of service at congested periods (e.g. a price in terms of bits of throughput). This would potentially offer even more advantages where it applied to several different CoSes. At present technology does not enable congestion pricing or CoS pricing to occur. As has been outlined above, arguably the main for this is that QoS is often not maintained for traffic passing between networks, thus undermining the development of accounting and billing systems, and the development of customer interface software that would support congestion pricing or a system of several categories of service quality.

Under present Internet technology packets are accepted by connected networks without specific guarantee (although SLAs typically provide compensation where statistical 'guarantees' are breached) and on an essentially "best effort" basis. As such, packets carrying e-mail are treated the same as packets carrying an ongoing conversation.²⁷ From the perspective of demand management, this equal treatment of packets according to best efforts is problematic on at least two counts:

²⁷ After the adoption of ATM by large IP backbones it became feasible for them to offer QoS statistics in their transit contracts.

1. There is a lack of incentive for networks to provide a service that would enable real-time service provision, as there is no combined business and technical model that would enable them to meet the higher costs involved in providing real-time service quality, and
2. There no account is taken of end-users' different demands for service quality; even for a certain type of application (e.g. video conferencing), customers will have quite different demands at any particular time.

Both of these problems can in part be addressed through *congestion pricing*. In the remainder of this section we discuss in more detail what implications the lack of congestion pricing might have for the future development of the Internet, and what if any are the policy implications. In the section that follows this one we discuss optional categories of service quality (CoS).

4.2 Prices and cost structures

To be economically efficient the structure of the prices offered to users should ideally match the structure of the costs users cause; that is, the way costs are caused should be reflecting in the way liability is incurred by the customer.

The main classes of relevant costs involved in the provision of Internet services are explained as follows:

- (i) Building an Internetwork involves fixed costs that do not vary with network usage. There are also development costs which are not incremental to single customers (such as software development). Flat rate pricing is the efficient way of recovering these costs. However, such costs can not be said to be incremental to any single customer, and so the most efficient flat rate pricing would involve different prices being charged to each subscriber, with those with strong demand paying more than those with weak demand.²⁸ The idea here is that no one should be excluded by the prices which are intended to recover the costs of providing the basic network and software.²⁹

²⁸ Indeed, if the seller's overall prices are constrained so that she makes only a reasonable return on her investments, the most cost effective access prices involve prices being set according to the inverse elasticity of demand of each subscriber, with the condition that no person should be charged a subscription price more than their willingness to pay. For obvious reasons this has rather more theoretical than practical application. This type of pricing is commonly referred to as Ramsey Pricing, a full discussion of which can be found in Brown and Sibley (1986).

²⁹ Remember that the cost of customer access (the local loop) should already be met by connection and subscription charges levied by the access operator. Where leased lines or xDSL is used there will be some additional costs that are caused by the subscriber.

In practice such prices do not normally vary in order to include customers with weak demand, and such customers may therefore have to find other ways of accessing the Internet.

- (ii) There is also an initial cost for an ISP in connecting a customer to the Internet. These are mainly administrative costs, and as they occur on a one-off basis they should be charged in the same way if pricing is to be most efficient.

As there is a positive incremental cost involved with each person's subscription, this will make up a one-off connection fee per subscriber, together with a return on any incremental capital associated with these costs.

- (iii) As the Internet becomes congested there is a marginal cost incurred when extra packets are sent, and this includes the delay experienced by all users at that time. To avoiding the congestion externality cost this implies, marginal cost pricing would have the following attributes. It would:

- a) encourage users who have relatively weak demand during congested periods (e.g. a low willingness-to-pay) to shifting their demand to uncongested periods) and,
- b) send a signal in the form of additional marginal revenues to ISPs to invest in more capacity when there is significant congestion.³⁰

If a price can be charged which is greater than the margin cost of delay, the general rule is that it will pay the network to increase capacity.³¹

At present the structure of prices paid by Internet subscribers bears little relationship to the way costs are caused. Obtaining Internet service today will involve payment for one or more of the following elements:

- A periodic subscription fee which may include an upper limit in terms of Gbit per month delivered, before the subscription charge jumps to a higher level;
- DSL / leased line installation charge;
- Leased line / DSL rental, and
- Per minute / second charge for dial-up calls to the ISP, levied by the access network provider (typical the incumbent telecoms operator).

30 Crucially this is dependent on pricing structures between ISPs, a matter we have assumed thus far to be unproblematic.

31 Because capacity investment tends to be 'lumpy' the rule often needs to be qualified by the following consideration: the incremental revenue earned must be more than the cost of building a certain incremental increase in capacity.

Clearly the later bullet is the only one that implies marginal usage costs for Internet users.³²

There is political pressure in many countries to make Internet access available which does not entail any marginal usage charges. Two types of service that presently provide this feature are broadband services (mainly ADSL), and FRIACO. Already some governments have invested large sums in order to expand the availability of ADSL services. It seems likely that in terms of absolute numbers or in terms of the proportion of all Internet subscriptions, unmetered Internet usage will become more common in the future.³³ During periods when the Internet is congested this is the opposite to what is needed if the Internet is to converge with the traditional platforms, such as the PSTN, CATV and free-to-air TV (FTATV).³⁴ The migration of small businesses and residential users to ADSL for which there is no usage sensitive pricing³⁵ will increase traffic on the ATM access connection and also on the Internet, and *ceteris paribus* will also tend to increase congestion on the Internet.^{36,37}

Where there is no limitation on subscriber numbers and users face no marginal usage costs, the Internet is being treated much like a public good. Pure public goods are not depleted with use; i.e. my usage of it does not effect the enjoyment you get from using it. This is clearly not the case with the Internet and we should therefore expect it to exhibit similar problems that plague those services that are treated as public goods, but are in fact not. These problems are popularly referred to as *the tragedy of the commons*, a problem which occurred when livestock farmers were allowed free access to common land on which they could graze their animals. The farmers took advantage of this offer with each one of them failing to recognise the effect that (apparently free) grazing of their animals was having on the ability of the others to do likewise.³⁸ The

32 ISPs typically only charge a subscription fee to Internet end-users, although in some cases the ISP does not charge the end user at all, but rather shares with the access provider the per call-minute revenues received from dial-up sessions, i.e. the price levied on the customer for the 'telephone' call to the ISP. This is sometimes referred to as the 'free' ISP business model. The reason the ISP can share in these revenues is that the call price is in excess of the access provider's costs in providing that service, which on average has lower costs than a normal telephone call, although it is charged at the same rate.

33 Businesses that have leased line access to their ISP already avoid usage sensitive prices.

34 The ADSL modem splits the information into voice and data, with only voice being allowed to enter the access provider's switch. Internet data is being directed to the ISP normally over an ATM access connection between the ADSL modem pool "DSLAM" and the first point-of-presence (PoP) of the Internet. In this regard ADSL provides an "always on" Internet access service.

35 Some ISPs have a step function in the price charged between 2 or more levels of usage (e.g. bits per month), but the level of usage in each category is so large that no marginal usage costs exist for the vast majority of users.

36 The core Internet is protected against overloading due to the limitation in the ATM access connection where most regional operators do not guarantee more than 10% of the peak capacity of the ADSL speed. This means that as more users share the ATM access connection actual capacity experienced declines.

37 Under such pricing arrangements end-users tend to be grouped together such that pricing tends to result in low users subsidising high users.

38 This phenomenon is also known as an externality cost. The tragedy of the commons is the most common form of market failure. For example, it explains pollution, global warming, and natural resource depletion.

result was over-grazing such that the commons became quite unsuitable for grazing by anyone.^{39,40}

In the case of the Internet the lack of an economic mechanism for congestion management results in a degradation of service quality for everyone. Improvements in software, hardware, and the declining cost of capacity have, however, provided all of us with level of service for traditional Internet services that is tolerable on most occasions. But by applying the same network resources to all packets, Internet networks are unable to provide real-time services over the Internet which approach the quality needed for mass market take-up.

Usage based pricing can in principle be designed to shift some demand from peak periods to other times, and can also signal to ISPs when demand is such as to make it economic for them to increase the capacity of their networks. The idea is that customers should ration their own usage during periods of congestion according to the relative strengths of each user's demand. For users with very weak demand (say, a willingness to pay for service during a congested period of zero, assuming they can use it during uncongested periods at no marginal cost to themselves), there is little benefit obtained by the user compared to the costs imposed. At times of congestion, however, the cost of sending extra packets would include the additional delay, packet loss and QoS degradation imposed on other users.⁴¹

When the Internet is uncongested, usage-based pricing is not helpful at all; it actually has a detrimental effect on economic welfare. At these times the cost of sending an additional number of packets is virtually zero. We say that the marginal cost of usage is zero, and it is a demonstrable economic axiom that under these circumstances a usage sensitive price is inefficient – it reduces economic welfare – flat rate pricing is optimal.

With the telephone network, time-of-day has been used as an element of pricing. Subscription charges are the flat rate charge, with usage being charged on a per minute (or per second) basis, and typically varying according to the time of day. The idea with time-of-day pricing is to dissuade callers with low demand from usage during the most congested period, encouraging them to shift their usage to a period when per minute charges are much lower. This is optimal because the capacity investments costs

39 A similar problem has occurred with global warming, the loss of biodiversity, and the depletion of natural resources such as fish stocks. If usage remains 'free' and the resource can be depleted or over-used, then the rule is that either the numbers of users have to be rationed, or the total amount of usage must be restricted if the resource is to remain viable.

40 Quotas are a common approach to addressing these types of problems, and where trading in quotas is permitted, this tends to result in improved efficiency within the industry. Unfortunately quota numbers tend to be difficult to police, resulting in illegal over-usage. Quotas can also be systematically over-provided where quota numbers are not strictly set according scientific data, but are subject to political compromise.

41 Moreover, where there is no system that enables end-users to purchase a QoS they demand, it could be argued that there is also a cost associated with the absence of a market for real-time services, given that these would develop if a marginal cost pricing schemes operated.

required to handle the traffic from subscribers with weak demand during peak usage, are higher than the present value of their willingness-to-pay (WTP) for the capacity needed to satisfy that demand.

Both time-of-day and call minute/second charges appear to have less relevance for the Internet than for the PSTN. One reason for this is that peak usage of the Internet is thought to be less stable in time, especially in regard to end user ISPs, and thus time-of-day pricing may not provide a fully effective means of congestion management. It suggests that an attempt to raise session or usage prices at a particular time of day when the Internet is most congested may miss periods of congestion.

Another problem with time-of-day pricing is that the Internet is made up of a great many networks, and even in the same time zone peak usage may well occur at different times in different places. Moreover, an ISP providing transit to several ISPs, some of which have rather different peak usage times, suggests that different prices would apply at the same time of day to ISPs that are in the same market competing (on the margin) with other ISPs, even if their traffic / time patterns are not the same. This may raise competition neutrality concerns.

A further problem is that unlike a switched circuit, which is rented exclusively by the paying party for the duration of a call, packets of data on the Internet share capacity with other packets such that costs are packet more than time related. Some account could be made for this by pricing Internet usage on the time by bandwidth basis. Compared to a packet-based marginal pricing system, a proxy based on time-of-usage prices by access bandwidth will have significantly depleted efficiency advantages relative to a packet-based marginal pricing system.

An elegant and potentially highly efficient solution to the marginal cost pricing problem with the Internet has been described by Mackie-Mason and Varian (1995) (M-V), and referred to as the "smart market". From our perspective, the main attribute of M-V's contribution is its pedagogic value in setting out the economic problems, in part through the solution M-V propose. Theirs is more a description of what a near ideal solution would look like, rather than being a practical solution to congestion management (at least not practical under present technology).⁴² It is a solution discussed for one CoS, but it could operate for any number of CoSes.

M-V's scheme would impose a congestion price during congested periods which would be determined by a real-time *Vickrey auction*. The way this would work is that end-users would communicate a bid price for their packets just prior to beginning their session. The Vickrey auction design is known to provide a strong incentive for all end-users to communicate their maximum willingness-to-pay for the item, (in this case

⁴² The ideal solution would involve dynamic price determination, in which prices differed across subscribers and changed continuously to reflect the 'state of the system'. The issue of optimal pricing and near optimal pricing is addressed in Annex II.

outgoing and more importantly, returning packets), i.e. it provides "the right incentives for *truthful revelation*".⁴³ This is because under the Vickrey auction design the price actually charged to any end-user is not the price each person bids, but is the price bid by the marginal user – the last person to be granted access to the Internet under the congestion restriction, i.e. the market clearing price.⁴⁴ All users admitted to the Internet during this period pay the same price. Those end-users with a willingness to pay which is less than the market clearing price would not obtain access at that time, and would have to try again later. When the Internet was uncongested all bidders would be admitted and the price charge would be zero.

An additional attraction of the "smart market" is that under competitive conditions it provides correct economic signals for ISPs to increase capacity. This would occur when the marginal revenues from admitting further users at peak usage are greater than the marginal cost of adding capacity, thus communicating a profitable investment opportunity. Network capacity will thus be maintained so that marginal revenue equals marginal cost, which is the most economically efficient outcome.

The smart market may lack practically, but its economic attributes need to be understood by the Internet network design community in order that economically informed choices are made between technological options.

The M-V solution was published in the mid 1990s, and while this type of auction still has relevance for congestion management on the Internet, there have been many technical developments which have diverted interest away from the smart-market solution. Perhaps most significantly, technological developments will enable packets to be treated differently such that there may be multiple virtual Internets each with different QoS attributes, with packets being tagged according to the class of service (CoS) they would receive.

4.3 Class of service pricing

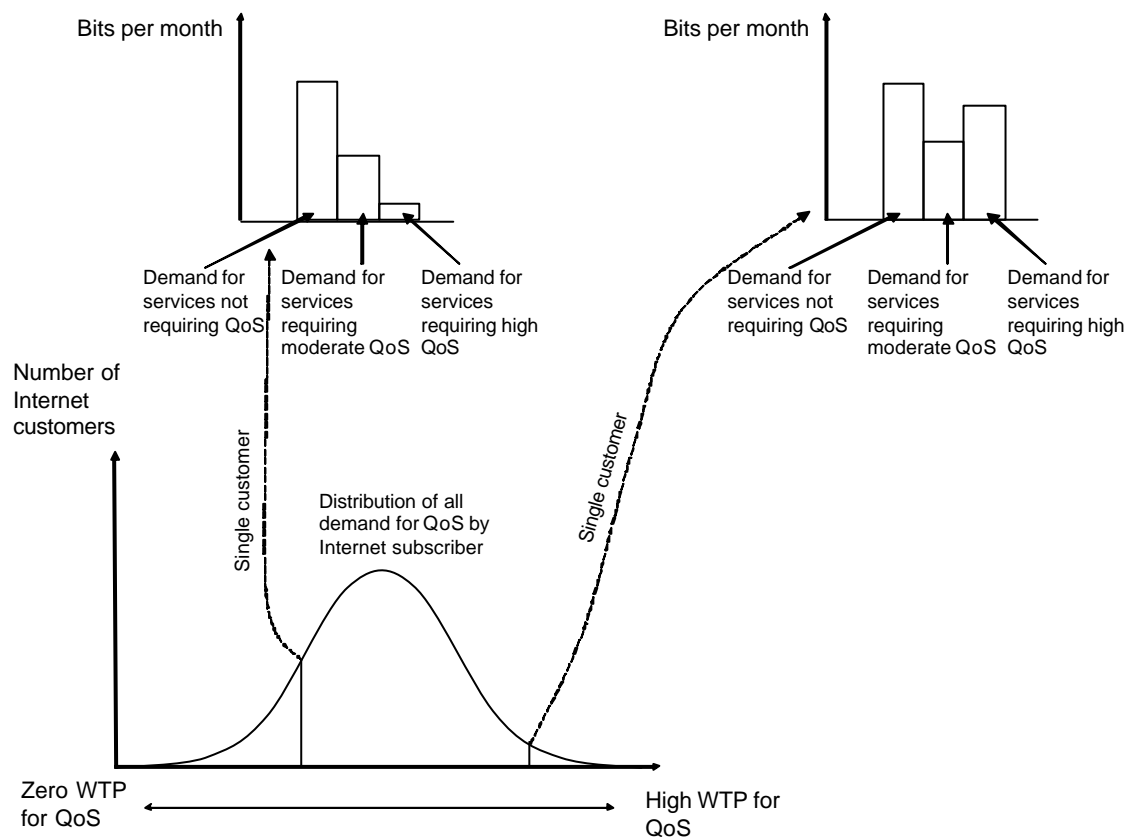
While M-V do not discuss different CoSes, their analysis is applicable within a CoS environment. Users would simply specify a CoS in addition to their bid price. Further conditions controlling the passing of demand between different CoSes could be specified by bidders if they preferred to be dropped to a lower CoS in the event that their bid for service with particular QoS characteristics was lower than the market clearing price. This may involve users having to input a couple of additional parameters if they wished to be considered for other CoS categories.

⁴³ In 1996 William Vickrey received the Nobel Prize in Economics for his early 1960s work on the economic theory of incentives under asymmetric information.

⁴⁴ The one exception to this statement is the marginal user whose WTP equals the market clearing price.

In a CoS internet World, ISPs might tag packets according to the CoS indicated by their customers. A system that enabled end-users to select among several different CoSes for their session datagrams would result in their ISP billing them according to the numbers and types of tagged packets sent. By itself such a system would not be ideal if it meant that ISPs would still have to work without actual marginal cost prices. A subscriber may pay a premium to use the higher CoS possibly based in bits of usage as well as a monthly subscription, but she could then be able to send all her packets during peak usage, or perhaps at other times – there would be no difference in the price she would pay. This means that for each CoS there would be no explicit mechanism aligning the demand for the service at congested periods with ISPs' incentives to invest in capacity, making congestion avoidance by the ISP difficult, even though average revenues on such a network may be more than high enough to cover average costs. As the higher CoS service would be sold as a real-time premium quality service, there would be an incentive to maintain that quality, but over-provisioning would still be required due to the lack of marginal cost price signals. Such solutions to the congestion problem would not therefore be optimal, although they may offer sufficient refinement to enable CoS development and the widespread provision of real-time services on the Internet.

Figure 4-1: Demand for Internet service deconstructed



Notes: WTP → willingness to pay.
Source: WIK-Consult.

The bulk of revenues that pays for the Internet come from end-users (organisations and households). Where for a certain communications purpose only one service class is offered and all demands are treated equally. But each end-user's demand will in fact be made up of untapped demand for various service classes, depending on such things as the communications purpose, application requested, and preferences that may only be known to each end-user. One truth we can be sure of is that the demand for multiple service classes will be *derived* from the demands of end-users. ISPs will be keeping this in mind as these service class developments materialise.

We show this situation in Figure 4-1. The lower distribution captures all Internet customers, from those specialised organisations that have mission critical services that require both high admission probability and QoS values as well as needing traditional e-mail and browsing services, to those who only use the Internet to send non urgent messages. However, most customers who make up this distribution can be expected to use the Internet for several different purposes, and to consume several different services, such as e-mail, file transfer, WWW, and video streaming. Differences in an individual's demand for CoSes will depend in part on the application and purpose of the communication, as indicated by the bar graphs in Figure 4-1.

The main problem currently with VoIP and real-time interactive services is not that networks can not provide the QoS statistics needed, but rather, it appears that only a limited number do and these tend to be private IP networks i.e. intranets. ISPs providing public Internet services have too little incentive to develop VoIP services, especially where packets containing voice conversations pass over other networks.

4.4 Pricing and QoS between ISPs and their transit provider

Pricing between the transit provider of an ISP serving a local VoIP provider, will be crucial in determining the pricing between the ISP and VoIP provider, and the pricing of the VoIP providers and its customers. As well as the importance of the level of prices, the structure of those prices charged by a transit ISP will tend to be reflected in the prices charged by the others. In this section we discuss transit price structures, and QoS guaranteed contracted by transit providers.

4.4.1 The structure of settlement prices

As has been noted already, with rare exceptions interconnection settlement between ISPs that have a peering arrangement do not involve payment. The settlement model is a sender-keeps-all arrangement. Peering arrangements exist right through the hierarchy of the Internet but for most ISPs peering will make up a much smaller proportion of their interconnected traffic than will transit.

As with peering, transit contracts are confidential. However, by speaking to numbers of

relatively senior people in various ISPs enough information has been obtained to provide a picture of the structure of settlement prices for transit, although not surprisingly, the information is fairly general and says nothing about price levels.⁴⁵

There are several possible charging arrangements for transit. The end-user's ISP (or online service providers – OSPs) could pay as the receiver of transit traffic), the Web hosting ISP (the firm sending the requested data) could pay, or both ISPs could pay the transit provider.

In practice, transit is typically charged on a *return traffic* basis, i.e. on the basis of the traffic handed over to the ISP whose customer requests the information. ISPs that provide transiting (mainly large ISPs and IBPs) charge on the basis that traffic flows from themselves to their ISP customers. Transit providers do not pay anything to their ISP customers even though they receive traffic from them, albeit much less than the traffic flowing from transit providers to customers. While this may not seem very equitable at first glance, present transit charging arrangements have some economic advantages. Not least of these is that it is the largest flow which tends to dictate the network capacity needed, especially at points of interconnection. As most transited packets flow from Web hosting ISPs through transit to another ISP to online service providers, it is this traffic that appears to give rise to congestion and governs the investment needs of ISPs that provide transit.

In analysing transit pricing arrangements it is useful to do so in terms of the quality of the economic signals the prices send to the parties involved, especially concerning investment, congestion management, usage, and competition between transit providers. In practice, however, this is made difficult and prone to error as the information is not publicly available and information that is provided verbally tends to be quite general.

The available information suggests that there is no accepted industry model that governs the structure of these prices. Some larger ISPs are able to negotiate a price structure with the transit provider, while others choose from a menu. There appear to be three basic dimensions around which transit price offers are structured:

- A fixed rate for a certain number of bits per month;
- A variable rate for bits in excess of this amount, and
- A rate based on peak throughput, which may include:
 - pipe size, representing the *option* for peak throughput, and
 - some measure of *actual* peak throughput ('burstiness').

⁴⁵ The ISP I had contact with include: MCI/WorldCom, Cable & Wireless, and Genuity.

Two part-tariffs appear standard where the fixed charge may be relatively low per bit compared to the variable component.⁴⁶ To the extent that ISPs can accurately estimate their future monthly usage, such arrangements allow ISPs to pay transit charges in the form of a predetermined monthly charge, any extra bits being charged a premium. Premiums may be high but quite possibly in keeping with the transit providers costs in making this extra capacity available for peak demand.

However, it is understood that some transit buyers pay a flat rate only option. The rationale for flat rate option is that it provides certainty to network customers who have an annual budget to spend on communications and who prefer the riskless option of paying a certain amount known in advance for all their traffic requirements. We would expect that for such customers their overall transit costs to be higher than they would be under a two part tariff, as they have effectively rejected any pricing component that would restrict their peak demand.

It is common for larger customers to negotiate specific details according to their particular requirements. Large content providers that maintain their own router, and many ISPs (all of whom do likewise) will frequently have interconnection arrangements with more than one transit provider.

Where the non-usage price makes up a low proportion of an ISP's monthly transit bill this price structure can enable the transit buyers' to bargain for better deals from transit providers. Pricing like this can make multi-homing a more effective policy for ISPs and large content providers, as in addition to a small pipe-size based charge, ISPs and content providers will only pay for the packets they send to their IBP transit provider. Thus, the ISP could choose to send all of its traffic via the transit provider that is providing the best price/QoS, but retains the *option* to switch its traffic to the other IBP should its price/QoS offer become superior, or should an outage occur on the IBP's network the ISP is currently using for transit. In short, this arrangement appears to provide a valuable option to switch between IBPs which multi-homed ISPs or content providers may not be paying for.⁴⁷

The flat rate price structure is thus a take-or-pay arrangement which detracts from the ability of ISPs and content providers to play off transit selling ISPs against each other over the period of the contract. For some firms that take the flat rate option, however, it can meet their needs for revenue certainty over the duration of the contract.

It seems to the author that there are reasons for ISPs selling transit to prefer a prominent role for base-load and optional capacity pricing, with the inclusion of some type of payment for the peak capacity option, like pipe size. A price that is also based on the variability of traffic throughput would enable those transit buyers who send a

⁴⁶ Routers keep a record of traffic statistics, i.e. there are counters in the router (port).

⁴⁷ In many markets such options are purchased directly. Indeed, in some cases there are markets in which options are bought and sold.

relatively constant bit rate to receive a lower price in keeping with their relatively greater reliance on base load capacity rather than peak load capacity.⁴⁸

There is no indication of any pricing for CoS, presumably as these services appear not to operate over the public Internet, an issue we address in more detail below

4.4.2 QoS guarantees for transit

Transit selling ISPs tend to offer QoS guarantees which usually address three QoS dimensions: latency, packet loss, and service availability. Transit selling ISPs keep the statistical data necessary to verify their own QoS and provide periodic reports to clients. Any breach of QoS parameters must be confirmed by the transit provider's own data. Contracts that require 100% availability are apparently the norm today due to competition, although obviously it will not be met in reality, so very occasionally transit providers will have to pay agreed compensation in cases of outage.

The ability of transit providers to start offering service level agreements (SLAs) with stated QoS parameters coincided with operators' use of ATM in their transport networks. With this technology the corresponding IP packets are transmitted over different 'virtual tubes', referred to in ATM terminology as virtual paths (VP). However, QoS guarantees only apply if the flow of cells received conform to the traffic parameters that have been negotiated. Such conditions require networks to shape their own traffic at the border, just before handing over for delivery to the transit provider.⁴⁹

The reason transit services do not include CoS options are as follows:

- There is presently no workable business model that enables end users to select different CoSes depending on the application they wish to use, or more generally, on their individual demands for different QoSes.
- There are no billing systems operating that would enable higher prices to be charged for higher CoS options.
- There are QoS problems at borders such as those concerning standards, equipment, and management interfaces which restrict the ability of CoS options to operate across different networks .

⁴⁸ One European ISP said in an interview that transit prices had dropped by 90% in the three years to March 2000. Another said that in Eastern Europe they had dropped by 50% between March and October 2001.

⁴⁹ Annex III contains a description of the QoS tributes of ATM networks.

4.5 QoS and the Next Generation Internet

The Internet presently provides a number of different services to end-users and the range of services seems likely to become greater in future. The Internet is converging with traditional platforms over which services like point-to-point telecommunications services, point-to-multipoint conference services and multicast and broadcast distribution services like video streaming and TV, have been provided. Note, that two-way CATV networks are already providing the integration of traditional point-to-point call services like voice telephony with TV broadcast distribution and pay-per-view video services and classical Internet services like e-mail and WWW access.

In an IP network all of this information can potentially be organised into packets or cells (datagrams) and transmitted over the Internet, although as provided through the Internet, the experience of consumers with at least some of these services is typically of relatively low quality in comparison with the service quality provided by relevant legacy platforms.⁵⁰

The provision of different CoS is considered one of the key ingredients needed for the next generation Internet to become a reality. By this it is meant the ability to provide the speed and reliability of packet delivery needed for services like VoIP and interactive services, to be provided over the Internet to a quality that enables mass market uptake. Indeed, the Next Generation Internet is defined as a network of networks integrating a large number of services and competing in markets for telephony and real-time broadcast and interactive data and video, in addition to those services traditionally provided by ISPs.

Although many of the specific technologies that are discussed in this report are not yet fully developed, several offer the prospect that high quality real-time services could be commonly provided over the Internet in the medium term. Actual business solutions that rely on these technologies are yet to fully develop,⁵¹ however, due in part to the highly diverse nature of the Internet, and the service quality and lack of seamlessness problems which have been discussed earlier in the report.⁵²

Before convergence can be said to have occurred, there will be a transition period during which real-time services such as VoIP, begin to put real competitive pressure on legacy PSTN providers, and in this regard it is interesting to think about how QoS on this transitional Internet will differ to what is provided today.

50 Exceptions do arise, such as on intranets where network designers are better able to address end-to-end QoS.

51 See the various articles in the special issue "Next Generation Now", of *Alcatel Telecommunication Review* (2001).

52 See Keagy (2000) for more details. Ingenious developments exist, however, which take advantage of the present state of Internet QoS. ITXC, for example, provides VoIP service using software that enables them to search different ISP networks for the best QoS being offered at that time. Where no QoS is available that would provide acceptable quality, calls are routed over the PSTN. See <http://www.itxc.com/>

One of the main transitional problems over the next few years may have less to do with the quality of real-time sessions, but with service admission control which can enable over-loading of packets in the IP datagram network to be avoided during congested periods. In traditional networks like the PSTN or switched frame relay networks where capacity is assigned during the connection admission phase, the blocking probability for a service is described by known probabilities and these define the GoS that customers will receive. As the main capacity bottleneck inside the network lies in the access part of the network, service admission control is likely to be mainly limited to these areas of the network.

In multi service networks, as the NGI service admission control under GoS values are not described in the same way as for traditional networks, but through more sophisticated models and algorithms.⁵³ Without the application of effective demand management considerable over-provisioning will be required if large numbers of users of real-time services (especially VoIP) are not to experience instances of network unavailability that are too frequent for them to tolerate. What may happen in this transitional Internet is that subscribers who are sensitive to service availability will remain with the PSTN for much longer than subscribers who are more price sensitive and who do not mind facing a significant probability that when they attempt to make a call they will be denied admission and will have to try again after a short period.

Currently there are different possibilities for provide QoS on IP Networks and these include MPLS, *DiffServ*, and Ipv6, or perhaps most easily the ToS octet in the Ipv4 header for the definition and recognition of a traffic hierarchy. For the Next Generation Internet five traffic levels are envisaged which are shown in Table 4-1.

Table 4-1: Traffic hierarchies in next generation networks

Traffic level	Traffic type	Service example
NJ4	Traffic for OAM and signalling functions	Network or connection Monitoring
NJ3	Real time bi-directional traffic	Voice and video communication
NJ2	Real time uni-directional traffic	Audio Video streaming, TV distribution
NJ1	Guaranteed data traffic	Retrieval services
NJ0	Non guaranteed data traffic	Best effort information service

Source: Melian et. al. (2002)

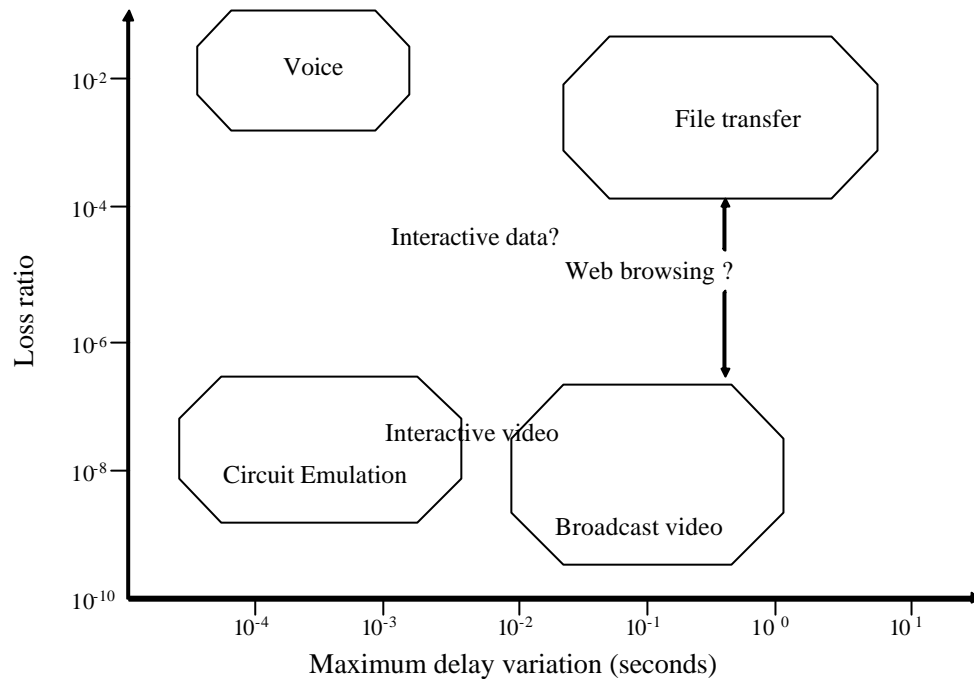
Excluding NJ4, which is intended for internal network use, the four different QoS identified in Table 4-1 should allow all the services identified in Figure 3-1 to fit relatively

⁵³ See Ross (1995).

well into at least one of the four QoS options shown. **Error! Not a valid bookmark self-reference.** suggests what these might look like.

According to the options identified by Table 4-1, when a user initiates a session she would have to pay a tariff corresponding to the service class. The price will decrease, from NJ3 to NJ0. In cases where the network does not have sufficient capacity for the required service the user may chose to default to a lower CoS.

Figure 4-2: Fitting CoSes within service QoS requirements



4.6 Conclusion regarding QoS and CoSes

There are several factors presently holding back the development of the Internet into an integrated services network. These can be grouped into overlapping categories:⁵⁴

- Congestion management *on* IP networks is not well developed, and often results in inadequate quality of service for some types of service, e.g. VoIP;⁵⁵

⁵⁴ As this study concerns the Internet backbone, we do not address issues relating per se to customer access.

⁵⁵ According to McDysan (2000), a number of resource types may be a bottleneck in a communications network, the main ones being the following: transmission link capacity; router packet forwarding rate; specialised resource (e.g. tone receiver) availability; call processor rate, and buffer capacity.

- QoS attributes are often not retained *between* ISPs due to technical reasons, such as software and even hardware incompatibility (an ISP's software/hardware may not support the QoS features provided by another ISP);
- There is a lack of accounting information systems able to provide the necessary measurement and billing between networks;
- There is no interface with end-users that enables different CoSes to be chosen in a way that provides value to users, and
- The quality of access networks is in many cases insufficient to enable the provision of next generation Internet services.

These issues remain largely unresolved, although considerable effort is being undertaken to overcome them. ⁵⁶

Even between ISPs, interconnection pricing is not configured so as to enable pricing to work as a congestion management tool. Peering arrangements do not involve explicit pricing such that there can be no congestion prices or prices that reflect marginal costs. The situation is less inefficient in the case of transit interconnection as prices are typically quoted according several parts, these being:

- a fixed rate for a certain number of bits per month,
- a variable rate for bits in excess of this amount, and
- possible also an amount concerning peak throughput.

This price structure will assist transit ISP in deciding on the level of investment in network capacity in order to avoid chronic congestion problems. Prices structured in this way do not, however, enable pricing to be used as a more active instrument for congestion management, i.e. customers are all free to use the Internet at it most congested period for no *additional price*.

56 Note that jitter is the primary impediment to transmitting VoIP over the Internet. A typical VoIP call over the Internet would traverse many different networks, with widely varying latency and QoS management. As a result, VoIP over the public Internet results in poor quality and is typically discouraged by VoIP vendors. Nevertheless, many software applications exist to provide free voice services over the Internet. The common characteristic of these Voice over Internet systems is very large receive buffers, which can add more than 1 second of delay to voice calls. Free voice is attractive, but to business users, the poor quality means that these systems are worthless. However, some residential users are finding them adequate—especially for bypassing international toll charges

5 Present and future of “real-time” IP service

5.1 Existing VoIP

Today there are firms offering VoIP service. In general the quality is poor and the service unreliable. However, reasonable quality is obtainable on some occasions and where this occurs it seems likely that VoIP providers are using some combination of the following:

- interoperability with the PSTN through SIP and H.323 terminals and protocol groups, and the use of compression technologies;⁵⁷
- technical methods which keep datagrams on-net;
- dynamic assessment of the QoS on different parts of an ISP transit network, or perhaps assessment of QoS on several transit networks which the VoIP service provider has contracts with, such that calls are routed where QoS is best at that moment, and
- for computer to phone VoIP, computers connected via the PSTN to UDP ports, which imply some QoS differences compared to TCP (greater packet loss but lower mean delivery times) which is favourable to VoIP.

Firstly, we look at real-time IP services today – principally VoIP.⁵⁸ Broadly speaking, there appear to be three categories of business venture that are relevant:

1. Those that are selling an integrated IP solution to the internal electronic communication needs of individual companies;
2. Those that are selling phone to phone VoIP services to the public, and
3. Those that are selling computer to phone VoIP services to the public.

The first point does not appear to be directly relevant to this study. For one thing these services are not priced to the end-users.⁵⁹ The discussion which follows is therefore oriented toward the second and third options.

⁵⁷ An H.323 terminal is an end-user communications device that enables real-time communications with other H.323 endpoints. A gateway provides the interconnection between a H.323 network and other types of networks such as the PSTN. SIP was developed by the IETF in 1999. It is a protocol for establishing, routing, modifying and terminating communications sessions over IP networks.

⁵⁸ Note that video-streaming requires end-users to have access bandwidths that are much greater than those available to a majority of subscribers. Moreover, video can be buffered without significantly affecting the service quality, and thus it has less than real-time QoS requirements.

We begin by discussing the mechanics of a VoIP service and then move on to analyse the commercial case which is essentially an international Accounting Rate by-pass service.

Note that most of the discussion and documentation about real-time IP services is directed toward the first of these two categories. The second category takes up relatively little of this discussion and documentation.

5.1.1 Public IP telephony on Proprietary IP networks

ISPs that provide transit for IP communications tend to stretch across international boundaries, i.e. they are not defined by country borders as is common with incumbent PSTN operators. ISP networks are not based on a traditional PSTN configuration where national operators in a country connect with the outside World in international “no man’s land” i.e. using the concepts of half circuits. This enables an ISP in Country *A*, e.g. Sprint, to connect through a gateway directly with the PSTN in Country *B* – say, somewhere in Africa. In this way the ISP can operate a Wide Area Network (WAN) between two countries, or due to cost and reliability issues it may be a Wide Area Ethernet Network that is deployed. With the placement of interface devices and software⁶⁰ in countries *A* and *B* which transform messages originated over the PSTN into IP, and for incoming messages, convert the information from IP back into a form that can be handled by the PSTN, the ISP is potentially able to offer a phone to phone VoIP service. The situation is shown diagrammatically in Figure 5-1.

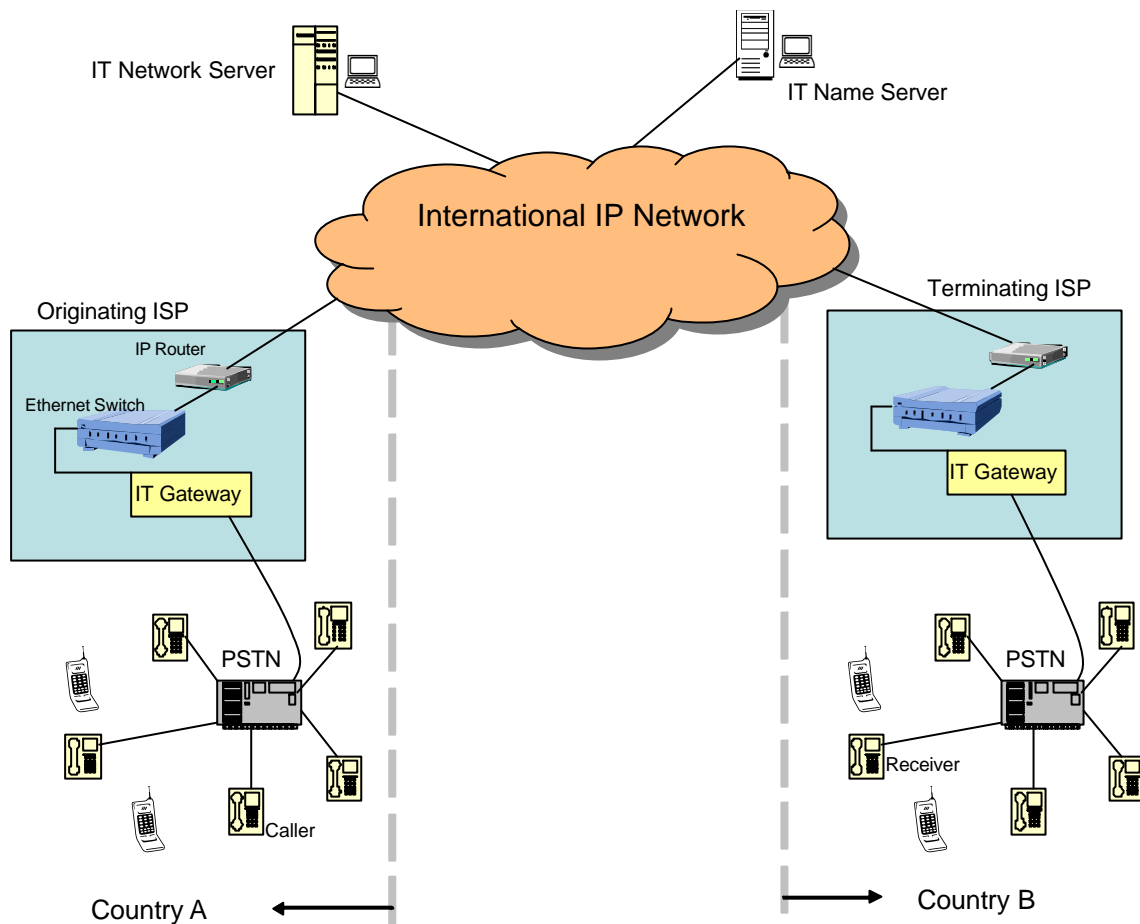
It might be argued that this service falls slightly short of what we might call “voice over the *public* Internet” as there is most likely only one IP network between the two countries over which the conversation is sent. It is also possible for firms to rent parts of the network which are configured to provide the renting firm with its own virtual network. As we have explained, where IP packets cross networks service quality (e.g. packet loss and delay) is frequently too poor for a VoIP service to operate. The upshot is that most VoIP is provided using a single transit ISP’s network.

The type of technical arrangement shown in Figure 5-1 is known as a Wide Area Ethernet. Note that these are a very recent invention. Where this IP model operates we should expect it to involve large mainly urbanised populations as they can access the ISP’s gateway without having to make an expensive long distance call.

59 On integrated services intranets smart terminal equipment can order network resources at each moment of usage. An IP telephone will order real-time resources without the end-user being involved.

60 Gateway devices are sophisticated computers and software that connect PSTN calls to and from the IP network. The software encodes and compresses calls coming from the PSTN, allowing voice to be carried more efficiently than over the PSTN network.

Figure 5-1: International VoIP using a Wide Area Ethernet Network



A variation on this structure is for the origination end to involve a PC and not a telephone. In this case the IT gateway as outlined in Country A is not needed, although the connection between the caller and the ISP is likely to still involve the PSTN.

5.1.2 International Accounting Rate bypass

What the transit ISP and service provider in the above case have enabled is for originating callers to bypass the incumbent operator and thus bypass the Accounting Rate System.⁶¹ Given the poor quality of service, International VoIP is essentially an

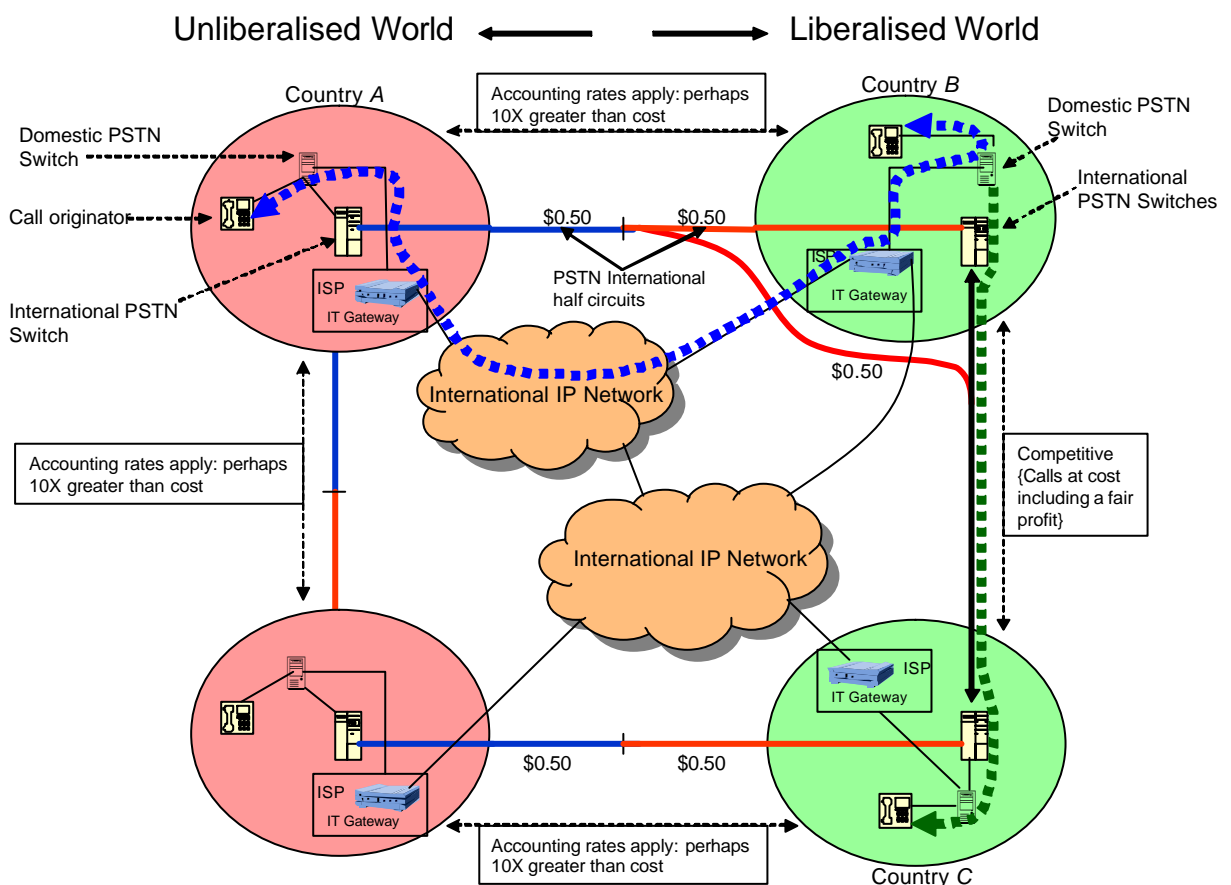
⁶¹ The *Accounting Rate System*, was developed many years ago when most telecoms operators were self-regulating state-owned monopolists. The scene was designed for a World in which there was not competition to originate or terminate international calls. Originally, the accounting rate was supposed to provide compensation for the full cost of an international call from origin to completion. The *settlement rate* is almost always half the Accounting Rate and foresaw a distribution of the "costs" of a call between the two countries.

An operator in a country from which a call originates (call it A) receives a *collection rate* (the advertised charge for a call to a specified country) and pays to the operator in the country receiving and terminating the call (say, country B), an amount called the *settlement rate* which is expressed

Accounting Rate bypass service and is most attractive in countries that maintain high Accounting Rates. Broadly speaking, these countries constitute the unliberalised World. Once into Country B, say the USA, the call will pass through the ISPs gateway and into the PSTN. From there the call will be terminated, either in the USA or possibly in some other country. Therefore, at either end of a VoIP call, quite long distances may be covered on the PSTN between gateway servers and telephones.

Figure 5-2 shows this situation diagrammatically. The originating caller in Country A could make a PSTN international call to the person they want to speak with in Country B, but the per minute price of this call will include a settlement rate charge of \$0.50 per minutes (i.e. half the assumed Accounting Rate), and an additional amount charged by the incumbent in Country A for call origination. Normally this will include the other half of the Settlement Rate and a retail mark-up, providing a per minute call price of significantly more than \$1 per minute.

Figure 5-2: VoIP international Bypass



in an amount per minute of traffic. Periodically, operators in countries A and B will settle their accounts, the operator with more outward than inward call minutes from the other making a payment representing the difference.

The path of a VoIP call from a caller in Country *A* to a receiver in Country *B* is shown by the thick dotted blue line.

In the case of calls to Country *C*, because the particular international IP network does not connect directly to Country *C*, a VoIP call may break out in the PSTN in Country *B* and be sent to the called party in Country *C* via the PSTN. As the regulatory regimes in *B* and *C* are liberal (i.e. no Accounting Rate Charges are included enabling service prices to be more closely aligned with service costs), the extra cost of getting from *B* to *C* is only a few cents. This means that relative to a PSTN call from *A* to *C*, this option also provides considerable cost saving for callers in Country *A*. The thick dotted green line shows the path of such a call.

Box 1: Accounting Rate bypass using traditional services

Where firms have traditional international private networks some level of bypass becomes possible. Single-ended breakout occurs when either the receiver or originator of the call connect to the private network over the PSTN. In practice, the level of bypass resulting from such calls will be fairly minimal.

Double ended breakout occurs when both sender and receiver connect to the private international network via the PSTN. This is also known as simple International Simple Resale (ISR) and is generally illegal except in fully liberalised countries, and even here it is sometimes illegal on unliberalised routes.⁶²

International VoIP is also an international Accounting Rate bypass service. As in the case of ISR it bypasses the PSTN international switches and thus avoids incurring any Accounting Rate liabilities. As the cost plus competitive mark-up for the service is typically only a fraction of the amounts charged for PSTN calls from unliberalised countries that include International Accounting Rate charges to similar destinations, bypass service providers can charge prices that are a fraction of those charged by the incumbent and still earn very healthy profits.

Note that if an efficient PSTN operator in *A* can purchase termination into country *B* competitively and not be required to pay \$0.50 per minute - perhaps paying \$0.10 or less - it would be able to earn a fair rate of profit charging only a fraction of the present retail price (also known as the collection rate).⁶³ Country *B*, however, does not permit the PSTN operator in *A* to purchase termination in *B* competitively. This is because

⁶² Call-back is not an accounting Rate by-pass service. Rather it reverses the direction of the call. See M. Scanlan, (1998), "Using call-back to demonstrate the discriminatory nature of the proportionate return rule", *Telecommunications Policy*, Vol 22, 11 December {reprinted in its corrected version in the subsequent issue}.

⁶³ One operator in Europe prices calls to New Zealand, a distance of Km20,000, at between \$0.05 (off-peak) and \$0.06 (peak) per minute. Clearly it expects this service to be profitable.

operators in *B* can not obtain competitive rates for termination into *A*. Indeed, *A* insists that calls from *B* pay a \$0.50 per minute termination charge (i.e. the Settlement Rate). Although Country *B* is generally liberalised, it imposes a regulation on its international carriers that requires them to charge the same termination price to *A*'s operator as *A*'s operator charges those in *B*.

Although the dollar numbers mentioned above are not averages or means of actual costs and prices, differences in prices and costs of this magnitude will be common in practice. Such differences provide great incentives for people to find ways of bypassing Accounting Rate charges. Herein lies the present commercial motivation for international VoIP services.

5.1.3 Private IP telephony on corporate IP networks

One of the main growth areas for IP telephony over the next few years is expected to be on corporate intranets. Corporate IP networks do not suffer from the network to network quality of service problems outlined in Chapter 3. Moreover, the congestion problems that tend to undermine VoIP on IP networks that are used by the public,⁶⁴ are more manageable on private intranets. Indeed, for integrated services intranets it is essential that congestion is minimised as businesses are generally much less price sensitive as far as the trade off between price and quality of service is concerned.

Most of the VoIP solutions that are being discussed commercially at present appear to involve the integration of IP telephony on corporate IP networks, and these do not provide VoIP service to the public.

5.1.4 Proprietary IP routing technology

Another voice over the internet model involves the service provider using specialised software that routes session packets over parts of the Internet that are performing best at that time. In the event that no route exists at that moment which meets required quality of service standards, calls are apparently routed on the PSTN. This solution has therefore been described as a 'hybrid' solution. The caller will not have any control over this selection. The PSTN option appears necessary due to demand by customers that the service be available to them when they require it. As voice over the Internet is not available some of the time, even with the help of specialised congestion avoiding software, the option to switch to the PSTN may be necessary for the service to be commercially viable.

⁶⁴ As noted already, the reason is largely due to the absence of a price mechanism that can match individual demands with different service qualities.

The service would appear to require that relatively liberal regulator environments exist in each end of an international call. This is suggested by the service provider being able to originate and route calls over the PSTN in the case that no IP network can provide the service at that time. As many developing economies do not fit this description, this VoIP model may not be applicable to many developing economies. This solution may be most likely to work with a Wide Area Ethernet Network as discussed in Section 5.1.1.

5.2 Future “real-time” IP networks

5.2.1 Technological aspects

MPLS appears to point the way toward future solutions to off-net QoS problems and the provision of multiple service classes over the Internet. Much of the present Internet uses ATM at layer 2 to transport IP level 3 data. These two layers operate independently of each other. For the provision of class of service options using IP over ATM, separate end-to-end VCs have to be configured for each class of service for each VPN. Such an approach lacks scalability and implies inefficient use of network resources.⁶⁵

With MPLS, on the other hand, there is a partial integration of the two layers, resulting in layer 2 becoming layer 3 aware. In this regard MPLS has both scalability and network resource efficiency advantages over ATM; it does not require end-to-end VCs to be configured for each class of service. This advantage is especially beneficial when integrating MPLS class of service support in conjunction with an MPLS VPN service. The way it works is through a label switching router.

There are two mechanisms provided by MPLS which operate when packets pass through a router or switch which are QoS enabling. These are:

- The classification of packets into different service classes, and
- The controlling of QoS characteristics (e.g. jitter, packet loss, and bandwidth) to be applied to particular packets

It is thus easy for packets to be marked as belonging to a particular class after they have been classified the first time. Initial classification uses information carried in the network layer or higher-layer headers (e.g. in the ToS field). A label corresponding to the resultant class can then be applied to the packet. Label switched routers can handle labelled packets having to be reclassified.

MPLS does not provide a cure for all QoS problems. It goes some way toward

⁶⁵ As note above, where ATM is employed the Internet presently operates using one type of channel only.

addressing some of the existing problems including the off-net QoS problems.

By providing a broader definition of labelling, the application of MPLS can be applied to wavelengths which act as their own labels. The extended MPLS protocols are referred to as Generalised MPLS, or GMPLS.⁶⁶

Other develops will also be needed, such as to enable accounting and billing based on such things as congestion prices. Also needed will be end user interfaces to enable users to choose the class of service that packets from a particular session will receive. Non of these systems are yet developed. It seems likely that some coordination will be needed between all of these areas, with the prospect that the next generation Internet – the one that provides for convergence with other platforms – is going to take some years to evolve. Indeed, there is no guarantee that it will evolve in the next 10 years, although for many years progress has been it that direct.

5.2.2 Pricing and settlements

As we have seen above, there are three main types of costs incurred in providing network services, and this same structure should ideally be reflected in the structure of retail and wholesale prices.

This price structure implies:

- One-off charge to connect to the Internet;
- A periodic (monthly) subscription charge, and
- Usage costs which highest during the period when demand is strongest.

In the future some ISPs may freely offer a different structure of prices, but those who take this offer can be expected to pay the ISP to bear the risks and efficiency costs entailed in price structures that are at odds with cost structures. End-user ISPs that adopt price structures for end-users which vary fundamentally from the wholesale price they must pay the transit provider, are taking on additional risk and would normally only agree to do this for a 'manageable' proportion of their overall traffic costs, and of course for an increased return.

Where there are several classes of service the situation potentially becomes a lot more complex, especially if we are looking for really *optimal* pricing.⁶⁷ The likelihood is however, that pricing will be kept relatively simple and stable, and that the same

⁶⁶ The Optical Internetworking Forum has extended several GMPLS components and defined a set of UNI protocols explicitly. The protocols are known as Optical User-Network Interface, O-UNI.

⁶⁷ Optimal pricing of services on a network of computers is the topic explored by the academic papers that are discussed in Annex II.

structure of prices will apply within each class of service, although as the attributes of the service classes differ so will the level of prices.

Such changes are likely to apply to both retail and wholesale services, and that sender-keeps-all peering arrangements may not survive in the long term.

Although pricing will not be completely optimal for network management, second best solutions can operate fairly efficiently, as has been experienced with traditional PSTN telecommunications networks.

6 Analysis of research on Internet service pricing

The main feature of the research papers discussed in this section (and analysed in more detail Annex I) is that they address issues of price setting as an element of network management. This is a feature which is not presently available on IP networks, and it appears to be some time off before such a system will appear in an operational form. The research papers are thus mainly academic. They address mathematical models of networks which include pricing and demand modules within the models. Such models are constructed according to the following considerations:

- the relationships between variables that the authors consider are most important, given the issues they wish to address, and
- the assumptions which are chosen to enable the model to focus on specific research issues.

The approach suggested by these two bullets is standard practice in all areas of scientific research. Referees of research papers submitted to academic journals for publication will check that the mathematical logic is correct, assess whether the design of the model and the choice of assumptions are such as to lead to insights that have importance for 'our' understanding of the research field, and where empirical techniques are also employed (not in the papers assessed) whether techniques chosen are the appropriate ones, and that the conclusions are supported by the data.⁶⁸

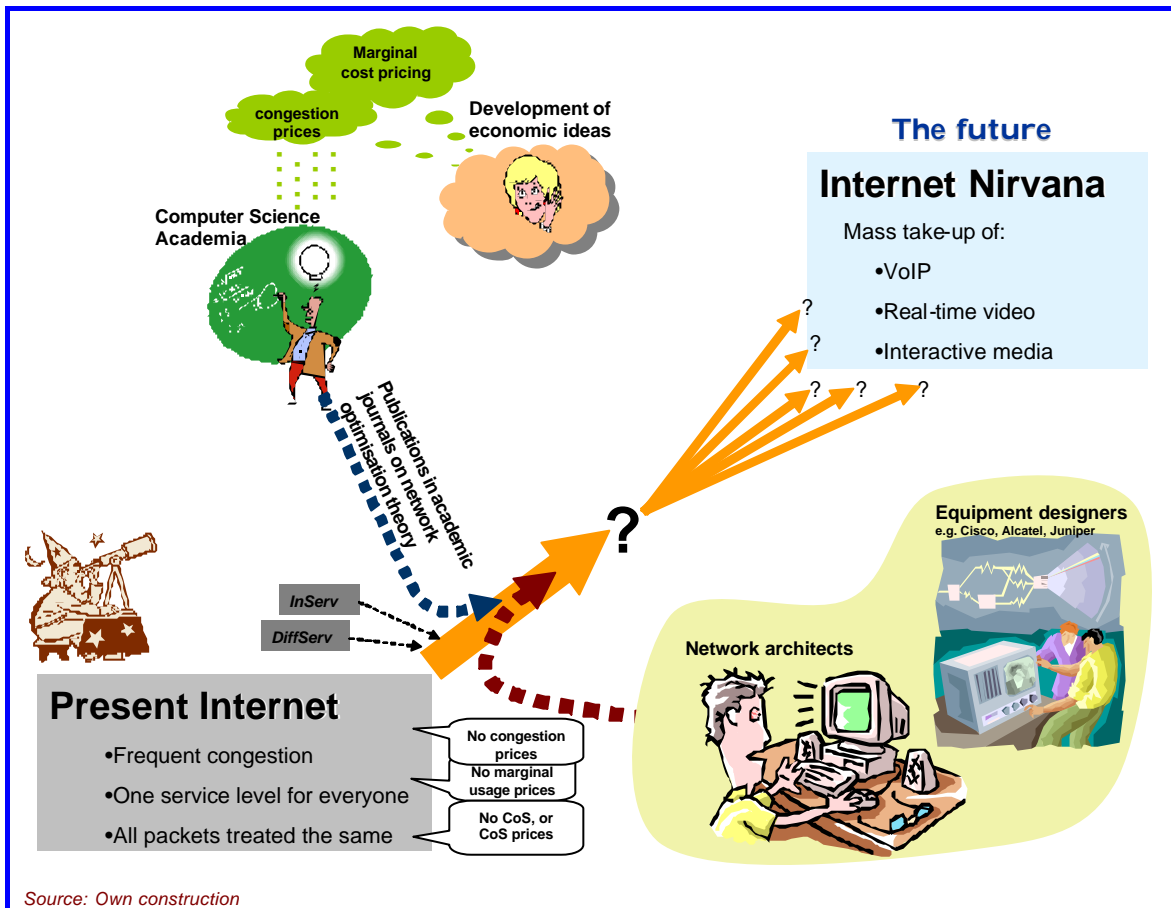
The problems the models we refer to in this section, are trying to address are complex. There are engineering design issues, such as those that give rise to cost causation, there is normally more than one class of service involved, and they include price demand relationships in the model. Especially in regard to dynamic optimisation, some types of problem are computationally intractable, and some simplifying assumptions must be adopted. [Directly or indirectly](#), the research papers discussed in Annex II address these problems.

In terms of multiple classes of service, the papers all include pricing as an active element in network management. However, the research does this in order to look at the interaction between optimum pricing and more than one class of service. It is not apparent that these papers have a direct practical implication as far as the future introduction and design of class-of-service architectures – features that may become part of the Internet in the future. This research is, however, likely to assisted in the education of those computer network architects who are involved in finding practical solutions to the lack of class-of-service options and the pricing of such options on the

⁶⁸ Referees will be peers of authors (i.e. other experts in that field of research) who decide whether the research paper should be published in the form submitted, whether it should be modified before being published, or perhaps that it should not be accepted for publication.

Internet. Figure 6-1 provides a visual depiction of the spread of economic ideas among Internet engineers in their search for solutions that will bring about the net generation Internet.

Figure 6-1: Depicting the cross-fertilisation of economic ideas to computer network design



What all of the research papers have in common is an acceptance that prices can and should be used to help manage congestion and provide for quality of service improvements on the Internet. Where a model includes the possibility of real-time services making up the bouquet of available services, optimisation problems become very complex, implying a dynamic pricing arrangement which mirrors in some way the state of the system, i.e. prices continuously evolve to manage (optimise) the demand for network services (e.g. the inflow of datagrams). Such optimisation of network resources would require prices to vary at each entry point and to change continuously over each user's session period. Unless the model is very simple, models that pursue full network optimisation are computationally intractable. Thus, researchers tend to look for simplifying adjustments which provide for near optimisation, but are at the same time computationally tractable.

From an economic perspective, the encouraging thing about these papers is that they

all entail a good appreciation that demand management should be an integral part of network investment planning and traffic optimisation. This has not always been true among computer network designers. Indeed, some Internet industry commentators continue to argue against this approach in written papers and at conferences.⁶⁹ (The issues are discussed in Section 4 and in Annex I).

As time goes by institutions that are not part of the community of economists learn more about economics and more and more non economists take on board economic ideas. This is part of institution and knowledge building processes. It happens in government administrations, in other professions, such as law and accounting, and it happens among academics in very diverse disciplines. Indeed, academia is normally where this *cross-fertilisation* takes place first.⁷⁰

The academic research papers are discussed in more detail in Annex II

69 As we noted above, the architects of *IntServ* were apparently unaware of the economic incentive problems that would arise if this architecture was to become widely available for Internet users.

70 Differences between countries in the degree to which their institutions make progress in understanding and picking up ideas even help explain why some countries are less well-off than others.

7 IP regulatory issues

Future regulatory issues concerning networks using IP appear to be mainly confined to three: ISPs; organisations that we describe today as traditional telecommunications operators, and next generation mobile operators.⁷¹

The regulation of content such as in regard to decency, security, and intellectual property is not addressed in this report.

7.1 Regulation and ISPs

Market failure issues relating the possible regulation of ISPs concern market power and externalities. While the ISP model tends to be competitive, there are potentially there are two main problem areas:

1. In some regions, there may be too little competition between transit providing ISPs such that end-user ISPs are being required to pay excessive prices for transit, or are receiving a QoS which is less than is provided in more competitive environments. The main reasons for the lack of competition will be one or more of the following:
 - (i) Low levels of demand (ability to pay), such as can occur in developing economies. This can result in very few transit providing ISP providing service in a region;
 - (ii) Onerous licensing regimes which:
 - unnecessarily limit the numbers of competitors, or
 - charge too high fees, or imposes other costly obligations which discourage entry;
 - (iii) Other regulatory problems exist that make investing risky, the main one being poorly developed regulatory institutions, which have the same result as (i), and
 - (iv) Because the incumbent operator is imposing excessive charges for granting ISPs access to its network, and/or is preventing access to its network.

These are country regulatory concerns. In the case of (ii) and (iii) there is no ex

⁷¹ The regulation of content, or privacy issues are not addressed in this report.

ante regulatory cure to these problems. Indeed, where they exist regulation can be the cause rather than the solution to the problem. What is recommended is a more liberal regime, although developing institutions able to support this can take many years.

In the case of (iv) above, the issues concern regulation of the Incumbent, and this topic is address more fully below.

2. Another potential problem area is with core ISPs, of which there are perhaps five or six, although more than one of them is in financial difficulties at present. Core ISPs (also called Tier 1 ISPs) are the only ISPs that have virtually complete routing tables. They sit at the top of a loose hierarchy that is the Internet. They peer with each other and virtually no one else.⁷² Core ISPs do face competition from Tier 2 and 3 ISPs, as well as from firms that offer transit substitutes, principally those that store content close to the Internet's edges. Perhaps the most commonly known of these are firms that provide *caching* services. The potential for competition concerns about core ISPs is an *international issue* which at present appears to be adequately addressed by the US and EU competition law authorities, and most especially rules that prevent firms from attaining market dominance through mergers or takeovers. Therefore, this issue is not addressed further in this report.

3. Another potential area of regulatory interest with the Internet is with naming, numbering and addressing. Numbers can be a scarce resource mainly due to the costs resulting from changes having to be made to a numbering scheme when numbers within that scheme become exhausted. This is the situation regarding the replacement of IPv4 by IPv6 which will likely be necessary in the next 4 to 7 years. Some countries are apparently requiring ISPs to switch to IPv6 by a certain date in the future. Others, such as the EU, are pushing for an early switch to IPv6 by injecting large amounts of public funds. However, this is an issue that the Internet Community is keeping a close eye on, and it is debatable as to whether any administration needs to do more than observe developments at this stage. However, as it is not essentially a country specific issue IP numbering management is not discussed further here.

Access to names numbers and addresses is a related issue which has attracted regulatory interest due to the need for networks that are based on different addressing schemes (e.g. telephone number and IP numbers and names) to interoperate where both networks provide the same service. The main example is the need for VoIP providers and PSTN networks to be able to have calls terminated on the other network. The policy issues focus on the **ENUM**⁷³ debate for which there is are two distinct groups: those that support a

⁷² See Section 2.2 (including Figure 2.2) for further discussion about peering.

⁷³ ENUM is a global addressing scheme linking PSTN and IP networks which was standardised by

role for the ITU in setting certain rules, and those who argue that no centralised administration or rule making is required. As this is not a country specific regulatory issues, it is not addressed further in this report.

The availability of numbering and addressing data to competitors is essential in order for competition to develop between competing platforms and for this reason countries will need to keep abreast of developments and make sure that competitors are not closed out due to their inability to get access to the required databases.

4. **Standardisation** or the lack of it in the Internet is another area of interest to regulatory authorities, particularly given the off-net QoS problems that have been outlined in this report. This is an enormously complex issue as it is tied in with technological development and there are unquantifiable costs and benefits involved. However, as this also is not a country specific concern standardisation is not discussed further in this report.⁷⁴

The Internet is not directly regulated and its Internationality and border-less structure would make regulation very difficult to implement and operate. As a general rule, where Internet networks and other firms are starting to compete with each other and do not receive equal regulatory treatment, it is recommended that regulatory authorities begin addressing the problem by first looking at ways to remove regulations from the PSTN to bring about competitive neutrality, rather than starting by looking at way to regulate the internet or ISP so as to maintain or restore competitive neutrality.

7.2 Regulation and incumbent PSTN operators

Most of the regulatory issues for NRAs that involve the Internet are concern with traditional telecommunications operators. As convergence occurs between the PSTN, CATV and IP networks – wire and wireless – regulations need to be reviewed to make sure that firms which are competing with each other are equally treated under the law, i.e. that regulations do not favour one technology over another, or one type of delivery platform over another. The overriding principle is that there should be a level playing field. There are several areas where PSTN operators tend to be regulated that appear to be in need of reform over the next few years in order for **competitive neutrality** to be retained. Perhaps the main entity likely to loose out from a failure to address these issues is the incumbent operator. Arguably the main issues are covered briefly below.

- *RPI-X or CPI-X price capping*

An important part of price capping involves forecasts about the level of

the IETF in September 2000.

⁷⁴ For a general discussion of the economic issues, see David and Greenstein (1990).

competition that will evolve for price capped services over the period of the price cap. Forecasting the convergence rate of the Internet with the PSTN is likely to involve considerable error compared to the pace of actual convergence.

- *The structure and units of retail prices*

Retail PSTN call services are priced in seconds or minutes according the time-of-day. VoIP may not be priced in terms of session times, although if it did it would presumably have to be time by bandwidth. In any even it is unlikely to involve the same gradient between off-peak and peak rates.

- *The level of prices - International bypass*

Many operators, especially those in countries that are not fully liberalised, continue to earn high profits on International calls. Already many of these countries are facing significant bypass due to VoIP. Tariff reform is needed, although this can prove difficult due to political resistance.

- *Interconnection prices and price structures*

Interconnections costs are capacity related. Using minutes as a unit over which capacity costs for interconnected traffic are distributed has been important to the growth of competition to provide PSTN services as it has kept entry costs for new entrants lower; i.e. it has enabled them to rent circuit minutes rather than buy the capacity which is needed to terminate their calls on the incumbent's network. One explanation for requiring interconnection charges to be levied on a per minute basis is that circuits are dedicated for the entire period of a telephone call – they can not be used by anyone else. This is not the case for a VoIP call which involves statistical multiplexing such that peak usage costs are bit related. The potential for difficulties when levying per minute charge for Internet usage were discussed briefly above.⁷⁵ In order for PSTN operators not to face a regulatory disadvantage, the regulated price structure of PSTN interconnection tariffs may well need review in the near to medium term, perhaps with one outcome being that PSTN interconnection would be priced in terms of 'busy hour' capacity costs. As per minute charges are in principle built up from capacity costs, part of the work needed in order for such changes to be implemented has already occurred.⁷⁶

⁷⁵ See Section 7.2.2.

⁷⁶ Other areas of possible market failure caused by regulation are less obvious although potentially important, and include reduced levels of new entry, competition, and investment, caused by investor shyness due to regulatory uncertainty and the risk of regulatory opportunism. These are, however, real problems and arise especially in utility industries where there are long-lived investments prone to being stranded by regulatory or political decisions. (US tariffs on steel imports, for example, strand investors' assets in countries where steel producers export to the US.) We do not address this type of market failure here but direct readers to the study we did jointly with a partner: Cullen International & WIK (2001), "Universal service in the Accession Countries", especially pages 82-96 in the Main Report, and 8-13 in the Country Report.

- *levying special taxes*

The levying of special taxes such as universal service contributions. Where these are based on some measure of market share they tend to push up the operating costs of the taxed entity. Clearly where these taxes are significant and levied on some competitors (e.g. PSTN operators) and not others (e.g. VoIP operators), a competitive non-neutrality can arise.⁷⁷

7.3 Regulation and next generation mobile operators

Next generation mobile network that work with IP will be developed. Indeed, they already have a name – UTRAN (Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access Network – although the technology is still confined to the laboratory. While mobile network operators (MNOs) are largely unregulated (as is the case with ISPs), they will compete increasingly with incumbent PSTN networks. At present GSM networks are mainly complimentary to incumbent PSTN operators, i.e. their existence boosts the incumbent's profitability compared to a situation where they did not exist. There are however, both complimentary and substitution effects presents, and while complimentary effects are far larger than substitution effects at present, substitution effects may be increasingly evident once UMTS network operators begin service.⁷⁸ Just as convergence with fixed wire data networks provides a reason for regulators to review the structure and units of measure regarding regulated prices on fixed wire networks (e.g. interconnection), convergence will also involve MNOs. The arguments involving UMTS networks are thus very similar to those that can be found in the section above which discusses ISP.

http://europa.eu.int/information_society/topics/telecoms/international/news/index_en.htm

⁷⁷ An analysis of the tax issues can be found in WIK (2000) and in Scanlan and Neu (2001). Indeed in some countries net USO costs have been recovered through interconnection charges. A detailed analysis of why this is not an advisable form of cost recovery can be found in chapter 3 of WIK (2000). An analysis of the problems of trying to recover access subsidies in similar way can be found in Scanlan and Neu (2002).

⁷⁸ The complimentary nature of GSM networks can be viewed in terms of increased numbers of calls originated on the incumbent's fixed wire network, and increased interconnection revenues from calls originated on GSM networks and terminated on the incumbent's fixed wire network.

References

- Alcatel Telecommunication Review (2001), Special issue "Next Generation Now", vol. 2.
- Anania, L., and R. Solomon, (1997), "The Minimalist Price", in L. McKnight and J Bailey (eds.) (1997), *Internet Economics*, MIT Press, Cambridge, Mass.
- Brown, S. and D. Sibley (1986), *The Theory of Public Utility Pricing*. Cambridge University Press.
- David, P. and S. Greenstein (1990), "The economics of compatibility standards: An introduction to recent research", *Econ. Innov. New techn*, (1): 3-41.
- Gupta, A., D. O. Stahl and A.B. Whinston, (1995) "A stochastic equilibrium model of Internet pricing", Mimeo, University of Texas at Austin.
- Hwang, J., Weiss, M. and S.J. Shin (2000), "Dynamic Bandwidth Provisioning Economy of A Marked- Based IP QoS Interconnection: IntServ - DiffServ", paper presented at the 28th Telecommunications Policy Research Conference, September 23-25, Alexandria, Virginia.
- Kercheval, K. (1997), "TCP/IP over ATM", Prentice Hall PTR.
- Korilis, Y.A., T.A. Varvarigou, and S.R. Ahuja, (No date), "Incentive-compatible pricing strategies in non-cooperative networks", Mimeo, Bell Laboratories.
- Marbach, P., (No date), "Pricing priority classes in a differentiated services network", Mimeo, University of Cambridge.
- Marcus, J. S. (1999) *Designing wide area networks and Internetworks – A practical guide*, Addison-Wesley, Reading (Mass.)
- McDysan, D. (2000), *QoS and Traffic Management in IP and ATM Networks*, McGraw-Hill.
- McKnight, L. W. and J.P. Bailey (eds.) (1997), *Internet economics*, MIT Press, Cambridge, MA.
- Odlyzko, A. (1998), "The economics of the Internet: Utility, utilisation, pricing, and quality of service", mimeo, AT&T Labs – Research.
- Paschalidis, I.Ch., and J.N. Tsitsiklis, (2000), "Congestion dependent pricing of network services", *IEEE/ACM Transactions on Networking*: Vol 8, 2, pp 171-184
- Ross, (1995), *Multi service loss models for Broadband Telecommunication Networks*, Springer Berlin.
- Scanlan, M. and Neu, W. (2002 forthcoming), "Universal service policies and institutions in emerging economies", Diskussionsbeitrag Nr. 120, *Wissenschaftliches Institut fuer Kommunikationsdienste*.
- Semeria, C. (1996), "Understanding IP addressing: Everything you ever wanted to know", <http://www.3com.com/nsc/501302html>
- Smith, C. and D. Collins (2002), *3G Wireless Networks*, McGraw-Hill Telecom Professional, New York.
- Squire, Sanders & Dempsey LLP and WIK (2002), "Market definitions and regulatory obligations in communications markets", A study for the European Commission.

- Wang, Q., Peha, J.M. and M.A. Sirbu (1997), "Optimal pricing for integrated service networks", in: McKnight and Bailey (1997), p.353-376).
- Wang, Q., Peha, J.M. and M.A. Sirbu (1995), "The design of an optimal pricing scheme for ATM Integrated-Service networks", paper presented at MIT Workshop on Internet Economics (March).
- WIK (2002), *The economics of IP networks: market, technical and public policy issues relating to internet traffic exchange*. A study for the European Commission.
- WIK (2000), *Study on the re-examination of the scope of universal service in the telecommunications sector of the European Union, in the context of the 1999 Review*. A study for the European Commission.
- Yuen, C. and W. Tjioe, (No Date), "Modeling and verifying a price model for congestion control in computer networks using PROMELA/SPIN", University of Toronto.
-

Glossary

AAL	ATM Adaptation Layer
ABR	Available Bit Rate
AS	Autonomous System
ATM	Asynchronous Transfer Mode
CATV	Cable TV
CBR	Constant Bit Rate
CDN	Content Delivery Network
CDV	Cell Delay Variation
CER	Cell Error Ratio
<i>Ceteris Paribus</i>	all other things being equal
CLR	Cell Loss Ratio
CMR	Cell Misinsertion Ratio
CoS	Class of Service
CTD	Cell Time Delay
<i>DiffServ</i>	Differentiated Services (Protocols)
DNS	Domain Name System
DSCP	Differentiated Services Code Point
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Access Multiplexer
DWDM	Dense Wave Division Multiplexing
ECI	Explicit Congestion Indicator
ENUM	Extended Numbering Internet DNS
FRIACO	Flat Rate Internet Call Origination
FTP	File Transfer Protocol
GMPLS	Generalised MPLS
GoS	Grade of Service
GSM	Global System for Mobile Communications
IBP	Internet Backbone Provider
IETF	Internet Engineering Task Force
<i>IntServ</i>	Integrated Services (Protocols)

IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISO	International Organisation for Standardisation
ISP	Internet Service Provider
LAN	Local Area Network
LAIN	Local Area IP Network
LSP	Label Switched Path
MC	Marginal Cost
MinCR	Minimum Cell Rate
MPLS	Multi Protocol Label Switching
MPOA	Multi Protocol over ATM
NA	not available
NAP	Network Access Point
NGI	Next Generation Internet
nrt	near real-time
OAM	Operation, Administration and Maintenance
OC	Optical Carrier
OSI	Open Systems Interconnection
PCR	Pick Cell Rate
PIR	Packet or Cell Invention Ratio
PLR	Packet or Cell-Loss Ratio
PoP	Point of Presence
PoS	Packet over Sonet
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RFC	Request For Comments
RSVP	Reserve ReSerVation Protocol
rt	real time
RTCP	Real-Time Control Protocol
RTP	Real-time Transport Protocol
SCR	Substantial Cellrate
SDH	Synchronous Digital Hierarchy

SECBR	Several Error Cell Block Ratio
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SONET	Synchronous Optical Network
TCP	Transfer Control Protocol
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
VBR	Variable Bit Rate
VC	Virtual Circuit, Channel or Connection
VLSM	Variable Length Subnet Masking
VoIP	Voice over IP
VP	Virtual Path
VPIP	Virtual Private IP Network
WAN	Wide Area Network
WAIN	Wide Area IP Network
WDM	Wave Division Multiplexing
WIK	Wissenschaftliches Institut für Kommunikationsdienste
WTP	Willingness to Pay

Annex I:

QoS and the limitations of cheap bandwidth ⁷⁹

In recent years several people have pointed out that with the rapidly declining cost of bandwidth and the rapid increase in computing power, "throwing bandwidth" at congestion problems can be a cost effective way of addressing QoS problems.⁸⁰ Indeed, it has been claimed that this option negates proposals to introduce pricing mechanisms such as class of service options, to control congestion, and may also negate the proposals that would provide mainly technical means to discriminate between higher and lower priority packets, such as *IntServ* and *DiffServ* architectures, which we addressed above. Mainly because of falling costs of transmission and processing, and the rapid growth in processing power, the suggestion is that congestion on the Internet will be a temporary phenomenon, implying that there is no need to change the structure of existing prices.

Evidence in favour of the "throw bandwidth at it" solution to congestion includes information that shows that bandwidth has grown much faster than traffic volumes⁸¹, with the inference being that after several more years of divergence in growth rates it will not matter that the priority of treatment of packets on the Internet is according to the order of arrival, and that low priority emails get the same QoS as do VoIP packets – all packets will get a QoS which is so high that the hold-up of messages where perceived QoS is very sensitive to latency and jitter, by those that are not, will have no material effect on the QoS experienced by end-users. In general, the argument is that the rapidly declining cost of bandwidth and processing will mean that more "bandwidth" will be the cost effective means of addressing QoS problems. In short, the claim is that all services will receive a premium QoS.

While the report's author would tend to concur that on many occasions apparent over-engineering could be an appropriate option, I do not see that in general throwing bandwidth at congestion problems is the cost effective way to address QoS problems that stand in the way of VoIP and other applications that have strict QoS requirements. Indeed, even if the issue of the opportunity cost of this approach was put to one side, I am sceptical that this approach can sufficiently address the problem of congestion to enable an all-services Internet to effectively compete with other platforms like the PSTN. One reason for this is that demand for bandwidth is likely to increase enormously due to the following factors:

⁷⁹ This annex comes from WIK (2002).

⁸⁰ See for example Ferguson and Huston (1998); Odlyzko (1998); and Anania and Solomon (1997) where the claim is less explicit.

⁸¹ See Odlyzko (1998).

- Increased access speeds for end-users (e.g. xDSL) in the short to medium term (and access speeds several times greater than effective xDSL speeds in the next 10-20 years);
- If a service quality arrives that becomes capable of delivering high quality VoIP, it will likely result in many customers (perhaps a majority of existing PSTN subscribers) moving their demand for voice services onto the Internet as in many cases it will likely have a significant price advantage;
- When customer access speeds reach levels that enable HDT quality streaming video, the Internet will have converged with CATV and broadcasting, and likely demand for content (including from different parts of the world) will result in an enormous increase in the volume of Internet traffic, and
- 3G and 4G mobile Internet access may also result in large increases in demand for the Internet, be it for voice, WWW, e-mail, file transfer, or streaming video.

I have intimated in Section 3 above that without a marginal cost pricing mechanism there is no thoroughly accurate means of providing the proper incentives for ISPs to invest in a timely way in upgrading capacity. The pricing mechanism is the ideal way of connecting investment incentives with demand and in the flat-rate pricing world of the Internet where marginal congestion costs are far from zero, no such pricing mechanism presently operates.

However, perhaps the most important issue is not whether it is possible to address QoS problems for real-time services by throwing bandwidth at the problem, but, whether there is not a more cost effective option to the combination of flat-rate pricing and over-engineering the Internet: and if this option exists, whether it provides for a pricing mechanism which will have a more realistic chance of meeting the claims made for it (one that is able to better match marginal costs of capacity upgrades with marginal revenues, when QoS is degraded by congestion).

In the view of the author, a flat-rate "one-service-fits-all" Internet is very unlikely to be the arrangement that ushers in the next generation "converged" Internet i.e. an Internet where WWW, streaming video, file transfer, email, and voice services, are provided to a price/quality that makes these services highly substitutable with those provided over other (existing) platforms. In short, I do not accept that falling capacity costs will result in the Internet being able to avoid "the tragedy-of-the-commons" problem⁸²; i.e. the claim that supply will in practice outstrip demand. This is not in keeping with our experience with policies that make things that are not pure public goods, free at the point of delivery.⁸³ Where this has occurred, experience shows that overuse / congestion

⁸² "The tragedy-of-the-commons" is a problem of market failure which we discuss further below.

⁸³ In cases where there are subscription fees but users face no marginal usage costs, outcome have been much improved, but without there being a large over-investment in capacity, some congestion is typically still experienced.

typically occurs.

Over-provisioning requires networks to be built which cope with an expected level of peak demand.⁸⁴ This tends to result in lower levels of average network utilisation and thus higher average cost per bit. It is well known that Internet traffic tends to be very 'bursty' (demands high bandwidth for short periods). In larger ISP networks, the 'burstyness' of end-user demands tends to be somewhat smoothed due to the large number of bursts being dispersed around a mean.⁸⁵ In order to provide a service that is not seriously compromised at higher usage periods by congestion, average peak utilisation rates on backbones of roughly 50% may be the outcome, with very much lower average utilisation rates over a 24 hour period.

In the last 3-4 years there has been progress in setting up standards for IP networks that address QoS, e.g. Real time protocol (RTP), "Resource reSerVation Protocol" (RSVP), *DiffServ*, and *IntServ*. Services provided by these protocols are not yet commonly available on the Internet but may be implemented in the routers of some corporate networks or academic network structures like TEN 155, and even in some larger ISPs, although not yet between larger ISPs.

84 In practice even in PSTN networks blocking occurs during congested periods. In the Internet world this is done with admission control algorithms.

85 This effect is called stochastic multiplexing but it should be dealt with carefully. Some studies on Internet traffic suggest that the length of web pages and the corresponding processing and transmission time are not according to an exponential distribution but are better approximated by a distribution with large variance e.g. by a Weibull distribution. Some authors have claimed that the distribution is Pareto resulting in a near infinite variance and cancelling any stochastic multiplexing effect. But these studies are based generally on data traffic in academic networks, which is not representative of traffic on the commercial Internet.

Annex II:

Research papers on pricing and CoS

Note that the research papers reviewed in this annex have no obvious practical application at present to the Internet or IP networks. The papers are academic in nature and mainly address computational and modelling problems concerning the optimisation of network design.

As a practical matter class-of-service pricing to users on IP networks is not presently available and many problems will need to be solved in order that a workable scheme can be designed and implemented

Gupta, Stahl and Whinston (1995)

The research paper by Gupta et al is based on a spot market with priority queues, where session packets receive different priority depending on the priority class that customers buy. It is a model based on expectations and stochastic values (i.e. values are not known with certainty). There are K classes of non-interruptible service with users positioning themselves into these classes according to the strength of their demand. The price for k=1 (the highest priority class) would be least congested and the most expensive. The model allows prices to be decentralised and charged “for each machine in the Internet”. Rental prices are adjusted to optimise the trade-off between greater through-put of data and longer waiting times. Customers expected service costs depend on the expected load imposed by their use⁸⁶, the particular priority class, and the cost of machines (which we can generalise as the investment cost of the network elements that are used due to the customer request(s)). Session prices are the sum of the k priority price charged at each network computational machine based on expected units of processing work. Prices across classes are adjusted iteratively to allow expectations to adjust given experience values of variables (e.g. delay times). For users, costs are a function of price and delay times.

Customers are allocated a class of service depending on the rental price they pay, which depends on the customer’s cost of delay, and which may be different for each individual. Those who pay the most might expect to get a service that enables real-time services at virtually any time. One of the attractions of this paper is its use of *adaptive expectations*. It has a high degree of appreciation of economic theory.

⁸⁶ i.e. the model works with processing costs rather than packets of data

Marbach (no date)

Marbach develops two models in his research paper. The first is set up as a non co-operative game, where the prices paid by users differ depending on the priority they have chosen. In the model all packets that are accepted into the network receive equal treatment, so that the priority choice governs the senders expectation about the probability of having his packet accepted. This is curious model as non co-operative games are normally used when the participants are able to exert some influence on the outcome, which is not realistic for most end-users of the Internet, although on smaller networks the relevance of this modelling approach is apparent. The model also enables the provider to capture a larger share of the consumer surplus through price discrimination. Customers choose the priorities they attach to their packets and thus have control over QoS to the extent that various CoS schemes operate.

Prices paid by users are on a per packet basis. This has better incentive properties than one based on workload (which is the approach of Gupta et al) where processing power is increasingly rapidly, and workload is to some degree an endogenous decision of each network designer.

Where existing capacity is allocated in order of priority classes, at any particular time there will be a priority class for which not all packets can be sent, i.e. demand is greater than capacity. Some of the packets in this class will thus be lost. All packets marked with a priority above this class will be sent successfully. Where there is but one QoS for sent packets the model provides a congestion price u^* which is similar to the market clearing price in M-V's auction model. Information to assist users in making their choices is provided by way of a control link (a signalling channel containing network "intelligence" data).

Marbach's second model extends the this first model so that it becomes a congestion pricing model. It is per packet-based and is applicable where several classes of service quality are available in addition to certainty of packet delivery these. In this case there would be a different u^* for each CoS. It provides a price at which the network can cope with the traffic submitted without QoS being degraded, but any lower price would result in degraded QoS (lost packets) as demand would increase.

Yuen and Tjioe (No Date)

The research paper by Yuen and Tjioe includes a report about the use of simulations a computer based model to inspect the properties of the Marbach model. As with M-V the model relies on network users to adjust their behaviour according to the prices they are being charged. The model includes different classes of service, with the one tested by Yuen and Tjioe having a "premium priority class", and a "best effort" class. All packets in the premium class are guaranteed to be transmitted. The first service provided is a signalling channel, and through this each user is informed of the probability of

transmission based on the total demands from users. If demand exceeds capacity then the probability of transmission is less than 1. The transmission probability is set until the next cycle and so will become less accurate as time proceeds. Yuen *and Tjioe* then add an administrator process to the model which periodically monitors network usage and adjusts prices so as to provide for dynamic network resource management.

Korilis, Varvarigou, and Ahuja (No date)

Korilis et al are interested in the optimal management of a network of a certain size involving a limited number of users who have some knowledge about the behaviour of other users. The instrument used for this purpose is pricing, where users pay in terms of a unit of flow, the price of which varies from link to link in order to manage congestion within the network. The model is thus of a non-cooperative game where each user adapts his or her strategy depending on the routing decisions of other. "Optimisation" is initially obtained on a link by link basis, with Consumers (users) doing the best they can given their demands, the network resources available and prices charged.

The authors then extend the model to allow for the endogenous determination of price on each link per throughput per of data. They employ a real-time WWW market where users seek to acquire capacity at that time on necessary resources. Information provided to the user is the residual capacity and price of each resource. The idea here is for price to work so that the average congestion in the system is minimised.

The behaviour of the network provider in terms of dimensioning the network i.e. the supply-side of the picture, is not addressed. The model could provide the means by which a monopolist would maximise his profits, or it could be used to maximise social welfare. Adaptation of the model from a single supplier to allow for a limited number of competitive suppliers (i.e. each being able to deliver the traffic as demanded) may provide useful insights for managing a network of networks.

Paschalidis and Tsitsiklis (2000)

Paschalidis and Tsitsiklis (P&T) model a service provider where capacity is fixed (i.e. pricing does not take account of the need to increase capacity in a rapidly expanding network). The authors mainly look at pricing for revenue maximisation, although in places a brief analysis of welfare maximisation is also provided where service providers charge on a volume by time basis.

P&T compare what they refer to as near optimal dynamic pricing, with static pricing. Dynamic pricing occurs where prices continually adjust to reflect the state of the system (i.e. the loading on the network). The optimal dynamic pricing scheme must be based on the state of the network, with pricing being determined in real-time at each and every node. P&T allow for several classes of service, although in this regard the

model is quite restrictive in that the classes are assumed to have identical characteristics. The idea here is that the network would separate customers depending on their different demand strengths, and price to them differently. This is known as price discrimination, and while it can be welfare enhancing, this is not the case with the P&T paper.

The computation of optimal dynamic prices soon becomes unmanageable as the number of classes and the capacity increases. Fully dynamic pricing therefore lacks practicality and P&T enquire whether, and under what conditions, static prices (i.e. those that apply over a significant period of time and this do not exactly track the state of the network) offer a satisfactory proxy. The pricing policy investigated is one that is fixed over the duration of the session, but varies periodically between sessions.

Although P&T's work does show us that the static modelling of prices may be satisfactory, further work is likely needed using slightly different model design, relaxed assumptions (such as non identical classes of service), allowing for customer substitution between classes, and with more focus on the welfare implications. Where we are interested in the applicability of this work for understanding how pricing might work on the Internet, there is also a need for applied econometric work in order to provide data about the demand for internet services.

Wang, Peha, and Sirbu (1995)

The paper by Wang et al addresses price setting in a two service class ATM network; a guaranteed service relating to some pre-specified QoS (which is not absolute but stochastic), and a best effort service class.

ATM networks perform *traffic policing* to control the admission of cells onto the network according to a range of predetermined parameters that are contracted with each customer. Tags in the ToS field of IP packet headers are not recognised by ATM networks. Thus, before traffic policing occurs, sending networks shape their traffic so that it fits their customer profile. Moreover, with IP operating at either end of an ATM network, IP packets would have to be separated before going into the ATM network, and channelled into an ATM VC according to the class of service that was to be provided.

This level of interpretability between ATM and IP (layer 2 and 3 of the ISO⁸⁷) is not in general provided for, and nor is it possible for end-users to have any control over such a process. Thus, as with the other research papers, the Wang et al paper is not concerned with providing a practical solution to existing internet problems, but constitutes theoretical research that will be added to the body of knowledge about pricing and network management. which will find its way into the design of practical solutions to pricing and QoS service that are implemented in the future.

87 See Figure 2-1.

For the best effort service customers are charged on a per cell basis. Given information about demand, network capacity and the size of the buffer, a cut-off price is determined, and cells for which customers' willingness to pay is less than this price are dropped from the system. How this price is determined is a central feature of the paper, along with price determination for the guaranteed class.

The paper address price setting for profit maximisation and not price setting to maximise social welfare. It focuses on a three stage procedure as an approximation for the first best maximisation which is mathematically intractable. These stages are: the investment decision; optimal pricing for guaranteed service, and spot pricing for best effort service. The procedure is iterated until a stable state is obtained.

Annex III

QoS attributes of ATM networks

In ATM networks arriving cells fill a logical bucket which 'leaks' according to specific traffic parameters, and these parameters form the basis for QoS contracts. The parameters can include: cell loss rate (CLR); cell time delay (CTD); cell delay variation (CDV); peak cell rate (PCR), substantial cell rate (SCR), minimal cell rate (MinCR), and explicit congestion indication (ECI).

Operators have recently begun to implement a form of WAN IP-switched architecture under Multi Protocol Label Switching (MPLS). The adoption of this technology will result in some changes in SLAs that ISPs have with their transit provider. MPLS is discussed in Section 3.5.

Table 0-1: Suitability of ATM Forum service categories to applications

Applications	CBR	VBR-rt	VBR-nrt	ABR	UBR
Critical data	Good	Fair	Best	Fair	No
LAN interconnect	Fair	Fair	Good	Best	Good
WAN data transport	Fair	Fair	Good	Best	Good
Circuit Emulation	Best	Good	No	No	No
Telephony	Best	Good	No	No	No
Video conferencing	Best	Good	Fair	Fair	Poor
Compressed audio	Fair	Best	Good	Good	Poor
Video distribution	Best	Good	Fair	No	No
Interactive multimedia	Best	Best	Good	Good	Poor

Source: McDyson (2000)

CBR: constant bit rate,
 VBR-rt: variable bit rate, real-time,
 VBR-nrt: variable bit rate, non-real-time,
 ABR: available bit rate, and
 UBR: unspecified bit rate.

The suitability of ATM service categories to applications is shown in Table 0-1. For services that require higher quality of service features like real-time voice and interactive data and video, ATM networks can be configured to provide sustained bandwidth, and low latency and jitter, i.e. to appear like a dedicated circuit.

Note that for Internet transit traffic, QoS guarantees typically apply, but only in regard to

a standard service. Outside of private networks, i.e. on the public Internet, the QoS guarantees are those associated with ATM AAL5 and UBR, and are not such as would enable reliable real-time service quality. Other classes of service that can be provided by ATM networks, such as AA1 and CBR or VBR-rt are apparently not recognised by ISP networks.

Where *IntServ* or *DiffServ* architectures are employed by sending and receiving networks, a CoS system may operate off-net although this appears to be vary rare or non existent in practice.