Come the Revolution – Network Dimensioning, Service Costing and Pricing in a Packet Switched Environment

Gareth Davies, Michael Hardt and Frank Kelly[1]

Abstract
The telecommunications industry is going through a technological revolution, involving a transition from multiple voice and data networks to integrated service networks using packet switching technologies. The paper describes a new framework (the IPCP Framework) for capacity planning and costing in an IP network environment. Capacity Planning is carried out in four stages, the first of which is based on the concept of Effective Bandwidth, with the other three addressing the needs of different classes of traffic (Real Time, Interactive Data and Streaming, and Delay-Tolerant Services). The Framework enables network operators to reduce costs by optimising their capacity requirements and, with the aid of the IP Costing Module, to determine the costs of the various services that use integrated IP networks, each of which may have a different set of service quality requirements. Illustrative results show that capacity requirements and costs per Mbit vary significantly by service type, a finding with far-reaching implications for business planning and pricing. The approach has been extended to support the design of the new pricing models that will be required in an IP environment and to provide a basis for decisions on traffic management and product portfolio optimisation. The paper also describes the policy implications of the migration to multi-service, packet-based networks, and emphasises the need for regulators, as well as operators, to gain an understanding of the new economics of service provision.

Keywords: IP networks; packet switching; capacity planning; service costing; pricing, regulation

Introduction

The telecommunications industry is going through a technological revolution, involving a transition from networks based primarily on circuit-switched technologies to those based on packet switching. At the same time, multiple networks, each designed to meet the needs of a particular service or traffic category, are giving way to integrated networks, able simultaneously to support the requirements of a wide range of services.

These trends are evident in both mobile and fixed networks. The development plans of mobile operators revolve around the growth of 3G networks, designed to carry a wide variety of traffic types, from conventional voice, to interactive data, video and messaging. Many of the major fixed network operators are also busy planning the migration from a circuit-switched, PSTN environment to integrated, multi-service networks based on packet switched technologies. The switch is likely to take 5-10 years to complete, but it is clearly happening.

The casualties of this revolution will be many, and they will include the approaches currently used within the industry to dimension networks, and to cost and price the services carried over them. New methods are required, first to enable network planners and commercial managers to quantify the relationship between traffic

[1] Gareth Davies was until recently a Partner in IBM Business Consulting's Telco Strategy Group, based in London. He can be contacted on tel. +44 (0) 7802 917975 or by e-mail at garethddavies@btconnect.com. Dr Michael Hardt is a Managing Consultant in IBM Business Consulting's Telco Strategy Group, based in London, tel. +44 (0)7808 904351, fax. +44 (0)20 7928 4464, e-mail: michael.hardt@uk.ibm.com. Frank Kelly is Professor of the Mathematics of Systems at the University of Cambridge, tel. +44 (0)1223 337963, fax. +44 (0)1223 337956, e-mail: f.p.kelly@statslab.cam.ac.uk.

volumes, service quality and capacity in the networks they operate and then to determine the costs that are causally attributable to the various services they carry.

This is a more difficult challenge in a multi-service, packet switched world because the relationships involved are much more complex than in a circuit-switched environment. A book of Erlang tables is no longer enough to enable the planner to say how much capacity is required to meet a given service quality standard. An integrated packet network will often handle a wide range of traffic types, each with its own service quality requirements. Part of the difficulty is that service quality is a more complex concept in a packet-switched network, with a number of dimensions which, depending on the service involved, might include packet loss, blocking probabilities, transmission speed, jitter and delay.

The IP Capacity Planning Framework

Although a considerable amount of research has been done into the statistical issues raised by these challenges, it has not yet yielded much in the way of effective tools available to practitioners in the industry. With the aim of going some way towards filling this gap, IBM has recently been working with Professor Frank Kelly of Cambridge University on the development of an IP Capacity Planning (IPCP) Framework.

The IPCP Framework is designed to enable industry practitioners to quantify the relationships between traffic volumes, service quality, capacity, and cost in a multi-service, packet-based environment. It is designed to be sufficiently flexible to be applicable to a range of packet-based technologies and protocols, including 3G mobile networks, ATM, MPLS and IP/TCP.

The IPCP Framework has a number of potential applications, including the following:

- Capacity planning
- Network management e.g. defining rules for traffic prioritisation or trunk reservation
- Capital budgeting
- Service costing
- Cost related pricing
- Portfolio planning and revenue maximisation.

Service Categorisation

The IPCP Framework is underpinned by the view that traffic types, or services, fall into three main categories:

> **Type A: Real-Time Services** – which have strict latency requirements e.g. voice;
> **Type B: Interactive Data and Streaming Services** – which can tolerate a limited amount of delay, but where transmission speed is still a key user requirement e.g. web browsing; and
> **Type C: Delay-Tolerant Services** – which can tolerate more significant delays, without materially affecting the Quality of Service (QoS) perceived by the customer e.g. e-mail, file transfer.

This categorisation is designed to reflect important differences in the way the various services make demands on the network. Real-time traffic is from connection-oriented services that require immediate transmission, such that queuing at network routers must be avoided. Assuming packet transfer delay and delay variation are adequately controlled, the critical QoS parameters are packet loss and network availability (or the probability of blocking). Efficiencies can be achieved by carrying such traffic over a packet switched network, as opposed to a circuit switched network, through statistical multiplexing over very short timescales (measured in milliseconds).

With interactive data traffic, immediate transmission is no longer required, but the user's willingness-to-pay will still depend critically on the time taken for a document (e.g. a web page) to be downloaded. In this case, the relevant QoS measure will not be latency, as defined in a narrow, technical sense (i.e. the round trip delay associated with delivering a data packet) but transmission speed. This in turn will depend on the amount of spare capacity available in the network, measured over periods that are still relatively short (a number of seconds, depending on the size of the documents being sent).

Streaming services are considered in the same category as interactive data, because their capacity requirements can be modelled using the same basic approach. The QoS requirements of both types of service can be defined with reference to the probability that a given transmission speed will be achieved for a certain proportion of the time.

Delay-tolerant services can accept more uncertainty about the transmission time, and significantly slower delivery, without significantly impacting the QoS perceived by the user. Provided the network is capable of managing it, traffic of this type can be delayed for several minutes, until capacity is available, without having any significant impact on service quality.

This service categorisation is broadly similar to the QoS classes defined by the ITU for IP communication. The ITU identifies four service categories, namely: Real Time, Interactive, Non-Interactive and Unspecified[2]. It should be noted, however, that the QoS standards specified by the ITU for each service class are defined from a technical, engineering perspective, whereas the approach adopted here is based more on the user's experience.

More specifically, the ITU defines service quality in terms of the following technical parameters:

- one-way delay, or latency;
- delay variation, or jitter; and
- packet loss.

In the IPCP framework, on the other hand, QoS is measured primarily in terms of:

- packet loss (the percentage of packets that are 'dropped', i.e. not transmitted due to congestion);
- blocking probability (the probability that an admission control mechanism refuses a reservation request for a given flow); and
- transmission speed (the amount of data transferred in a given time interval (kb/s)).

The inclusion of blocking probability and transmission speed allows examination of the impact of congestion on a user's perceived QoS in a way that would not be possible with a narrower, more technical approach.

To illustrate the point, the one-way delay standard of 400 milliseconds for interactive data traffic, which has been set by the ITU, relates to the end-to-end delay inherent in the network, because of its configuration and the technologies used in its provision. Our approach focuses on the delivery time experienced by the user[3], which is a very different concept. For example, the time taken to download a web page may be much greater at peak times because the network has responded to congestion by slowing down transmission speeds. For the user, the important issue is the amount of time it takes to download a document of a given size, which will be inversely related to transmission speed. A slow download may be referred to as "delay", but it is the subjective delay perceived by the user rather than the kind of delay addressed by the technical standards.

[2] See ITU Draft Recommendation Y.1541, Network performance objectives for IP based services, 2002.

[3] This issue has recently been considered by ITU-T Study Group 12, and in the US by Committee T1, Working Group T1A1.3: see for example Technical Report on Performance Parameters for IP-based Applications, ftp://ftp.t1.org/T1A1/T1A1.3/3a130521.doc (2003).

The 3G Partnership Project (3GPP) has also defined four QoS categories for end-to-end service delivery in a 3G network, namely: Conversational, Streaming, Interactive and Background[4]. As with the ITU, the 3GPP QoS standards are defined from a technical viewpoint[5]. In other respects, however, the categorisation is similar to our own, with "Conversational" corresponding to "Real Time" and "Background" corresponding to "Delay-Tolerant."

Four-Stage Approach to Network Dimensioning

Within the IPCP framework, the capacity requirements of these three traffic types are considered in four stages, as illustrated in Figure 1. The first stage of the analysis is a preliminary stage, aimed at assessing the capacity needs of individual services using the concept of Effective Bandwidth, whilst the remaining three stages examine in turn the capacity requirements of each of our three classes of traffic. Each stage is described in more detail below.

Figure 1: The four-stage approach to capacity planning

	Preliminary Stage **Effective Bandwidth Calculation**	**Stage One** **Real Time Services**	**Stage Two** **Interactive Data Services**	**Stage Three** **Delay Tolerant Services**
Service focus	▪ Real Time ▪ Individual services	▪ Real Time ▪ Multiple services	▪ Interactive Data ▪ Streaming	▪ Delay tolerant services
QoS focus	▪ Packet loss	▪ Blocking probability	▪ Transmission speed	▪ None
Output	Effective bandwidth function for each service	Capacity requirements for multiple real-time services	Absolute or incremental capacity requirements of interactive data services	Absolute or incremental capacity of delay-tolerant services

Preliminary Stage: Effective Bandwidth Analysis

The purpose of this stage is to determine the number of connections of a real-time traffic source (e.g. the number of voice calls) that can be carried simultaneously on a network facility with a given capacity, with a given probability of packet loss. This is done using the concept of *Effective Bandwidth* [6], which allows capacity requirements to be calculated on the basis of information on:

- Traffic characteristics:

[4] See 3G Partnership Project Technical Specifications 23.107 and 23.207 on Quality of Service concept and architecture.

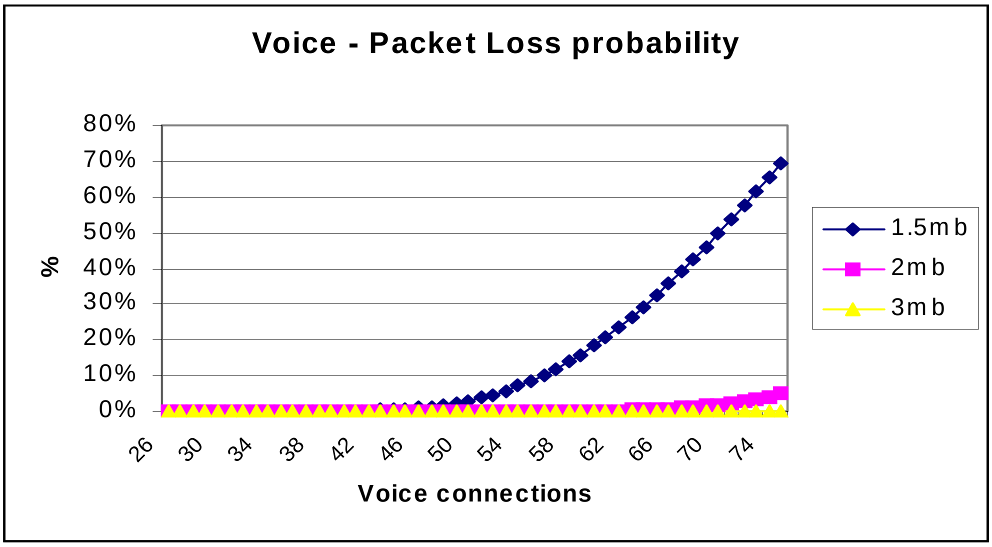
[5] Transmission speed and related traffic parameters have also been addressed in recent ITU work, see for instance ITU-T Recommendations Y.1221 – which is a companion to ITU recommendation Y.1541. Blocking probability is being addressed in a planned new Recommendation, Y.1530.

[6] For discussion of this concept, see Kelly, 1999; and Chapter 8 of Courcoubetis and Weber, 2003.

    - Mean traffic flow;
    - Peak traffic flow; and
- Quality of Service constraints:
    - Packet loss rate

As an example, Figure 2 shows the relationship between the number of simultaneous voice calls and packet loss on links with a capacity of between 1.5 Mbit/s and 3.0 Mbit/s.

Figure 2: Relationship between voice connections and packet loss[7]



This relationship between packet loss probability and the number of connections implicitly defines the effective bandwidth of the service in question: a given packet loss probability is compatible with a maximum number of (in this example, voice) connections. The effective bandwidth of a connection is then given by the capacity of the link, divided by the maximum number of connections compatible with the required QoS level (specified in terms of the allowed level of packet loss).

The ratio of effective bandwidth to mean (or total) traffic can vary considerably by service, depending on traffic peakiness and QOS requirements. This point is illustrated in the table below, which shows the traffic characteristics and Effective Bandwidth of voice connections, and various types of videoconferencing service, when carried on an 8 Mbit/s link. As indicated in the final column, the ratio of Effective Bandwidth to mean traffic can be more than three times as great for a high quality videoconferencing service than for voice.
In other words, per Mbit of traffic, some real-time services impose much greater demands on the network than others. This has important implications for service costing and pricing that will be discussed further in later sections of this article. It should also be noted that this finding holds true, regardless of the QOS management capabilities of the network.

[7] The calculations underlying this chart are based on a peak rate of 64 kb/s and a mean rate of 24.8 kb/s for each voice connection.

Figure 3: Effective bandwidth of voice and videoconferencing connections on an 8 Mb/s link

	Peak traffic per connection (kb/s)	Mean traffic per connection (kb/s)	Required Packet loss (maximum)	Effective Bandwidth (kb/connection)	Ratio EB/Mean
Voice	64	25	0.01%	30	1.2
Videoconferencing:					
MPEG-4 High quality	2000	400	0.1%	1600	4.0
MPEG-4 Low quality	1000	90	0.1%	286	3.2
H.263 – High quality	1400	256	0.1%	1000	3.9
H.263 – Med. quality	320	64	0.1%	105	1.6
H.263 – Low quality	84	16	0.1%	20	1.3

Stage One: Capacity Requirements of Real-Time Services

As noted above, real-time traffic is associated with connection-oriented services. The purpose of Stage One is to determine the capacity required to ensure that an arriving connection faces an acceptable level of blocking probability[8]. This is akin to the Erlang calculation in a circuit-switched network, and involves the use of similar equations.

Amongst other things, this stage of the analysis can be used to assess the efficiency improvement associated with the move from circuit-switched to packet-switched technology. Once the required QoS has been specified, the capacity savings associated with the adoption of packet-switching can be calculated. This is illustrated in Figure 4, which shows the amount of voice traffic (measured in Erlangs) that can be carried over a 2Mbit/s link, with various rates of packet loss and blocking probabilities. The efficiency improvement achievable with packet switching ranges from 128% to 180%, depending on the grade of service required.

Figure 4 – Efficiency improvement associated with the adoption of packet-switching

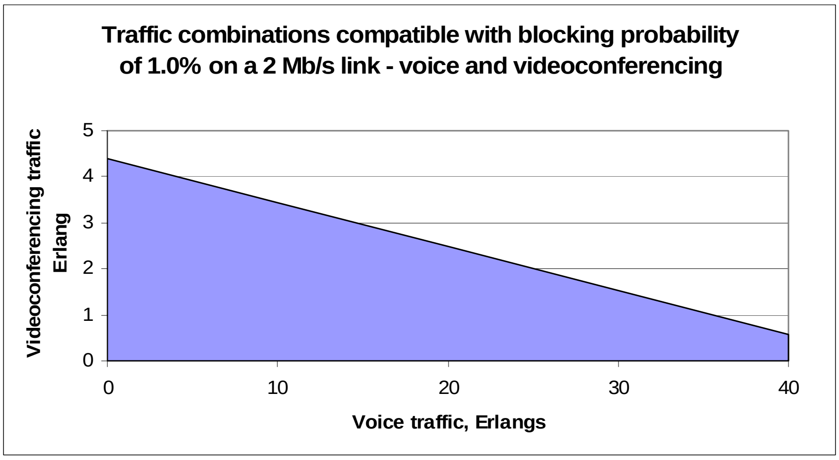
	Blocking Probability 0.1%	Blocking Probability 1.0%
Capacity of 2 Mbit/s link (Erlangs)		
▪ Circuit switched	16.7	20.1
▪ Packet switched:		
-- 0.1% packet loss	39.9	46.0
-- 1.0% packet loss	46.7	53.3
Efficiency improvement with packet switching		
-- 0.1% packet loss	139%	129%
-- 1.0% packet loss	180%	165%

An important feature of this stage is that it can deal simultaneously with the requirements of more than one variety of Type A (**Real-Time Services)** traffic. It can be used, for example, to examine the various

[8] The packet loss probability is taken as given, from the Preliminary Stage.

combinations of voice and videoconferencing traffic that can be carried on a resource with a given capacity. An example of this kind of output is shown in Figure 5.

Figure 5: Relationship between voice and videoconferencing[9] capacities of a 2 Mbit/s link



Stage Two: Capacity Requirements of Interactive Data and Streaming Services

Stage Two of the IPCP Framework considers the capacity requirements of Interactive Data and Streaming Services, in either absolute or incremental terms. These two modes of operation are discussed in more detail in the next section. The volume of this Type B interactive data traffic is specified in terms of:

- the mean data volume during the peak period; and
- an Access Limit which restricts the amount of capacity that can be used by any single traffic source.

The key QoS parameter for Interactive Data Traffic is considered to be transmission speed[10], as this will determine the amount of time taken for a document of a given size to be delivered. Once the required data speed has been specified for the peak period, the capacity needed to handle Type B (**Interactive Data and Streaming Services)** traffic can be calculated.

Separate calculations are made for long and short documents. For long documents, the transmission speed will converge on the average transmission speed achieved on the facility. For short documents, the probability of achieving a transmission speed in excess of a given level can be determined. QoS requirements can be defined independently for long and short documents.

Figure 6 illustrates the sorts of relationship that can be quantified at this stage of the analysis. In this case, the graph shows the incremental capacity required to handle various volumes of Interactive Data traffic (e.g. web browsing) at transmission speeds of 150kb/s and 120kb/s for long and short documents respectively. The amount of Type A traffic (**Real-Time Services** - voice / video) is taken as given. Starting capacity is set just high enough for the QoS requirements to be met for Type A traffic (**Real-Time Services** i.e. no excess

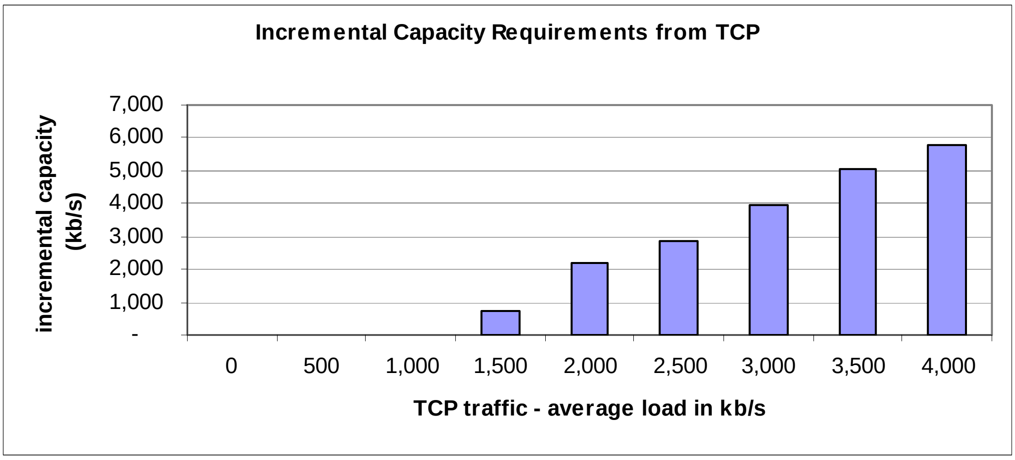
[9] Based on medium quality H.263 videoconferencing service, with peak and mean data speeds of 320 kb/s and 64 kb/s respectively. The required packet loss rate for each service is assumed to be 0.1%, and the required blocking probability for each service is assumed to be 1%.

[10] We use the approach to statistical bandwidth sharing developed by Ben Fredj *et al*, 2001, and Bonald and Roberts, 2003, to calculate transmission speeds and to assess their dependence on load.

capacity). A certain amount of Type B (**Interactive Data and Streaming Services)** traffic (e.g. TCP traffic) can be carried without adding extra capacity, as it can use any spare capacity[11] that is not needed by Type A traffic. Once a threshold amount of TCP traffic is exceeded (here 1,000 kb/s), the incremental capacity required by the TCP traffic increases in proportion to the amount of additional traffic.

.

Figure 6– Incremental capacity requirements from TCP traffic



Stage Three: Capacity Requirements of Delay-Tolerant Services

Once again, the requirements of Type C (**Delay-Tolerant Services**) traffic can be specified in either absolute or incremental terms, and the calculations are in this case comparatively straightforward. It is assumed that the traffic is carried on a best-efforts basis, such that no service quality standards are applicable. Where capacity is shared with other categories of service, the Type C d**elay-tolerant service** traffic uses the spare capacity left over when the requirements of those other service have been met[12]. If this is insufficient, extra capacity is added until all of the Type C traffic can be delivered. If delay-tolerant Type C services are being considered in isolation, the associated traffic is spread evenly throughout the day, or some part thereof.

Nodes and Links

The approach described above needs to be applied separately to each node (e.g. router) and transmission link in the network. It is therefore necessary to specify the rules that will dictate how traffic between any two end-points will be routed through the network. These rules are used, within the IPCP Framework, to convert end-to-end traffic matrices for each service into an offered traffic load for each node and link in the network. Alternative routings can be specified, for use when the preferred route is congested.

QoS Management Capabilities in the Network

[11] The amount of spare capacity (as a percentage of total capacity of the resource) is given by: 1 – mean utilisation rate of the link. This corresponds to an assumption that the relevant timescales are shorter for type A traffic than for type B traffic.

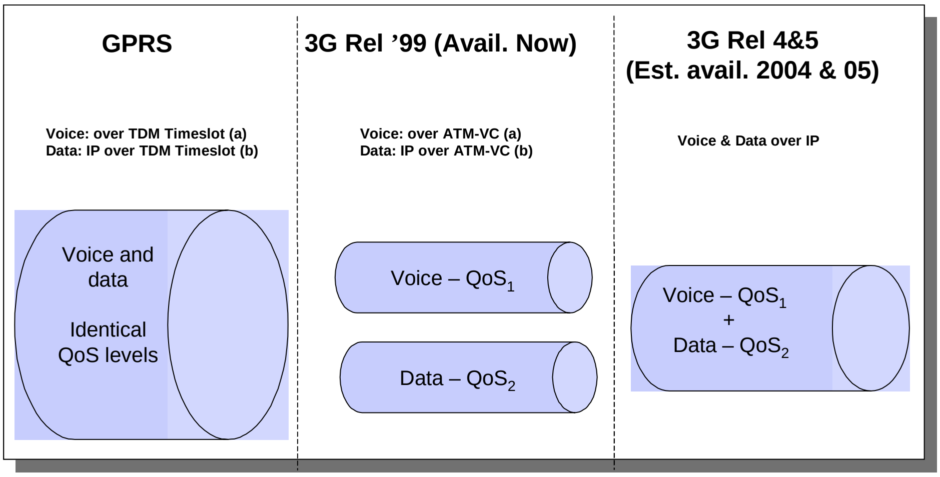
[12] This sort of priority system can be built into a packet switched network through network control functions such as access control or queuing functionality - see Chapter 3 of Courcoubetis and Weber, 2003, and Gevros *et al*, 2001.

When applying the IPCP Framework, account clearly has to be taken of the extent to which the network is capable of offering different levels of service quality to different traffic streams. At one extreme, the traffic streams from a range of services may all be treated in the same way. In these cases, the parts of the IPCP Framework that are applicable will depend on the QoS standard the operator wishes to provide to the services in question. For example, an operator with an IP/TCP network that recognises only one service class may wish to achieve a particular transmission speed target for all traffic, in which case Stage Two of the IPCP Framework would be applicable.

At the other extreme, the network may be capable of supporting different QoS standards for all three of the traffic classes, in which case all four stages of analysis are likely to be required. In this situation, Stages Two and Three of the analysis are likely to focus on the incremental capacity requirements of Types B and C traffic, rather than on their absolute needs.

In practice, the QoS management capabilities of today’s networks are evolving rapidly, from the crude to the sophisticated. The changing capabilities of packet-based mobile technologies are illustrated in Figure 7, which is based on the situation of one mobile operator and may not be applicable to others. In a GPRS environment, voice and data traffic share the capacity available in the Radio Access Network, but have dedicated channels once they reach the core network. Only one QoS standard could be delivered in the shared access network, and the operator would need to decide what that standard should be. Whatever the standard chosen, the IPCP Framework can be used to determine the resulting capacity requirement.

Figure 7: Evolution of traffic management in mobile IP networks



With the migration to Release ‘99 of the 3G technology, the partitioning of capacity extends into the Radio Access Network, allowing different QoS standards to be applied to voice and data traffic. In this situation, the capacity requirements of the different service categories can be considered in turn, using Stages One, Two and possibly Three as well. The Framework would still be operating in “Partition Mode” in the sense that it would be assessing the absolute capacity requirements of segregated traffic streams.

Once Release 5 of the 3G technology is available, the position will change again, as the protocols associated with this Release will be able to support a range of QoS classes, for a number of different services sharing a single facility. At this point, the Framework would be applied in “Shared Use Mode”, and the focus in Stages Two and Three will be on incremental rather than absolute capacity requirements.

Work that IBM has been doing with several of the major European fixed network operators indicates that they too are planning for a migration to integrated, multi-service packet switched networks. Whilst the timescales and migration paths are less certain than in the case of some mobile operators, and may involve time horizons of 5-10 years for full implementation, partial moves in this direction can be expected in the next year or so.

How Much Cost Saving?

Given the pressure currently on both fixed and mobile operators to minimise their capital outlays, one of the main applications of the IPCP Framework is to assess the potential efficiency benefits of moving to an integrated IP network, and then to help ensure that they are achieved.

It is important to note that capacity savings can arise in several different ways. First, they may result simply from the move from circuit switched technology to packet-switching: it has already been seen that savings of this type can be very significant in the context of a single service (see Figure 4 above).

In addition, efficiency benefits are potentially achievable through the statistical multiplexing of several different services on a single packet switched resource. The extent of these savings will vary considerably, depending on the balance of service types involved, their traffic characteristics and the approach taken to service quality management. Figure 8 provides some illustrative data, based on the capacity requirements of given quantities of voice, videoconferencing and interactive data traffic.

Figure 8: Illustrative efficiency benefits of carrying multiple services in an integrated packet switched network[13]

	Voice	Videoconferencing	Interactive Data
Traffic volume:			
- Erlangs	46	4.45	-
- Mean Kb/s	1141	285	1680
Stand alone capacity requirement (SACR) in Kb/s	2048	2048	2048
Incremental capacity requirement (ICR) in Kb/s	1300	590	1098
Total capacity requirement for all three services (Kb/s)	4100		

In this case, each service, if provided by itself, would require a capacity of 2 Mb/s, implying a total capacity of over 6 Mb/s. If all three services are provided together, in an integrated network, the required capacity falls to 4.1 Mb/s, a capacity saving of one third. The achievement of this saving depends on the network's ability to

[13] The characteristics of the voice and videoconferencing services are as in previous figures, with service quality requirements of 0.1% and 1.0% for packet loss and blocking respectively. For interactive data, we assume an access limit of 200 kb/s per data source, a transfer rate of 150 kb/s for longer documents, and a 90% guarantee of a transfer rate of 120 kb/s for short documents. The incremental capacity requirement of each service is measured on the assumption that the other two services are already being provided.

manage the QoS delivered to each service, which in this case means giving priority to the real-time services, and providing some form of trunk reservation, so as to achieve the desired blocking probabilities.

If such QoS management capabilities are absent, achievement of the target QoS for the real-time services would involve providing the interactive data service with a level of service quality in excess of its requirements. In the current example, this would increase the ICR of the interactive data traffic from 1098 kb/s to 1798 kb/s, and reduce the overall cost saving from 33% to 22%.

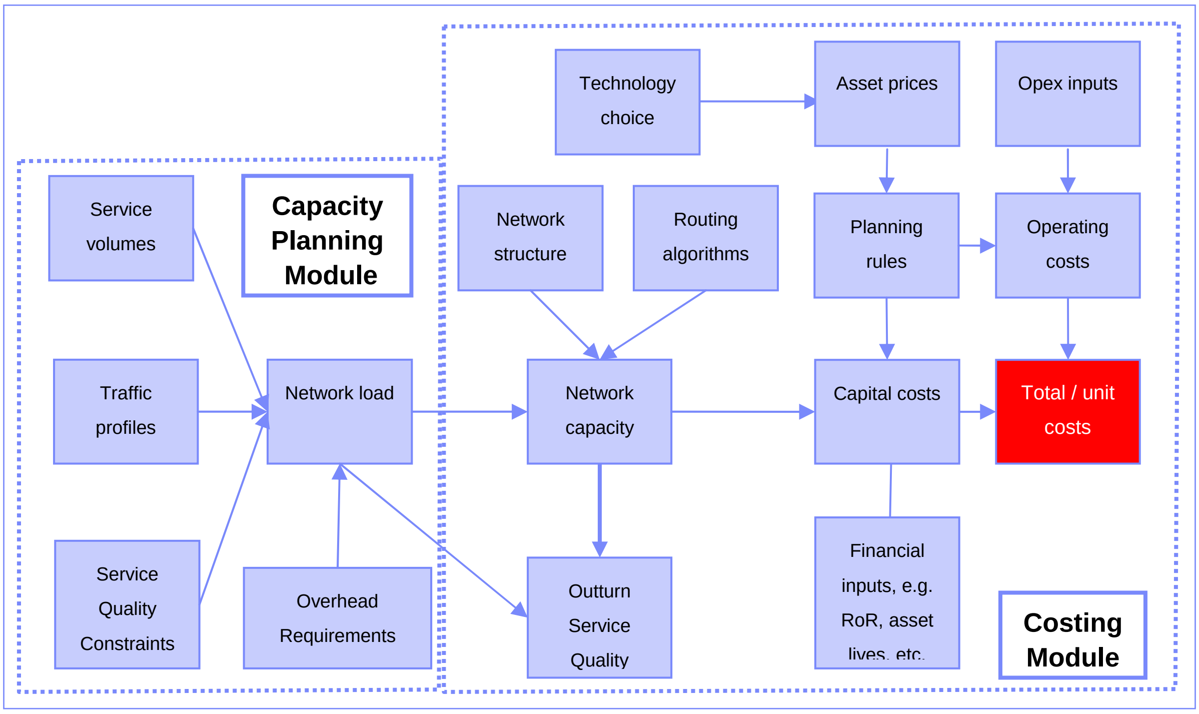
IP Costing

IBM has also developed an IP Costing Module, for use in conjunction with the IPCP Framework. The Costing Module is essentially a "bottom-up" cost model, which takes the capacity requirements determined by the IPCP dimensioning tool and calculates the costs of meeting them, based on the technologies specified by the planner.

The module can be used in either static or dynamic mode, that is to assess provisioning costs at a point in time, or cash flows over a period of years. At an aggregate level, the Module can be used as a capital budgeting tool, to determine the capital investment required to meet projected traffic volumes and QoS targets at the least possible cost to the operator. At a disaggregated level, it can be used to analyse the costs of service provision, on a marginal, incremental, stand-alone or fully-allocated basis.

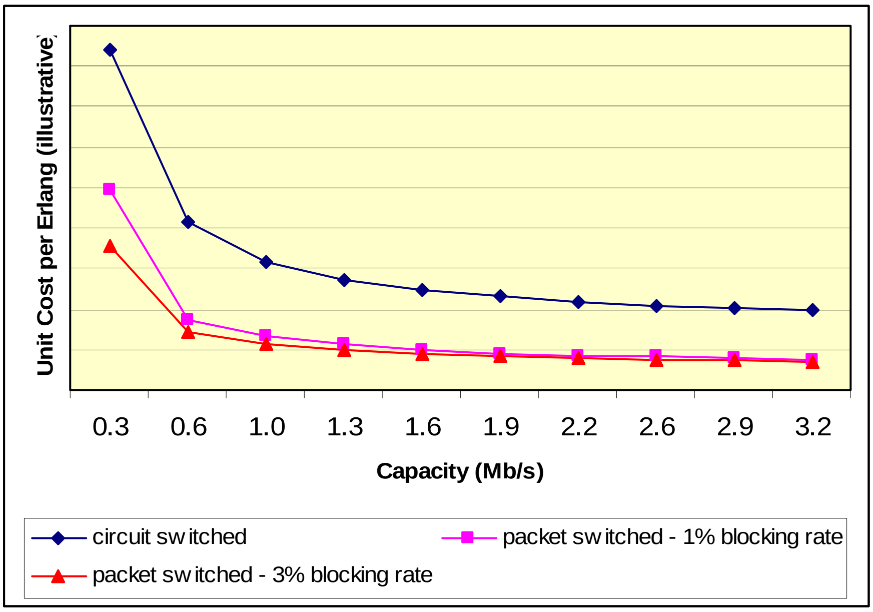
As bottom-up models are already quite widely used for service costing in a circuit-switched environment, particularly in the context of setting regulated interconnect tariffs, it is probably not necessary to go into detail about the operation of the Costing Module. Figure 9, however, provides an overview of the logical structure of the model, and a brief description of each stage of analysis is set out in the Annex.

Figure 9: Overview of IP Costing Module



Illustrative outputs from the Costing Module are shown in Figures 10-12. Figure 10 shows how the average cost of voice traffic varies by volume, in a circuit switched and packet switched environment. Two lines are shown for packet switching, with blocking probabilities of 1% and 3% respectively; the circuit switched line is based on a blocking probability of 1%. The curves have the conventional shape associated with the presence of economies of scale in handling larger traffic volumes. By quantifying the cost of improved service quality, the IP Costing Module can inform decisions about the QoS levels that should be offered to customers in different market segments, and the way in which different grades of service (e.g. gold, silver, bronze) should be priced.

Figure 10: Average cost of voice traffic – circuit switched v. packet switched

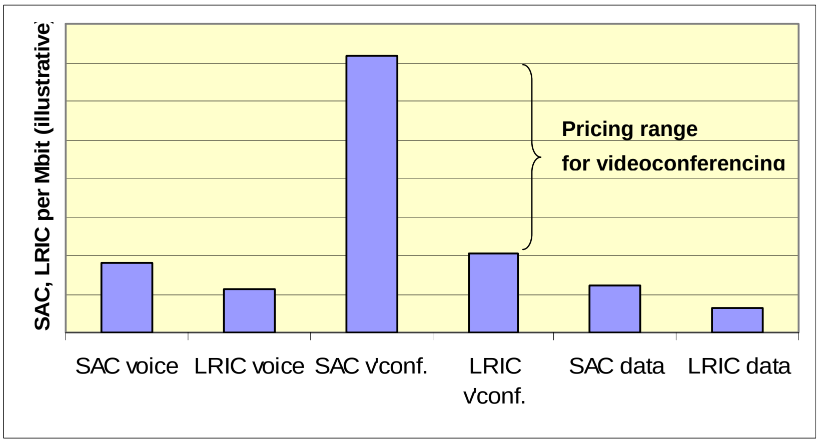


The Long Run Incremental Cost (LRIC) and Stand Alone Cost (SAC) of voice, videoconferencing[14] and interactive data services are illustrated in Figure 11. This chart is based on the traffic volumes shown in Figure 8 above, with the results expressed as unit costs per Mbit[15]. It is noticeable that, not only do the costs per Mbit vary considerably by service, but that the ratios of stand-alone to incremental cost do as well. This sort of information can be of use in making pricing decisions, as sustainable, profit-maximising tariffs will generally lie between the incremental cost floor and the stand-alone cost ceiling of the service concerned.

[14] H.263 medium quality videoconferencing.

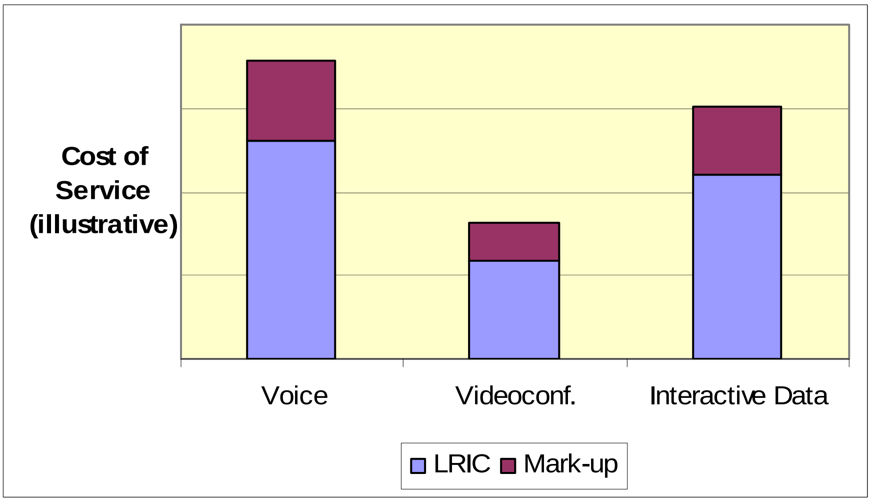
[15] With packet loss rates of 0.1% and blocking probabilities of 1% for the voice and videoconferencing services.

Figure 11: Price floors and ceilings based on incremental and stand-alone costs



In an industry characterised by economies of scale and scope, such as telecommunications, the sum of the LRICs of each service will be less than the total costs of the firm, the shortfall being the fixed common costs of the business. The difficulty with these fixed common costs is that they cannot be attributed to individual services on a causal basis, although they still need to be recovered if the business is to remain adequately profitable. In these circumstances, it can be useful (or necessary in order to meet regulatory requirements) to derive estimates of the Fully Allocated Costs of each service. These are obtained by adding an allocation, or mark-up, of fixed common costs to the incremental cost of each service. Various methods can be used to make this allocation, the most common of which is perhaps an equal percentage mark-up on incremental cost. Figure 12 illustrates this approach, for voice, videoconferencing and interactive video services[16].

Figure 12: Fully allocated costing



Pricing with Uncertainty and Congestion

The evolution of integrated IP networks is likely to be accompanied by the development of hybrid pricing structures with the following elements:

- Flat rate e.g. per month
- Per minute e.g. for voice

[16] The traffic and QoS data used for this chart are the same as for Figures 8 and 11.

- Per Mbit e.g. for file transfer
- Per event e.g. per message

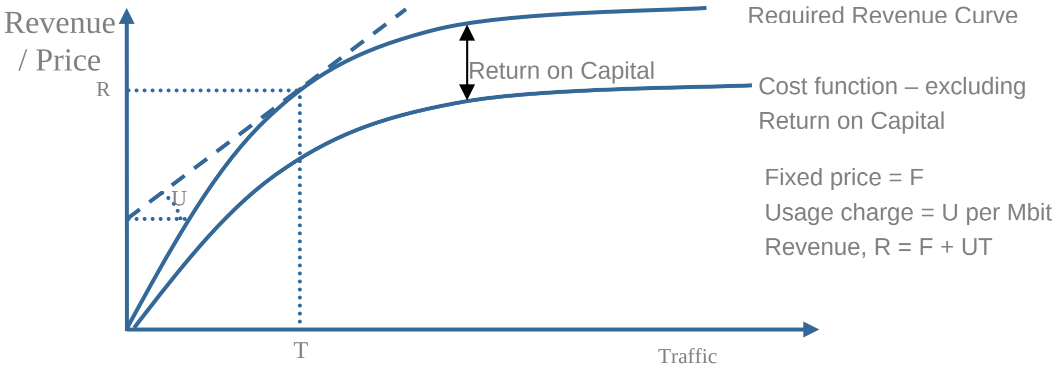
The biggest growth area is likely to be per-event charging, because it enables the suppliers to set prices which for many services are (i) more easily understood by customers than a charge per Mbit (or per minute) and (ii) more closely aligned to the market value of the service being provided. Per event charging for text messaging has proved very lucrative for the mobile operators, who are now looking to repeat the experience with picture messaging over their 3G networks. Per event charging is also well suited to the downloading or streaming of data, music or video, where most of the value resides in the content of the deliverable rather than the manner in which in it is delivered.

The growth of these hybrid pricing structures will depend on a number of factors, including (i) the development of the more sophisticated data capture and billing systems needed for their implementation, and (ii) the establishment of alliances and revenue sharing arrangements between network operators and content providers. Rapid progress is currently being made on both of these fronts.

While prices in competitive markets should normally reflect what customers are prepared to pay, cost information provides a useful starting point for the determination of a profit-maximising charging structure[17]. With this in mind, the IP capacity planning and costing framework has been extended to produce pricing options for the services being carried. The options are generated using a tool called the IP Price Planner, and are designed to meet the revenue requirements of an individual service or group of services.

The basic idea behind the approach is illustrated in Figure 13. The figure shows a Required Revenue (RR) curve for an individual service, which has been obtained by combining the cost function for the service with a target rate of return (simply by adding the required profit or return on capital to the cost function that, in turn, shows the sum of operating costs and depreciation incurred, by level of output). The shape of the curve reflects economies of scale and scope in the delivery of the service. Whilst the RR function is initially determined at an aggregate level, it can be converted into a revenue requirement per user by dividing across the expected number of users. A range of tariff options can then be defined by drawing tangents to the disaggregated RR curve, with each tangent representing a different combination of flat rate and usage charges.

Figure 13: Defining tariff options for an individual service

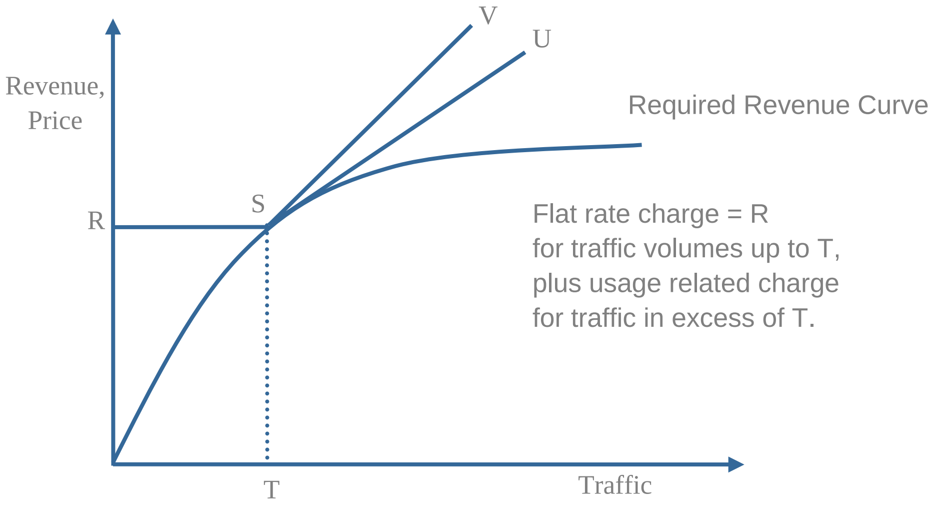


[17] For a pricing approach focusing on the willingness to pay for various levels of QoS see, for example, Walker *et al* 1997.

Typically, the approach would be used to design price packages for different market segments within the user community. The power of the Price Planner lies in its ability to extend this basic method to address (i) the QoS needs of the various market segments, as well as their traffic volumes, and (ii) to produce pricing packages which cover the full range of services being delivered over the integrated IP network. By using the multi-dimensional capabilities of the IPCP Framework and the associated cost analysis, the Price Planner can be used, for example, to design a pricing package tailored to the needs of young mobile customers who generate only modest amounts of voice traffic but a large number of text and picture messages, and periodic bursts of interactive data, associated for example with game-playing on the internet.

Figure 14 shows a variant on this sort of price package, where the customer is charged a flat fee (R) for traffic volumes up a specified maximum (T), with usage charges being applied if the maximum is exceeded. Once again, the Price Planner can be used to determine an appropriate flat fee and usage charge for any individual or group of customers.

Figure 14: Flat fee for a specified traffic volume


V
U
Revenue,
Price
Required Revenue Curve
S
R
Flat rate charge = R
for traffic volumes up to T,
plus usage related charge
for traffic in excess of T.
T
Traffic

One benefit of such pricing options is that they give users an incentive to reveal accurate information to the service provider about their expected traffic volumes, as by choosing the right tariff option they will minimise their expenditure. This in turn may provide information that is useful to the service provider, in managing and if necessary augmenting its network capacity. This sort of information will be especially useful during periods of great uncertainty over future traffic volumes, as is currently the case, for example, with 3G data services.

In the IP networks of the future, one of the key network management issues will be how to deal with congestion. Broadly, the three available options are (i) to block the excess traffic, (ii) to deliver it with a lower QoS (e.g. slower transmission speeds, or higher packet loss), or (iii) to charge a higher rate for carrying it. At some stage, dynamic charging structures may well adjust tariffs on a second by second basis in order to choke off congestion-causing traffic (see Henderson, Crowcroft and Bhatti 2001). In the nearer term, cruder methods are likely be required, such as applying higher usage related charges for traffic above the expected levels. In terms of Figure 14, this would be equivalent to increasing the slope of the line SU (which is a tangent to the RR curve), and charging instead on the basis of line SV.

[19] See for example, Oftel, *Frequently asked questions on the regulation of Voice over Internet Protocol services*, issued by the Director General of Telecommunications, 2 April 2002.

If congestion surcharges are to be applied, they should be based on the costs imposed by the excess traffic on other users of the network. One of the capabilities of the IP Price Planner is that it provides a basis for estimating congestion costs, in terms of the opportunity cost of the capacity that would be occupied by the excess traffic. The opportunity cost is determined in this case as the value of the displaced traffic and/or the consequential decline in the QoS of other services.

Regulatory Policy Implications

Not surprisingly, the migration from circuit-switched to packet-based networks will have profound implications for the regulation of the telecommunications industry. Perhaps most obviously, it will be a key enabler for growth in the market for broadband services. As a result, there will be increasing importance attached to regulation of broadband access, which to some degree is likely to remain a bottleneck, at least for the foreseeable future. For fixed network providers, this will mean continued, and perhaps increasing regulatory intervention to facilitate Local Loop Unbundling and ensure the availability of wholesale DSL-type services. For mobile operators, it is likely to involve a requirement to demonstrate that 3G mobile access is being made available to Mobile Virtual Network Operators (MVNOs) and service providers on non-discriminatory, and perhaps cost-related, terms. Current moves towards the introduction of accounting separation in the mobile sector (e.g. in Ireland) represent a clear trend in this direction.

The second implication is that a number of the existing regulatory structures will need to be recalibrated using the sort of IP-based service cost models described in previous sections of this paper. We are referring here to the regulation both of retail prices and interconnect charges, using price caps or other controls, designed to cover the costs of service provision, including a reasonable return on capital employed. The increasing use by network operators of packet-based networks to originate and terminate PSTN (Public Switched Telephone Network) traffic, and to provide leased line or special access services, will often be invisible to the end user. However, the costs involved in delivering interconnect services in this way will be different, and new methods will be needed to measure them.

One example of this effect concerns 3G mobile termination. It seems likely that the recent regulatory assault on mobile termination rates will at some stage be extended to 3G networks, as the economic arguments for regulatory intervention are essentially the same. Moreover, operators may otherwise be able to circumvent the existing regulations by using 2G capacity to originate calls and 3G capacity for call termination. Similar arguments will apply to the origination and termination of calls on the fixed network. In all of these cases, regulators and operators will need to settle on new techniques for assessing, in particular, the Long Run Incremental Cost of PSTN traffic and VPN (Virtual Private Network) services in an integrated, multi-service, packet-switched environment.

Thirdly, important changes in market structure are likely to follow from the convergence of traditional PSTN and VOIP voice services. Up to now, regulators at both EU and national level have used differences in QoS and other factors (e.g. numbering arrangements) to justify treating the two as separate markets[19]. This has fitted well with their desire to avoid regulating the internet and to promote the take-up of IP services. But with the deployment of IPv6, and other approaches which support service differentiation, the QOS gap can be expected to narrow, and the principle of technological neutrality (a key element in the new EU framework) is likely to mean that the two markets will eventually be treated as one.

At the retail level, this is likely to mean less regulation rather than more. The convergence of the two markets will mean that established network operators such as BT will face increasing competition from new generation VOIP service providers, reducing the need for the perpetuation of retail price caps on the voice services of incumbent operators.

At the interconnect and wholesale level, however, the likelihood of regulatory intervention is likely to grow, rather than diminish. Where network operators are considered to enjoy significant market power, they may be required to offer cost-related VOIP interconnect services, in parallel to their conventional PSTN offerings. This could apply to mobile operators, who may find VOIP call termination rates subject to regulation, as well as to fixed incumbents.

This point links fairly directly to the fourth effect, which concerns the proliferation of services in the new environment. The growth of new service categories, each with their own QoS and cost characteristics, will inevitably lead to more complex charging and accounting arrangements for IP traffic. At one level, this means that the commercial arrangements for interconnection between ISPs will become more sophisticated, with usage related pricing and service level agreements replacing peering arrangements. Faced with this trend, signs of which have been evident for several years, regulators are likely to continue to leave the setting of interconnect charges to commercial negotiation, but they may be called upon to intervene in the event of a dispute.

At another level, service proliferation will put increasing pressure on the Flat-Rate Internet Access Call Origination (FRIACO) regimes which have been introduced by a number of regulators, as a means of promoting the take-up of internet services. At some point, the economics of provisioning will require the charging structure for internet access to be modified to take account of QoS differences and traffic volumes, at least at the wholesale level, even if these differentials are not reflected directly in the retail market.

Last but not least, the migration to integrated packet-switched networks is closely linked to the emergence of a more complex industry structure, which will inevitably attract a significant amount of regulatory attention. This point can be illustrated with reference to Figure 15, which shows layer models for what Fransman calls the “old” telecommunications industry and the “new” infocommunications industry[20].

[20] See Chapters 1 and 2 of Fransman, 2002.

Figure 15: Layer Models of the Old Telecoms Industry and the New Infocommunications Industry

Old Telecoms Industry	New Infocommunications Industry
Layer - Activity	*Layer* - Activity
	VI *Customers*
	V *Applications Layer, including contents packaging* – e.g. web design, on-line information services
	IV *Navigation & Middleware Layer* – e.g. browsers, portals, search engines
3 *Services Layer* - voice, fax, 0800	III *Connectivity Layer* – e.g. internet access, web hosting
	IP Interface
2 *Network Layer* – circuit switched network	II *Network Layer* – e.g. optical fibre network, DSL, radio access network, ISDN, ATM etc.
1 *Equipment Layer* - switches, transmission systems, customer premises equipment	I *Equipment Layer* – e.g. switches, transmission systems, customer premises equipment, routers, servers etc.

Source: Fransmann 2002, pages 18 and 37

The key feature of the new industry model is the greater complexity of the upper layers, in which competition will occur between:

- vertically integrated network operators, using alliances to expand their activities in the provision of middleware, applications and content; and
- vertically specialised suppliers with no network facilities of their own.

In this situation, regulators will, as always, be keen to ensure that suppliers with significant market power at one point in the value chain do not use that power to distort competition in other, related, markets. This is a large and complex subject, and of course it is not possible to say how the structure of the industry will evolve. But it is reasonable to suggest that those at most risk from such regulatory intervention will include the fixed and mobile operators who control network access, as well as perhaps the owners of unique and valuable content. An alliance which gave a powerful network provider unique access, say, to Premiership football highlights, would be near the top of the regulatory hit-list, fuelling the demand for regulatory intervention to promote open, non-discriminatory access.

Under the new EU regulatory regime, national regulators have strong investigatory powers and the ability to impose heavy penalties on firms found to have behaved in an anti-competitive manner. Notwithstanding the objective of 'light-handed' regulation, there will be a risk of tough, ex-post regulatory intervention that firms should take into account when developing their broader commercial strategies.

All of these effects mean that policy-makers and regulators, as well as operators, will need to develop a good understanding of the economics of service provision in integrated IP networks. The need for this knowledge is likely to be at least as great, if not greater, in the new industry environment, than it has been in the past.

Conclusion

The transition to integrated IP networks will fundamentally change the economics of telecommunications service delivery. Our analysis, though illustrative, has shown that in a fully integrated IP network, costs per Mbit of traffic are likely to vary significantly by service, in accordance with a range of factors including: the volume of traffic per source relative to the capacity of the facility; the peakiness of the traffic; and the quality of service required. The merit of our approach is that it enables these relationships to be quantifed.

Network operators and service providers who fail to appreciate the significance of these changes, and the implications for their own activities, are unlikely to succeed in the new environment. Conversely, those who stay ahead of the curve will be well placed to take advantage of the many commercial opportunities available as the new networks and services evolve towards maturity. Amongst other things, operators will need to develop networks capable of delivering different levels of service quality to a variety of traffic classes, and tools capable of planning and managing them. These network capabilities will need to be mirrored in the business planning and marketing arena, through new methods of service costing and pricing, in the BSS/OSS arena e.g. via the development of new systems for data capture and billing, and in the regulatory arena, through the adoption of policies designed to reflect the economics of the new networks. With the help of tools such as the IPCP Framework, the IP Costing Module and the IP Price Planner, operators and regulators will be better placed to rise to this challenge.

References

Kelly, F.P. 1999, *Charging and Accounting for Bursty Connections*, in Internet Economics, Lee W. Knight and Joseph P Bailey, Eds., Cambridge, Massachusetts, MIT Press, pp253-278.

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

Fransman, M. 2002, *Telecommunications in the Internet Age: From Boom to Bust to…?*, Oxford OUP.

Ben Fredj, S., Bonald, T., Proutiere, A., Regnie, G. and Roberts, J. 2001, *Statistical Bandwidth Sharing: A Study of Congestion at Flow Level*, ACM Computer Communications Review 31 (4) 111-122.

Walker, D., Kelly , F. and Solomon *J.*, 1997, *Tariffing in the new IP/ATM environment, Telecommunications Policy (*21) 283-295.

Henderson, T. Crowcroft J. and Bhatti, S., 2001, *Congestion pricing: paying your way in communication networks*, IEEE Internet Computing, Vol. 5, Issue 5, pp85-89, Sep/Oct 2001.

Gevros, P., Kirstein, P., Crowcroft, J. and Bhatti, S., 2001, *Congestion Control Mechanisms and the Best Effort Service Model*, IEEE Network, Vol.15, no.3, pp16-26, May/June 2001.

Bonald, T. and Roberts, J.W., 2003, Congestion at Flow Level and the Impact of User Behaviour, Computer Networks Vol. 42, pp521-536.

Annex: Brief Description of IP Costing Module

The main stages in the analysis are as follows:

A. Service Volumes

If the model is to be used to determine network capacity requirements, the modeller must specify the number of network termination points (NTPs) and the traffic volumes between each pair of NTPs, for each service type. Traffic volumes are specified in terms of mean and peak traffic loads. Alternatively, the modeller can specify the capacity of the network, the number of NTPs and the required levels of service quality, and then use the model to determine the carrying capacity of the network.

B. Traffic Profiles

The variability of traffic by time of day must also be considered. The modeller will normally specify traffic flows separately for (i) the network peak period and (ii) the day as a whole. Capacity requirements will be driven by volumes during the network peak period, which may be defined as a busy hour, or as some other period if preferred. All-day (or total annual) traffic volumes are used in the calculation of unit costs.

C. Service Quality Constraints

The modeller will typically specify the service quality requirements of each service type, as described in the main body of the paper. These will act as one of the drivers of capacity. Alternatively, if capacity and traffic volumes are given, the model can be used to assess the resulting service quality.

D. Overhead Requirements

Network planners will normally wish to increase network capacity in order to allow for factors such as future traffic growth, internally generated traffic (e.g. for testing) and resilience. The model allows for uplift factors to be applied, to take account of these requirements.

E. Network Load

Elements A-D above define the traffic load which the network should be designed to accommodate, in terms of the traffic flows between any two NTPs. This a key input into the capacity planning process described in the main body of the paper.

F. Network Structure

As well as specifying the number of NTPs, the modeller is required to define the structure of the network, in terms of the number of nodes and links it contains. The information on nodes and links is combined with data on routing patterns (see G below) and network load, to determine the load on each part of the network. Capacity requirements are then calculated separately for each link and node in the network.

G. Routing Algorithms

The routing algorithms define the path through the network, to be taken by traffic between any pair of NTPs. Provision is made for the traffic between any two points to be split between two or more paths if required. If capacity is fixed, the modeller can also specify alternative routings, to be applied in the event of network congestion.

H. Network Capacity

The required capacity of each node and link in the network is calculated on the basis of the network load, the network structure and the routing algorithms, using the steps described in the main body of the paper.

I. Outturn Service Quality

Once the capacity and traffic load for each part of the network has been defined, the model will calculate the service quality that will actually be achieved, by service type, again using the steps outlined in the main body of the paper.

I. Technology Choice

The modeller is then required to specify the type of equipment used to provide each node and link in the network, so that the relevant asset prices can be identified. This might include, for example, the manufacturer and model of the routers to be used at network nodes.

J. Asset Prices

The prices of the relevant equipment types must then be specified by the modeller. For a network operator, these will normally be the prices currently being paid, or expected to be paid, for such equipment.

K. Planning Rules

The rules referred to here are those used by the network facility planners to translate network load and capacity requirements into equipment lists. These rules can either be applied off line i.e. outside the modelling framework, or simplified for inclusion in the model. The latter approach can provide much faster turnaround times, though at the expense of some degree of accuracy. Non-network and fixed capital costs are specified separately.

L. Financial Inputs

These inputs relate to the cost of capital, the asset lives of the various plant and equipment categories and the method to be used to calculate depreciation. They are required in order to convert the asset prices and equipment lists into annual capital charges.

M. Capital Costs

Capital costs are expressed as annual charges for specified service traffic increments. These charges cover both the cost of capital and depreciation, and can be calculated using a variety of methods e.g. tilted annuity.

N. Opex Inputs

Operating costs are handled more crudely than capital costs. Causally attributable network and non-network operating costs may be expressed as a percentage of the capital cost or the annual capital charge. Non causally attributable overhead costs are expressed as lump sums.

O. Operating Costs

Annual operating costs are calculated on the basis of the opex inputs, taking account of equipment volumes and asset values or annual capital costs as appropriate.

P. Total/Unit Costs

In the final stage of analysis, capital costs and operating costs are brought together to give total costs, and these are divided by the relevant traffic volumes to give costs per unit. Incremental, stand alone and/or fully allocated costs can be calculated, depending on the definition of the service increment and the treatment of fixed common (overhead) costs.