Market Managed Multiservice Internet

Huw Oliver, Dave Songhurst[1]
Internet and Mobile Services Laboratory
HP Laboratories Bristol
HPL-IRI-2000-3
29th June, 2000*

internet service providers, application service providers, mobile service providers, network charging

There are currently many proposals for pricing and charging models but, in the absence of real experience of providing commercial communication services for a wide variety of applications, the proposals are generally based on plausibility arguments. Within M3I we intend to carry out large scale simulations and experiments and modelling analyses to gather information about the applicability of differentiated pricing through usage-based and dynamic charging schemes.

Our key goal is to gauge the effectiveness of market forces in allocating network resources to end-systems according to their application requirements and utilities.

1 BT Research, Ipswich, UK
* Internal Accession Date Only
Approved for External Publication

1 Introduction

Internet technology is becoming the infrastructure of the future for any information that can be transmitted digitally, including voice, audio, video and data services of all kinds. The traditional industries that supply such services are undergoing a major disruption. The new suppliers of such services are operating in a maelstrom. The problem has derived, in part, because of the historic roots of the Internet in the military and, later, academia. Neither of these communities had any motivation to build the infrastructure needed to price, charge and collect payments for the provision of Internet services.

Here is a simple overview of the problems some of these industries are facing.

1.1 Internet Service Providers

Internet Service Providers (ISPs) currently make money from

- Subscription charges (perhaps monthly for an individual; for a larger enterprise an annual fee plus initial connection charge plus customer premises equipment charge etc.)
- Advertising
- E-commerce (we are seeing the beginnings of this)
- Technical Support

In some countries (for example, the UK) ISPs receive a "kickback", that is a portion of the charge for telephone access from the telecommunication operator. The telcos often view this as an abuse of data service lines (set up for telephone banking, for example) but have not received a very sympathetic ear from the regulators who want to see Internet usage increased.

But the ISPs only rarely make their money from the actual services they provide. Currently ISPs provide access and some basic services but those are becoming commoditised. Above the network service, ISPs would like to provide advanced services (VoIP, Intranet, Extranet, integrated e-mail/voice messaging, content, relationship management, etc.) and would like to build sophisticated business models around those services.

Another disruption will occur when network quality of service mechanisms start to appear. Then there will be many customers with business critical needs who will willingly pay for the better quality of network service. ISPs are, in general, unprepared for this event.

Finally, the current peering models of carrying peer ISP's traffic free are breaking down. The asymmetric volumes of traffic are leading larger ISPs to refuse to carry the traffic from smaller ones for free.

1.2 Application Server and Service Providers

The computing industry is being transformed by the ability to outsource computing services over the Internet. Five years ago running a large piece of business software meant making large investments in installation, maintenance, upgrading and deployment and support staff. Now services are being provided on a pay-per-usage basis. Application service providers own the application (complex data analysis tools or sophisticated business software) and provide it on-line to a wide range of different enterprises. Sometimes a separate Application Server Provider will host the application and provide network access to it.

Their problem is that charging for these services has to be done in an ad-hoc manner. Quality of service is being provided, but in a way that is exclusive to the customers of the ASPs.

The ASPs would like to provide their customers with flexible ways to pay for the services. At the moment the model is similar to traditional software licensing - pay a fixed fee whatever your level of use. More desirable is usage charging both for datasets and for level of speciality of software. This would allow smaller suppliers to enter the market more easily and would allow customers to tailor their use to their budget.

1.3 Telecommunication Operators

For the telecommunication operators the move to an IP based infrastructure is an opportunity as much as a problem. Current billing systems are

- Centralised (data is transported from around the network to a central processing farm via physical tape or X.25 or FTAM).
- Inefficient (bill production takes months)
- Expensive (up to half the charge in a typical telephone bill is for putting that bill together)

The PSTN billing systems have typically taken hundreds of man-years to construct. Re-implementing that software for the Internet would be a nightmare. For the telcos, the Internet philosophy of intelligence at the edges applied to billing is intriguing.

Finally, PSTN Telephony Traffic Theory has been good for half a century. The single service connection-based telephone networks and basic human behaviour have not changed significantly during that period. This all changes with the multiple voice/video/data services of the packet switched Internet (and will get worse when multiple qualities of service arrive).

1.4 Mobile Telephony Operators

Enormous investments are being made in the future of the packet switched mobile services (GPRS and UMTS). In May 2000 the UK auctions for UMTS bandwidth raised a total of 22.48 billion pounds. How will such investments be recouped?

Current charging models for mobile communications are inappropriate for mobile packet networks. The reason is that these are connection-oriented and only take into account the duration of the connection through the access network. Such schemes do not account for the actual traffic that goes through the connection and work as if the connection was always fully utilised. The wrong incentives that these pricing schemes present to the users are to keep connections only during the time data are being transmitted, and hence incur a high signalling overhead for setting and tearing down connections at the time scales of the bursts of the data.

While traditional connection charges may be used in the initial absence of other mechanisms, this position is unsustainable.

The problem, common to all these industries, is that an inappropriate pricing system will convey the wrong incentives to the end users and lead to inefficiency, reduced profitability and ultimately lead to congestion.

2 Solutions through network charging

The introduction of network charging aims to tackle these problems. It has (at least) three inter-related objectives:

- to differentiate services, giving users the option to pay more for specific services or for better service quality
- to ensure network efficiency through suitable incentives so that users will constrain their demands appropriately
- to ensure fairness, so that network resources are shared according to need

There are three types of charging mechanism that could be used, possibly in combination, in order to compose a sound market-driven resource allocation mechanism: differentiated service charging, usage-based charging, dynamic charging.

2.1 Differentiated service charging

Customers are able to subscribe to different services at differentiated prices depending on the facilities being offered or the level of service quality. Pricing may be flat rate or combined with usage charges. Differentiated flat rate pricing is a way to achieve market segmentation but still suffers from the problems with regard to efficiency and incentives. It may be difficult to offer high quality services economically unless some form of usage charging is used to constrain demand.

2.2 Usage-based charging

Telephony networks have traditionally used duration-based charging with distance-related tariffs. The development of multiservice packet-switched networks based on asynchronous transfer mode (ATM) led to proposals for a range of new services, including variable bit-rate services with varying levels of service guarantees. Research into new charging schemes for variable bit-rate traffic led to the European ACTS project CA$hMAN, which implemented and trialled a range of novel charging schemes (Songhurst, 1999). These included tariffs based on a combination of duration and volume charges for each connection, designed to approximate the “effective bandwidth” of the connection. Effective bandwidth (Kelly, 1997) is a measure of the resource that the network needs to reserve for a variable bit-rate connection in order to ensure the required service quality.

In the Internet usage-based charges have taken the form of charges for access time and charges for total data volume. However, unlike ATM the Internet protocol is not connection-oriented and so does not facilitate the concept of reserving resources for individual connections This makes it difficult to relate usage charges to service quality for individual connections. Developments such as resource reservation protocol (RSVP) are an attempt to counter this, but there are other reasons for seeking a different approach in the Internet:

- The Internet is inherently decentralised – the concept is a simple network with intelligence existing in the end-systems.
- There is an endless possible variety of user applications that will use the Internet. It is not feasible to second-guess these and offer differentiated network services that will match their requirements.
- Bandwidth is cheap. Intense competition has driven ISPs to flat-rate, or free, Internet service, at least for low-rate access.

These factors lead towards a decentralised and dynamic approach to charging and service quality.

2.3 Dynamic charging

The marginal cost of carrying data over the Internet when there is capacity available is essentially zero. However with no usage charging there is no way to control congestion and ensure that users can get good quality service. This leads to the concept of charging for usage only when there is contention for resources. The charge that users should pay is a *shadow price* that reflects the cost to other users through increased congestion.

This principle has been recognised in telephone networks through the simplistic approach of peak and off-peak pricing. However, considerations of competition and network efficiency should lead to a genuinely dynamic charging system where charges for usage vary in real time in relation to congestion in the network. The Internet does offer the possibility of a simple means to achieve this through the use of packet marking (Gibbens and Kelly, 1999).

3 M3I – Market-Managed Multiservice Internet

In January 2000, Hewlett-Packard Ltd, Athens University of Economics and Business, Technical University of Darmstadt, BT, ETH Zurich and Telenor began a two year project called M3I to build and test network service charging approaches.

3.1 Objectives

In the M3I project we propose the use of pricing mechanisms for controlling demand for scarce resources, in order to improve the economic efficiency of the system. Standard results in economic theory suggest that increasing the value of the network services to the users is beneficial to both the users and the network operator (since he can charge them more and get back a bigger percentage of their surplus). Using pricing mechanisms helps in that respect. When demand is high, prices are being raised and hence deter the users with low valuation for the service to use it. This leaves resources to be available for the users that value them more, and hence are ready to pay more.

As noted earlier, prices can work in many time scales. *Dynamic prices* work in short time scales and reflect the instantaneous state of the network in terms of excess demand for resources. These are in many cases called *congestion prices* since they reflect the cost due to performance degradation that a user imposes to the other users that share the same network resources. The work in Gibbens at al. (1999) is an attempt to define congestion prices in an implementable way for packet networks (the Internet), where these prices achieve the economically fair allocation of the resources of the network to the competing connections, see Kelly et al. (1998). Prices that work in larger time scales (time-of-day pricing) can also be used in our framework. The idea is that differential pricing makes users "self-select" and choose the service quality they prefer for the posted price.

In general, the difficulty of designing an appropriate pricing and charging system comes from the need to solve three problems simultaneously. The solution must make a good economic solution, technical solution and end-user solution. Taking these separately:

3.1.1 Economic Issues

The solution must provide the correct incentives, be fair and be the basis of a sustainable business. It must also be sufficiently flexible to allow new business models to be created around it.

We have seen the high cost of putting the current telecommunications billing systems in place. We must be sure to avoid doing this again. The issue is particularly acute for any measurement-based approaches. While Internet measurement techniques have advanced in recent years, capturing Internet traffic behaviour analytically is notoriously difficult. The speed, computational and storage requirements of measuring traffic at a low level and recreating high level statistics are largely unknown.

If dynamic pricing is used there needs to be a clear and auditable relationship between the advertised current price of service use and the final customer bill.

When end-systems choose their own rate control algorithms to suit a wide variety of different applications, will the network operate in a stable way? We need a solution whereby rates converge to stable values rather than fluctuate randomly or cyclically.

3.1.2 Technical Issues

Internet congestion effects take place at many different levels of timescale. Many periods of congestion have a duration of just a few milliseconds. These are far too short to be dealt with by any changes in users' behaviour. There is also the issue of duration of user flows. These are often of too short a duration to include any user reaction to network state. Studies have been done, however, to show that the presence of such short flows do not lead to instability [12].

The primary thrust of a market managed approach is to use price as an admission control to the network. It needs to be provably effective if it is to replace the traditional approaches. These two forms of admission control will also live side by side for the foreseeable future. What will be the global effects of having both price and technical admission control for different flows?

Flow measurement indications and price signals will lead to extra control traffic. How significant will this be? We must ensure that we have a scalable wide area measurement solution.

Finally, we have to be clear about what congestion means. Flows typically cross multiple hops across multiple domains. They are not connection oriented and different packets within a flow may take different routes. The term "network congestion" is an oversimplification when the reality is that we have a widely distributed set of transiently congested resources.

3.1.3 User Issues

Perhaps the most widely levelled criticism of usage charging and dynamic pricing is that the user cannot predict the final bill. Internet usage is increasing dramatically and hence users worry that a usage based charge might lead to a similar dramatic increase in their bills. Studies have shown that predictability has a high value [11].

A second issue is that of relating user level service utility, such as sending an email or attending a teleconference, to lower level service charging. A packet charge, megabyte charge or even flow charge means little to the average end user. The end user needs to be able to see the overall price attached to their high level service in order to decide whether to use it.

Pricing and charging schemes aim to shape the end user's behaviour to one that leads to high network efficiency and maximum social utility. There is, however, often too great a disconnect between the user's control and how an application generates traffic. We need a solution that puts bandwidth consumption control in the user's hands simply and effectively.

Finally, we need a solution that will work for the many different kinds of user. The main categories are that of residential user and corporate user. While a residential user might be very sensitive to charges, a corporate user will only be indirectly motivated to limit costs. We need a solution that allows a corporate policy to be enforced without overly restricting the individual corporate users.

3.2 A mechanism for congestion pricing

Ramakrishman and Floyd (1999) have proposed the implementation of Explicit Congestion Notification (ECN) in TCP. Resources that are nearing congestion can signal this fact to end-systems by marking packets before it is necessary to start dropping packets. The authors of this proposal suggest that end-systems should back-off in response to marked packets. However Gibbens and Kelly (1999) propose that congestion marks should be interpreted as charges, and end-systems should be free to vary their rate in accordance with their utility and the charges received. Various algorithms for marking packets are possible – a simple method is to mark all packets arriving at a resource when its input buffer exceeds a given threshold.

With congestion charging, end-systems will no longer use the standard TCP rate control algorithm. Instead they will use algorithms that depend on the needs of the user application with parameters that could be set by an end-user. One can distinguish at least three broad types of application having different rate control requirements:

3.2.1 Adaptive applications

Many applications are able to adapt their rates over a broad range without a major loss of utility to the end-user. For such adaptive applications Kelly et al (1998) propose a variant of TCP rate control in which one parameter can be set by the application, in response to network charging signals, so as to ensure convergence to a rate that maximises the user’s net utility. This ‘willingness-to-pay’ parameter identifies the end-system’s desired level of payment per unit time, and available bandwidth is shared between flows in proportion to these parameter values.

3.2.2 Real-time non-adaptive applications

Applications of this type, such as high-quality audio and video, are unable to adapt below a certain rate. They will exercise rate control to maintain at least this minimum rate regardless of pricing signals unless the charge becomes so large that the flow decides to terminate. An application initiating a new real-time flow will exercise a form of self-admission control by testing the network bandwidth price before deciding whether to commence. Gibbens and Kelly (1999b) also discuss the use of gateways between a group of users and the network,

able to use congestion pricing information to perform a distributed acceptance control function.

3.2.3 File transfers

File transfers require to transfer a file of a given size within as short a period as possible, generally with no benefit to the end-user before it is completed. Examples of this type of connection include traditional file transfers (transfers of data files using ftp applications), email, and many components of Web pages (image files etc). File transfers do not benefit from shared bandwidth – their optimum rate control is to transmit at peak rate when the price is low enough, otherwise to suspend transmission. Existing TCP rate control does completely the wrong thing for this wide class of application. Gibbens and Kelly (1999) and Key and Massoulié (1999) discuss possible file transfer algorithms with congestion charging.

The key issue is that congestion charging enables rate control, and hence network resource allocation, to be determined intelligently by end-systems according to their application requirements and utilities.

3.3 Architecture

The architecture consists of the following main pieces:

- Pricing and Admission Control. Here are the mechanisms for setting prices for services, communicating those prices to the end user and the APIs that allow the user to react to that pricing information.
- Charging and Accounting. Here are the mechanisms to measure service usage and to charge according to the prices communicated as above. The aim will be to ensure that the accounting system is as generic as possible, with the ability to plug in appropriate meters that are specialised to measure different aspects of the network and of services provided over the network.
- Below these the Internet is providing raw services. The mechanisms for providing those services at different qualities are still under development in the IETF and one of the challenges for the architecture is to be neutral toward ultimate choices of QoS mechanism as far as possible.

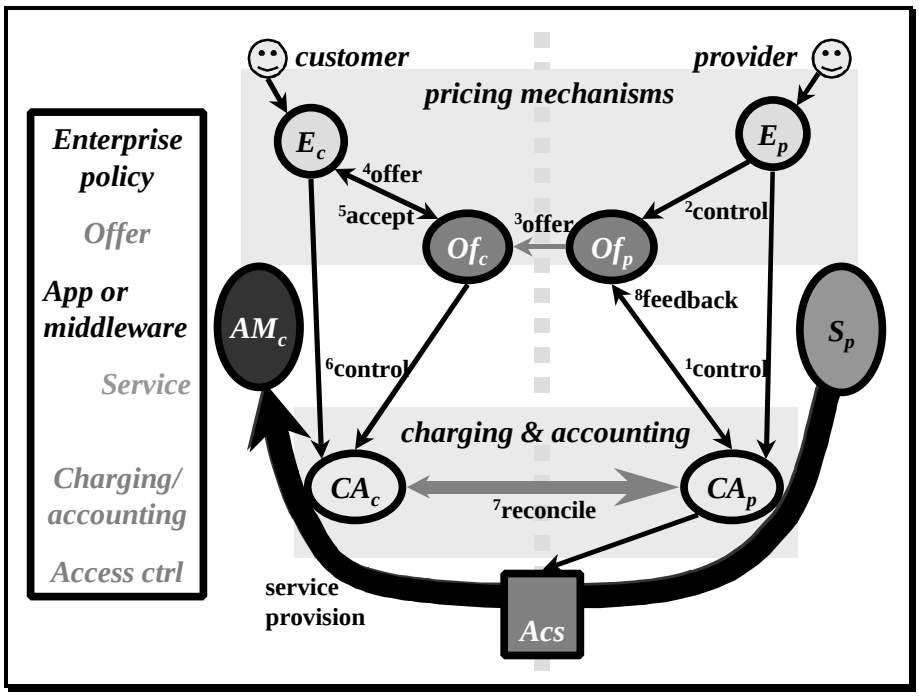

Figure 1

Another challenge is that of scalability. We propose that edge providers charge their local customers for both sending and receiving (more accurately for each class of service in each direction at a separate price). Thus, we extend charging recursively to apply at the boundary between any pair of providers. We also claim that as long as all customers are usage-charged for both sending and receiving by the network providers at all edges, both multicast and unicast charging can be achieved very scalably. While Figure 1 represents the customer-provider relationship at the network edge, it can also represent the relationship *between* network providers. In this case the 'customer application', AM_c, would be just another network service in a chain ending eventually at an edge-customer, but otherwise the architecture is recursive as shown in Figure 2.

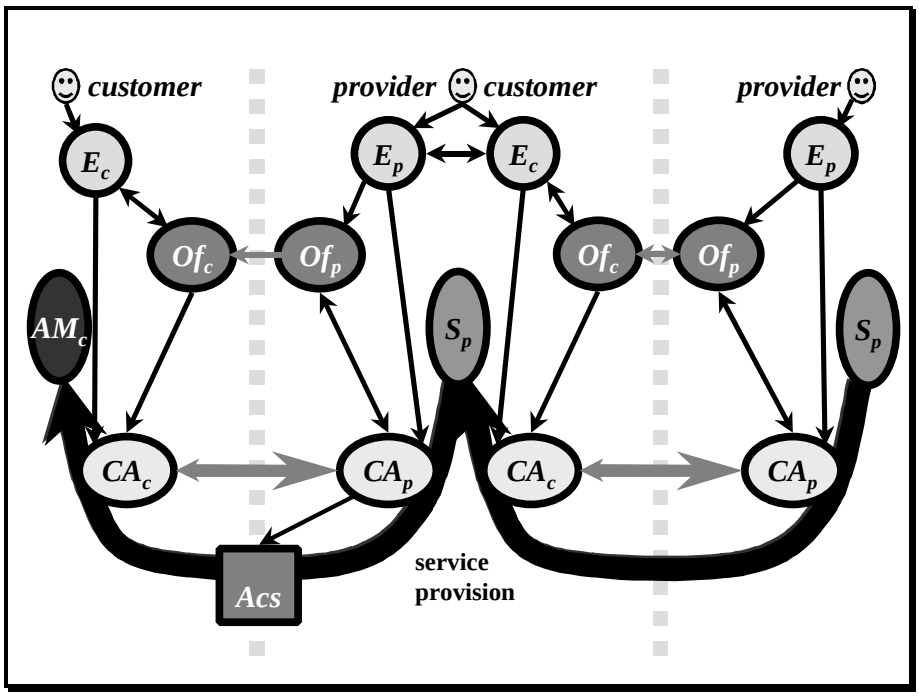

Figure 2

3.4 Planned approach

M3I is tackling the following tasks in order to pursue the objectives described above.

3.4.1 Test platforms

The project will develop test platforms that will be used to evaluate the above architecture, including the implementation of dynamic charging schemes.

3.4.2 Trials

The test platforms will be used to carry out trials, including user trials, which will investigate issues such as

- User sensitivity to quality and price, and user/network interface required to communicate this.
- Use of agents to automate user reaction to dynamic pricing.
- How corporate customers can manage price reaction policies for groups of users.

3.4.3 Modelling

Modelling is a major part of the project, supplemented where possible by results from trials. It includes the following:

- Development of cost models for Internet service providers.
- Formulating business models for ISPs who offer differentiated services through charging.
- Analysis and simulation of networks using dynamic charging in order to evaluate stability.

- Economic models of competitive markets involving many ISPs having differing business models and using different charging schemes. An important aim is to analyse the competitive advantages of a dynamic market-based approach to charging.

4 Summary

There are currently many proposals for pricing and charging models but, in the absence of real experience of providing commercial communication services for a wide variety of applications, the proposals are generally based on plausibility arguments. Within M3I we intend to carry out large scale simulations and experiments and modelling analyses to gather information about the applicability of differentiated pricing through usage-based and dynamic charging schemes.

Our key goal is to gauge the effectiveness of market forces in allocating network resources to end-systems according to their application requirements and utilities.

Partners

The M3I project is a joint undertaking by key members in the Internet Industry: Hewlett-Packard Ltd, Athens University of Economics and Business, Technical University of Darmstadt, BT, ETH Zurich and Telenor.

References

[1] R.J. Gibbens and F.P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, 35, 1999.

[2] F.P. Kelly, A. Maulloo and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, 49, 1998.

[3] C.A. Courcoubetis and M.I. Reiman, "Pricing in a large single-link loss system", *Proc. 16th International Teletraffic Congress*, pp 737-746, Elsevier (1999)[4] MacKie-Mason and H. Varian, "Pricing the Internet," in *Public Access to the Internet*, B. Kahn and J. Keller (eds.), Prentice Hall, 1994.

[5] A. Odlyzko, "A modest proposal for preventing Internet congestion", available in http://www.research.att.com/amo/doc/modest.proposal.ps.

[6] D.J. Songhurst (1999), editor. *Charging Communication Networks: from Theory to Practice*. Elsevier, Amsterdam. ISBN 0-444-50275-0.

[7] F.P. Kelly, "Charging and accounting for bursty connections". In L.W. McKnight and J.P. Bailey, editors, *Internet Economics*, pp 253-278, MIT Press, Cambridge MA, 1997.

[8] K. Ramakrishnan and S. Floyd, "A proposal to add Explicit Congestion Notification (ECN) to IP", RFC2481, The Internet Society, January 1999.

[9] R.J. Gibbens and F.P. Kelly, "Distributed connection acceptance control for a connectionless network", *Proc. 16th International Teletraffic Congress*, pp 941-952, Elsevier (1999).

[10] P. B. Key and L. Massoulié, "User policies in a network implementing congestion pricing", Workshop on *Internet Service Quality and Economics*, MIT, December 1999, http://www.marengoresearch.com/isqe/index.htm

[11] www.index.berkeley.edu

[12] F.P. Kelly, "Models for a self-managed Internet", Royal Society discussion meeting on Network Modelling in the 21st Century, London, December 1999, http://www.statslab.cam.ac.uk/~richard/research/topics/royalsoc1999/