Chapter 1

Charging in communication networks

1.1 Multiservice networks

Charging, accounting and billing are crucial features of telecommunication services. How should the network provider design tariffs for the range of services offered? This is partly a marketing decision - tariffs must be attractive to customers - but network providers are also concerned with efficiency and cost-recovery. Charging schemes should encourage efficient use of the network and should generate revenue in a fair way according to the relative usage of customers.

In multiservice networks, tariffs might depend on a number of parameters defining the traffic and quality of service characteristics of a connection, in order that charges should reflect network resource usage. The way that a customer uses the network depends on the tariffs and also on how the customer values each type of connection (the customer's utility, in the language of economics). This interplay between tariffs, network resource usage, and customer incentives is a fertile area for economic and mathematical models.

Multiservice networks need to include facilities for charging, accounting, and billing. In this context, charging designates the calculation of a charge for a connection. This is calculated based on some characteristics of the connection, according to a charging scheme, which in turn is part of a tariffing policy. Accounting involves gathering the information necessary so that total charges can be itemised against tariffs and usage measurements. Billing involves collecting charge information over a given period and communicating this to the customer in the form of a bill.

Another important concept is advice of charge, where a customer can be given on request the charge for a specified call (whether intended, ongoing, or just completed).

A tariffing policy may include both connection charges and subscription charges. We will say that charging is usage-based when connection charges are included. Usage-based charges may also include subscription charges that are not related to usage (some authors have used the specific term usage-sensitive to refer to charges which could have both components).

Paying for what you get, and getting what you need

The role of charging is not only to cover the costs of service provision and generate income for the service provider, but also to influence the way that customers use network services. This happens as each individual customer reacts to tariffs and seeks to minimize charges. The tariff structure should provide the right incentives for users to use network resources efficiently. This is
the key idea of incentive compatibility. Tariffs should guide customers to select services and use the network in ways that are good for overall network performance.

Tariffs which are not incentive compatible give wrong signals and lead users to use the network in very inefficient ways. One example of this is the Internet which faces intense congestion problems due to its ineffective pricing structure, which is based primarily on flat rate pricing. Under flat rate pricing, charges depend only on the rate of the access pipe which connects customers to Internet service providers. Such a pricing scheme provides no incentives for users to use less bandwidth than the rate of their access pipe. Furthermore, flat rate pricing does not enable users to adequately reveal their preferences for network usage. All users are treated the same, even though different users might have a different value for the same service. Both of these limitations result in a congested network where resources are not used according to the actual needs of users.

Usage-based charging is necessary for incentive compatibility, and economic theory suggests that usage-based charging will be employed where there is perfect competition. This is important in view of the worldwide process of deregulation which is increasing the competitive nature of the telecommunications market. However charging schemes will also be determined by marketing and strategic decisions, customer preferences, and the cost and complexity of implementing and operating these schemes.

The difficulty of charging in multiservice networks

Traffic characteristics and user behaviour in telephone networks have been widely studied. An important feature of telephone networks has been that once a user is granted a connection, the resources associated with the connection remain reserved throughout its duration. Furthermore the quality-of-service for each connection is the same. Unlike such networks, in broadband multiservice networks the full bandwidth does not need to be reserved for the whole duration of the connection, but is used on-demand. Since in broadband networks there will be a large number of connections carrying bursty traffic which share network resources (link bandwidth and buffer space), there are significant gains in statistically multiplexing such connections. Multiservice networks are intended to carry traffic with different characteristics, which may vary in time, and support connections with different quality-of-service (QoS), expressed in terms of loss probability and delay. Clearly the quality of the network service should reflect the charge for the service since the network, in order to guarantee it, must reserve a corresponding amount of resources for the connection.

The amount of network resources used by a connection and the QoS experienced by the user also depend on the statistical properties of the traffic generated by the connection. Within the telecommunications and computer industries it is possible to discern two extreme approaches to this issue. One (impractical) approach is to expect the user to provide the network with a full statistical characterization of traffic, in advance, which is then policed by the network. Another approach stresses the difficulty for a user of providing any information on traffic characteristics, and expects the network to cope nevertheless. The correct balance will necessarily involve trade-offs between the user’s uncertainty about traffic characteristics and the network’s ability to statistically multiplex connections in an efficient manner. A desirable property of a charging scheme would be to encourage the cooperative sharing of information and characterization effort between the user and the network. This can be realized if tariffs encourage cost-minimizing users to make a more accurate characterization of some statistical properties of their traffic. This information can then be used by the network to more efficiently multiplex user connections.

The cost of charging

An important requirement of a charging scheme is that it is efficiently implementable. By this we mean that the information required by the charging scheme should be easy to obtain and to manipulate. It is well known that accounting and billing are major parts of the total cost of telephone
1.1. MULTISERVICE NETWORKS

networks. Hence, a prerequisite for a realistic charging scheme is low cost of implementation and operation. Current technology allows sophisticated traffic measurements to be done in hardware, which greatly expands the spectrum of charging functions which are feasible to implement. However, detailed statistics are costly to manipulate and store. Hence, there is a trade-off between the amount of statistical information gathered (which would allow more accurate characterisation of a user’s traffic) and the cost of gathering, storing, and manipulating such information.

Services

Multiservice networks will, in general, offer two types of basic services which differ in the performance guarantees offered by the network: guaranteed services and elastic services.

With guaranteed services the network provides some form of quality-of-service (QoS) guarantees in terms of loss probability and delay. A key concept for such services is the traffic contract between the user and the network. From the part of the user, this contract specifies the constraints that the traffic must satisfy. These constraints are expressed in terms of filters (usually in the form of leaky bucket constraints), placed at the entrance to the network, which are used to police the user’s traffic. From the part of the network, the traffic contract specifies the QoS the network must guarantee.

Guaranteed services are subject to connection admission control (CAC) whereby a user sends a connection request to the network specifying desired QoS and a traffic description. The network uses this traffic contract information to check if it has enough resources to satisfy the request. Once a connection is accepted, the network will need to reserve some amount of resources in order to satisfy the connection’s QoS requirements. Such network control is also known as open loop control. Examples of such services include the Constant Bit Rate (CBR) and Variable Bit Rate (VBR) services defined by the ATM Forum. These services are very close to the Deterministic Bit Rate (DBR) and Statistical Bit Rate (SBR) services defined by the ITU-T for Broadband Integrated Services Networks (B-ISDN). The guaranteed service and controlled-load service recently defined by the IETF for the Internet’s integrated services architecture are also examples of services providing some performance guarantees.

If charges are to reflect resource usage, it is clear that tariffs must take into account the QoS and traffic description of connections, since the amount of resources the network needs to reserve depends on both. However the traffic contract parameters alone will not accurately determine resource usage. This is because the contract only confines the user traffic to lie within specific ranges determined by the traffic description. Since the user will not in general need to produce the worst case input allowed within the contract, charging according to this worst case traffic would not give the right incentives. From the above discussion, it becomes clear that in order to create tariffs with the right incentive properties we need to combine traffic contract parameters with actual measurements. How this is done is closely related to the statistical multiplexing capability of broadband multiservice networks.

Unlike guaranteed services, elastic services do not have specific performance guarantees. As a result, performance during overload periods deteriorates. Such services are intended for applications that can adapt their sending rate to varying network conditions. To support this, elastic services employ closed loop congestion control which provides, throughout the duration of a connection, feedback signals to the user. These feedback signals provide an indication of the congestion inside the network. In the case of congestion the user’s traffic rate must decrease, whereas in the absence of congestion the traffic rate is allowed to increase. Examples of elastic services are the traditional TCP/IP service in the Internet and the Available Bit Rate Service (ABR) in ATM networks.

The design of charging schemes for elastic services should take into account the feedback control mechanisms of such services, and address issues of network optimality and convergence. In this direction, it is important to identify a key role of charging, namely to lead the network to stable
operation and economically efficient utilization of network resources. This implies that charging should be closely integrated with congestion control, whose role is to minimize the extent and spread of overload conditions which lead to information loss.

### Service-level and content charging

Multiservice networks should have the capability to offer higher-level services that utilise the basic transport-level services. In ATM networks the transport-level services are provided by ATM Layer Transfer Capabilities such as DBR, SBR, and ABR. Services such as real-time video, Internet access, etc., may be provided over transport-level services by Service Provider organisations. These service providers must consider how to charge for higher-level services, and in particular how their charges relate to charges for transport-level services.

For some types of service (such as video on demand) the service provider also wishes to charge for content. Connection charges may be bundled in with content charges so that they cannot be distinguished by the customer.

This book is primarily concerned with charging at the transport level, and in fact the CA$hMan project was aimed mainly at charging schemes for ATM Layer Transfer Capabilities. However we will also discuss (in chapters 4 and 8) how service-level and content charging might be implemented over CA$hMan charging schemes.

### 1.2 Internet

The Internet started in the late 1960s as a United States Department of Defense ARPA (Advanced Research Project Agency) funded project, initially called ARPANET [78]. In the early 1980s a new family of protocols were specified for the ARPANET and associated networks; these are commonly referred to using the names of the two basic protocols TCP/IP (TCP:Transmission Control Protocol, IP: Internet Protocol). In 1987 the U.S. National Science Foundation (NSF) funded a network connecting the six U.S. national supercomputer centers. This network was called NSFNET and served the research and academic community. At the same time initiatives in Europe and worldwide developed networks based on the TCP/IP protocol suite. Today the Internet has grown to become a ubiquitous network used as a communication and information tool for researchers, students, teachers, business and the general public.

The remarkable success and growth of the Internet has convinced many agencies, organizations and researchers that it will play an important role (but perhaps not in its present form) in the evolving broadband network infrastructure. Even today, network operators are offering both Internet and ATM services. Both services are bit transport services, hence their pricing models are not independent. Work is underway in the IETF (Internet Engineering Task Force) to enhance Internet’s current best-effort only service with real-time services [43, 36]. The IETF’s service model is very close to the ATM Forum’s service model, suggesting that the two pricing models will share many commonalities. Internet’s changing nature [96] and the problems of congestion have made it clear that its economic model (in which flat rate pricing where tariffs depend on the customer’s access pipe is the most common) needs to evolve. This has prompted numerous related studies and workshops (cf. [44, 28, 89, 91]), which have shown, among other things, the importance and need for usage-based pricing. Because of its widespread use and availability, the Internet provides a valuable testbed for testing pricing models in order to understand the practical implications under large scale, real-world environments, and the users’ response to various pricing models [51].

We begin by describing the aspects of Internet technology that affect pricing, identifying the areas where it differs from ATM. Next we describe recent proposals for pricing Internet services. Our

---

1 ARPA is now called DARPA (Defense Advanced Research Projects Agency)
1.2. INTERNET

Objective is to present the underlying key ideas and understand their relation with the networking technology.

Internet technology and costs

The Internet’s success can be attributed to the effectiveness of statistical sharing of network resources (e.g., communication lines, routers), the positive network externalities from being connected to the network (i.e., the ability to effectively communicate with a large number of users), and the openness of the underlying communication protocols (namely TCP, UDP/IP) along with the interoperability of the different implementations. Specifically, the Transmission Control Protocol (TCP) deals with end to end issues (e.g., message segmentation and reassembly, flow control, retransmissions) while the Internet Protocol (IP) is responsible for routing packets to their destinations. Each packet contains all the necessary information for routing (source and destination addresses), and is routed independently. This forms the basis of IP’s connectionless communication paradigm. This characteristic, which is the cornerstone of the Internet’s success is what makes accounting difficult: since all packets are independent at the IP level, how can a source be accounted for the amount of resources it uses? On the other hand, in connection-oriented architectures, such as ATM and the telephone network, accounting is, in principle, simpler. Due to its connectionless nature, the Internet’s transport service resembles the postal service. However, this analogy cannot be extended to the economic aspects. In the Internet, the cost of sending an extra packet is essentially zero (when there is no congestion), which is not the case with the postal service where labour forms the largest percentage of the total cost.

In more detail, the cost of the Internet consists of the following [82]

- The incremental cost of sending an extra packet. In the absence of congestion this is essentially zero.
- The congestion costs, or social costs of delaying other users’ packets. As in the previous, this cost is zero when there is no congestion.
- The fixed costs of the network infrastructure (e.g., routers, communication lines), maintenance, and management.
- The incremental costs of connecting to the network. This involves the cost of the access lines and customer premises equipment needed to connect to the network. This cost represents the largest portion of the total cost for an organization to connect to the Internet [91].
- The cost of expanding the capacity of the network.

Fixed costs constitute the major percentage of the total costs. On the other hand, the marginal or incremental costs are non-zero only in the presence of network congestion. This observation leads to the proposition that under normal (i.e. uncongested) operation, network charges should include only fixed charges. However, in the presence of congestion, charges should also include a non-zero usage charge which depends on the level of congestion and the magnitude of the users’ contribution to it. The effect of congestion on prices can be expressed either with dynamically adjusted prices or time-of-day sensitive prices.

From the previous discussion, one understands that congestion is tightly related to the pricing. In fact, many believe that usage-sensitive pricing provides the only practical way for tackling the problem. The current Internet offers a single best-effort service where all packets are treated the same. However, this can be unfair since a user can obtain an increasing share of capacity by sending a continuously increasing number of packets, while ignoring packet losses, at the expense of conforming TCP connections. Some proposals for Internet pricing provide a way to differentiate packets, providing different service to different packets. Furthermore, this service differentiation
translates to monetary value: packets receiving better service are charged more than packets receiving lower quality service.

Besides charging, there are other methods for controlling congestion. One approach involves overprovisioning the network. However, experience has shown that the demand for bandwidth has always been ahead of the supply. In addition, router technology advances at the same rate as the technology used by the end-systems. This suggests that demand will continue to overtake supply in the future. Another approach for controlling congestion involves engineering solutions, e.g., priority queueing. However, such solutions fail to give users the incentive to use the offered services rationally. This incentive problem can be solved with pricing.

Internet’s integrated services architecture

Recent work within the Internet Engineering Task Force (IETF) has focused on enhancing the Internet’s service model, which currently consists only of best-effort services, in order to support services with some guarantees [43, 36]. This work has led to the development of two new service types: guaranteed service [98] and controlled load service [105]. With the guaranteed service type the network provides a deterministic delay guarantee, whereas with controlled load service type the network provides service close to that provided by a best-effort network under lightly loaded conditions. Hence the controlled load service offers a bandwidth guarantee but not a hard delay guarantee. It is important to note that for both the guaranteed service and controlled-load service types the user traffic is described using a leaky bucket, which is also used for traffic description in ATM networks. This traffic description is part of the traffic contract between the user and the network. Because of these similarities (traffic contract, leaky bucket characterisation) the same pricing structure used for guaranteed services in ATM networks can potentially be used for both guaranteed services and controlled-load services in the Internet.

Internet pricing: current status

The most common model for Internet pricing is to charge a flat rate which depends on the size of the customer’s access pipe. Many believe that such a model is suboptimal and unable to tackle the problems of congestion. What flat rate pricing lacks is the ability to give users the incentive to use network capacity rationally. Under flat rate pricing, a low volume user, who occasionally uses up to the peak bandwidth that his access pipe allows, is charged the same as a high volume user who continuously has his access line working at a high utilisation. With usage-based pricing, customers with high-capacity access are not charged the same, rather their charge depends on the level of usage. In this way, a provider can offer more flexible and competitive tariffs. Interestingly enough, some Internet Service Providers have started to offer usage-based pricing for high capacity access pipes. Flat rate pricing is not without advantages. Predictable network charges and essentially zero accounting overhead are among the most important.

Work on Internet pricing

Bohn et al [30] present a scheme to differentiate user traffic based on the precedence field in the IPv4 header. Specifically, end users set the precedence field depending on the level of service they require. Intermediate routers maintain more than one queue for packets with different precedence values and implement a priority service discipline rather than simple first-in-first-out (FIFO). Packets with a higher precedence are placed in a queue with higher priority, hence experience better service (lower delay) in periods of congestion. A quota system can be used to discourage users from always selecting a high precedence. If one relates quotas to monetary units, the resulting scheme charges based on (precedence) priority level independently of the level of congestion. An
advantage of this proposal is that it can be gradually implemented in the parts of the Internet where congestion is a problem.

MacKie-Mason and Varian [82] propose a “smart market” approach to pricing. According to the approach, each packet contains a “bid” which indicates how much a user is willing to pay for the transmission of the packet. Routers queue packets in the order of decreasing bids, hence packets with a higher bid experience less delay. All end users whose packets were transmitted pay the cutoff price which is the value of the bid of the last packet the router served. This price equals the equilibrium supply equals demand price. Advantages of the scheme are that optimal prices can be computed as soon as packets, with their respective bids, arrive at the switch. In addition, users do not have any advantage in lying, hence their best strategy is to truthfully declare their bids. The disadvantages of the scheme include the additional overhead in the packet header, and most importantly the complexity of the service discipline in the switches which has now become tightly interconnected with the pricing scheme.

Gupta et al [65] model the Internet as a collection of servers (e.g., providing entertainment, news, database services) offering a number of priority classes. Servers post prices and expected waiting times for each priority class they offer. Based on these values, clients (users) are free to select the priority class that maximizes their benefit. The aggregate user demand depends on the prices, hence the server can control it (hence the average delays) through the posted prices. It is proved that there exists a unique welfare maximizing allocation of resources, and with simulation it is demonstrated that approximately optimal prices can be computed using a decentralised algorithm. The algorithm is motivated by the classic tatonnement process where the price is increased when the delays at a priority queue are excessive and vice-versa. A user selects the priority class that maximizes his benefit, and the network adjusts prices using exponential averaging\(^2\). The price updates use online measurement of the average flows and time average estimates of expected waiting times and do not require knowledge of the demand functions.

Both the scheme in [82] and the scheme in [65] present a method to obtain (or at least approximate) optimal prices. The optimal prices depend on the level of congestion, and provide necessary feedback required for network administrators to decide whether investment for capacity expansion is justified. In [82], optimal prices are computed as soon as packets with their respective bids reach the router. On the other hand, in [65], an iterative approach was used to obtain optimal prices. In this case the convergence time of the algorithm is important. This time tends to be large for high latency networks. Neither of these schemes take account of interconnection of networks, which is a predominate characteristic of the Internet.

Odlyzko [90] presents an approach to pricing the Internet called Paris Metro Pricing (PMP), due to its resemblance to the pricing structure of the Paris Metro. The basic idea is to partition the Internet into several logical networks, each with separate and nonsharable resources. The (fixed) price of bandwidth would be different for each logical network, and the expectation would be that higher priced networks would be less congested than lower priced networks. Hence, the scheme allows a user who requires better performance to switch to a higher priced, and less congested logical network.

Clark [42, 41] takes a different approach to pricing the Internet. The main objective of the approach is to discriminate users at times of congestion. The scheme allows different users to obtain a different share of capacity at times of congestion, by purchasing a profile or expected capacity. One possible way to express user profiles is the two parameters of a leaky bucket. Based on his profile, packets are marked as in or out depending on whether they are within the user’s profile or in excess of it. In the absence of congestion all packets (both those marked in and those marked out) receive the same service. On the other hand, in the presence of congestion routers preferentially drop packets marked out since those packets have exceeded the profile purchased by the user. A user’s expected capacity is not a guarantee from the network to the user. In fact, even under

\(^2\)If \(\pi\) denotes the posted price then \(\pi \leftarrow (1 - \alpha)\pi + \alpha \bar{\pi}\), where \(\bar{\pi}\) is computed from the equilibrium conditions using online measurements of the average flows and time average estimates of the expected waiting times.
overload conditions, the network does not guarantee this capacity. The expected capacity represents the capacity a user expects to be available to him, and provides a method allowing different users to obtain a different share of network capacity during periods of congestion. An important advantage of the scheme is that internally the network switches are required to implement a simple scheme where in periods of congestion, packets marked out are preferentially dropped. Buying expected capacity rather than peak rate of the access link allows a customer to vary his profile depending on his demand without changing or upgrading his access link. At the same time, the network provider can better dimension his resources based on the expected capacity he sells, rather than the sum of the peak rates of his customers’ access links. However, the author does not present specific methods or approaches on how this can be achieved. Similarly to the scheme in [30], prices do not depend on the state of congestion. Finally, the author does not address the issue of how user profiles (expected capacities) are assigned a price value.

Related to the above is the Premium Service Model of Nichols et al [88]. This scheme guarantees that a user’s contracted capacity is there when he needs it. When the capacity is not used, it may be used by best-effort traffic. At the entrance of the network, traffic belonging to premium service is marked with the premium service bit set. Inside the network there is a separate queue, with higher priority, for premium service data and for normal best-effort data (which has the premium service bit cleared). The newly formed Differential Service for the Internet Working Group with the IETF [62] is investigating the policy and architectural issues for supporting differentiated services in the Internet, based on servicing packets with different priority or with different dropping preferences.

Rather than present a specific pricing scheme, Shenker et al [97] focus on structural and architectural issues such as the local control of pricing policies, multicast charging, and receiver charging. In the proposed architecture, charges are determined locally at the access point, hence the name edge pricing. This is important because more detailed pricing schemes can be implemented and tested at the edge of the network, while maintaining a simple and generic network core. Rather than try to compute the congestion costs, which the authors argue are inherently inaccessible, they propose to approximate it with prices that depend on the quality of service (QoS), the time-of-day, and the expected path from the source to the destination. In the case of multicast traffic, there is a potentially large and varying group of destination users. One approach for accumulating accounting information to the source would be to have receivers periodically send accounting messages. While traveling to the source, each node (router) would add the cost of its downstream link to the accounting message. Branching nodes would add the sum of the cost of all downstream links. Hence, when the accounting packets reach the source they will contain the accounting information for the whole multicast tree. In the case of receiver charging, charges can be computed at the “exit” point (the receiver’s access point). For multicast connections each receiver can be assigned a fraction of the cost he is charged with. Policies for assigning these fractions are discussed by Herzog et al [68].

The problem of pricing the best-effort Internet does not have a single solution. Practical evidence supports this claim: in New Zealand, usage sensitive pricing has been successfully used to finance the high cost link to the Internet [37]. On the other hand, usage-based pricing was unsuccessful in Chile [27], and when the U.S. Department of Defense (DoD) attempted to implement it in its interagency Defense Data Network (DDN), [70, 28]. Furthermore, a usage-based pricing scheme was too complex to be implemented for New Zealand’s internet Frame Relay network. This experimental evidence suggests that there is no single solution for pricing best-effort services, and methods to be adopted will necessarily depend on factors such as the relative cost of the link, the topology (e.g. single high capacity link or meshed network), and the nature of the customers (commercial organisations, academic sites, residential users).

Consider finally that we are in a new technological and commercial environment brought about by the deregulation of the telecommunications industry, the growing integration of services, and the ability to successfully use best-effort services for the transmission of voice, audio and videoconferencing (e.g. using MBONE). In this environment the problems of pricing Internet services, pricing ATM services, and pricing telephony can no longer be approached independently.
1.3 ATM

The Internet Demand Experiment (INDEX)

Whereas for networks offering a limited range of services, such as telephone networks, a large amount of empirical data exists, this is not the case for networks that have the ability to offer a wide range of service qualities. Such data is particularly important in order to identify the market structure for network services. The Internet Demand Experiment (INDEX) [101] at the University of California at Berkeley aims to answer exactly this question, namely to understand the structure of user demand for Internet access for different price/quality combinations. Among the objectives of the INDEX project are

- to measure user demand for Internet access as a function of quality of service, pricing structure, and application
- to demonstrate an end-to-end system that provides access to a diverse group of users at attractive price-quality combinations
- to develop a prototype system that can be scaled to serve the demand for remote network access from a whole community of a large university.

Internet settlements

A different but related problem to that of charges by Internet providers to end users and organizations is that of inter-carrier financial settlements [38]. A settlement agreement is an agreement between two or more service providers which specifies an agreed set of metrics and the corresponding charges. It is current practice that competing carriers usually exchange traffic free of charge. It is expected (and even inevitable) that this situation will change in the future. Carpenter [38] attempts to stimulate a discussion on Internet settlement metrics. Metrics identified include the access capacity, connect time, total traffic, peak traffic, number of announced routes, mean RTT (round trip time), mean loss rate, and others. The author identifies the need for developing measuring these metrics and collecting results.

Related to this is the lack of existence of metrics to quantify the quality of best-effort service and methods to measure it [63]. Such metrics necessarily involve the cooperation of service providers and are required if customers want to know what kind of service to expect for the prices they are charged by their providers. Paxson [92] provides a general framework for the particular metrics to be developed by the IETF.

1.3 ATM

Asynchronous Transfer Mode (ATM) has been under development for a number of years as the technology to support Broadband Integrated Services Digital Networks (B-ISDN). ATM is designed to combine the benefits of packet switching with connection-oriented transport, in order to offer integrated services with appropriate quality of service guarantees and efficient use of bandwidth. Standards for ATM, defined by the ATM Forum and by ITU-T, are approaching a mature and stable state.

ATM is designed to provide shared capacity for multiple traffic classes having different traffic characteristics and quality of service requirements. Several key features of ATM enable this to be achieved efficiently:

- The use of packet switching to enable statistical multiplexing at the cell and burst levels
- Connection-oriented operation with connection acceptance control to ensure quality of service guarantees.
• Rate control protocols which allow elastic traffic services to use spare capacity efficiently.

Standards

The ITU standard for the B-ISDN protocol reference model [5] defines three layers:

1. The ATM Adaptation Layer (AAL) assembles data from user connections into ATM cells for transportation and reassembly at their destination.
2. The ATM Layer is responsible for the end-to-end transfer of user cell streams, including functions such as flow control, cell routing and switching.
3. The Physical Layer transfers cells between ATM switching nodes. This layer will typically use Synchronous Digital Hierarchy (SDH).

The ITU has defined several ATM Layer Transfer Capabilities (ATCs) [13, 20]. These are tailored to distinct demands of likely application groups. The ATC used by a connection should fulfill the general requirements of that connection, for example whether bandwidth varies over time at the initiative of the network, and what actions the network may take if cell rates are exceeded by the source. Detailed characteristics of the ATC are captured in the associated traffic contract, and in the QoS class for the connection. The defined ATCs are:

• Deterministic Bit Rate (DBR): throughput at the peak cell rate (PCR) is guaranteed throughout the connection. DBR is intended for constant bit rate traffic with strong delay requirements. Examples could include voice and videoconferencing.

• Statistical Bit Rate (SBR): throughput at the sustainable cell rate (SCR) is guaranteed as a long term average, with bursts at up to the PCR for periods constrained by the maximum burst size (MBS). Sub-categories of SBR vary in how they treat the Cell Loss Priority (CLP) indication in each cell. SBR is intended for variable bit rate sources. Delay guarantees can be provided for real-time applications. A typical application would be video with variable rate encoding.

• Available Bit Rate (ABR): the network uses a rate control protocol to signal changes in the allowed rate of each connection; the rate may vary up to the PCR, and there may also be an agreed minimum cell rate (MCR). ABR is intended for sources which are able to adapt their rate according to network conditions but have low cell loss requirements. Examples include Web browsing and file transfer.

• ATM Block Transfer (ABT) allows transfer characteristics to be negotiated for ATM blocks (groups of cells).

The ATM Forum has specified service capabilities analogous to these ATCs, with one addition [54]:

• Unspecified Bit Rate (UBR): there are no throughput or quality of service guarantees. A typical application might be email.

Ensuring quality of service

The network aims to achieve maximum efficiency by multiplexing many connections with varying bit rates. However, if utilisation is high then statistical fluctuation in the overall cell rate can cause cell losses. This is controlled in several ways:
- Buffering is used to store cells until links or switches are free to handle them. Buffer lengths are constrained by the need to meet delay guarantees.
- UBR cells can be dropped or delayed.
- Rate control signals can be used to vary the rates of ABR connections.
- Connection acceptance control (CAC) is used to limit the acceptance of new connections having throughput guarantees (essentially DBR and SBR connections, and ABR connections with non-zero MCR) so as to keep the total bandwidth demand within the resource available.
- Usage parameter control (UPC), or policing, is used to monitor connections to ensure that they stay within their traffic contract. Cells exceeding this contract may be tagged, and could be discarded.

Quality of service classes

The ATM Layer Transfer Capabilities have associated network performance parameters - these are cell loss ratio, cell transfer delay, and cell delay variation. In addition the traffic descriptor defines the cell delay variation tolerance where appropriate.

To complement the connection profiles described by the ATCs, and to describe the Quality of Service parameters of the connections offered by a network, ITU-T Recommendation I.356 [12] describes classes of Quality of Service. There are at present four QoS classes. Their main characteristics are outlined below. Values for the limits on cell delay and cell loss are given in I.356:

**QoS classes**

Class 1, the ‘stringent’ class, offers stringent limitations on both cell loss and cell delay, regardless of the CLP (cell loss priority) indication. The stringent limitations for delay apply both to the average cell delay on the connection and to the Cell Delay Variation (CDV).

Class 2, the ‘tolerant’ class, likewise offers guarantees regardless of the CLP indication of each cell. This QoS class specifies limited cell loss, with limits that are slightly more tolerant than in QoS class 1. QoS class 2 offers no limitations on cell delay.

Class 3, the ‘bi-level’ class makes a distinction between cells with CLP indication CLP = 0, and cells with CLP indication CLP = 1. Guarantees apply only to CLP = 0 cells. The guarantees are limitations on cell loss, and they follow the values of QoS class 2.

Class U is the ‘unspecified’ class. Neither delay nor loss have specified limitations for connections of this class.

I.356 defines appropriate combinations of ATC and QoS class as follows:

<table>
<thead>
<tr>
<th></th>
<th>DBR</th>
<th>SBR</th>
<th>ABR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: stringent</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2: tolerant</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3: bi-level</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>U: unspecified</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

Quality of service, resource usage, and charging

In practice, there are no fixed rules to determine which services and applications will use each ATC. For the user, there is a trade-off between cost and quality of service. Consider voice traffic as an example. Voice has traditionally been carried at constant bit rate in narrowband telephony
networks, and would naturally use DBR in an ATM network. However with variable rate encoding it could use SBR with equally acceptable quality of service, or ABR with less reliable quality. The choice should depend on the pricing structure and on the quality of service achievable (the quality of service provided by an ABR connection would depend on network conditions, varying with time of day).

In a competitive environment one would expect that the usage-based component of charges for different ATCs should reflect the relative costs of resource usage in the network. This cost is not determined simply by the average bandwidth requirement of a connection - it depends also on the bandwidth variability, the peak rate, and the QoS requirement. For example for an SBR connection with a highly variable cell rate and a low delay requirement the network might need to reserve bandwidth considerably greater than the mean rate of the connection in order to meet the QoS guarantee. In contrast the network might need to provide very little capacity for elastic traffic (ABR and UBR) since these connections are rate-adaptable and can use the spare capacity left over by guaranteed services. Hence it has been suggested [103] that charges for different ATCs might differ by several orders of magnitude.

The problem for the network operator then is to devise a pricing structure for the ATCs offered. Competition should ensure that prices reflect relative network resource usage costs. The charging algorithms should convey appropriate incentives to users and should not be too complex. If these aims are achieved successfully then users will be able to make efficient choices of traffic contract with a good correspondence between the utility to the user and the cost imposed on the network. It remains an open question as to whether a realistic pricing structure can support the full range of ATCs defined for ATM.

1.4 Requirements for charging schemes

Charging for telecommunications services has two main aims. The first is to generate revenue which enables network and service providers to operate profitably. The second aim, met by usage-sensitive charging, is to generate incentives for users to constrain their traffic demands appropriately, resulting in efficient network operation. This second aim is particularly important in broadband multiservice networks where user applications have the ability to generate high data rates - often much higher than the user needs to assure an acceptable quality of service. However it is likely to be difficult to characterise and predict resource usage in a multiservice environment and the requirements for charging schemes should therefore be evaluated carefully.

The requirements for a charging scheme are imposed mainly by the two key roles - provider and customer. In the context of ATM networks we will focus on charging for ATM-level services provided by a network operator. Other roles can be distinguished (service providers, application developers, regulator) - their impact on requirements is considered later in this section.

We consider here high-level requirements that apply to the charging scheme as a whole. Network operators also have more detailed requirements for the charging system implementation - these will be discussed in chapter 5.

Customer requirements

Predictability

Charges should be broadly predictable, therefore the major part of usage charges should be based on measures that customers can understand and control.
1.4. REQUIREMENTS FOR CHARGING SCHEMES

Ease of use

Customers need to be given simple information through a clear interface to enable them to make correct choices of tariff and traffic contract.

Traceability

Bills should be clear and understandable to customers, so connection charges should have supporting information that indicates choice of service and tariff and measured usage parameters.

Network operator requirements

Revenue generation

Charging should generate the required revenue for the network operator. Telecommunications charging has traditionally been based on a combination of subscription and usage charges. Subscription charges provide a steady and predictable revenue. Usage charges provide a more variable revenue but one that is linked to variable network costs.

Cost-effective implementation

Usage-sensitive charging imposes several overheads on the charging system, in particular the collection and processing of usage measurements and the presentation of detailed data in customer bills. These overheads can add significantly to the cost of implementing and operating the charging scheme. The requirements for usage-sensitive charging, and the design of the charging system, must be carefully assessed to ensure that the cost is not too great.

Flexibility

The charging scheme should be adaptable to meet the requirements of all types of customer, to work effectively with user applications, and to serve as a basis for charges that are defined at the service level.

Usage incentives

Charges should generate incentives to users to constrain appropriately the traffic that they generate, so ensuring efficient network operation.

Interconnect

Different network operators may use different charging schemes, but the usage parameters on which charging is based should be standardised to facilitate interconnect charging. Chapter 8 includes a discussion of interconnect charging.
CHAPTER 1. CHARGING IN COMMUNICATION NETWORKS

Requirements of other parties

Service provider

A service provider can offer to customers high-level services that operate over ATM connections provided by a separate network provider. The service provider is therefore a customer of the network provider in addition to being a supplier of services to the end-customer, and therefore has many of the requirements listed above. A more specific requirement of the service provider is to be able to charge at the service level in a way that relates logically to the ATM-level charges imposed by the network provider. Service-level charging is discussed more fully in chapter 8.

Some service providers might also provide and charge for content. Usually the service provider would like to offer an integrated charge to the end-customer, who would not then be able to distinguish between connection charge and content charge.

Application developer

The characteristics of ATM traffic will to a large extent be determined by software applications under the control of end-users (for example, applications such as videoconferencing, Web browsing, file transfer). The developers of these applications may be able to maximise their efficiency (in terms of bandwidth usage, for example by using sophisticated video coding methods), and to include controls which allow the user to trade data rate against perceived quality. Whether, and how, application developers will provide these facilities depends to a large extent on the charging schemes that users are faced with.

Regulator

Many countries now have regulatory bodies specifically responsible for overseeing the telecommunications industry and regulating the transition to a fully competitive environment. Charging, and in particular the level of charges, is naturally a key area for regulation. Within this area there are two requirements which might influence the charging schemes adopted by network operators:

- Charges should be cost-orientated. This should occur naturally in a fully competitive environment, and it is therefore a reasonable aim for regulation in markets where one operator is dominant [19]. This is a key issue for determining interconnection charges.

- Tariffs should be clear and understandable to users, and should enable comparison between different providers.

Discussion

It should by now be apparent that the desires of different parties are sometimes in conflict and it will be difficult, perhaps impossible, to design charging schemes that fully satisfy all of these requirements. Perhaps the most important difference is between the desire for simplicity and predictability and the need to generate appropriate usage-related incentives. Charging schemes that are closely usage-related are less likely to be simple, understandable and predictable. In the end it is a commercial decision for network providers to choose charging schemes weighing usage-sensitivity against customer acceptability.

One way to resolve or lessen this conflict is to ensure a good user interface - one that can help the user to make good choices of tariff and traffic contract and to understand the resulting charges. This issue came to the fore in the CA$hMAN project, which directed some of its work to the design of the user/network interface and of intelligent agent software to help users.
Another important issue concerns charging for elastic traffic (such as TCP/IP, and the ABR service in ATM). Incentive-compatible charging schemes for elastic traffic are congestion-sensitive - the busier the network the more the user must pay to achieve a desired throughput. The schemes investigated in CA$hMAN and described in chapter 3 are certainly of this kind. This property conflicts with many of the requirements listed above (including simplicity, predictability, revenue generation), and customer acceptability must be a major concern. In fact this form of charging is not designed for predictable revenue generation but rather as a means of providing control signals leading to efficient network operation. The charges are an incentive to users to cooperate in this control, and may well be acceptable to users provided they are substantially lower than charges for guaranteed services.

1.5 The CA$hMAN Project

The ACTS Programme (Advanced Communication Technologies and Services) was established under the Fourth Framework Programme of European activities in the field of research and technological development and demonstration (1994-1998). The Programme aimed to support research and development in advanced communications in order to facilitate economic development and social cohesion in Europe. Under the Programme, individual companies, public sector organisations, research institutes, schools and universities agreed to work together as individual project consortia, pooling their knowledge and resources in pursuit of specific research objectives covered by the ACTS workplan. All ACTS research was conducted in the context of usage trials to ensure relevance of the results and to encourage a broadening of awareness of the benefits that advanced communications may bring. Twenty two National Host organisations supported project experiments and acted as a window to the many trials conducted.

ACTS Project AC-039 was named CA$hMAN - “Charging and Accounting Schemes in Multi-service ATM Networks”. CA$hMAN ran for three years, from September 1995 to August 1998. Its objectives were to study and develop, implement, verify and compare charging and accounting schemes for ATM networks. It was a multidisciplinary project involving participants from university departments, telecommunications hardware and software manufacturers, and network operators:

- Intrasoft SA, Greece
- Ascom Monetel, France
- ATecoM GMBH, Germany
- Ericsson AS, Norway
- ICS-FORTH, Greece
- ISS University of Aachen, Germany
- Lucent Technologies, USA
- Lyndewode Research, UK
- Royal KPN, Netherlands
- Telenor Research, Norway
- Telscom AG, Switzerland
- University of California Berkeley, USA
- University of Cambridge, UK
CA$hMAN’s main achievements were to develop a range of simple but effective pricing models for both guaranteed and elastic services, to design and implement a platform for charging and accounting management based on TINA principles and advanced technology, and to use this platform for trials that were able to demonstrate CA$hMAN charging schemes in operation and to explore user-network interface issues and the use of intelligent agent software. The multidisciplinary approach was very successful, combining economic and mathematical models, hardware and software development, and operational network issues.

This book does not set out to describe all of the work done in CA$hMAN. Instead we aim to present an approach to multiservice network charging that reflects the CA$hMAN experience.