computer communications

Computer Communications 00 (2001) 000–000

www.elsevier.com/locate/comcom

A survey of pricing for integrated service networks

Xinjie Chang[a,*], David W. Petr[b]

[a]*SBC Technology Resource Inc., 4698 Willow Road, Pleasanton, CA 94588, USA*
[b]*ITTC, University of Kansas, 2335 Irving Hill Road, Lawrence, KS 66045, USA*

Received 24 October 2000; revised 3 March 2001; accepted 23 March 2001

Abstract

Advances in technology have greatly increased the demand for a single integrated service network that can provide multiple service classes for different user requirements. For such a multiple-service network, congestion control is one of the key issues to be addressed. However, without an appropriate mechanism to encourage end users to use the network properly, over-utilization and congestion are unavoidable. For this problem, it is widely accepted that pricing is a proper tool to manage congestion, encourage network growth, and allocate resource to users in a fair manner. However, how to charge for the traffic and at what price is still under study. In this paper, we first briefly review the state of the art and technological growth of congestion control for integrated service networks (ISN). Subsequently, we present a summary of the recent developments on various pricing policies and different charging and billing schemes that have been proposed for ATM and Internet Differentiated Services. Some architecture and implementation issues are also discussed. Finally, some future trends are identified.

1. Introduction

An integrated service network (ISN) or multiple-service network is a single network that can support a variety of applications and services, each with different traffic characteristics and different service requirements. ATM is a prime example, in which different applications are categorized into different traffic classes. Traditionally, the Internet has been viewed as a single service class network, that is, 'best-effort' only. However, considerable work has been done recently for providing Quality of Service (QoS) features over Internet Protocol (IP) networks. With emerging concepts such as RSVP [1] and DiffServ [2] serving as the building blocks, it is widely expected that the Internet is also going to be able to provide various classes of service.

However, there is a hot debate on the solution for network congestion and QoS problems. Some researchers argue for the 'big bandwidth' solution [45], that is, just throw bandwidth into the network to satisfy peak demand. However, this underestimates the coming bandwidth-hungry applications. Although the ever-increasing bandwidth and ever-decreasing cost means more and more available bandwidth (Fig. 1), new bandwidth-hungry applications always seem to appear with the emergence of high-bandwidth networks (Fig. 2), and it seems that users' desire for additional capacity is unlimited. Therefore, many researchers have realized that proper congestion management mechanisms or protocols should also be considered.

Recently, congestion control based on the priority scheme has been extensively studied; various strategies have been developed for different networks with different traffic types. However, one major flaw of these schemes is the failure to recognize that users place different subjective values on their own traffic streams. As users make their individual decisions on whether and how to use the network, it is not sufficient to hope that users will try to act in a cooperative way and be aware of achieving network efficiency by themselves. This is so because network performance is a function of offered load while the offered load is in turn a function of the incentives individual users encounter when using the network [5]. It is argued by many researchers that a well-designed usage-based pricing scheme for ISN will be a proper mechanism to offer such user incentives so that they will adjust their behavior and try to achieve efficiency [6,7]. Moreover, a proper pricing strategy can also induce users to implicitly reveal information about the traffic they are sending, which may help the network to further optimize resource allocations.

The remainder of the paper presents a detailed survey of the recent development on pricing for integrated service networks. In Section 2, we present a thorough description of various static and dynamic pricing schemes proposed

* Corresponding author. Tel.: +1-925-598-1219; fax: +1-925-598-1322.
E-mail addresses: chang_xj@yahoo.com (X. Chang), dwp@ittc.ukans.edu (D.W. Petr).

0140-3664/01/$ - see front matter
PII: S0140-3664(01)00327-9

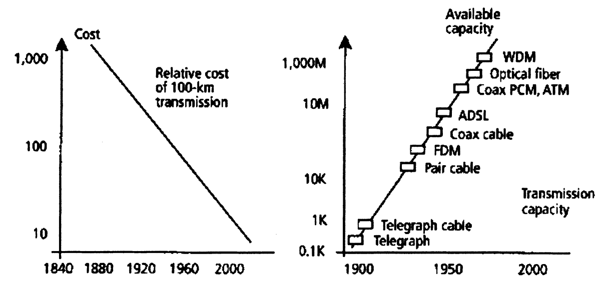

Fig. 1. Transmission capacity increase with cost decrease [55].

during the past several years. The basic ideas, advantages and flaws are analyzed and compared. Section 3 describes the wholesale and retail pricing problem, which is an emerging topic due to the rapid development of the network infrastructure. We also discuss in Section 4 a wide range of architecture issues, such as user interface, billing and accounting. Finally, in Section 5 we point out several future research topics.

2. Usage-based pricing

Since the early 1990's, a number of papers addressing the topic of pricing have been published. These pricing schemes are based on different principles such as: static or dynamic pricing, value-based or resource-based, usage-sensitive or contract-based, and technology-oriented or marketing-oriented. In this section, we classify these pricing schemes into two general categories: static pricing and dynamic pricing, as in Ref. [8]. It also should be noted that some of the pricing policies are proposed for ATM and some are proposed for the Internet. However, in our opinion, the next generation Internet borrows heavily from ATM Quality of Service concepts, and Internet developments such as RSVP, MPLS and DiffServ are quite similar to corresponding components in ATM. Although there are some differences in implementation details, they are similar in underlying models and basic ideas. So we argue that those pricing policies proposed could be used for both types of networks.

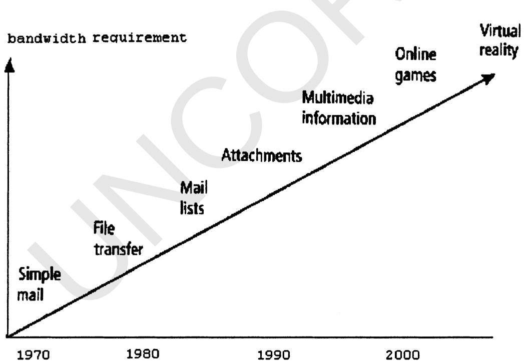

Fig. 2. The continuous increasing of bandwidth requirements [55].

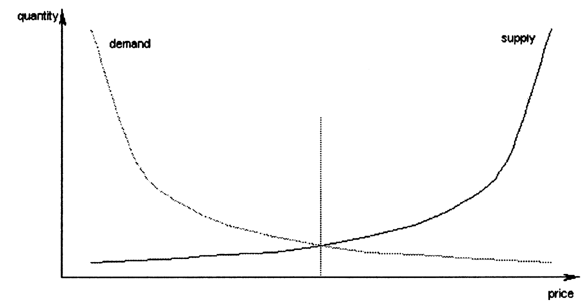

Fig. 3. Demand and supply relationship.

Most usage-based pricing schemes are based on the supply–demand relationship analysis in economics [8], as illustrated in Fig. 3 below. Generally, users' demand will decrease with the increase of price while supply will increase and vice-versa. There is an equilibrium point between the demand-price and supply-price curves where the demand and supply will be equal under that price.

2.1. Static pricing

'Static' pricing means that the price is set by the network provider based on observation and estimation from some historical data and is independent of real-time network utilization. Advantages of static pricing are simplicity of implementation and predictability from the customer's point of view.

Based on different granularities of the usage measurement, prices can be set as per-byte, per-packet and per-connection based. In Refs. [5,9], Cocchi et al. first compare a very simple per-byte flat-pricing with graduated-fee charging for prioritized networks. The authors present an abstract formulation of the service disciplines and propose a priority-sensitive pricing policy based on competitive game theory. The concept of utility function borrowed from economics is introduced to evaluate user satisfaction (the trade-off between performance incentives and monetary incentives). Simulation studies show that the pricing mechanism can improve user satisfaction and spread the benefits of ISN over all users. Although the model studied is quite simple and only two priority levels are considered, this provides an initial effort to grapple with user incentives for ISN. Expanding on this concept, DaSilva et al. [10] perform a thorough analysis of the case in which users are offered more priority levels. The existence and uniqueness of the Nash Equilibrium for the non-cooperative game is analyzed and a non-linear programming model is used to find the Nash Equilibrium. The authors argue that by appropriately selecting the different prices for various priority levels, network providers will be able to induce an optimal point that can maximize both revenue and aggregated utility.

In Ref. [11], Parris et al. study a per-packet-pricing scheme for prioritized networks, in which users are charged on the basis of the number of packets transmitted, regardless of service class. It is shown that utilization is high when price per packet is low, and revenue shows a trend to first increase and then fall as price increases. This is due to the limited budget of users considered in the model. ATM and RSVP are connection-oriented; i.e. an end-to-end connection is required to be set-up before the transmission of each session. In this case, the ‘set-up’ charges should also be considered. Parris et al. [11] also study this case by adding an additional set-up charge on the per-packet charging. Simulation studies show a similar trend as the results for per-packet charging. However, it also shows that by adding a set-up charge, the same revenue can be achieved at a lower per-packet price. This means a set-up charge will benefit network providers. Breker’s simulations corroborate these results [8,12].

The time-of-day pricing scheme, which is frequently used in telephony networks, has also been studied. Honig et al. [13] present a simple pricing policy containing two different entries: ‘Day price’ (or ‘peak price’) and ‘night price’ (or ‘off-peak price’) in an attempt to achieve traffic smoothing. A similar policy is also studied in Ref. [11]. Since users who want to transmit data during high network utilization periods will be charged more, some of them may choose to wait until a low network utilization period. By implementing this mechanism, network utilization can be distributed evenly over all time periods and very high peak utilization can be avoided. Parris et al. [14] also compare this scheme with the per-packet pricing scheme. By comparing the call blocking probability and peak utilization, it is argued that time-of-day pricing is a useful tool for congestion avoidance.

For per-connection pricing schemes, Lindberger [15] proposes a scheme in which the usage charge is calculated in proportion to the bandwidth required, distance and call duration. Songhurst and Kelly [16] also study a similar policy in which the charge is proportional to both the volume of the traffic and the duration of the connection. A connection charge is also imposed. Recently, a new topic for static pricing has emerged, that is, pricing for Permanent Virtual Circuit (PVC) vs Switched Virtual Circuit (SVC) services. This problem comes from ATM networks. Most of the currently installed ATM networks provide only PVC connections, but with the introduction of SVC services, network providers want to find how to provide incentives to encourage some users to transfer from PVC to SVC service. Liu and Petr [17] consider the effect of tariffs for connection set-up (denoted as variable ‘s’) and bandwidth allocation (denoted as variable ‘a’) on user choices. The On–Off traffic model is used for characterizing the arrival pattern of user data. It has been shown that by properly adjusting the relationship between s and a, the network provider can provide enough incentive to encourage such a transfer.

In Ref. [18], Odlyzko proposes a so-called Paris Metro Pricing (PMP) scheme. The basic idea is borrowed from the Paris Metro system, in which different classes of cars are provided with different prices. For example, first class cars are more expensive but less congested and second class cars are less expensive but more congested (this is also widely used in airline systems). In the PMP scheme, the total link capacity is spliced into several channels, each with a different price. In each channel only best effort service is provided. It is argued that fewer users will use the more expensive channels so that they are less congested and provide better quality. This can provide expected service quality levels, but not service guarantees. The resource (channel) segmentation also sacrifices the multiplexing gain that can be achieved, especially as the number of channels increases. In the initial proposal only a static scheme is considered, in which each channel has a fixed fraction of the total capacity and fixed price. However, it can be extended to a dynamic scheme with dynamic channel capacity and price. Moreover, the PMP scheme may be able to be used together with such protocols as RSVP to provide guaranteed quality of service.

A static priority pricing policy for four different service classes (low priority best effort, high-priority best effort, soft guaranteed service and hard-guaranteed service), each with different tariffs, is considered in Ref. [19]. Charging is based on the throughput and connection time. Moreover, a penalty coefficient is attached with each connection, which is proportional to the difference between real-time throughput and reserved bandwidth. If a connection’s usage is higher than what it reserved, it has to pay a higher price (priority price times penalty coefficient).

Morris et al. [20] describe a detailed charging scheme for ATM networks. This scheme adopts a static price strategy (with set-up charge) in which volume (the product of traffic rate and duration) is used as the measurement of the usage. Since it is relatively easy to get the information on duration, a major concern of the scheme is on how to obtain the traffic rate information for each type of traffic. For CBR (constant bit rate) traffic, the PCR (peak cell rate) can be obtained at connection set-up time. For VBR (variable bit rate) traffic, the PCR, SCR (sustainable cell rate) and burstiness (leaky bucket size) are used to approximate the variable traffic rate. It is suggested that the effective bandwidth should be used as a proper measurement of traffic rate for this type of traffic, which can simplify the charging calculation. For ABR (available bit rate) traffic, the reservation-based charge is first calculated based on MCR (minimum cell rate) information and a partly usage-based charging scheme is used for the VAR (value above reservation) traffic rate. Finally, for UBR (unspecified bit rate) traffic, real-time throughput measurements are needed to get the traffic rate.

2.2. Dynamic pricing

Since bandwidth is scarce especially during congestion, efficient prices must reflect the current availability of resources. Dynamic pricing allows more formal optimization by taking into account the fluctuations in network utilization. Most of the literature discusses dynamic pricing

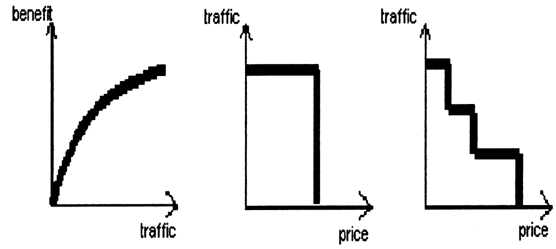

Fig. 4. User benefit function.

schemes based on the computation of the marginal congestion cost (or opportunity cost) and the optimal points have the charge equal to the marginal cost. There is no doubt that the major concerns for network providers are cost recovery, revenue and profit. However, since network access is viewed as a kind of service industry, customer satisfaction is most important due to the current intense competition among network providers. Currently, most research work focuses on the optimization of social welfare (aggregated utility).

2.2.1. Best-effort traffic

The most often used dynamic pricing scheme is a bidding price scheme because many researchers argue that users should have the freedom to send traffic and show their willingness to pay for it. Breker [8] refers to a so-called 'Transport Auction' scheme. In that scheme, each user first sends his traffic and his willingness to pay to a software agent installed on his workstation. The agent then checks if the user's bid is higher than current network price. If it is, the agent will offer the traffic to the network and admission control software is activated to check if the required resources are available. This scheme can provide relatively high revenue for the network provider. However, it can not prevent the user from 'fooling' the network by misrepresenting the traffic characteristics or by bidding maliciously. Mackie-Mason and Varian [7] propose a per-packet bidding price scheme called a 'smart market' scheme. In this scheme, each user assigns a willingness to pay for each packet he sends to the network. The network will accept the packets that have a bidding price higher than the current cutoff price, which is calculated from the marginal congestion cost. The dynamic lies in the fact that price for each packet will vary with time reflecting the current state of network load. The authors also argue that this scheme can force users to bid by their true values of willingness to pay.

In Ref. [21], Peha analyzes pricing strategies for three types of traffic: guaranteed, packet-oriented best effort, and stream-oriented best effort. The author states that for the second type, a per-packet-pricing strategy such as smart-market or spot pricing is proper because packets are independent and demands fluctuate randomly. For the stream-oriented best effort traffic, applications can declare their arrival process and performance objectives (or priority levels) in advance, so that the network can use this information to achieve better service and satisfy user requirements. From this analysis and some numerical results, Peha claims that the stream-oriented best effort service is an important service class that should be provided in integrated service networks. This service class also can provide an incentive for users to offer traffic characteristic information and helps to improve performance and in turn lower the charges for users.

For similar traffic types, Fankhauser et al. [22] propose a bandwidth reservation and auction scheme. The difference from the earlier auction based scheme is that link capacity is logically segmented into smaller units and the auction is based on these units rather than each packet. In Ref. [23], Lazar and Semret present a similar scheme. Clearly, these can also be viewed as an extension of the PMP scheme.

2.2.2. Elastic traffic

Murphy's dynamic pricing model [24,25] borrows the flow control concept from the ABR traffic type for ATM networks. The authors emphasize that users should be free to make their own choices and that they will do some local optimization (maybe implicitly). They first analyze different types of users and focus on the study of the so-called 'adaptive users', who can and will respond to feedback information from the network by changing his traffic offered to the network. Here, price is a proper candidate for the feedback signal. Some users with stringent traffic requirements will pay a premium price to get guaranteed service while some adaptive users prefer flexible pricing and can tolerant different levels of congestion. The dynamic pricing analysis is based on supply and demand relationships. Each adaptive user will place a benefit (similar to 'willingness to pay') on the resource he is allocated. Given the current price, each use can determine how much traffic to send to the network based on his current benefit function value (Fig. 4). The network provider thus decides the price based on current network conditions and tries to equate the marginal congestion cost with the marginal benefit of users. Obviously, when the network is lightly loaded, the price should be low, and during high-load period the price should be higher. Based on this idea, they also propose a dynamic iterative algorithm to achieve optimal pricing and show that the system can reach an equilibrium state where the total requirement and price will not vary much (if at all) over time. By simulation, Breker [12] argues that this is not always the case if all the users have the same benefit function. Fortunately, this situation is quite rare. Another shortcoming of the scheme is that the initial price of the iteration will influence the convergence rate [8,12]. Since adaptive users cannot predict their traffic characteristics beforehand, they cannot provide useful information to help the network allocate resources optimally. We suggest that this scheme could be used together with some real-time traffic measurement schemes in order to obtain improvements in performance.

Kelly has published a series of paper studying the relationship between flow control and dynamic pricing. In Ref. [26], Kelly proposes the concept of 'elastic traffic' and a proportional fairness criterion. The 'elastic traffic' concept is similar to the 'adaptive users' but is more general in that users or applications are able to modify the data transfer rate according to the available bandwidth and current network pricing. In essence, this is very similar to the ABR traffic category in ATM networks. Each user will choose a charge that he is willing to pay. The network then determines the rate that can be allocated to the user based on the proportional fairness criterion, that is, rate is allocated in proportion to the how much the user will pay for his share. This dynamic process consists of user's choice of charge and network's choice of rate. It has been illustrated that this system can achieve an equilibrium point, which is a system optimum point with regard to proportional fairness.

In Ref. [27], Kelly et al. present a thorough analysis of these concepts. They first construct an overall system optimization problem, which attempts to maximize the aggregated utility function. Then they decompose the problem into two classes of sub-problems. For each user there is a net benefit optimization problem and for the network there is a single profit maximization problem. A primal algorithm and a dual algorithm are proposed, respectively. In order to study the effectiveness and stability of the flow control problem with pricing input, Gibbson and Kelly [28] develop a distributed multi-user game model played among users to find the optimal solution. Key and McAuley [29] expand this model into a more general framework, where users play against the 'Network' which represents a resource system. A TCP-like algorithm is thus proposed and the issues of protocol and possible candidates for objective function are also discussed.

One problem of the proportional fairness allocation is scalability since the network has to know all users' choices of charge. As the network size grows, it becomes more and more complex and time-consuming to calculate each user's rate. For this problem, Biddiscombe et al. [30] propose an iterated estimation algorithm. At the very beginning, it is assumed that every user has an equal share of the bandwidth. From this starting point, each user can change his choice of charge and the network can recalculate the new price and reallocate the capacity among users. It has been demonstrated that for a logarithmic type utility function for each user, this scheme can maximize the aggregate user utility.

2.2.3. Guaranteed service

In contrast to Kelly's assumption of elastic traffic, some researchers have studied the pricing problem for inelastic traffic, e.g. traffic with stringent performance requirements that need to be guaranteed. Wang et al. [31] study dynamic pricing for best-effort service and performance guaranteed services. For the best-effort service, price is computed with regard to current buffer occupancy and predicted willingness to pay. The network constantly updates the cutoff price on a per-cell basis and only accepts those having a higher willingness to pay than the cutoff price. For guaranteed services, price reflects the opportunity cost (similar to congestion cost) of providing the service while taking into account the service characteristics and shadow prices. Ji et al. [32] develop a charging scheme in which the price for each type of service is based on the QoS degradation caused to other users sharing network resources.

For inelastic traffic, Jiang et al. [33] present a pricing scheme based on effective bandwidth [34] of user traffic. However, it is assumed that the network knows the user's benefit function in addition to current trunk capacity and virtual path routing. In Ref. [35], Low et al. consider the dynamic pricing problem in which each user has some budget constraints. Thus, user requirements are limited by their budgets. The objective of each user is still maximizing his utility while the objective of the network is to maximize the social welfare. In this paper, effective bandwidth is used as the proxy of usage charge. Courcoubetis and Siris [36] also study the pricing problem for inelastic traffic. In this case, each user negotiates a Service Level Agreement (SLA) with the network, which describes the user's traffic characteristic and QoS requirements. Similar to the pricing problem discussed above, the goal of the network is to maximize social welfare, e.g., aggregate utility functions. However, what makes the problem more complicated is the constraint of user performance guarantees. By using the effective bandwidth, Courcoubetis and Siris convert this constraint into the form of limited effective link capacity. For the resulting constrained optimization problem, shadow price is viewed as the LaGrange coefficient. And it is argued that the optimal point is achieved when the user's marginal benefit of higher resource requirement equals the cost for additional resource required. A simple case with two service classes is discussed based on this idea. Considering the complexity in computing effective bandwidth, approximated effective bandwidth is used and price is related to this approximated effective bandwidth. The issues of user incentives and fairness are also discussed. Although this policy is proposed for the Internet, it clearly can also be implemented in ATM networks by changing the SLA to Traffic Contract. In Ref. [37], Siris et al. state that by using the effective bandwidth concept, users will also implicitly indicate some useful information to the network, which can be used to optimize admission control and resource allocations. The operator's incentive in using this scheme is also explained. Moreover, they demonstrate an evolution model where two network providers are involved. They show that through competition and evolution, the operator who migrates from the original peak-rate based tariff to an effective-bandwidth-based tariff will attract more and more users while the one that continues with the peak-rate tariffs will only attract users with CBR type traffic. From this analysis, the connection time based tariff and volume and duration based tariff strategies are compared. The results illustrate that the former strategy is only proper for CBR

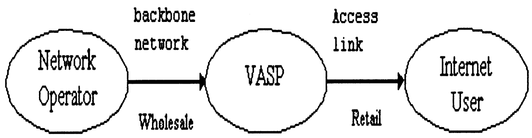

Fig. 5. 3-tier model.

traffic while the latter is proper for VBR, ABR and best effort traffic.

Gupta et al. [38] present a priority pricing mechanism in which the congestion is expressed and measured as delay. They abstract the Internet to consist of servers (content providers) connected by a backbone, network (access) providers, and users. Access providers measure the load on the backbone and set the price according to the congestion status. The optimal price is achieved as a 'stochastic equilibrium'[39], which can maximize net social benefits. Thus users can choose a different way or time to get their service based on their individual decisions on the value of service and charge, which distributes the load evenly. This scheme requires knowledge of the traffic characteristics and the true congestion level at the equilibrium point. Again, an iterative algorithm for estimating the dynamic pricing is provided. Further more, Gupta et al. also discuss some interesting issues about billing systems and cost recovery.

Wang et al. [40] design a model for the analysis of optimal pricing for both best-effort and guaranteed services. For best-effort service, the bidding price is used and for guaranteed service classes the network provider will set the charge in a way to maximize some utility function. In this paper, two types of utility functions are analyzed, i.e., profit maximizing and social-welfare maximizing. It has been shown that for these two different objectives, the resulting pricing schemes are quite similar. For best-effort service, the optimal pricing is essentially computed from the shadow price of the marginal cost, which has been widely used. For guaranteed service classes, the cost for reserving resources should also be taken into account in addition to the usage-based charge. They also state that the proposed scheme considers all the factors including performance guarantees, resource usage, time of service and duration of connection, which have been covered only partly by many other schemes.

Korillis et al. [41] first argue that the Nash Equilibrium approach widely used in many dynamic pricing schemes is not efficient and can only achieve an optimal point for individual users but not the network system. Based on this argument, they propose an improved model with routing games. The network scenario they study consists of a group of users and a set of links between one source-destination pair. Users will spread their traffic among a proper subset of links to get their expected throughput while attempting to minimize the cost at the same time. The network provider will thus set the price for each link to force the system into an efficient 'target operating point', where users follow a nominal flow distribution. Similar to other schemes, price is set in proportion to the congestion level of each link. In contrast to many other schemes, here it is the network provider who makes the optimization, which therefore requires that the network has perfect knowledge of all users' demands and cost functions. This does not scale well. In order to be used for large-scale networks, an iterative pricing algorithm is also presented.

Carle et al. [42] propose a charging scheme that considers error control issues. The main concern is that with the wide use of wireless communications and its merging with the Internet (Wireless IP) and ATM (wireless ATM), there is the distinct possibility of errors or losses. This is especially true for low-cost, low-quality-requirement traffic. Errors will cause retransmission and for most usage-based charging systems that only consider throughput, the retransmitted packet will also be charged to the user. This is seen as unfair to the customers. From this concern, they state that charging should be based on goodput rather than throughput. The FEC (forward error correction) and ARQ (automatic repeat request) mechanisms are used in their scheme to distinguish the first time and retransmitted packets.

In Ref. [47], Karsten et al. present an analysis of an optimal pricing mechanism for Internet Integrated Services (IntServ), which include controlled load service, guaranteed service and guaranteed rate service. A virtual resource mapping between the IntServ rate parameters and resource parameters for cost calculation is presented.

3. Wholesale and retail pricing

Most literature on usage-based pricing studies the relationship between users and the network providers. However, current data networks in fact have a 3-tier structure consisting of the backbone network operator (NO), Value-Added Service Provider (VASP), and users. The NO provides network links to the VASP and charges for them in a wholesale model, and the VASP in turn provides network service to users and charges for that in a retail model, as shown in Fig. 5 [43]. Mackie-Mason et al. [7] also mention using the 'smart-market' as a wholesale price. It is assumed that there is a third party between the users and network providers who will buy network capacity in a wholesale way and then resell it to users.

Botvich et al. [43] discuss trial results of a charging system for IP over ATM networks in which ATM is used as the backbone. For the retail part, a static pricing scheme related only to the connection time and volume is used. For the wholesale part, a simple volume-based charge based on different traffic types (CBR, VBR, ABR and UBR) is used. It seems a little strange that the charge for VBR is higher than that for CBR. Even so, from the measurement and evaluation criterion, this scheme is a good candidate for implementation in the real world.

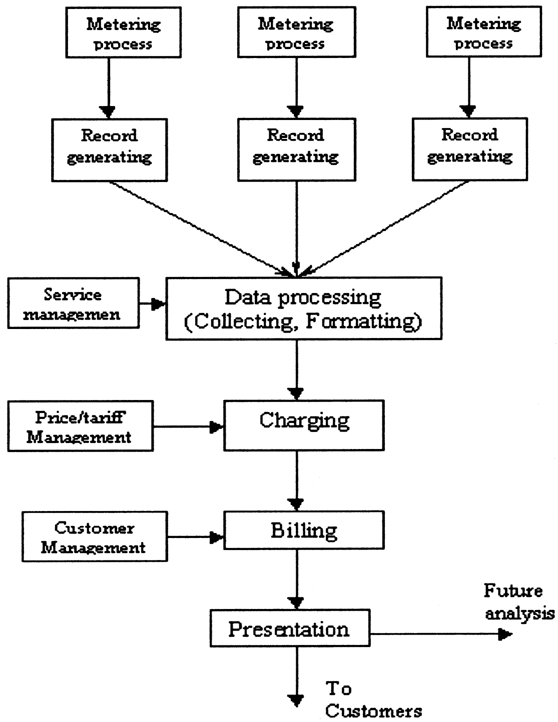

Fig. 6. A general structure of billing systems.

Semret et al. [44] present a theoretical analysis of this 3-tier scenario (although with different terminology). The paper first studies the relationships between VASP and NO and between peer VASPs. Generally, there are two kinds of players: seller and buyer. The NO is always a seller and users are always buyers. The goal of the user is to maximize utility within the constraint of his budget. The goal of the NO is to recover its cost. However, a VASP will be a buyer as to the NO, and it can get a constant amount of capacity from the NO which limits the number of users it can support and the quality it can offer to the users. Moreover, one VASP's capacity also depends on his peer VASP's capacity. At the same time, VSAP is also a seller as to the users. By expanding the bidding price strategy analysis into a two-dimensional space (one dimension for price and the other for quantity), a so-called progressive second price auction scheme is proposed. Game theory is used to analyze the optimal strategies for users and VASPs. It has been shown that with a proper stabilizing mechanism, the proposed static game can achieve a non-zero equilibrium point. The basic idea is that the competition among VASPs and user demands will result in a dynamic and efficient partition of the network resources among services being offered.

The accounting issue is discussed in Hazlett's paper [45]. The difficulties include the large traffic volume of the current Internet, the complex routing information, and lack of measurement tools. MacKie-Mason et al. [7] argue that 'congestion accounting' may be a possible solution. However, this needs 'global accounting' that can track the packet through its path, which is not feasible for an organization. Based on these arguments, Hazlett proposes an interim solution consisting of a hierarchical priority scheme. In essence, this is similar to the 3-tier models. Each organization gets access to the Internet with some priority level in a wholesale mode from the network provider; within the organization, each member shares the access link according to a sub-priority number. The priority level and the sub-priority number decide the real priority of the packet of this user in the network. Those who want higher speed can make a contribution to increase his priority level. Anyone can make contribution to increase the priority level of an organization or the sub-priority of a member of an organization (himself or someone else). Hazlett states that this solution can provide revenue for the network provider to recover cost, and users can make their individual optimization. However, there are still some open issues. For example, since anyone can contribute to increase the priority level of the organization, this in fact gives benefits to all the members of this organization. In this case, someone may just wait to get the benefit for someone else's contribution.

4. Architecture

The arguments for a usage-based pricing and charging system appear to be irresistible. However, there are still many issues concerning real world implementation that need to be considered carefully. Recent focus has been more on architecture issues rather than pricing mechanisms. Stiller's paper [46] presents definitions for some terminology.

Charging and billing are considered the core business processes of a network service provider and some of the most proprietary ones [54]. Billing and charging are becoming increasingly complicated due to the continuous emergence of new applications and service classes. A charging and billing system consists of at least three parts: metering of traffic, data recording and formatting, and charging and billing. A general structure is shown in Fig. 6. Obviously, metering usage information of data traffic is the first step. Due to the great variety of traffic types and characteristics, different records can be generated. There is a loose format requirement since this recorded information may also be used for other parts of network management. For every connection, there may also be several records generated and stored in different network nodes. For the purpose of charging and billing, these records should be collected and transmitted to the charging center. First, they will be filtered and formatted for the convenience of charging. Based on the pricing or tariff information, charges will be calculated and records will be stored. Sometimes this may also include such information as taxes and discounts. Finally, bills will be generated in a periodic manner or at any time as required.

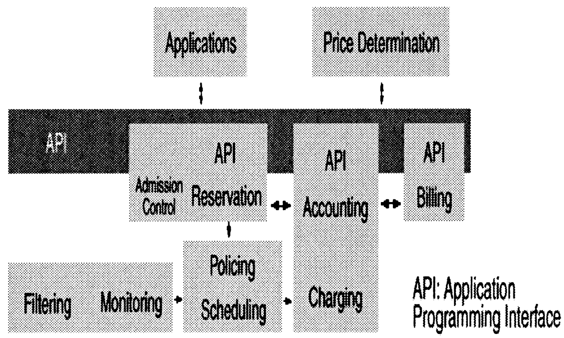

Fig. 7. API model for billing.

It is also argued that real-time billing will help users to make better decisions.

4.1. Connection detail record

The first important part for the real-world implementation is the record of all (or some) of the packets. A connection detail record (CDR) can be generated and stored by the edge switches or routers and can provide an important input to the accounting and billing system. This concept is borrowed from telephony networks, in which a call detail record is generated for each call. The record includes source/destination number, time, and duration information. For data networks, the record is more complex.

Reference [20] describes a list of required fields of the CDR for ATM networks: source/destination address, connection set-up time and tear down time, duration, pricing and usage information. However, such fields as reserved bandwidth, required QoS, and experienced QoS should also be included. If the experienced QoS is lower than that stated in the traffic contract, there should some compensation for customers, for instance a price discount.

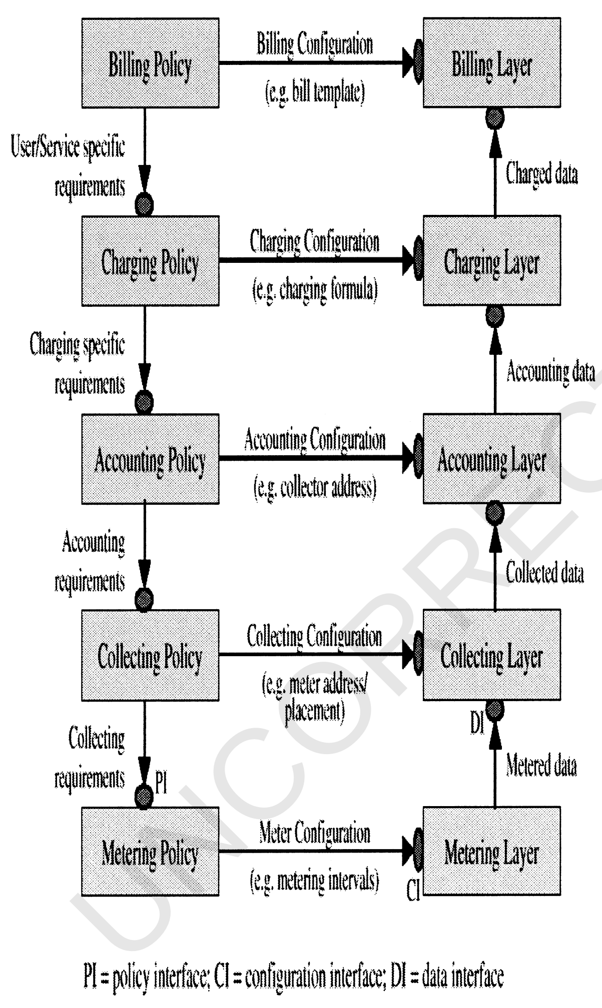

Fig. 8. Layered billing system structure.

4.2. Accounting and billing

The billing and accounting system has been widely recognized as the most important issue in implementing pricing mechanisms in the real world. First, it is important for network providers who want to make a profit by providing network access or services. Many new applications and protocols have been proposed for various purposes. However, if the network provider cannot find a way to charge for the services and bill the customers, it is hardly possible for them to implement it. Next, it is also important to users. Users need to know their exact costs in order to make intelligent service decisions and to balance their budget.

As shown in Fig. 7 [19], Frankhauser proposes an API (application programming interface) model for charging and accounting. Applications and price determinations are based on different APIs that can provide various functions for reservation and scheduling, accounting and billing. Information can be exchanged among different APIs for calculating charges and feeding back pricing information to the applications or users. This model can be integrated into current computer communication networks and can be implemented on routers and hosts.

A layered billing system structure [48] is illustrated as in Fig. 8. The metering layer is the underlying component to track and store traffic information. The collecting layer can relay charging related events and information back and forth between the accounting layer and meter layer. The accounting layer is responsible for the consolidation of connection and charging records, which are in turn the input for the calculation of charge for each connection in the charging layer. Finally, the billing layer collects charging information periodically and generates the bill to the users. Clearly, with each layer associated with specific data and functionality, this provides a framework for a real-world implementation. In Ref. [48], a billing architecture with PIP-NAR (Premium IP Network Accounting Record) used for information exchange and RTFM (Real Time Flow Management) architecture used for meter collection is given as an example. It is also argued that this framework can be integrated with the policy-based billing strategy to provide more intelligence for the network.

4.3. Pricing architecture

Shenker et al. [49,50] discuss the 'pricing architecture' from several aspects. First, they criticize the method of using marginal cost for the calculation of price. Then they propose a new concept of 'edge price', which can be determined locally (at the edge of the network) and is based on the approximate congestion level and routing information. This is proper for large-scale networks where multiple network providers and/or content providers are involved. They mention some interesting points such as nonlinear-pricing ('per-unit price depends on the quantity purchased') and the relation between ISPs (access providers) and ICPs (content or server providers). Secondly, charging of multicasting traffic is discussed. Most current pricing mechanisms consider only the uni-cast case, i.e. one link or one source-destination pair. With the increasing use of applications such as video-conferencing and telemedicine, multicast is becoming an important issue. Carle et al. [51] present a framework for charging of video multicast over ATM with heterogeneous traffic, multiple interconnected IP service providers, and non-negligible losses. The tradeoffs among video quality, error tolerance, delay and costs are discussed.

4.4. Charging the receiver

Charging the receiver is an interesting and challenging problem [45]. In the current network, data is often sent at the request of the recipient, so it makes sense that the recipient should be charged rather than the sender. This is true for contexts such as video-on-demand and multicasting applications. For some emerging access media such as ADSL (advanced digital subscriber line), the link for incoming and outgoing traffic has asymmetric bandwidth. How to charge for incoming traffic presents similar problems as charging for outgoing traffic. The central problems in this context are: how can the receivers express their willingness to pay, and how can they be billed. Another problem lies in the fact that it is difficult to identify which packet should be charged to the sender and which should be charged to the receiver. In the real world, a user could even be a sender and a receiver simultaneously during an interactive session. Stiller et al. [46] discuss some possible approaches to these problems, especially dealing with accounting implementations. However, this is still an open issue.

5. Future trends

5.1. PVC vs SVC

More and more service providers are going to implement the switch virtual circuit (SVC) in their ATM networks. Pricing can be used to provide incentives for users to migrate part of their traffic from PVC to SVC service. Since each time there is a need to set up and tear down of the connections, connection fees need to be added. Siris et al. [37] pointed this out during his study of the pricing strategy of British Telecom. Liu and Petr [17] did some initial exploration on this topic. An important consideration is that the QoS of the connection should be associated with the pricing for the service, for example by using the effective bandwidth to estimate how much resource is needed to support a service. Work to date has been based on static pricing schemes. Further research on dynamic pricing for the PVC vs SVC scenario will be more interesting.

5.2. Multiple-provider scenario

Currently, most research work focuses on understanding the behavior of a group of users under some given pricing structure from one network provider. However, this is not the case in the real world, in which there are many network providers competing for customers, with the number of network providers continuing to grow. Understanding user behavior is only one part of the problem; studying the network provider's behavior under a multiple-provider scenario should be the next research topic. Messerchnitt and Hubaus [52] illustrate an interesting scenario in which network providers (resource managers) have to bid in order to attract more users due to competition.

6. Conclusion

In this paper, we reviewed the recent research efforts on the pricing for integrated networks. One thing need to be pointed out is, in spite of the compelling arguments from the academic researchers for more complex pricing schemes such as usage-based pricing and dynamic pricing, the industry currently seems to be moving more and more towards extremely simple pricing schemes. For example, Internet service and wireless phone service are often priced at a flat amount per month, usually with a usage cap. Traditional pricing distinctions such as usage, time of the day, and distance seem to be evaporating. These simple pricing policies can be very attractive to service providers (simplified billing) and users (predictable costs) alike.

We believe that pricing for the network service will follow a model similar to the current cable TV industry, in which customers pay a flat fee for the basic service (basic network access in the network world) and have the options to pay higher, usage-based fee for value-added service (like, VoIP, Video on demand, security…).

6. Uncited References

[3]. [4]. [53].

References

[1] L. Zhang, et al., RSVP: A new resource ReSerVation protocol, IEEE Network 5 (1993) 7.

[2] R. Geib, Differential service for internet and ATM, available at: http://www.internet2.edu/qos/qbone/qsg/i2qos-geib-difs-atm-02.html.

[3] Fang Lu, Raj jain, ATM congestion control, http://www.cis.ohio-state.edu/~jain/cis788-95/atm_cong/.

[4] Nanying Yin, Michael G. Hluchyj, On closed-loop rate control for ATM cell relay networks, Proceedings of IEEE INFOCOM, Vol. 1, 1994.

[5] Ron Cocchi, et al., A study of priority pricing in multiple service classes networks, SIGCOMM, Zurich, Switzerland, 1991.

[6] R. Edell, P. Varaiya, Providing internet access: what we learn from the INDEX trial, Keynote talk, IEEE Infomcom' (1999) 99.

[7] J.K. Mackie-Mason, H.R. Varian, Pricing the internet, in: B. Kahin, J. Keller (Eds.), Public access to the internet, Prentice-Hall, 1994.

[8] L.P. Breker, A survey of network pricing schemes, Proceedings of the 8th Symposium on Computer Science, University of Saskatchewan, 1996

[9] Ron Cocchi, et al., Pricing in computer networks: motivation, formulation and example, IEEE/ACM Trans. on Networking (1993) 196 (Dec.).

[10] Luiz A. Dasilva, David W. Petr, Nail Akar, Equilibrium pricing in multiservice priority-based networks, IEEE GlobeCom'97, Phoenix, AZ, 1997.

[11] C. Parris, S. Keshav, D. Ferrari, A framework for study of pricing. I integrated networks, Technical Report, Tenet Group, ICSI, UC Berkeley, 1994.

[12] L. Breker, Evaluation of a dynamic bandwidth pricing scheme, available at: http://www.cs.usask.ca/grads/lpb133.html.

[13] M.L. Honig, K. Steiglitz, Usage-based pricing for packet data generated by a heterogeneous user population, IEEE Infocom'95, Vol. 2 1995.

[14] C. Parris, C., D. Ferrari, A resource based pricing policy for real-time channels in a packet-switching network, Technical Report, Tenet Group, ICSI, UC Berkeley, 1994.

[15] K. Lindberger, Cost based charging principles in ATM networks, in: V. Ramaswami, P.E. Wirth (Eds.), Teletraffic contributions for the information age, Elsevier, 1997.

[16] David Songhurst, F.P. Kelly, Charging schemes for multi-service networks, in: V. Ramaswami, P.E. Wirth (Eds.), Teletraffic contributions for the information age, Elsevier, 1997.

[17] Liu Yuhong, David W. Petr, The influence of pricing on PVC vs. SVC service preferences, Technical Report, ITTC, University of Kansas, July, 1999 (ITTC-FY2000-TR-12960-03).

[18] A. Odlyzko, A modest proposal for preventing internet congestion, Sep, 1997, Available at: http://www.research.att.com/~amo.

[19] G. Fankhauserm, B. Stiller, B. Plattner, Arrow: a flexible architecture for an accounting and charging infrastructure in the next generation internet, NETNOMICS: Economic Research and Electronic Networking, Vol. 1, No. 2, 1999.

[20] D. Morris, V. Pronk, Charging for ATM services, IEEE Communications Magazine, May,1999.

[21] J.M. Peha,, Dynamic pricing as congestion control in ATM networks, IEEE Globecom-97, Phoenix, AZ, USA, Nov, 1997.

[22] G. Fankhauser, B. Stiller, C. Vögtli, B. Plattner: Reservation-based charging in an integrated services network 4th INFORMS TELECOMMUNICATIONS Conference, Boca Raton, Florida, U.S.A., March 11–18, 1998, Session MC-2.

[23] A. Lazar, M. Semret, Auctions for networking resource sharing, CTR Technical Report, Columbia University, 1997.

[24] Liam Murphy, John Murphy, Feedback and pricing in ATM networks, Third Workshop on Performance Modelling and Evaluation of ATM Networks, IFIP TC-6, Ilkley, UK, 2–6th July 1995.

[25] Liam Murphy, John Murphy, Pricing for ATM network efficiency, Third International Conference on Telecommunication Systems, Modelling and Analysis, Nashville, USA, 16–19 March 1995.

[26] F.P. Kelly, Charging and rate control for elastic traffic, European Transactions on Telecommunications 8 (1997) 33–37.

[27] Frank Kelly, Aman Maulloo, David Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998) 237–252.

[28] R.J. Gibbens, F.P. Kelly, Resource pricing and the evolution of congestion control, Automatica (1999) 35.

[29] P.B Key, D.R. McAuley, Differential QoS and pricing in networks: where flow-control meets game theory, IEE Proceedings Software, March, 1999.

[30] M.D. Biddiscombe, J.E. Midwinter, S. Sabesan, Application of free-market principles to telecoms resource allocations, IEE Electronic Letters, 1999.

[31] Q. Wang, J. M. Peha, M. A. Sirbu, The design of an optimal pricing scheme for ATM integrated service networks, Journal of Electronic Publishing, Special Issue on Internet Economics, 1995

[32] H. Ji, J.Y. Hui, E. Karasan, GoS-based pricing and resource allocation for multimedia broadband networks, IEEE Infocom' 96 (1996) 1020–1027.

[33] H. Jiang, S. Jordan, A. pricing, model for high speed networks with guaranteed quality of service, IEEE Infocom' 96 (1996) 888–895.

[34] F.P. Kelly, Notes on effective bandwidth, in: F.P. Kelly, S. Zachary, I. Zeidins (Eds.), Stochastic networks: theory and applications, Oxford University Press, 1996.

[35] S. H. Low, P. Varaiya, A new approach to service provisioning in ATM networks, IEEE/ACM Trans. on Networking, 1(5), Oct, 1993.

[36] C. Courcoubetis, V.A. Siris, Managing and pricing service level agreements for differentiated services, IEEE/IFIP IWQos'99, UCL London, UK, May, 1999.

[37] V.A. Siris, et al., Usage-based charging using effective bandwidths: studies and reality, TR 243, ICS-FORTH, January 1999.

[38] A. Gupta, D.O. Stahl, A.B. Whinston, An economic approach to network computing with priority classes, Working Paper, CISM, University of Texas at Austin, 1994. Available at: http://cism.bus.utexas.edu/alok/nprice/netprice.html

[39] A. Gupta, D.O. Stahl, A.B. Whinston, Pricing of services on the internet, in: Fred Phillips, C.C. Cooper (Eds.), IMPACT: how ICsup4(2) research affects public policy and business markets, Greenwood Publishing, Connecticut, 2001.

[40] Q. Wang, M.A. Sirbu, J.M. Peha, Pricing for ATM network services, in: G.W. Brock, G.J. Rosston (Eds.), The internet and telecommunications policy, Selected paper from 1995 TPRC, Lawrence Erlbaum Accosciates, 1996.

[41] Y.A. Korillis, T.A. Vararigou, S.R. Ahuja, Incentive-compatible pricing strategies in noncooperative networks, IEEE INFOCOM'98, San Francisco, CA, March–April 1998.

[42] Georg Carle, Thomas Schmidt, Jochen Seitz, Fair ATM charging with consideration of traffic characteristics and QoS parameters, Colloquium on 'Charging for ATM' (CA$hMAN Accounting Workshop), London, U.K., 12 November 1996, Institution of Electrical Engineers (IEE), London, 1996, pp. 5.1–5.7 (ISSN 0963-308).

[43] D. Botvich et al., On charging for internet service provided over an ATM network, EEE Workshop ATM'97, Lisbon, May 1997.

[44] N. Semret et al., Market pricing of differential internet services, CU/CTR/TR 503-98-37, Columbia University.

[45] D. Hazlett, An interim solution to internet congestion, Social Science Computer Review 15 (1997) 2 (Summer).

[46] B. Stiller et al., Charging and accounting for integrated internet services—state of the art, problems, and trends, INET 98, Geneva, July 1998.

[47] M. Karsten et al., Cost and price calculation for internet integrated services, IEEE Infocom'98.

[48] F. Hartanto, G. Carle, Policy-based billing architecture for internet differentiated services, IFIP 5th Intl. Conference on Broadband Communciation '99, Hong Kong, Nov, 1999.

[49] S. Shenker, Service models and pricing policies for an integrated services internet, in: G.W. Brock, G.J. Rosston (Eds.), The internet and telecommunications policy, Selected paper from 1995 TPRC, Lawrence Erlbaum Accosciates, 1996.
[50] S. Shenker, et al., Pricing in computer networks: reshaping the research agenda, in: G.W. Brock, G.J. Rosston, et al. (Eds.), The internet and telecommunications policy, selected paper from 1995 TPRC, Lawrence Erlbaum Associates, 1996.
[51] G. Carle et al., Charging for ATM-based IP multicast services, Interop/Networld 98, Las Vegas, May 1998.
[52] D. Messerchmitt, J.-P. Hubaux, Opportunities for electronic commerce in networking, IEEE Communications Magazine 9 (1999) 37.
[53] A.M. Odlyzko, The economics of the internet: utility, utilization, pricing, and quality of service, available at: http://www.research.att.-com/~amo/doc/
[54] Henry Haojin Wang, Telecommunications management networks, Chapter 11, Mcgraw-Hill, 1999 (TK 5105.5.w355).
[55] Mark Norris, Understanding networking technology: concepts, terms and trends, 2nd ed, Artech House, 1999 (TK5102.N67).