Understanding Internet Pricing: An Evaluation, A Classification and An Integrated Volume-Based Proposal

by

Jun Wang
B.Eng, Shanghai Tie Dao University, 1995

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Supervisor: Dr. Kin F. Li

ABSTRACT

The flourishing development of the Internet and network applications in communications, business, and entertainment, has resulted in resource allocation difficulties among a seemingly unlimited number of users and more complex charging rate determinations. As an important part of the global economy, some of the currently used Internet pricing schemes, such as flat rate charging, lack economic and network efficiency. Furthermore, many cross-disciplinary issues have yet to be resolved in the Internet Pricing research area.

After an overview of the representative charging schemes, this work introduces a well defined structure for a complete and comprehensive understanding of Internet pricing schemes. Moreover, an evaluation cube representing eight different perspectives is proposed to examine charging schemes. Based on the evaluation results, classifications of the surveyed charging schemes from the economics and technology perspectives are presented. A new charging scheme is proposed that integrates the advantages of existing schemes. Distance, application type, and congestion are considered in the charging rate determination. The scheme is very flexible as an Internet Service Provider can adjust the priorities of the charging factors to satisfy its own objectives. The new scheme is analyzed and evaluated quantitatively and qualitatively. The results show the design objectives are met, and the scheme could be readily adopted by the Internet Service Providers.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADSL	Asymmetric Digital Subscriber Line
CAC	Call Admission Control
ATM	Asynchronous Transfer Mode
ABR	Available Bit Rate
BE	Best-Effort
CPS	Cumulus Pricing Scheme
CP	Cumulus Points
DiffServ	Differentiated Services
EBW	Effective Bandwidth
INDEX	INternet Demand EXperiment project
ISDN	Integrated Services Digital Network
ISP	Internet Service Provider
IVB	Integrated Volume-Based
IntServ	Integrated Services
GP	Guaranteed Performance
PMP	Paris Metro Pricing
RSVP	Resource Reservation Protocol
QoS	Quality of Service
UBR	Unspecified Bit Rate

Acknowledgement

I would like to express my greatest appreciation to my supervisor, Dr. Kin F. Li for his encouragement, support and guidance throughout my graduate studies and thesis work. Thank him for his great patience and understanding which are really outstanding and impressive. Not only knowledge, research methodologies which I learned a lot from him, but also his enthusiasm and dedication to the work make us moved.

I also would like to thank my parents, all my family member for their consistent support, understanding and immense love to me.

Dr. Fayez Gebali, Dr. Hausi Müller and Dr. Eric Manning, their kindness and help to me throughout the study and research. Dr. Hajime Shibata, and all of the LAPIS group for their help and the talks during my graduate studies.

Finally, thanks to all my friends both in China and here for their invaluable friendship and concern for me.

Dedication

I dedicate this work to my beloved parents, ZhunZheng Wang and Jingyao Ye.

Chapter 1

Introduction and Background

1.1 Introduction

The Internet, one of the most successful information infrastructures, has been advancing remarkably and continuously in recent years. Without a doubt, the number of Internet users and the traffic load over the network have been growing dramatically. At the same time, with the advent of new processor technology, high speed networking, and special purpose high performance VLSI chips, an increasing number of resource consuming network applications, such as video conferencing, has become technically feasible and welcomed by users. All these point to a need of finding effective solutions for resource allocation in the Internet. Many congestion control and traffic management mechanisms, as well as methods to provide QoS guaranteed service over the best-effort-based Internet, have been proposed by network engineers.

On December 23rd, 1992, the National Science Foundation (NSF) announced that it would cease funding the ANS T3 Internet backbone [16]. Since then, Internet has entered a commercial era. Now, almost everyone is paying directly or indirectly for his/her Internet access. Since 1993, more and more attention has been focused on the subject of Internet pricing by economists and even engineers. It is believed that pricing, which is about how to set prices for Internet users according to their usage, could provide incentives that influence users' behavior. Economists found Nash equilibrium and Pareto efficiency can be achieved through a careful allocation of resources according to user's willingness to pay. Meanwhile, from network engineers' points of view, it is more exciting as charging schemes could be used as a soft, but effective congestion control mechanism comparing to other traditional technical solutions. By raising or lowering a charging rate, an Internet Service Provider (ISP) can balance the network traffic load based on the assumption that some users would reduce their traffic generation when the price rises. In the past decade, many academic researchers from economics and technical fields have devoted their efforts to this endeavor and have proposed different Internet access charging schemes with diverse objectives.

These charging schemes include Volume-Based charging, Time-Based charging and

QoS-Based charging. These charging schemes are defined by the information collected and used for charging: number of transmitted packets, duration of transfer, and required QoS level, etc. Moreover, different objectives are reached by the different characteristics of these charging schemes. For example, the Time-Based Charging Scheme is appreciated by users and ISPs because of its simplicity and ease of implementation. A QoS-Based Charging Scheme, on the other hand, provides different QoS levels according to the application's requirements, and the user is charged accordingly.

Furthermore, with the soaring number of Internet users, some economists have proposed more complicated charging schemes to achieve economic objectives. For example, the Smart-Market scheme which uses a bidding mechanism to sort out the winning packets to be transmitted, raises the important economic concept of marginal cost for transmitting packets over a congested network. Other innovative charging schemes have been proposed to improve the network efficiency for the whole society.

Additional relevant issues are raised for the various charging schemes, such as whether the ISP's revenue is maximized, or whether the scheme is socially fair to the users. Section 2 examines related issues in Internet pricing. The objectives of this work are set in Section 3. Finally, economics terms used throughout this work are highlighted for readers' convenience in Section 4.

1.2 Issues in Internet Pricing

With the expansion of Internet applications, it is believed that Internet pricing will play an important role in resource allocation and congestion control in the network. Related issues to be considered are:

- Understanding of currently available charging schemes and classifications.
- Evaluating charging schemes from different aspects and design objectives.
- Proposing future charging scheme development direction.

Our research work is built upon the above issues.

1.3 Objectives of This Work

Our work on Internet Pricing investigates the charging rate agreement between the ISP and users for the usage of Internet services provided. The users are conceptually considered to be at the edge of the network, and are charged for the usage of the resources allocated including routers, links and ISP supporting software along the transmission route.

The objectives of this research are:

- To survey previous work and the prevalent charging schemes (Chapter 2).
- To propose a well defined structure for a complete and comprehensive understanding of charging schemes (Chapter 3).
- To evaluate charging schemes from eight different perspectives (the evaluation cube in Chapter 4).
- To classify charging schemes according to how well they achieve the design objectives (the classification function in Chapter 4).
- To propose a new charging scheme that integrates the advantages of the charging schemes surveyed (Chapter 5).
- To analyze and evaluate the proposed scheme quantitatively and qualitatively (Chapter 6).
- To provide directions for future work (Chapter 7).

1.4 Economics Terms

There are certain economics terms used throughout the thesis. For readers' convenience, these terms are explained below. The audience can also refer to the excellent economics text [18]:

- Utility: Utility is the total benefit or satisfaction that a person obtains by consuming a good or service.
- Utility Function: A utility function is a way of assigning a number to every possible consumption bundle such that more-preferred bundles get assigned larger numbers than less-preferred bundles.
- Utility Maximizing: For individual consumer, utility maximizing means the consumer chooses the consumption possibility that maximizes total utility, given his/her income and the prices of all goods or services. In terms of the whole society's utility, utility maximizing refers to maximizing the society's total utility.
- Nash Equilibrium: If there is a set of strategies in a game playing scenario with the property that no player can benefit by changing his/her strategy, while the other players keep their strategies unchanged, then that set of strategies and the corresponding payoffs constitute the Nash Equilibrium.
- Economic Efficiency/Pareto Optimality: Economic Efficiency is a situation in which the system's value (utility) is maximized. Given certain fixed resources, technology, and preferences, no changes will increase this value. Economic efficiency is also called Pareto Optimality.
- Marginal Cost: Marginal cost can be interpreted as the additional cost of producing just one more ("marginal") unit of the output.
- Shadow Price: In solving an optimization problem, the Shadow Price is the amount that the objective function value would change if the named constraint changed by one unit. The Shadow Price is valid up to the allowable increase or decrease in the constraint. A simple example of Shadow Price is for a given network system with certain resources of bandwidth, buffer size, CPU processing speed, and certain user traffic pattern, after the system finishes resource allocation according to some objective function (such as maximizing the total revenue of the ISP), certain Quality of Service for each user is provided. If now the bandwidth (or any other resource)

has been increased by one unit, let say 1M bps, how much the objective function will change (such as how much more revenue the ISP will get) is the Shadow Price.

Chapter 2

A Survey of Internet Pricing Schemes

2.1 Introduction

In the Internet Pricing research area, there are many charging schemes proposed in the past decades by academic and industrial researchers from engineering and economics fields. Some of these charging schemes are very different as they were developed from totally different perspectives. Some of the schemes are similar as one is an extension or improvement of another. The extensions and improvements have the same objectives and are suitable for the same specific situations. For example, Paris Metro is developed from Flat Rate and they are similar in that users will be charged a fixed sum irrespective of their usage. These are simple schemes desired by users.

In this chapter, almost all currently in-use and proposed charging schemes are discussed. It is intended to give the readers a basic understanding of what Internet Pricing is and what the current practices are. There are existing surveys of network charging schemes examined from different perspectives. Dasilva focused on charging schemes for QoS-enabled networks [7]. Falkner et al. examined charging schemes in broadband multiservice networks [8]. Using pricing schemes to control congestion was emphasized in Henderson et al.'s survey [10], while Stiller et al. investigated pricing schemes based on cost recovery models [25]. Our survey introduces the representative charging schemes from all perspectives and objectives. As a general survey, each of these charging scheme principles, as well as its advantages and disadvantages are analyzed.

Similar charging schemes are grouped into a section. For example, Flat Rate and Paris Metro are described in the same section. The features of the charging schemes and their differences and similarities will be examined. Within each section, the charging schemes are introduced in the order that the latter ones are providing some solutions for problems raised by the previous ones. This chapter concludes with a list of the characteristics of the charging schemes examined.

2.2 Flat Rate Charging Schemes

Though most economists emphasized their inefficiency in network resources allocation and unfairness among users, Flat Rate schemes are still the prevalent charging schemes in North America because of their simplicity and attractiveness to users [23], [21], [22], [9]. Flat Rate schemes are not favored by economists because of their potential "tragedy of the commons" results [2], where a shared resource is over-consumed by selfish individuals who only consider their personal benefits but not the cost to the society (other users) as a whole.

There are two different Flat Rate charging schemes:

- Single Service Flat Rate
- Paris Metro

2.2.1 Single Service Flat Rate

In the Single Service Flat Rate pricing scheme, users are charged a fixed amount per time unit irrespective of usage. All users are treated equally with no service differentiation, hence the name Single Service Flat Rate. A simple example of Single Service Flat Rate is $25 for monthly access with no usage charge.

In [23] and [9], it is argued that a Flat Rate charging method is simpler than other schemes, and it is shown that this simplicity is more suitable for less expensive but frequently used communication services. It presents a pattern of the charging methods in the history of communication technologies repeating in mail, telegraph, telephone, and Internet: quality increase, then price decrease, followed by usage increase resulting in revenue increase, and finally a more simple pricing structure.

The Single Service Flat Rate is the most simple charging scheme for both ISPs and users. There is no charging overhead or billing measurement needed. Users can easily control their budget. On the other hand, for those users who would like to pay more for better services, the Single Service Flat rate could not meet their requirement as there is no service differentiation provided. Furthermore, without a penalty of using more resources than required, users tend to overuse the resources.

2.2.2 Paris Metro

The name of the Paris Metro charging scheme came from the tariff in the Paris Metro. In this method, users have a choice of first or second class service. Services differ in price, and therefore their degree of congestion. Users sort themselves into groups according to their own preferences.

In [20], Odlyzko applied the same principle to Internet pricing. The backbones would be divided into several logically separate channels, each with a different price per byte. Users are free to select the channel the packets are to be sent on. The expectation is that the more expensive channels would attract less traffic, and therefore would be much less congested.

The Paris Metro scheme is a big improvement over the Single Service Flat Rate as it provides service differentiation to users so that for those who are willing to pay more could get better service. Also, as a Flat Rate charging scheme, Paris Metro is simple and no extra billing cost is needed. Nevertheless, Paris Metro's robustness in differentiating services by dividing the network into several parts still may not meet each individual's specific requirements. As it is claimed by most economists [8], economic efficiency that maximizes users' total utility cannot be met by Paris Metro. Moreover, as Single Service Flat Rate, the "tragedy of the commons" effect also exists in the Paris Metro scheme.

2.3 Simple Usage-Based Charging Schemes

There are three Usage-Based charging schemes: Time-Based, Volume-Based, and Distance-Based. There also exist some sophisticated Usage-Based schemes, for example, charge according to mean traffic rates, or burst rates etc. They will be introduced separately in later sections.

2.3.1 Time-Based

With a Time-Based charging scheme, users are charged according to their connection time to the ISP's access point. A simple example of a Time-Based charging scheme is $0.01 for each minute of connection, with no volume charge applied and irrespective of the distance.

Time-Based charging is still a very commonly used charging method for Internet access in Asia and Europe. For example, in Shanghai, one of the biggest cities in China, Internet users are charged 80 yuan (Approximately US $10) for 60 hours of ADSL (Asymmetric Digital Subscribe Line) access connection. This Time-Based charging scheme is compatible with the telephony practice in China and many Europe countries, where local telephone calls are charged based on duration.

As users are charged according to their connection time, the effect of "tragedy of the commons" is avoided in a Time-Based charging scheme. Since Time-Based is compatible with the telephony charging mechanism, the ISPs are assured of the costs and revenues as the costs for telephony can be covered through Time-Based charging scheme. A disadvantage of Time-Based charging is that there is no service differentiation provided. Users' specific preferences are not taken into account, accordingly economic efficiency is ignored here. Additionally, as volume and distance are not considered, it is unreasonable that users who send larger volumes over farther distance are being charged at the same rate as others, since they actually occupying more resources. Also, charging users higher for longer distance transmission route could drive users to use shorten route if possible, thus increasing the network efficieny.

2.3.2 Volume-Based

With a Volume-Based charging scheme, users are charged according to their traffic over the Internet. A simple example of Volume-Based charging is $0.0003 for 1 MByte of traffic.

Some market researchers have noticed that Volume-Based charging could be a very strong competitor among charging schemes for wireless Internet service. A very good example of this is NTT DoCoMo's famous i-mode. "In 1999, NTT DoCoMo launched the world's first wireless Internet service, redefining what mobile communication was all about. i-mode now boasts over 10 million subscribers, out of a regular cellular subscriber base of 50 million" [26]. In the report, it is said the biggest successful step of NTT DoCoMo is its decision to "switch the network to packet-switching and, more importantly, CHARGE on a packet-basis. You are charged for the amount of data that you download, not for the time that you use. Logon time is not charged. You are charged for the download but once the download is done, your viewing time is not charged" [26]. This is essential for the currently slow-speed 2G (GSM or CDMA) wireless connection.

Comparing with the Time-Based scheme, the Volume-Based scheme takes the characteristics of the Internet's statistical multiplexing into consideration, where resources are allocated to channels only when receiving or sending data. From this point of view, the Volume-Based scheme is more suitable for pricing Internet services. Moreover, same as the Time-Based charging scheme, as users are charged according to their usage, users will only send requests which are more valuable to them. The drawback of Volume-Based charging is that there is no service differentiation taken into account. Therefore, users' diverse requirements cannot be met and high economic efficiency cannot be achieved. Furthermore, distance is not considered, which is unfair for users who send packets to closer locations to pay the same price since less resources are allocated to their services. Moreover, by using distance as a charging factor, network efficiency will be improved.

2.3.3 Distance-Based

The Distance-Based charging scheme originated with traditional telecommunications services, such as telephony. Charging factors for traditional telecommunications services reflect the topology of the circuit-switched network, and are based on both distance and point-to-point route considerations.

Distance determination could be done in many ways: kilometers, number of hops along the transmission route, or even more specifically the destination's IP address. A simple example of Distance-Based charging is $0.015 for every minute from Victoria, Canada to Shanghai, China, while it is $0.008 for every minute from Victoria, Canada to Montreal, Canada.

As the resources required to carry the transmission are quite often proportional to the distance, Distance-Based charging is a fair scheme for both the ISP and the users. However, without other charging factors such as time and volume, the pure Distance-Based charging scheme does not fit well with both the ISP's and users' requirements. In practise, Distance-Based charging is usually combined with either Time-Based charging or Volume-Based charging, such as a metered minutes-of-use or packets-of-use rate structure based on distance.

2.4 Priority Pricing

Back in 1993, an article by Roger Bohn et al. [4] had already foreseen the potential resource allocation problem in the Internet as there is little control over possible high-volume incoming traffic, especially from the emerging bandwidth-hungry real time applications. The author claimed that this problem could be resolved by setting a packet's priority so that packets which require low latency and high volume will be set to higher priorities. This could be done by the end-users and applications, using the precedence bits in the header of an IP datagram. In order to prevent users from always setting their applications to the highest priority, a quota to allocate resources could be assigned to each user.

Nowadays, the idea of setting the precedence bits in the header of the packets to allow users to indicate the value of their traffic is applied in Priority Pricing. With the commercialization of the Internet, the previous quota has been replaced by payments where users pay more for higher priority level [14]. An example of Priority Pricing would be: P_{level1} is $ 0.0002/packet, while a higher priority P_{level2} is $ 0.0003/packet.

The Priority Pricing scheme [8] [6] allows users to inform the ISP of the priority of their service requests. Traffic is managed by sending only those packets with higher priority when facing traffic congestion. The price for those traffic requests which are marked as lower priority are cheaper than the higher priority ones. It is shown in [6] that it is possible to set the prices to reach Nash equilibrium that maximizes the sum of users' satisfaction with network performance. Using Priority Pricing alone cannot guarantee QoS, but by combining with RSVP (Resource Reservation Protocol) and CAC (Call Admission Control) technologies, QoS could be provided for real-time applications. The downside of this scheme is the overhead associated with the priority management. Moreover, no social fairness is considered – poorer users can only set their traffic to lower priorities which could be delayed or even dropped when the network is congested.

2.5 Smart-Market

As pioneers in introducing the concept of incremental cost in sending additional packets into a congested network, MacKie-Mason and Varian claimed that users should be charged by a marginal cost (the additional cost of producing just one more unit of the output). This cost can be determined through a Vickrey auction (winner would not pay the highest, but the second-highest bid). The network administrator collects and sorts all the bids, and then determines a threshold monetary value, the marginal cost of congestion, based on the network's capacity. Only the packets whose bids exceed the threshold are transmitted [8], [10], [17], [16].

The Smart-Market concept has attracted the attention of economists and engineers to the economic side of Internet. It is claimed that if the auctions are designed appropriately, the Smart-Market mechanism can encourage both network and economic efficiency. While at the same time, it is also argued that the Smart-Market mechanism is only a conceptual contribution. It is not practical and is not compatible with current available technologies, as users have to bid for transmission at each router along the transmission route. The complexity and high communication overhead would add extra load to the already congested network.

2.6 Edge Pricing

Edge Pricing [8], [10] concerns itself with the cost structure and network architecture, and tries to solve the problem of settling payments between interconnected domains. It is a solution to the scalability problem of the original Smart-Market, where with the increase of network structure, the communication overhead for delivering packets increases dramatically. With Edge Pricing, instead of users making individual payment to the owner of each congested router, each network operator retains control over how it charges its users at the edge of the network. Charges are locally computed based on expected network congestion, such as the time of the day, short-term congestion history, and so on.

Unlike Smart-Market where users have to bid for transmission at each congested router, Edge Pricing is much simpler and easier for users to pay for their transmission. Furthermore, Edge Pricing has been proven to be compatible with currently prevalent ATM (Asynchronous Transfer Mode) and RSVP (Resource Reservation Protocol) technologies that support a CAC (Call Admission Control) function at the edge of the network. Using these traffic engineering technologies, Edge Pricing could provide a QoS guarantee to each individual user's traffic flow. The major disadvantage of Edge Pricing is that economic efficiency is de-emphasized. In Edge Pricing, the prices are fixed over medium (day, week) or long (month, year) time frames. However, the network congestion status and users' utility functions are usually changing within a short time (second, minute), therefore it is not possible to set the prices at the utility-maximizing points in Edge Pricing. Thus, the overall economic efficiency is degraded.

2.7 Responsive Pricing

Responsive Pricing [8] is a dynamic price-setting strategy based on exploiting the nature of the users in the network: Elastic users will delay their transmission when price gets higher; and inelastic users cannot tolerate delay but can accept higher loss probability when the network is congested. Accordingly, the utility functions for these two groups of users are different. When the price is adjusted, both groups of users will change their traffic requests according to their own desires. The price is calculated according to the resource utilization level. When the network gets more congested, the price gets higher. Elastic users may drop their service requests, therefore the network traffic load is reduced.

The authors [8] stated that Responsive Pricing is designed for ATM ABR (Available Bit Rate) services, so it is compatible with existing technologies. The drawback of Responsive Pricing is its high overhead of measuring workload and accounting. Furthermore, it is claimed that the network could be unstable when both users' utility functions and the prices are changing. For example, when the network is getting less congested, the price for network resource usage decreases. Accordingly, more elastic users will start to transmit at the same time. This could cause network congestion. In Chapter Five, a charging scheme will be proposed, which has a similar characteristic exploiting the different users' natures while simplifying the network utilization measurement procedure.

2.8 Proportional Fairness

As shown in previous sections, some of the charging schemes, such as Smart-Market and Priority Pricing, have to reject services to poorer users in order to sustain services to users who are willing to pay more when the network is congested. Social fairness is not taken into account in these charging schemes. The Proportional Fairness Pricing [3], [8] tries to integrate the concept of fairness into the allocation of network resources. In this scheme, network resources are allocated in proportion to how much the user is willing to pay.

Proportional Fairness Pricing is derived from the following optimization problems seeking the optimal charging rates for the users' flows:

1. From the global system's point of view:
   Assumptions:
   - Network Resource J with Capacity $C_j, j \in J$.

- User r in the set of User R.
- $U_r(x_r)$ is the utility function of the transmission route to the user at rate x_r.
- The utility function $U_r(x_r)$ is assumed to be increasing, strictly concave, and continuously differentiable with respect to x_r.
- $\{A\}$ is an indicator matrix where $A_{j,r} = 1$ if resource j is required to carry the user's flow on route r.
- $C = (C_j, j \in J)$ and $x = (x_r, r \in R)$.

The system's optimal rates are then calculated by solving the following optimization problem:

$$\text{Max } \textstyle\sum_r U_r(x_r) \text{ (objective function)}$$
$$\text{subject to } Ax \leq C; x \geq 0. \text{ (constraints)}$$

The maximization problem could also be solved using the Lagrangian method. It is claimed by Kelly [13] that the Lagrange multiplier of μ_j of this problem implies cost of a unit flow through resource j, or in other words, the shadow price (the amount that the objective function value would change if the named constraint changed by one unit) of additional capacity at resource j.

2. From a user point of view: Based on the result from 1, user will be charged the shadow price of the resource on the route, i.e., the price $p_r = \sum_{j \in J} \mu_j A_{j,r}$. Furthermore, taking the cost into consideration, and maximizing each user's individual utility function after cost is the objective function in this case:

$$\text{Max } \textstyle\sum_r U_r(x_r) - p_r x_r \text{ (objective function)}$$
$$\text{subject to } x_r \geq 0. \text{ (constraints)}$$

3. From the network service provider's point of view: This is also based on the result that the charging price is equal to the shadow price of the resource on the route, $p_r = \sum_{j \in J} \mu_j A_{j,r}$. Maximizing total revenue is the objective function in this case:

$$\text{Max } \textstyle\sum_r p_r(x_r) \text{ (objective function)}$$

subject to $Ax \leq C; x \geq 0$. (constraints)

The Proportional Fairness charging scheme is economic efficient as it tries to maximize the users' utilities. It is socially fair as everyone is allocated some bandwidth according to their payment. Congestion control is implemented by allocating less resources proportionally to each user when the network is busy. Billing measurements are not required as users are charged according to their willingness to pay. However, it is still necessary to keep track of users' willingness to pay and to change the optimal resource allocation accordingly. The calculation of resource allocation is complex and is changing all the time which could bring burden to an already congested network. Furthermore, there is no individual QoS guarantee in Proportional Fairness Pricing as resources are allocated proportionally to users' willingness to pay. As the total requirements of resources are changing, there is no guaranteed amount of resources for each individual.

2.9 Effective Bandwidth

The Effective Bandwidth charging scheme [8], [15], [12] requires users to declare their mean and peak cell rate of their traffic during Call Admission Control (CAC). Accordingly, with the assumed Effective Bandwidth of the user's traffic, he/she is charged with a linear function which is tangent to the effective bandwidth curve.

The main idea of the Effective Bandwidth (EBW) concept is quantifying the relationship between a QoS grade and its bandwidth requirements. In more details, suppose there is an on-off traffic source transmitting over a link shared with other sources. The source transmits at a mean rate of M, and a peak rate of H, then the transmission probabilities of on and off are $P\{off\} = 1 - \frac{M}{H}$ and $P\{on\} = \frac{M}{H}$ respectively. The authors derived that the Effective Bandwidth required to guarantee QoS for this traffic flow over the link is as follows:

$$g_{on/off}(M, H) = \frac{1}{st} \log(1 + \frac{M}{H} (e^{stH} - 1))$$

There are two parameters which relate to the system capacity and buffer size:

- The space parameter *s*, which is measured in $(kbps)^{-1}$, is the degree of the multiplexing among the traffic flows. Parameter *s* depends on the size of the peak rate of multiplexed sources relative to the link capacity.
- The parameter *t* is the most probable duration of the buffer busy period prior to overflow.

For the above function, with a fixed capacity C, buffer B (which decides the parameters *s* and *t*) and peak rate H, the Effective Bandwidth function is a monotonically non-decreasing concave function of the mean rate M, as shown in Figure 2.1, which is a clarified version of the one shown in [8]. Consequently, the charging rate is designed as a tangent line at the point on the Effective Bandwidth curve where the mean rate is what the user claimed as M.

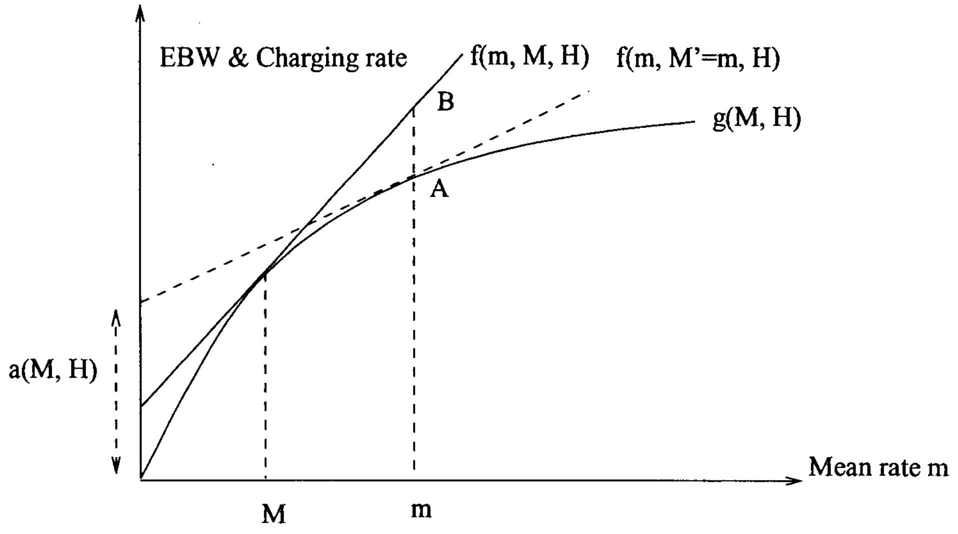

Figure 2.1. *Effective Bandwidth Pricing*

Assume now that the user is transmitting at a new mean rate *m*, then the charging rate function is f(m,M,H) = a(M,H) + m b(M,H), where

$$b(M,H) = \frac{e^{stH}-1}{st(H+M(e^{stH}-1))}$$

$$a(M,H) = g_{on/off}(M,H) - Mb(M,H)$$

The incentive of the scheme is that both over reservation and overuse of the bandwidth will result in punishment in the pricing rate. As shown in Figure 2.1, when a user delivers traffic at a mean rate *m* which is not equal to M as expected, the pricing rate (point B) would be higher than what it would be if M'=*m* was declared instead (point A). The function b(M,H) is shown as the slope of the tangent line f(m,M,H).

The advantages of Effective Bandwidth are discussed extensively in [8], [15], [19]. It is said that if the users act rationally, they will try to reduce their cost of the connection. Accordingly, the Effective Bandwidth scheme can be used to implement effective bandwidth-based Call Admission Control (CAC). Furthermore, individual user's QoS guarantee can also be realized. It is also claimed that network and economic efficiencies would be reached. The main disadvantage of Effective Bandwidth Pricing is its limited application to the use of ATM network technology. Also, the charging rate is changing with traffic status in a short time scale (second, minute) which potentially can bring heavy billing measurement and charging overhead to the network.

Under the more specific assumption of different QoS requirements of users' traffic flows, variations of the Effective Bandwidth Pricing were proposed in [19] for improvement.

2.10 Expected Capacity

The concept of Expected Capacity charging was examined in [5], [8]. The idea of Expected Capacity charging is that users will specify their expected required capacity, and the network service provider will then charge the users according to the expected capacity based on a long-term contract. Actual usage charge is not considered.

The Expected Capacity could be specified in various ways according to the user's usage pattern, such as minimum capacity required, maximum transfer time of data, or an effective bandwidth-based traffic characterization. Expected Capacity adopts policing mechanism to monitor extra traffic instead of the billing and measurement utilized in usage-based charging schemes. It makes an accept/reject decision according to whether the user has overused the resource as claimed in advance. This saves a lot of billing overheads for the network.

Expected Capacity has a number of advantages. First, it is based on a long-term contract which is desirable to users for its simplicity of billing. Second, it gives the service providers a more stable model of capacity planning. The main shortcoming of Expected Capacity is also because of its fixed charging rate over long time scale (month, year), its economic efficiency is not as high as other charging schemes that can adjust the charging rate within a short time frame according to the congestion status. Also, Expected Capacity is not convenient for those users who want to change their capacity requirements frequently.

2.11 Cumulus Pricing

In [24], a novel approach – Cumulus Pricing, proposed to charge differentiated Internet services efficiently. The main concept of Cumulus Pricing is to integrate different time scales into one mechanism. Some economic and network efficiencies are reached while users, like in Expected Capacity Pricing, only need to negotiate a long-term contract with network service providers for the expected resource requirements.

This long-term contract is related closely to the concept of a Service Level Agreement (SLA) in DiffServ. It includes a traffic specification declared by the user, a flat rate decided by the ISP and the user accordingly, and a couple of utilization thresholds that are used for traffic monitoring.

One of the innovative ideas in Cumulus Pricing is over- and under-utilization are represented by Cumulus Points (CPs). The CPs are set in red or green colors. The Red CP indicates the user has overused her capacities to some threshold and the green one means the opposite. Cumulus Pricing monitors user's traffic and sends this kind of feedback in CPs to user periodically (weekly or monthly).

The CP helps to implement the mapping of different time scales in the following strategy. In a Cumulus Pricing scheme, same as other complicated congestion-based charging scheme, the network provider monitors and measures each individual user's traffic (i.e., on a short time scale), while this measurement and monitoring information is only kept to the network provider's side. At the user's side, users only pay a fixed fee over a long time scale according to the contract signed. The accumulated CPs over time are only used to show the imbalance of the contract with real traffic, and leave the possibilities of further reactions later, such as re-negotiating a contract. In summary, in terms of monitoring and measurement at the network provider's side, Cumulus Pricing is on a short time scale which helps to increase network and economic efficiency, while at the user's side, the scheme is simple and is on a long time scale which is desirable to most of the users.

As shown in [24], the main advantage of Cumulus Pricing is that it achieves a better balance among the network, user and technology perspectives than any other proposed charging schemes. While at the same time, the authors pointed out that there are still a couple of theoretical issues to be resolved. One of these is how to derive the charging rate according to the QoS parameters defined in the SLA. The other is, similar to Expected Capacity, the network and economic efficiencies of Cumulus Pricing are not as high as other complicated Usage-Based charging schemes, such as Smart-Market or Responsive Pricing, that change the charging rate continuously.

2.12 The INDEX Project

The INternet Demand Experiment Project (INDEX) [11], [28], is an experiment designed to estimate how much people are willing to pay for various kinds of Internet Quality of Service. The INDEX designers configured the system to provide different QoSs to users and recorded the usage of each different QoS by each user. Users can change their requested QoS all the time and are billed monthly for their usage. From April 1998 to December 1999 there were about 70 users at the University of California, Berkeley participating in the INDEX project with residential Integrated Services Digital Network (ISDN) service as the access method.

In INDEX, users choose their desired network service from a menu of QoSs and are charged accordingly for every experiment. The service list usually includes different bandwidth-price choices. The menu of QoS choices changes every six to ten weeks in an experiment. For example, in the Minute Pricing Experiment, five bandwidths are provided for choosing: 16 kbps, 32 kbps, 64 kbps, 96 kbps and 128 kbps. In the Byte Pricing Experiment, only two bandwidths are available: 8 kbps and 128 kbps. Specifically, the following four experiments were preformed to understand users' preferences under different pricing schemes:

- Minute Pricing: Users are charged per-minute rates at each of the five different bandwidths.
- Byte Pricing: Users are given two choices: flat rate at low bandwidth or charge per byte at high bandwidth.
- Minute-Byte Pricing: This experiment is designed to show users' preference between minute charge or byte charge. Users can choose either minutes or bytes, or a combination of both.
- Flat Rate Buy-Out Option: Users are given the opportunity to buy out any of the five bandwidths which will be charged as flat rate with unlimited access. Users who do not buy any bandwidth or who want to use higher bandwidth which they have not bought out will be charged by Minute Pricing.

The result of the INDEX project shows the following characteristics:

1. Users have heterogeneous preferences.
2. Flat Rate Pricing can strongly influence the usage, and even cause over-usage.
3. Flat Rate Pricing is more attractive than Usage-Based Pricing to most of the users.
4. Users demand variable network service over time.

With the above conclusions, in [11], the authors supported the users' demands to call for more flexible pricing plans than the currently predominant flat rate and per-minute pricing plans. An interesting pricing scheme was also proposed to combine the benefits of Flat Rate and Usage-Based charging methods. Under this scheme, users can pay a low fixed rate for a basic network access service, and at the same time choose a higher QoS service according to their own requirements which will be charged by usage, such as per-byte, per-minute, or a combination of the two.

2.13 Chapter Conclusion

There are many Internet pricing schemes in use or proposed by engineers and economists. These charging schemes are different in their diverse objectives and so as their characteristics. A summary of their characteristics is shown in Table 2.1. In this table, short time scale refers to second, minute, or hour; medium time scale is based on day, or week; while long time scale indicates month, or year. These characteristics and the parameters for charging consideration will be examined in details in Chapter 3. Although the characteristics of Expected Capacity, Cumulus Pricing and INDEX project are identical as shown in Table 2.1, it will be shown that they are actually different when we study them in more details in Chapter 3.

Table 2.1. *Characteristics of Charging Schemes*

Charging Scheme	Characteristics
Flat Rate	Fixed rate over long time scale; No service differentiation; No QoS guaranteed; Suitable for Best Effort network.
Paris Metro	Fixed rate over long time scale; Some service differentiation; No QoS guaranteed; Suitable for Best Effort network.
Time-Based	Fixed time rate over long time scale; No service differentiation; No QoS guaranteed; Suitable for Best Effort network.
Volume-Based	Fixed volume rate over long time scale; No service differentiation; No QoS guaranteed Suitable for Best Effort network.
Distance-Based	Fixed distance rate over long time scale; No service differentiation; No QoS guaranteed; Suitable for Best Effort network.
Priority Pricing	For each priority, fixed rate over long time scale; Charge according to Priority level; Service differentiation; No individual QoS guaranteed; Suitable for Best Effort network.
Smart-Market	Dynamic charging rate over short time scale; Charging rate is per packet; Service differentiation; No individual QoS guaranteed; Users can input their preferred rate; Not compatible with current technology.
Edge Pricing	Fixed charging rate over medium time scale; Service differentiation; Individual QoS guaranteed; Compatible with ATM and RSVP technologies.
Responsive Pricing	Dynamic charging rate over short time scale; Charge according to resource usage; Some individual QoS guaranteed; Compatible with ATM ABR services.
Proportional Fairness	Dynamic charging rate over short time scale; Charge according to resource usage; Users can input charging rates which they are willing to pay; Service differentiation; No individual QoS guaranteed.
Effective Bandwidth	Dynamic charging rate over short time scale; Charge according to traffic characteristics; Users declare their mean rate and peak cell rate; Service differentiation; Individual QoS guaranteed; Compatible with ATM and CAC technologies.
Expected Capacity	Fixed charging rate over long time scale; Users can input their preferred charging rate; Service differentiation; Individual QoS guaranteed.
Cumulus Pricing	Fixed charging rate over long time scale; Users can input their preferred charging rate; Service differentiation; Individual QoS guaranteed.
INDEX Project	Fixed charging rate over long time scale; Users can input their preferred charging rate; Service differentiation; Individual QoS guaranteed.

Chapter 3

Examination of Charging Parameters

3.1 Introduction

As learned from Chapter 2, different Internet pricing schemes have different characteristics. In a well organized approach, diverse charging parameters can be used to examine charging schemes and to classify them into relevant groups. For example, there are charging parameters which are implemented as the charging unit of the tariff, such as call duration, call bandwidth, service level, etc. There are other charging parameters which take into account the network technology, for example, Best Effort network or Performance Guaranteed Network such as ATM. Consequently, charging parameters can be organized into charging components and sub-components according to their inter-relationship and their effects on the charging scheme.

We are not aware of, in the open literature, a complete listing of charging parameters considered in existing charging schemes. Moreover, there has not been any previous work on how to catalogue all these possible charging parameters into a well organized and useful picture.

In this chapter, we will propose a hierarchical structure that includes all the charging parameters encountered in Chapter 2. These parameters' relationship will be examined in detail (Section 2). Then, we will apply the charging parameters to current charging schemes to illustrate the hierarchical structure's effectiveness (Section 3). Finally, the advantages of cataloguing charging schemes using these parameters are discussed in the conclusion (Section 4).

3.2 The Charging Components

3.2.1 A Hierarchical Structure of Components

After a survey of currently available charging schemes, it was found that there are various charging parameters involved in each scheme. Moreover, there are special relationships among these parameters, and their effects on the charging rate are different. A hierarchical

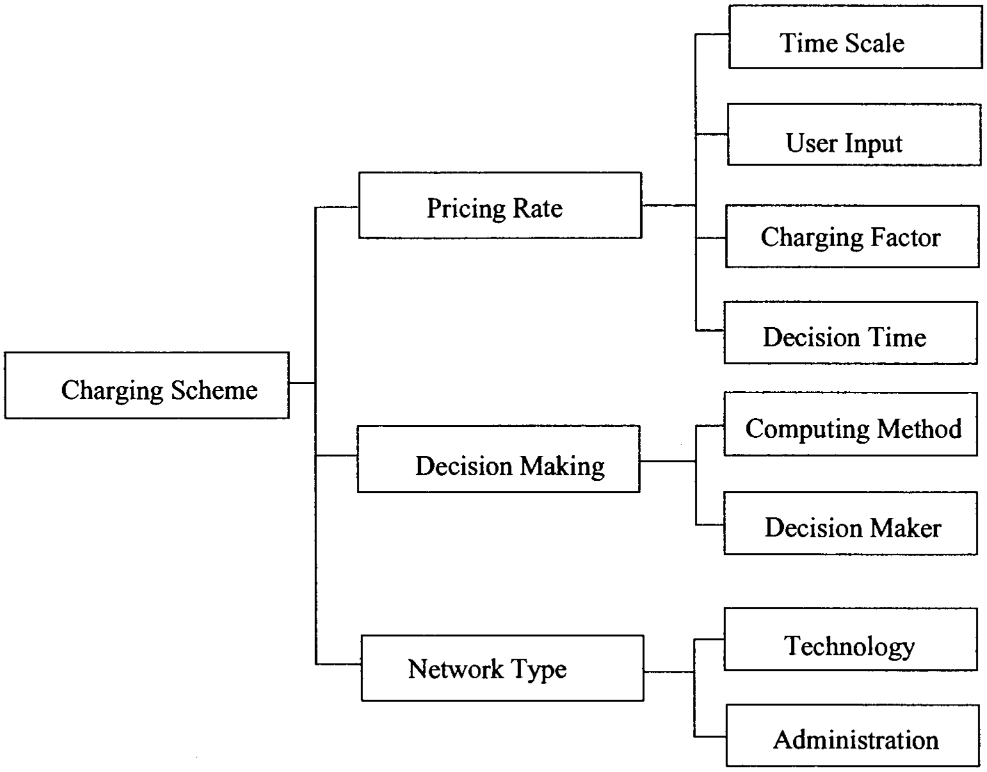

Figure 3.1. *Hierarchy of Charging Components*

structure could present these parameters in a clearer way and help us to understand the objectives of the particular charging scheme. Furthermore, a hierarchical structure shows how a charging scheme can be constructed from low level parameters. This proposed structure for charging components is shown in Figure 3.1. The hierarchy is composed of three levels, from top to bottom: Components, Sub-Components and Parameters.

We found that three components must be analyzed in order to understand a charging scheme: Pricing Rate, Decision Making and Network Type. Under each charging component, there are several sub-components which group charging parameters that have similar characteristics together. Charging parameters are at the lowest level of hierarchy and they form the basis for the charging scheme. These parameters and their typical values will be examined in the following sections [29].

3.2.2 Pricing Rate Component

The most important component in devising a charging scheme is the Pricing Rate Component. Under the Pricing Rate Component, sub-components related to the charging rate are defined: Time Scale, User Input, Charging Factor and Decision Time. The definitions of these sub-components are:

- Time Scale: The duration that the charge rate remains in effect is specified. There are three time scales: short, medium and long. Typical values for short time scale are second, minute and hour, and we consider charging rate changes within this time scale are dynamic. Typical values for medium time scale are day and week, which means charging rate may change within a day or a week. For long time scale, typical values are month and year, which the charging rate is considered comparably fixed.
- User Input: There are two possible inputs from users. One is: the user could have the authority to specify his/her desired charging rate. The other is: whether the user has the right to accept or reject the charging rate defined by the ISP.
- Charging Factor: Factors used for charging are defined, for examples, distance, connection time, Quality of Service, etc.
- Decision Time: There are two ways to calculate the charging rate, In Advance or Dynamic. In Advance means the charging rate is decided beforehand without taking real-time information into consideration, while Dynamic means the charging rate is dynamically calculated.

The parameters of the Pricing Rate Component is listed in Table 3.1.

We shall use the time-of-day charging scheme in the Edge Pricing category as an example to illustrate how the parameters under Pricing Rate can be applied: the rate is in effect within a medium time scale and is defined in advance based on historical and expected network consumption for that time period of the day; while the charging factor could be based on the connection time. Another example is the congestion-based Smart-Market charging scheme [8], [10], [17], [16]: the charging rate changes dynamically to network congestion

Table 3.1. *Parameters of The Pricing Rate Component*

Sub-Components	Parameters	Typical Values
Time Scale	Short Time Scale: The charging rate is calculated dynamically	Second, Minute, Hour
	Medium Time Scale: Charging rate is calculated over medium time scale, such as the time of the day or the day of the week	Day, Week
	Long Time Scale: Charging rate is decided for a long-term contract between the ISP and the user	Month, Year
User Input	Rate: This is a parameter to show whether a user can influence the charging rate by specifying her desired rate.	User Input Rate, No User Input Rate
	Choice: This is to show whether the user has the right to accept or reject a rate decided by other parties.	Yes, No
Charging Factor	Time: Charge according to how much time the consumer has kept the connection	Per Second, Minute, or Hour
	Space: Charge according to the distance between the sender and the receiver	Per Kilometer, Route, or Time Zone
	QoS/Priority: Charge according to the level of Quality of Service provided to the consumer or a specified priority	Throughput, Delay, Jitter and Loss Probability, Priority
	Data Volume: Charge according to how much traffic the consumer has created	Per Packet, Message, Call, MByte
	Resource: Charge according to the amount of resources the consumer has used	CPU Processing Time, Memory, Bandwidth, Number of Routers
	Traffic Characteristics: Charge according to the characteristics of the application. This is because different applications have different QoS requirements and transmission properties such as Peak rate and Average rate	Error and/or Delay, Sensitive or Insensitive, Peak rate, Sustained Rate, Burst Length, etc.
Decision Time	In Advance: The rate is decided before hand; this is usually the case of fixed rate pricing	$25/month
	Dynamic: The rate is decided dynamically according to some real-time conditions	$ X = function (traffic, QoS, route, ...)

situation within a very short time scale; the charging factor is based on data volume; and a user's input rate is considered by a bidding mechanism even though the final rate is lower than that user's input rate.

3.2.3 Decision Making Component

The other component at the highest level is Decision Making and its structure is shown in Table 3.2.

Table 3.2. *Parameters of The Decision Making Component*

Sub-Components	**Parameters**
Computing Method	Centralized: Information is collected, and computation is performed at a centralized location.
	Distributed: Computation is performed at local nodes based on limited network information.
Decision Maker	User: The charging rate is decided by the user.
	ISP: The charging rate is decided by the ISP.
	Application: The charging rate is decided by the application.

The Decision Making Component includes two sub-components. One sub-component indicates where the decision is made, either centralized or distributed. The centralized computing method is more accurate at measuring the status of the network since information is collected from other parts of the network. However, this mechanism is time and resource consuming, and network information collection generates considerable amount of traffic. Distributed computing has the opposite characteristics. It determines the rate at local nodes based on limited network information, but it uses less resources even though the determined rate may not reflect the complete picture of the network. Edge Pricing [8], [10] is an example of a pricing scheme that uses the distributed computing method.

The other sub-component of the Decision Making Component indicates who makes the final decision on the charging rate. It could be the User or the ISP or the Application. In the case that decision is made by the user, the charging scheme is more user-oriented while it could cause more transaction overhead between the user and the network. Other charging schemes let the ISP make the rate decision. This gives the ISP more flexibility to improve the network efficiency and implement other objectives, though the users have less control in the decision making process. It is also possible that decision making lies within the application, in which case the diverse traffic characteristics of different applications can be taken into consideration. This could improve the overall quality of the system, however, this would require new features to be added to current and legacy applications. As far as we know, there is no charging scheme that allows application to be the decision maker for charging rate. We consider this a possible direction for future development, and include it here for completeness.

3.2.4 Network Type Component

The third component at the highest level of the hierarchy is the Network Type Component, which is about what type of network and administration the charging scheme is developed for. From the technology point of view, the scheme could be suitable either for Best Effort (BE) network or for QoS-Guaranteed Performance (GP) network (including soft-QoS Guaranteed network). The Best Effort network is the original Internet. The Guaranteed Performance network is developed with Internet applications in mind. For some specific charging schemes, their characteristics are more compatible with certain network technology. The Responsive Pricing scheme for example, claims to be most suitable for ATM ABR traffic, but it may also be implemented in a Best Effort network.

From the administration point of view, a charging scheme could be developed for only a single network administrator, or multiple administrators. Though some of the charging schemes do not specifically deal with this issue, we can learn this factor through the scheme's complexity and the difficulties for multiple ISPs to collect statistics and to implement billing measurements.

3.3 Understanding Charging Schemes By Cataloguing

Based on the hierarchy of the charging components and using the lowest level parameters, a charging scheme can be represented as an 8-tuple:

Catalogue (charging-scheme) = <Time Scale, User Input, Charging Factor, Decision Time; Computing Method, Decision Maker; Technology, Administration>

This cataloguing process requires a detailed examination of the charging scheme under study, which enables a better understanding of the scheme with respect to the charging hierarchy and its components and sub-components.

Using the Smart-Market [8], [10], [17], [16] as an example to show the cataloguing process:

> "packets are sent at price zero when network is uncongested. However when network is congested, only packets which had a higher bid value will get through quickly. Users are not charged the price they bid, but rather are charged the bid of the highest priority packet that is not admitted to the network." [17]

According to the above information and other descriptions of the scheme, we can catalogue Smart-Market as:

Catalogue (Smart-Market) = <Short Time Scale, User Input Rate, Per Packet, Dynamic; Centralized Computing, Decision by ISP; Best Effort, Single ISP>

Other charging schemes introduced in Chapter 2 are catalogued as shown in Table 3.3.

Table 3.3. *Cataloguing of Charging Schemes*

Charging Scheme	Time Scale	User Input	Charging Factor	Decision Time	Computing Method	Decision Maker	Tech.	Admin.
Flat Rate	Long	Choice	Not Used	In Advance	Unknown	ISP	BE	Multiple ISPs
Paris Metro	Long	Choice	Not Used	In Advance	Unknown	ISP	BE	Multiple ISPs
Time -Based	Long	Choice	Time	In Advance	Distributed	ISP	BE	Multiple ISPs
Volume -Based	Long	Choice	Volume	In Advance	Distributed	ISP	BE	Multiple ISPs
Distance -Based	Long	Choice	Distance	In Advance	Distributed	ISP	BE	Multiple ISPs
Priority Pricing	Long	Choice	Priority	In Advance	Centralized	User	BE	Multiple ISPs
Smart -Market	Short	Rate	Packet	Dynamic	Centralized	ISP	Not Feasible	Single ISP
Edge Pricing	Medium	Choice	Various*	In Advance	Distributed	ISP	GP	Multiple ISPs
Responsive Pricing	Short	Choice	Resource	Dynamic	Centralized	ISP	GP	Single ISP
Proportional Fairness	Short	Rate	Resource	Dynamic	Centralized	ISP	GP and BE	Single ISP
Effective Bandwidth	Short	Choice	Traffic	Dynamic Characteristics	Distributed	ISP	GP	Single ISP
Expected Capacity	Long	Rate	Various*	In Advance	Distributed	ISP/User	GP	Multiple ISPs
Cumulus Pricing	Long	Rate	Various*	In Advance	Distributed	ISP	GP	Single ISP
INDEX Project	Long	Choice	Packet, Time	In Advance	Distributed	ISP	GP	Multiple ISPs

* The Edge Pricing, Expected Capacity and Cumulus Pricing schemes could be specified in various ways according to the user's usage pattern, such as usage time, minimum capacity required, maximum transfer time of data or an effective bandwidth-based traffic characterization.

3.4 Chapter Conclusion

In this chapter, we introduced a hierarchical structure to represent and catalogue Internet pricing schemes. There are several advantages in doing so as it helps us to:

- Understand the relevant issues involved in examining and analyzing a charging scheme.
- View the important aspects in evaluating and classifying a charging scheme.
- Know the prominent factors to consider in developing new charging schemes.
- Shape our objectives in developing new charging schemes.

With cataloguing, a clear and better understanding of charging schemes is attained in this chapter. In the next chapter, a further evaluation of charging schemes from different aspects will be shown by the introduction of an innovative evaluation cube.

Chapter 4

Evaluation and Classification of Charging Schemes

4.1 Introduction

Though a relatively new subject, Internet pricing research is multi-disciplinary and has taken technical, economic and social issues into consideration. As was mentioned in the introduction, Internet pricing has attracted more and more attention from engineers, economists and even regulators. With the proliferation of Internet applications for business, entertainment, communications, etc., it is imperative to have a more comprehensive understandings of the different aspects of Internet pricing schemes.

In this chapter, we will introduce an evaluation cube which includes eight dimensions representing various design and evaluation perspectives. Each of the dimensions and their relationships will be discussed in Section 2. Then, we will use the cube to examine and evaluate previously introduced charging schemes to get a better understanding of them. Finally, charging schemes are classified according to their characteristics in certain evaluation dimension in Section 3. As representative examples, classification of charging schemes based on the Economics and Technology Dimensions will be presented.

4.2 The Evaluation Cube

Through the survey in Chapter 2 and the cataloguing process in Chapter 3, we have concluded that there are eight specific dimensions with each representing some perspective in evaluating charging schemes. We introduce a cube structure consisting of these eight dimensions, as shown in Figure 4.1, as the evaluation criteria for charging schemes. Each dimension has its own objectives. Three planes of the cube can be identified with each plane consisting of four dimensions with similar or related objectives.

4.2.1 The Eight Dimensions

The eight dimensions are Individual, Practicality, Economics, Application, Admin-Network, Technology, Theory, Regulation. By taking a close look at each dimension and its objec-

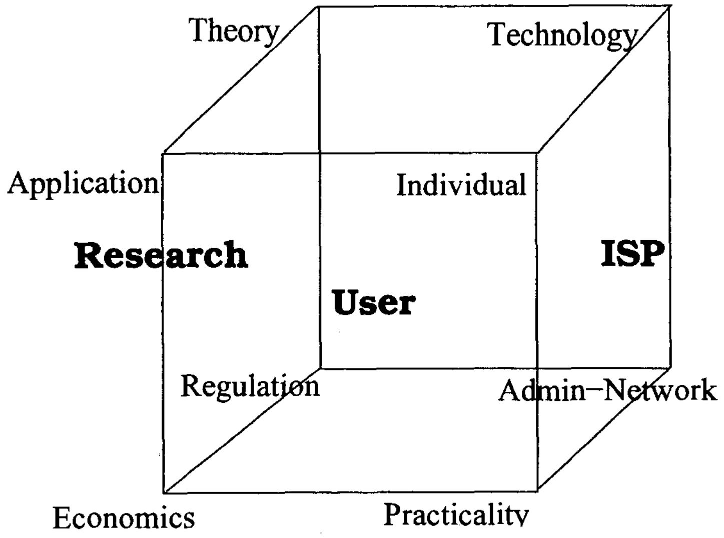

Figure 4.1. *The Evaluation Cube*

tives, we can get a broader, and at the same time deeper, understanding of the charging schemes.

For each charging scheme, it can be examined whether it has reached all, or only part of, the objectives associated with each dimension. For examples, for the Application Dimension, both Effective Bandwidth and Expected Capacity take application characteristics into consideration in determining the rate; both Flat Rate and Smart-Market consider the transmission from a packet point of view and no specific application requirement is considered; in Priority Pricing, users can set different priorities for their traffic according to the application's characteristics, however only relative priorities are implemented with no specific QoS guaranteed in the final resource allocation, therefore only part of the Application Dimension's objective is achieved.

4.2.1.1 Individual Dimension

The preference of each individual user is evaluated in this dimension. This includes the following:

- Simplicity of billing

- Stability of budget control
- QoS guaranteed service availability
- Flexibility of changing services and prices
- Nash Equilibrium: For each individual user, This is the situation where there is no better options which could be taken by the user under current circumstances. Each user is satisfied with his/her current situation.

4.2.1.2 Practicality Dimension

The Practicality and the Theory Dimensions are applied to examine from which point of view the charging scheme is designed, either implementation practicality or theoretical innovation.

The Practicality Dimension shows how practical and readily implementable the proposed charging scheme is. A good example is the INDEX project which provides detailed implementation of the network, user information, and the experimental results. As can be seen in [11], there is no innovative pricing scheme proposed. What it has done is implementing different combinations of charging schemes, and researchers can draw conclusions from the statistical analysis of the results.

4.2.1.3 Economics Dimension

The Economics Dimension shows how economic efficient a charging scheme is. Economic efficiency is a situation in which value (all users' utilities) is maximized. With the given resources, technology, and preferences, no changes will further increase utility beyond this efficiency value. Economic efficiency is also called Pareto optimality.

Economic efficiency has been emphasized in several proposed charging schemes. Economic efficiency is achieved in different ways in these charging schemes. In Smart-Market, economic efficiency is attained by setting the threshold value which represents the marginal cost of congestion, and only packets with bids exceeding the threshold value will be transmitted. In the case of Proportional Fairness, economic efficiency is reached by solving three optimization problems to get a price vector and a rate vector. If the users fix the traffic flow rates, the network is able to set the charging rate proportionally to the users' rates.

4.2.1.4 Application Dimension

The Application Dimension examines whether the charging scheme considers an application's characteristics. This is important for the coming QoS-guaranteed or -enabled networks.

As shown in Chapter 1, different applications have different characteristics, accordingly they also have different QoS requirements. By considering these characteristics, resources could be allocated more efficiently without degrading the total of qualities of services. An example of this is when the network is congested, network resources are allocated only to those applications which have more strict time constraints.

Older technologies, such as ATM, are designed to divide resources among different application groups. This also requires suitable application-based pricing schemes to be designed and implemented for these technologies.

4.2.1.5 Admin-Network Dimension

The Admin-Network Dimension takes the objectives of the ISP and evaluates a scheme from the network infrastructure point of view. In more detail, this dimension evaluates the charging scheme with the following criteria:

- Utilization efficiency of network resources
- Support for congestion control and traffic management
- Operating efficiency of the ISP
- Complete information on the network and users states
- ISP's revenue maximization

4.2.1.6 Technology Dimension

The Technology Dimension considers whether the pricing scheme complies with existing technology and how complex the scheme is. If technically infeasible, a charging scheme cannot be implemented even though the theory is very innovative. Feasible but complex charging schemes are also not desired by the ISP if there is a high volume of overhead for billing measurement, which makes the network more congested.

More specifically, the Technology Dimension considers the following objectives:

- Compliable with existing technology
- Compatible with current ISP administration infrastructure
- Low measurement and calculation requirements of billing and accounting

Same as the Admin-Network Dimension, the Technology Dimension also has close relation to the ISP. Nevertheless, they come from different perspectives. The main difference between these two dimensions is the Admin-Network Dimension focuses on improving network efficiency while the Technology Dimension centers on simpler and feasible technology.

4.2.1.7 Theory Dimension

As explained in the Practicality Dimension, the Theory Dimension is applied to examine whether the proposed charging scheme provides innovative theory or new idea to this research area. A pricing scheme designed with this dimension in mind is constructed specifically for research and it usually provides some heuristic methods and guidelines for further study, such as Smart-Market.

4.2.1.8 Regulation Dimension

The Regulation Dimension shows the degree of government intervention in order to maintain social fairness and other social and economic development objectives. These social goals would not be realizable without legislated regulations in a user-oriented market. A prime example is the de-regulation of the telecommunication industry in North America, to deter monopoly and to promote competition.

4.2.2 The Three Planes

Though the eight dimensions have different objectives, some of them have closer relationships with regard to some aspects. For example, the Individual Dimension considers user's preference while the Economics Dimension attempts to maximize the total utilities of users. Both of these two dimensions are user-related.

Based on the above principle, we can group the eight dimensions nicely into three higher evaluation dimensions or Planes, with each plane having four associated dimensions. These three planes are the User Plane, the ISP plane and the Research Plane.

Depending on which dimensions a pricing scheme is focused on, one can associate whether this scheme is user oriented, ISP oriented, or research oriented.

4.2.2.1 User Plane

The User Plane, at the front side of the cube in Figure 4.1, includes the Individual, Practicality, Economics and Application Dimensions.

From the objectives of each dimension, it is obvious that the Individual and the Economics Dimensions belong to the User Plane as they consider users' preferences and the total utilities of users. The Practicality Dimension is related to the users since users are not interested in theoretical research, and are more concerned with currently implementable charging scheme. The Application Dimension examines whether the charging scheme considers application characteristics. Without a fair consideration of applications' characteristics, the qualities of real time applications deteriorate seriously when the network is congested. This affects the users directly.

4.2.2.2 ISP Plane

The ISP Plane, at the right side of the cube in Figure 4.1, includes the Technology, Individual, Admin-Network, and Practicality Dimensions.

The Technology, Admin-Network and Practicality Dimensions are related to the network implementation or technology issues which are of concerns to the ISPs. In addition, each individual user has an explicit or implicit contractual agreement with the ISP.

In conclusion, the Technology, Admin-Network, Individual and Practicality Dimensions have close relationships with the ISP, and therefore they construct the ISP Plane.

4.2.2.3 Research Plane

The Theory, Regulation, Application and Economics Dimensions can be grouped into the Research Plane at the left side of the cube. The Research Plane focuses on further development and future directions of Internet pricing.

The Theory Dimension is based on the theoretical research point of view, so obviously it belongs to the Research Plane. In our survey of Internet pricing schemes, we noted that the Regulation, Application and Economics Dimensions are three prominent research directions of the charging schemes.

4.3 Applying the Evaluation Cube to Charging Schemes

Using the Evaluation Cube as a visual tool to consider the eight dimensions, one can evaluate a pricing scheme using the following function:

Evaluate (charging-scheme)= <Individual; Practicality; Economics; Application; Admin-Network; Technology; Theory; Regulation>

We shall apply this function to the Smart-Market charging scheme. Based on our Evaluation Cube, we may conclude the objectives of the Smart-Market charging scheme as follows: It is theory oriented as there are no implementation details; it is Admin-Network oriented since it does not differentiate traffic according to the application type; it is individual based as the user who pays more will get the services; it is economic efficient which is the intended objective of the charging scheme. Therefore,

Evaluate (Smart-Market) = <Individual; No Practicality; High Economic Efficiency; No Application; High Admin-Network Efficiency; Not Feasible; Theory; No Regulation>

A complete evaluation summary of the charging schemes is shown in Table 4.1. In the table, the evaluation results for Theory, Practicality, Individual and Regulation Dimensions are either "Yes" or "No", indicating whether the charging scheme has the corresponding objectives of the evaluating dimensions.

For the Application Dimension, the evaluation result could be "Yes", "Some", or "No". "Yes" or "No" represents whether the charging scheme considers the characteristics of each application. "Some" means there are some differentiations among the applications considered, but the quality of each application is not guaranteed. For example, users in Priority Pricing have the flexibility to set the priorities of their traffic. QoS stringent applications could be set to higher priorities. Although the priorities are considered, no specific bandwidth or resource is allocated to each application. Therefore, only "Some " of the application characteristics are considered.

For the Admin-Network Dimension, there are three evaluation results, "Low", "Some", or "High", representing the degree of network efficiency of the charging scheme.

There are several levels of technical feasibility for the Technology Dimension, "No Need", "Simple", "Medium", "Complex" and "Not Feasible". These show the complexity or infeasibility of the charging scheme. "No Need" means there is no billing measurement needed, no specific pricing mechanism implemented in the network. The particular technology the scheme is targeted for is also included in the table.

The Economics Dimension has five different evaluation results according to the economic efficiency of the charging scheme, "No", "Low", "Some", "Various" and "High". "Various" means the economic efficiency varies according to the detail implementation of the charging scheme.

According to the evaluation results, one can also learn if the charging scheme is user oriented, ISP oriented, or research oriented. For example, the evaluation results of Responsive Pricing for the four dimensions in the User Plane are: Individual (Yes), Practicality (Yes), Economics (High) and Application (Yes). It can be concluded that Responsive Pricing is a user oriented charging scheme.

4.4 Classification of Charging Schemes

Using the evaluation results, one could classify the charging schemes along any one of the dimensions. For example, charging schemes can be classified using the Economics Dimension according to their degree of meeting the economic efficiency objective. From Table 4.1, there are five categories of economic efficiency: No economic efficiency; Low economic efficiency; Some economic efficiency; High economic efficiency; and Various economic efficiency.

A classification on the Internet pricing schemes discussed in this thesis, by examining their characteristics in the Economics Dimension (degree of economic efficiency), is shown in Table 4.2. The characteristics shown in the middle column of Table 4.2 are shared by all the charging schemes which belong to the category. Within each category, we further divide the charging schemes into smaller groups according to their similarities, such as Congestion-Based or Credit-Based. For example, in the Credit-Based schemes, both Expected Capacity and Cumulus Pricing utilize the credit or policing mechanism to deal with user's usage instead of applying a real charge.

These charging schemes could also be classified using the Technology Dimension according to the scheme's technical feasibility and complexity, and compatibility with the network and administration infrastructure. Hence, charging schemes can be grouped into five categories: no need, simple, medium, complex, and not feasible. Classification of charging schemes according to their characteristics in the Technology Dimension (degree of Technical Feasibility) is shown in Table 4.3. Furthermore, within each category, the

Table 4.1. *Evaluation Summary of The Charging Schemes*

Charging Scheme	Individual -dual	Practi -cality	Econ -omics	Applic -ation	Admin -Network	Tech -nology	Theory	Regul -ation
Flat Rate	Yes	Yes	No	No	Low	No Need IP	No	No
Paris Metro	Yes	Yes	Low	No	Low	No Need IP	No	No
Time -Based	Yes	Yes	Some	No	Some	Simple IP & ATM	No	No
Volume -Based	Yes	Yes	Some	No	Some	Simple IP & ATM	No	No
Distance -Based	Yes	Yes	Some	No	Some	Simple IP & ATM	No	No
Priority Pricing	Yes	Yes	High	Some	High	Medium IP	No	No
Smart -Market	Yes	No	High	No	High	Not Feasible	Yes	No
Edge Pricing	Yes	Yes	Various	Yes	High	Medium ATM	No	No
Responsive Pricing	Yes	Yes	High	Yes	High	Complex ATM	No	No
Proportional Fairness	Yes	Yes	High	Some	High	Complex IP & ATM	No	No
Effective Bandwidth	Yes	Yes	High	Yes	High	Complex ATM	No	No
Expected Capacity	Yes	Yes	Various	Yes	High	Medium ATM	No	No
Cumulus Pricing	Yes	Yes	Various	Yes	High	Medium ATM	No	No
INDEX Project	Yes	Yes	Some	yes	High	Simple ISDN	No	No

Table 4.2. *An Economic Efficiency Based Classification of Charging Schemes*

Classification Based On Economics Dimension	**Characteristics**	**Charging Schemes**
No economic efficiency	No billing or measurement needed; Simple; No priority differentiation	Single Service Flat Rate
Low Economic efficiency	No billing or measurement needed; Simple; Some priority differentiation;	Paris Metro
Some economic efficiency	Some measure and billing needed No Priority differentiation	Time-Based
		Volume-Based
		Distance-Based
		INDEX
High economic efficiency	Need to measure and monitor user traffic; Complex; Differentiate users' priorities; Billing cost	Priority-Based
		Resource Based: Proportional Fairness
		Congestion-Based: Smart-Market; Responsive Pricing
		Application-Based Pricing
Variable economic efficiency	User will provide his/her own expected traffic load and characteristic; user will be awarded if he/she consumes within his/her own traffic bound, and penalized otherwise; Economic efficiency is achieved only with suitable pricing structure design and the cooperation of users	Credit-Based: Expected Capacity Pricing; Cumulus Pricing
		Rate-Based: Effective Bandwidth;

Table 4.3. *A Technical Feasibility Based Classification of Charging Schemes*

Classification Based On Technology Dimension	**Compatible Technology**	**Characteristics**	**Charging Schemes**
Technology Not Needed	IP	No billing measurement needed; No or low priority differentiation	Single service Flat Rate; Paris Metro Pricing
Technically simple	IP, ATM	Simple billing or measurement needed; No priority differentiation	Time-Based Volume-Based Distance-Based
Technically of medium complexity	IP	Some measure and billing needed; Priority differentiation	Priority Pricing
	ATM	User provide his/her expected traffic load and characteristic; User preferred medium/long time scale	Edge Pricing Expected Capacity Cumulus Pricing
Technically very complex	ATM	Need to monitor and measure user's traffic; Differentiate users' priorities; Extra billing cost	Effective Bandwidth; Responsive Pricing; Proportional Fairness
Technically infeasible	NA	Need to measure and monitor individual packet; Differentiate users' priorities; Extra billing cost	Smart-Market

charging schemes are divided into groups according to their compatible network technologies.

4.5 Chapter Conclusion

In this chapter, we introduced an evaluation cube with eight evaluation dimensions, which provides us with different aspects to examine Internet pricing schemes. Using the evaluation cube, a higher level and thorough understanding of the charging schemes can be reached. Furthermore, the evaluation cube can also guide our objectives when we design new charging schemes.

We also introduced a method of classifying charging schemes using the evaluation dimensions. Classification results according to the Economics and Technology Dimensions are shown. Of course, these charging schemes could also be classified using other evaluation objectives.

As we learn from this chapter, some of the charging schemes are more economic oriented, but are very complicated and even technically not feasible. Some other charging schemes are very simple, but their economic efficiency is not so satisfying. Thus, the ISPs have to consider all the related factors and make tradeoffs according to their objectives when they adopt a certain charging scheme.

In next chapter, we propose a new charging scheme which integrates some of the desirable features of the more complex charging schemes into a simple usage-based charging scheme.

Chapter 5

An Integrated Volume-Based Charging Scheme

5.1 Introduction

Chapter 2 to Chapter 4 introduce and examine charging parameters, and evaluate and classify Internet pricing schemes. It is noted that charging schemes are designed with certain objectives in some evaluation dimensions.

Many researchers have proposed charging schemes which try to integrate the advantages of both flat rate and complex usage-based schemes. For example, in the Expected Capacity case, by careful negotiation and pre-allocation of the network resources, both simplicity of billing and economic efficiency can be achieved. While for the INDEX project, users have the flexibility to choose a combined flat rate and usage-based charging when necessary. These integrated charging schemes can reach a well balanced tradeoff among the evaluation criteria and are expected to be the trend of future charging schemes.

Our conclusion from previous chapters is that the following three factors are the most important in Internet pricing and should be taken into consideration when designing a new charging scheme:

1. Distance: the longer the distance from source to destination, the more resources have to be allocated to serve the request. It is fair to both the users and the ISP if a user has to pay more for a longer distance transmission route. Moreover, network efficiency will be improved by making distance as a charging factor.
2. Application Type: different application types have different QoS requirements. For those applications which have higher QoS requirements, such as Video on Demand, more resources or exclusive use of resources are needed to meet their stringent delay and high bandwidth requirements. Also, as these applications with higher QoS requirements are more sensitive to network congestion, their qualities deteriorate easily and quickly as network congestion increases. Therefore, more strict congestion control mechanisms are needed for these applications.
3. Congestion: as it is shown in Chapter 2, recently proposed charging schemes have almost all taken this factor into consideration. This is due to the increase in resource

consuming applications and because it is not possible to keep on over engineering the network. Congestion control mechanisms could be provided by various means, including charging schemes. The underlying principle of a congestion control charging scheme is that with higher congestion along the transmission route, users should be charged higher. Under the assumption that some users will give up their service requests at higher charging rates, congestion control is reached.

Based on the above considerations, a charging scheme which is distance-based, application-based and has congestion control characteristic is desirable.

In this chapter, such a new scheme, the Integrated Volume-Based Charging Scheme, with the above characteristics, is introduced. This scheme integrates the advantages of complicated congestion-based charging into a simple volume-based charging scheme. Users see it as simple volume-based charging with the per volume rate determined by relevant factors. In addition, the scheme is also highly economic efficient and has traffic control capability. As was mentioned in Chapter 2, this new scheme is similar to Responsive Pricing from the congestion control perspective as both exploit the differences in users' nature: price-sensitive users will drop their requests when prices are higher. However, with Responsive Pricing, the charging rate is effective for a short time scale of seconds or minutes, while in our scheme, the charging rate is constant for each session which could last from seconds to hours. Using a longer charging time scale decreases the billing and measurement overhead dramatically.

The framework of the new charging scheme is introduced first in Section 2. The detailed design process of the charging rate formula is then shown in Section 3. Section 4 describes the charging scheme algorithm. Finally, the characteristics of the new charging scheme are listed in Section 5.

5.2 The Integrated Volume-Based Charging Model

The Integrated Volume-Based (IVB) Charging Scheme is developed from the general volume-based charging scheme. Like a Volume-Based Charging Scheme, the IVB counts traffic load in volumes that are sent to and/or received by the user, then multiplies it by the volume rate to get the final price of the service for the user. As an example:

A user wants to send a file to another site using ftp. Assume that the file size is 5MB, and the rate for each packet for this ftp session is \$0.1/MB, then the price for this session is:

$$\$0.1/\text{MB} * 5\ \text{MB} = \$\ 0.5$$

In a normal Volume-Based Charging Scheme, the charging rate per unit is always fixed, such as \$0.1/MB as shown in the above case. While in the IVB Scheme, the charging rate per volume of each session is calculated by a formula with the congestion status, destination address and application type as inputs. The charging rate is fixed for each individual session. Here, a session means a connection between a user agent/client and a server. For example, when a user browses a web site, it is the same connection or session when the user agent/client connects to the same server. If the user browses a different server, it is another connection or session. In this IVB scheme, the charging rate formula is the essential part, which is introduced in detail in the next section.

5.3 Charging Rate Formula of the Integrated Volume-Based Charging Scheme

The charging rate formula of the IVB scheme incorporates congestion status of the transmission route, distance and application type. The formula can be described as a function of three parameters:

Charging Rate = fn(distance, application type, congestion status of the transmission route)

5.3.1 Distance Parameters

Distance can be derived from the source and the destination IP addresses. The farther away the destination, the more resources are allocated for the service request. This means higher cost, and accordingly users should be charged with higher rates. Distance is an independent factor from the other two parameters and is represented as a charging coefficient "A1" and the propagation delay (the pure transmission time for a signal to reach the destination) coefficient "D" in the charging rate formula. The destination IP address can be used to derive both A1 and D. An example of this is shown in Table 5.1.

Table 5.1. *Destination IP Address and Distance Coefficients A1 and D*

Destination IP Address	A1	D
IP range_1	0.0002	0.05
IP range_2	0.0003	0.15
IP range_3	0.00035	0.20
IP range_4	0.0004	0.25
IP range_5	0.00045	0.3

The destination IP address ranges in Table 5.1 should be defined by the ISPs according to their network topology and their connection with other networks. The values of D can be obtained from experiments, such as the Henk Experiment [27]. Values of A1 are also defined by the ISPs according to their own preferences, such as the actual cost for the resources plus a certain percentage for profit margin.

5.3.2 Application Type Parameters

The IVB scheme also takes the application type of the service request into consideration. Application types can be viewed as three kinds:

- Time-insensitive service: such as email. Since these applications could be postponed till additional resources are available, they will not compete with time-sensitive or real-time services for the bottleneck resources in the network. In this case, existing technical solution (such as ATM) defines these services as Unspecified Bit Rate (UBR), which will only be delivered when there is additional capacity left after providing time-sensitive and real-time services. The charging rate for time-insensitive services should be fixed and should not take the congestion factor into consideration.

- Real-time service: such as Video Conferencing, Interactive Audio, Video On Demand. Since these applications have very stringent delay and jitter requirements, without some very strict congestion control mechanisms, their performances could degrade dramatically with very little extra traffic. It is considered that when the route is lightly loaded, the IVB scheme can work well. When the delay of the transmission route reaches some high threshold, it could mean that a congestion control charging scheme has not affected enough users to drop their service requests. In this situation, other congestion control mechanisms such as Admission Control and IntServ must be activated to ensure the overall quality of services.
- Time-sensitive service: such as FTP, Web Browsing and TELNET. These applications have higher time response requirements comparing to time-insensitive services, though they do not need real-time support. It is considered that the IVB scheme is most suitable for these applications with soft QoS requirements. Priorities can be set for these applications according to their requirement levels for delay and loss. Those applications with higher priority will be charged a higher rate. Furthermore, with an increase in the delay of the transmission route, their charging rates should increase more quickly than lower priority applications.

Applications with higher priority will be charged with higher rate per volume. This concept is represented by the coefficient “A2”. Services with higher QoS requirements are assigned with higher values of A2. Moreover, since applications with high QoS requirements are more sensitive to increase congestion along their routes, their charging rates should increase faster than less congestion-sensitive applications. This principle is represented by the coefficient “B”. This coefficient will take effect in conjunction with the congestion status parameter. Both A2 and B are set according to the application type. An example of this is shown in Table 5.2.

Table 5.2. *Application Type Coefficients A2 and B*

Application Type	A2	B
Email	0.00009	0
Browsing	0.00011	5
TELNET	0.00013	8
Voice over IP	0.00017	10
Video on Demand	0.00019	15

Since B governs the rate of increase, it reflects the application characteristics. For email service, B can be set to 0. As it is shown later, the charging rate formula for email service is fixed with no congestion status factor taken into account. As Video on Demand requires real-time support, its B value is set to the highest. Furthermore, the congestion control mechanism can be a combination of the proposed charging scheme and technical solutions such as Admission Control and RSVP.

5.3.3 Congestion Status Parameter

The congestion status of the transmission route is reflected in the round trip delay of some probing packets sent at the start of each session. There are two assumptions: First, the data packets of the session will be delivered using the same route as the probing packets; second, the congestion status is consistent throughout the session.

When the network gets more congested, the probing packets will be delayed along the route and their round trip delay will increase. The measured round trip delay on the transmission route is represented by the coefficient "C".

There are several issues which need to be further considered:

- As the proposed charging scheme is to have congestion control characteristic, it is required that the charging rate be increased with congestion on the transmission route.

Under the assumption that the number of users who will drop their service requests is proportional to the net increase of the charging rate, an exponentially increasing charging rate would mean the rate increases more quickly with a more congested transmission route, consequently more users will drop their service requests at a speedier pace. An exponentially increasing charging rate with respect to congestion is one of our design objectives.

- The coefficient "C" is the round trip delay of probing packets which not only includes the delay caused by congestion along the route but also the pure propagation delay which is not congestion related. Assuming processing delay at routers is negligible and using the propagation delay D, the real congestion factor is (C-D).
- Different applications also have different rate of increase with congestion along the route. This is represented by the application type coefficient B as explained in Section 5.3.2. In order to represent congestion along the route and an application's sensitivity to congestion, B*(C-D) is used as the exponentially increasing factor of the charging rate. The multiplication of B and (C-D) resolves any unit incompatibility issues.

5.3.4 The Charging Rate Formula

Based on the above analysis of each relevant charging parameter, the charging rate can be formulated as:

$$ChargingRate = A1 + A2 * e^{(B*(C-D))}$$

where

- A1: represents the charging coefficient for distance, derived from the destination IP address.
- A2: represents the charging coefficient for application, defined by the application type.
- B: represents the rate increase coefficient of the application.

- C: represents the round trip delay of the probing packet on the transmission route; C integrates both the delay caused by distance and the delay caused by congestion along the route.
- D: represents the propagation delay, derived from the destination IP address or by measurement.

The units for coefficients A1 and A2 are \$/Byte, B's unit is $(second)^{-1}$, and coefficients C and D use second as unit.

There are some special cases:

1. When A1 = 0, the charging rate is only application and congestion related; distance is not considered.
2. When A2 = 0, the charging rate is A1, and IVB becomes a distance-based charging scheme.
3. When B = 0, the charging rate is A1 + A2. Only distance and application type are considered, but not congestion status. This is used in situations where the transmission route is congested, and the application such as emails can be delayed for later services.
4. When C-D = 0, the round trip delay of the probing packets on the transmission route is caused by pure propagation delay only, and there is no congestion. The charging rate becomes A1 + A2. This is the same as case 3 above with congestion having no effect on the rate. In this case, it is because there is no congestion along the transmission route, while in case 3, it is caused by the application characteristic.

The above analysis shows that the proposed charging schemes has the desirable property that charging factors can be adjusted depending on one's objectives.

5.4 The Integrated Volume-Based Charging Scheme Algorithm

There are some assumptions to be made before the working procedure of the Integrated Volume-Based Charging Scheme is introduced:

- The values of the coefficients A1, A2, B and D are known before the session starts. They are stored in look-up tables similar to Table 5.1 and 5.2. Charging rate is calculated according to the formula once the round trip delay of the probing packets on the transmission route is known.
- The round trip delay of probing packets on the transmission route is measured by sending probing packets before the session starts. Each of the two ends has agreed to send the probing packets back to the source immediately.
- It is suggested that three probing packets are to be sent to measure the delay along the route. As congestion could cause packets to be dropped, it is not safe and accurate enough with only one probing packet. In the case that two of the three probing packets are dropped, it would be better to delay the session. The round trip delay of the route can be estimated by taking the average delay of two or all of the probing packets.

There are various ways to implement the IVB charging scheme. The above assumptions and methods illustrate one possibility. Accordingly, the working procedure of the IVB charging scheme suggested is as follows:

1. A user starts an application with a certain destination IP address.
2. The ISP sends out three probing packets which are timestamped to measure the round trip delay between the source and destination. After two or all three of the probing packets returned, an average round trip delay on the transmission route can be estimated.

3. The ISP calculates the charging rate based on the application type, the round trip delay, and the destination IP address.
4. The ISP provides the charging rate to the user for decision making. (As mentioned above, there are various ways to do this; one example is to have the users declare an upper bound of acceptable charging rates in advance, the interaction with users and thus the overheads could be avoided).
5. The user decides whether he/she will accept the charging rate:
    - If acceptable, the ISP will start the session, calculate the number of sending and/or received packets, and charge the user accordingly.
    - If unacceptable, the user can try again immediately or later.

5.5 Characteristics of the Integrated Volume-Based Charging Scheme

As the name implies, the Integrated Volume-Based Charging Scheme is not only volume-based, but also integrates the characteristics of distance-based, application-based, congestion-based, individual-based, session-based, and route-based charging schemes. Each of these characteristics is analyzed below. Some of these characteristics need further quantitative analysis which will be done in the next chapter:

1. Distance-based: With the knowledge of the destination IP address, distance as a charging factor is integrated into the proposed charging scheme. Further quantitative analysis will show that an increase in distance causes an increase in the charging rate.
2. Application-based: The IVB scheme takes application type into consideration. Different application types have different coefficients in the charging rate formula depending on their QoS sensitivity. Higher quality requirement means more resource are needed, and the charging rate should increase accordingly. Furthermore, the increasing rate of the charging rate with increased congestion on the transmission

route, for applications with higher QoS requirements, should be faster than applications with lower QoS requirements. It is assumed that the number of users who will drop their service requests increases linearly with charging rate. As more users whose applications require higher QoS requirements drop their requests, the service qualities will not deteriorated for those remaining applications whose users are willing to pay a higher rate.

3. Congestion-based: The IVB charging rate is affected by the round trip delay of probing packets on the transmission route. As congestion along the route causes packets to be delayed by queueing in the routers, the charging rate should change according to the congestion of the route. It will be shown that the charging rate increases with congestion, and the increasing rate of the charging rate is exponential, thus congestion control is achieved.
4. Individual-based: In this charging scheme, each user has his/her own individual billing measurement, and others' activities will not be interfered.
5. Session-based: The determined charging rate applies to all the bidirectional traffic of the same session.
6. Route-based: Using the round trip delay of the probing packets for each session with the same destination IP address, and assuming the same transmission route for the session, the rate for the particular route is established.

Characteristics 1, 2, and 3 will be shown in the quantitative analysis in next chapter.

5.6 Chapter Conclusion

In this chapter, the Integrated Volume-Based Charging Scheme is introduced. Based on our objectives, charging factors are represented by coefficients in a charging rate formula to achieve the desired effects the charging scheme is designed for. Working procedure of the scheme is described to show some implementation details. The IVB charging scheme is shown to be volume-based, individual-based, distance-based, application-based, session-based and route-based. The proposed charging scheme has a truly integrated nature which is one of the design objectives.

In next chapter, further quantitative analysis is performed to examine whether the proposed charging scheme meets our expectations. A qualitative analysis is also done to examine the scheme with respect to each evaluation dimension.

Chapter 6

A Quantitative and Qualitative Analysis of The Integrated Volume-Based Charging Scheme

6.1 Introduction

As mentioned in last chapter, distance-based, application-based and congestion-based characteristics are further examined in this chapter. Quantitative analysis is presented in Section 2 to verify whether the introduced algorithm and charging formula of the proposed Integrated Volume-Based Charging Scheme indeed meet the expected goals. Qualitative analysis is performed in Section 3 using the eight evaluation dimensions introduced in Chapter 4. Advantages and disadvantages of the proposed charging scheme are explored to gain a complete understanding of the scheme.

6.2 Quantitative Analysis

The quantitative analysis code was written in MATLAB (see Appendix A) for the charging rate formula proposed in last chapter:

$$R = A1 + A2 * e^{(B*(C-D))}$$

where:

- A1: represents the charging coefficient for distance, derived from the destination IP address
- A2: represents the charging coefficient for application, defined by the application type
- B: represents the rate increase coefficient of the application
- C: represents the round trip delay of the probing packet on the transmission route; C integrates both the delay caused by distance and delay caused by congestion along the route
- D: represents the propagation delay, derived from the destination IP address or by measurement

6.2.1 Objectives of the Quantitative Analysis

Based on the discussion of IVB's characteristics in last chapter, the objectives of the quantitative analysis are as follows:

1. Different distances (by changing coefficients A1 and D): This is to show that an increase in distance causes an increase in the charging rate. Since more resources are allocated to a service request of longer distance, accordingly the rate should be higher than a service of shorter distance. This is reflected in changing coefficient A1. The other distance coefficient D represents the propagation delay, which will be used to derive congestion delay by deducting it from C of the probing packets, that is (C-D).
2. Different applications (by changing coefficients A2 and B): It is needed to show an increase in transmission quality (according to the application type) increases rate. Assuming that applications with higher QoS requirements are allocated more resources, it is fair for the ISP to charge more for applications with higher QoS requirements. This is reflected in changing coefficient A2. Moreover, the charging rate of applications with higher QoS requirements should increase faster than the rate of applications with lower QoS requirements. This is reflected in changing coefficient B.
3. Different congestion status of the route (by changing coefficient C): It is necessary to show that with the increase of coefficient C, the charging rate should also increase so that the congestion control objective is achieved. The assumption here is that fewer users will use the service as price increases.

In this study, each of the five coefficients is varied to verify their effects on the charging rate. Furthermore, the charging rate in three situations with two changing parameters are examined: various round-trip delay and different applications; various round-trip delay and different distances, and various distance and different applications.

Table 6.1. *Typical Values of D*

Route	D
Henk Exp.	0.04
Intra Europe	0.13
Trans-Atlantic	0.17
Trans-Pacific	0.26
Europe to Asia	0.57

6.2.2 Typical Coefficient Values Used in the Study

Typical values of A1, A2, B, C and D are either valid assumptions or taken from open literature, in order to make our study realistic.

1. Distance Coefficients A1 and D:

   Values used for D are taken from reference [1] and the Henk Experiment [27]. In [1], the typical average propagation delay for Intra Europe, Trans-Atlantic, Trans-Pacific and Europe to Asia are provided. The values of D for these routes could be a range of values or certain typical values which are used here for illustration purpose without loss of generality. The Henk Experiment examined end-to-end performance on the Internet. The transmission route is from Bratislava, Slovakia to Munich, German, with an end-to-end delay of about 20 msec. We use a round trip delay of 40 msec assuming no significant delay at the destination. Typical values of D are shown in Table 6.1.

   The values of A1 are set using the result of the Henk Experiment as the reference point with a 40 msec round trip delay. The corresponding value of A1 is assumed to be \$0.00002 per packet which is about 7% of current volume-based charging rate, such as the \$0.0003/packet referred from [26]. The 7% is an assumed number since we intend to assign a comparatively small weight to the distance factor.

   The coefficient A1 is set to increase linearly with the increase of propagation delay D. This is based on the assumption that resource allocation increases linearly with dis-

Table 6.2. *Typical Values of A1*

Route	A1
Henk Exp.	0.00002
Intra Europe	0.00003125
Trans-Atlantic	0.00003625
Trans-Pacific	0.00004275
Europe to Asia	0.00008625

Table 6.3. *Typical Values of A2*

Application	A2
Email	0.00009
Web Browser	0.00011
TELNET	0.00013
Voice over IP	0.00017
Video on Demand	0.00019

tance, represented by the propagation delay D here. For other routes, A1 is assigned with an increase rate of $0.000005/40 msec as shown below:

$$A1(D) = 0.00002 + 0.000005 * \frac{D-0.04}{0.04}$$

The assumed values of A1 for each route are shown in Table 6.2.

2. Application Coefficients A2 and B

   The values of A2 are set such that applications with higher QoS requirements are charged higher as more resources are needed to ensure their QoS. Typical values of A2 used in the study are shown in Table 6.3.

   The coefficient B is set with the principle that more time stringent applications are assigned higher values of B. When the transmission route gets congested, higher values of B will cause the charging rate to increase more quickly as a multiplying factor into congestion status (C-D). Accordingly, more users of time stringent applications may drop their traffic requests when facing higher charging rate. Thus, the quality

Table 6.4. *Typical Values of B*

Application	B
Email	0
Web Browser	5
TELNET	8
Voice over IP	10
Video on Demand	15

of service of the remaining applications will not deteriorated severely, but this is at the expense of losing price-sensitive users. The values of B assumed in our study is shown in Table 6.4.

The value of B for email application is set to be zero. In this case, the coefficients C and D no longer have an effect on the charging rate:

$$R = A1 + A2 * e^{(0*(C - D))}$$

$$R = A1 + A2$$

Email application is time insensitive and therefore it could be postponed for later transmission. This postponement avoids the competition with QoS-stringent applications for bottleneck resources when the route gets congested.

3. Congestion Coefficient C

The congestion status along the transmission route is represented by C, the measured round trip delay of the probing packets on the transmission route. In this study, the values of C are assumed numbers which vary for different cases, and depend on the values of the other coefficients.

6.2.3 Effects of the Coefficients on the Charging Rate

There are five coefficients, A1, A2, B, C, and D, in the charging rate formula.

$$ChargingRate = A1 + A2 * e^{(B*(C-D))}$$

Table 6.5. *Typical Values of Coefficients*

Coefficient	Example	Typical Value	Range of Change
A1	Henk Exp.	0.00002	0.00002 - 0.00008625
A2	Web Browse	0.00011	0.00009, 0.00011, 0.00013, 0.00017, 0.00019
B	Web Browse	5	0, 5, 8, 10, 15
C	Henk Exp.	0.2	0.04 - 0.4
D	Henk Exp.	0.04	0.04 - 0.4

In this section, each of the five coefficients' effects on the charging rate are examined by using their typical values, while keeping the other coefficients fixed. The results are analyzed against expectations. It is worthwhile to point out that in real situations, it cannot be only one coefficient is changing while the others are kept to be fixed. For example, coefficients D and A1 should increase at the same time since A1 is a function of D as shown in Section 6.2.2. Typical values used for the coefficient are shown in Table 6.5.

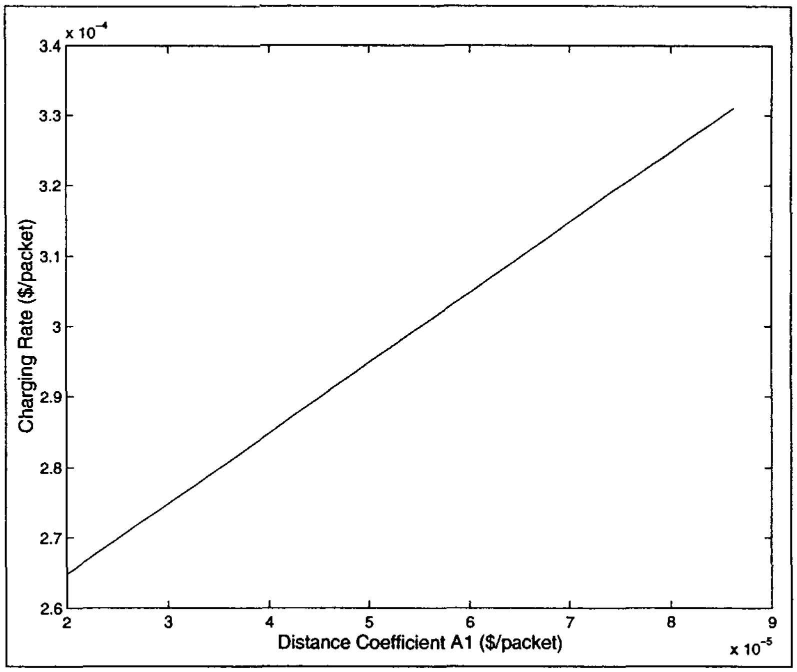

Figure 6.1. *Effect of Distance Coefficient A1 on Charging Rate*

1. Distance Coefficient A1:

The value used for A1 in Figure 6.1 is from $0.00002/packet to $0.00008625/packet, and coefficients A2, B, C and D are fixed at their typical values as shown in Table 6.5.

Under the assumption that resource allocation is linearly increasing with distance, it is designed that the charging rate increases linearly with distance coefficient A1 as shown in Figure 6.1. This is compatible with one of the objectives of the IVB charging scheme: It is distance-based and the charging rate increases linearly with distance.

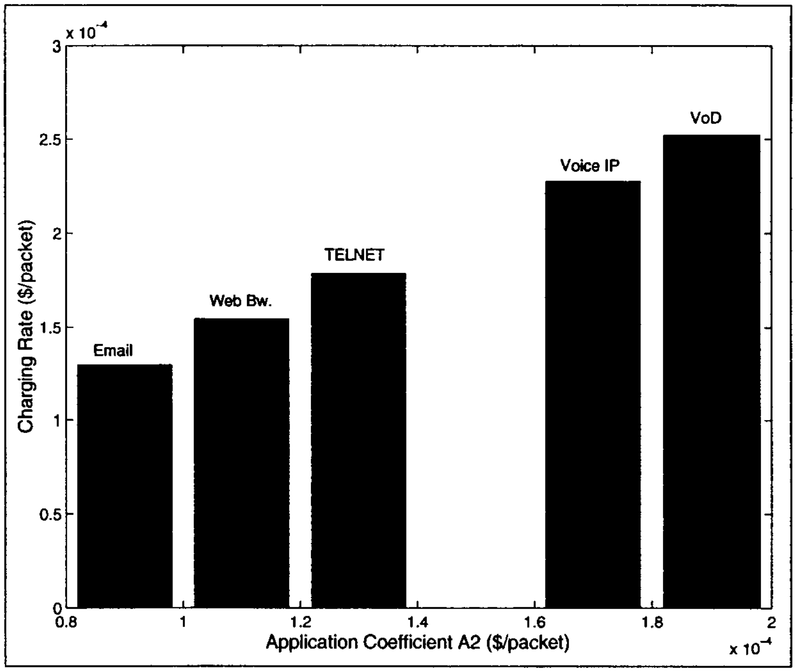

Figure 6.2. *Effect of Application Coefficient A2 on Charging Rate*

2. Application Coefficient A2:

The values used for A2 range from $0.00009/packet to $0.00019/packet according to the application type. Typical values of other coefficients A1, B, C and D used are as shown in Table 6.5.

As applications with higher QoS requirements require more resources, they should be charged with a higher rate. As shown in Figure 6.2, applications with various QoS requirements have different charging rates. The result is compatible with the design objectives. Through coefficient A2 and B, application-based charging is achieved.

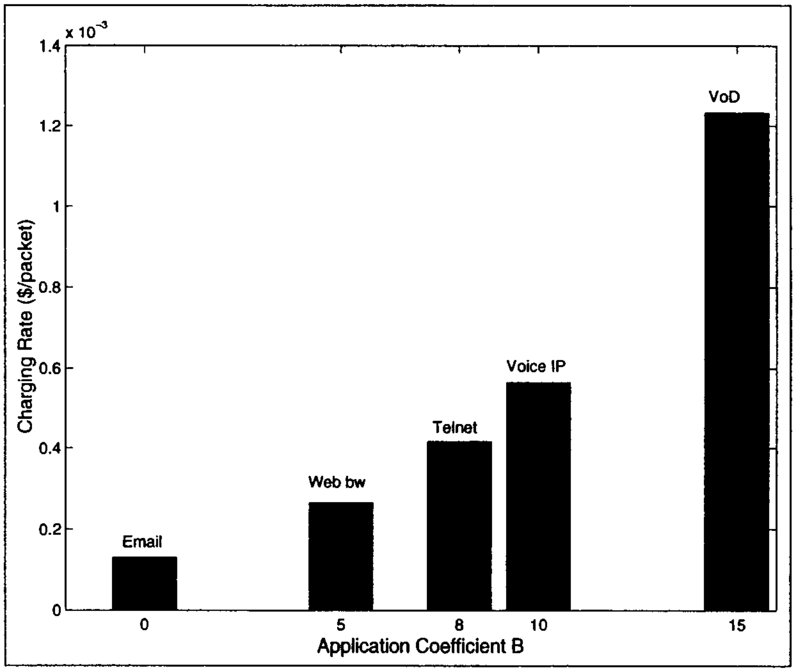

Figure 6.3. *Effect of Application Coefficient B on Charging Rate*

3. Application Coefficient B:

The values used for B are from 0 to 15 depending on the application type. Typical values of other coefficients A1, A2, C and D are fixed as shown in Table 6.5.

Coefficient B is essential in the charging scheme as it represents application characteristic. Congestion control is reflected in the use of B in the charging rate formula. Increasing values of B represent applications with higher QoS requirements. As shown in Figure 6.3, the charging rate increases exponentially with applications of higher QoS requirements as designed in the algorithm. With the assumption that more users will give up service requests at higher charging rates, the objective of congestion control is achieved.

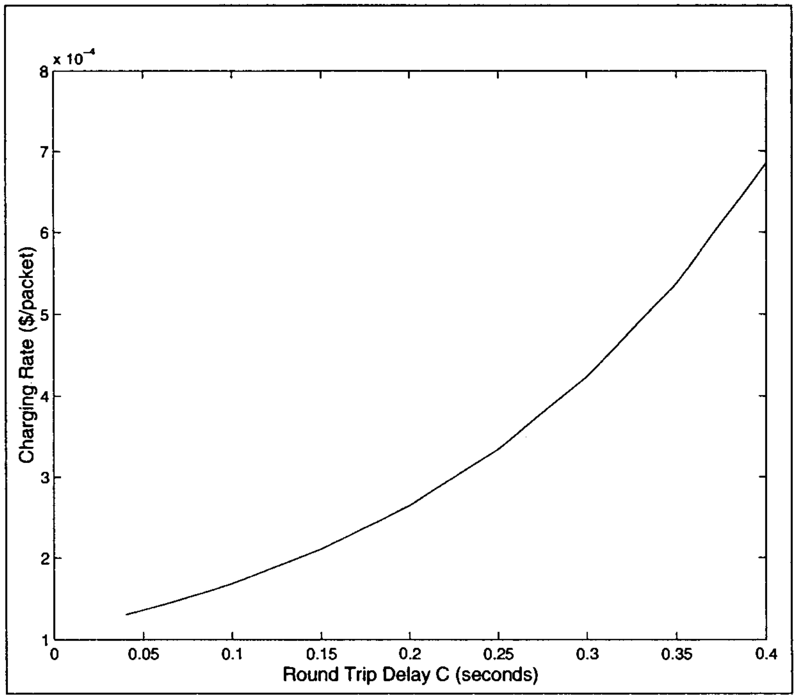

Figure 6.4. *Effect of Round Trip Delay C on Charging Rate*

4. Congestion Coefficient C:

   The values used for C in Figure 6.4 range from 0.04 to 0.4 seconds. Values of other coefficients A1, A2, B, and D, are fixed as shown in Table 6.5. When C is equal to 0.04, the value of the propagation delay D used here, there is no congestion on the transmission route. The round trip delay is completely caused by the propagation delay D.

   With the assumption that traffic requests decrease linearly with the increase of charging rate, the rate should increases exponentially with congestion so that more users will drop their requests. Consequently, it is designed that congestion control has more effect when the route gets more congested. The result of this study is shown in Figure 6.4. As expected, the charging rate increases exponentially with the round trip delay that reflects the congestion status of the route.

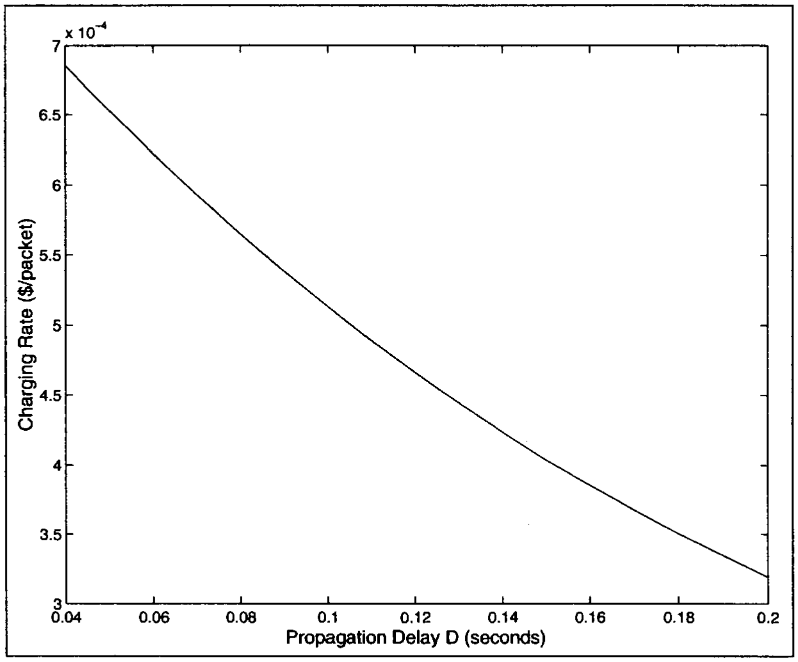

Figure 6.5. *Effect of Propagation Delay D on Charging Rate*

5. Propagation Delay Coefficient D:

   The value of coefficient D ranges from 0.04 to 0.2 seconds. The other coefficients A1, A2, B and C are kept constant at the values shown in Table 6.5. The upper bound of D (0.2 seconds) is the same as the value of C. This means that the round trip delay on the transmission route is completely caused by propagation delay. There is no congestion on the transmission route.

   The effect of increasing D is just the opposite of increasing C in the previous case. When D is increasing, it means that the round trip delay of the transmission route caused by propagation delay (i.e., distance) is increasing, and the delay caused by route congestion is decreasing. The charging rate is designed to decrease as congestion along the route decreases. As expected, the charging rate decreases with propagation delay D, as shown in Figure 6.5.

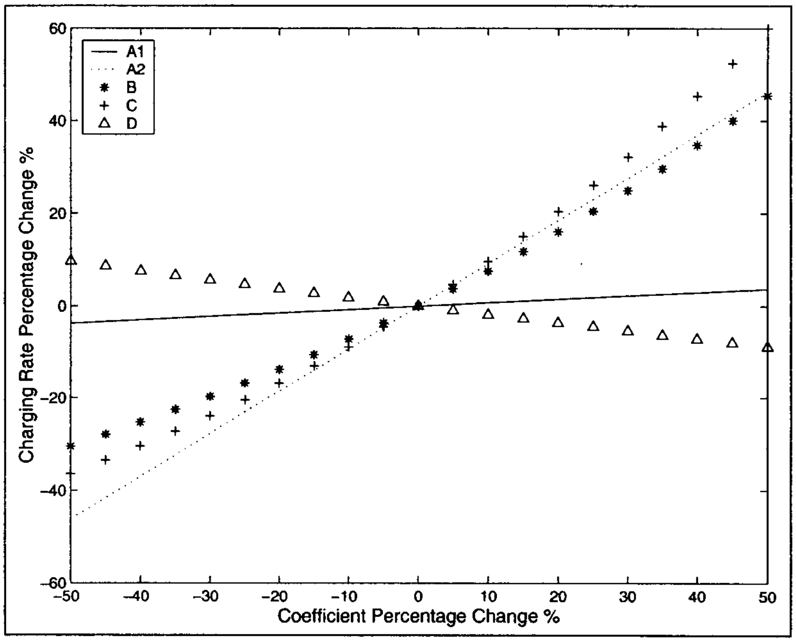

Figure 6.6. *A Rate Sensitivity Study*

6. Rate Sensitivity

As learned from the above analysis, an increase of A1, A2, B or C will increase the charging rate, while an increase of D will decrease charging rate. In order to have a clear comparison of the coefficients' effects on the charging rate, each coefficient is changed within ±50% of its typical value, and their effects on the rate change in percentage are examined as shown in Figure 6.6.

The IVB scheme is designed to have congestion control and application characteristics at higher priorities. This is clearly shown in the graph that the rate is much more sensitive to the change of coefficients A2, B, and C which are congestion or application related, as compared to coefficients A1 and D which are distance related. It should be noted that, the values of some of these coefficients could be changed according to each individual ISP's network structure and business strategy. This is an advantage of the proposed scheme, that it could be easily adjusted according to different situations.

6.2.4 Round Trip Delay versus Rate for Different Applications

In this section, how the charging scheme works in changing congestion status for different applications is examined.

Table 6.6. *Application Types and Their Coefficients A2 and B*

Application	A2	B
Email	0.00009	0
Web Browse	0.00011	5
TELNET	0.00013	8
Voice over IP	0.00017	10
Video on Demand	0.00019	15

There are three coefficients involved in this study: A2 and B that are associated with each application as shown in Table 6.6, and round trip delay coefficient C which ranges from 0.04 to 0.2 seconds. Distance coefficients A1 and D are kept fixed at $0.00002/packet and 0.04 seconds, respectively, as obtained in the Henk experiment [27].

The characteristics of application-based and congestion-based charging are shown in this study:

- The congestion-based characteristic is reflected by the congestion control mechanism (i.e., the rate) and it is stricter (i.e., higher rates) when the transmission route gets more congested. As it is shown in Figure 6.7, charging rates increase exponentially with the round trip delay (i.e., congestion) of the transmission route, for all applications. Under the assumption that the number of users will give up their requests increases linearly with the increase of the charging rate, a congestion control mechanism is evident in the IVB scheme.
- The application-based characteristic is reflected by the fact that applications with higher QoS requirements have more stringent congestion control comparing to applications with lower QoS requirements. This is because the quality of these high

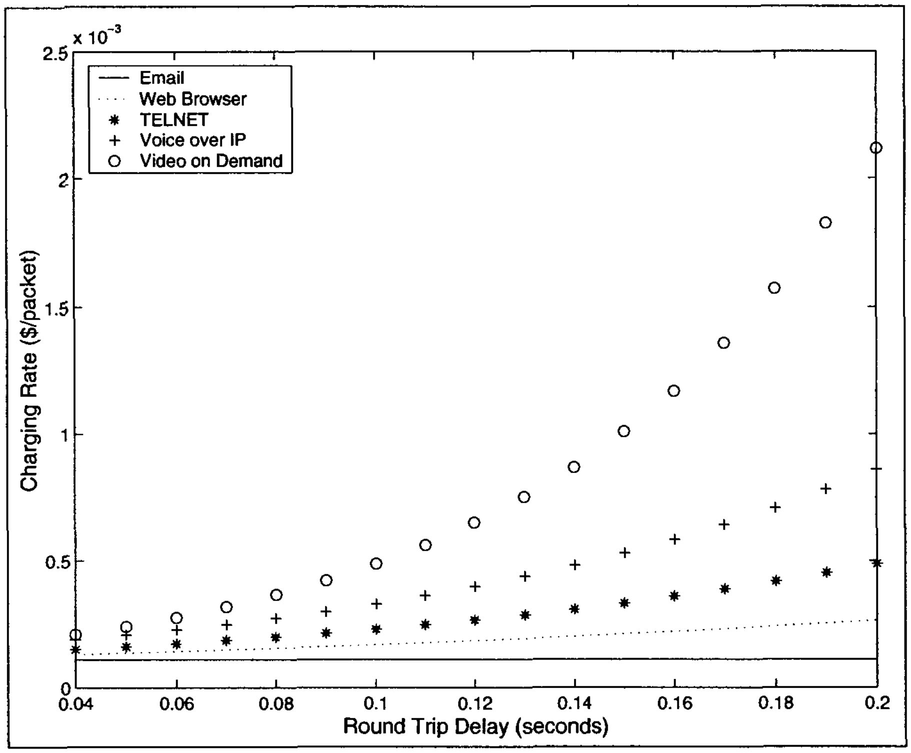

Figure 6.7. *Round Trip Delay versus Rate for Different Applications*

QoS applications can easily deteriorate when the transmission route is congested. As shown in Figure 6.7, charging rates for those QoS stringent applications increase more dramatically than those applications with no or low QoS requirements. With the assumption that the number of users give up their service requirements increases linearly with the increase of charging rate, more users of high QoS applications will give up their service requests when the round trip delay of the transmission route increases indicating congestion. This is compatible with the design objective that congestion control is stricter for applications with high QoS requirements.

Comparing the above results with the quantitative analysis objectives stated in Section 6.2.1, both 2 and 3 are verified in this case.

6.2.5 Round Trip Delay Versus Rate for Various Distances

Table 6.7. *Distance Coefficients A1 and D with the Corresponding Range of Round Trip Delay*

Location	**A1**	**D**	**C**
Henk Exp.	0.00002	0.04	0.04 - 0.4
Intra Europe	0.00003125	0.13	0.13 - 0.6
Trans-Atlantic	0.00003625	0.17	0.17 - 0.7
Trans-Pacific	0.00004275	0.26	0.26 - 0.8
Europe to Asia	0.00008625	0.57	0.57 - 1.0

This section shows how the IVB scheme works in changing congestion status for the same application as the destination (distance) changes. The distance factor is examined by changing distance coefficients A1 and D, as shown in Table 6.7. The round trip delay coefficient C takes different ranges which use the Henk Experiment as the reference point, and with the values of D considered. For example, from Europe to Asia, the distance D is 0.57, and therefore the range of C is 0.57 to 1.0. The application coefficients A2 and B are kept fixed at $0.00011/packet and 5, respectively.

The quantitative result of charging rate versus round trip delay for various distances of the same application is shown in Figure 6.8. The following conclusions can be drawn:

- The distance factors are incorporated as the longer distance routes have higher charging rates than the short distance routes. This is expected from our design objective (item 1 in Section 6.2.1).
- With the same round trip delay, a shorter distance route, such as "Intra Europe", has a higher charging rate than a longer distance route such as "Europe to Asia". Given the same round trip delay, a shorter distance has a higher delay that is caused by congestion, as compared to a longer distance route where the delay is contributed mostly from the propagation delay. This agrees with our design objective that congestion control is a more important factor than distance in the charging rate determination.

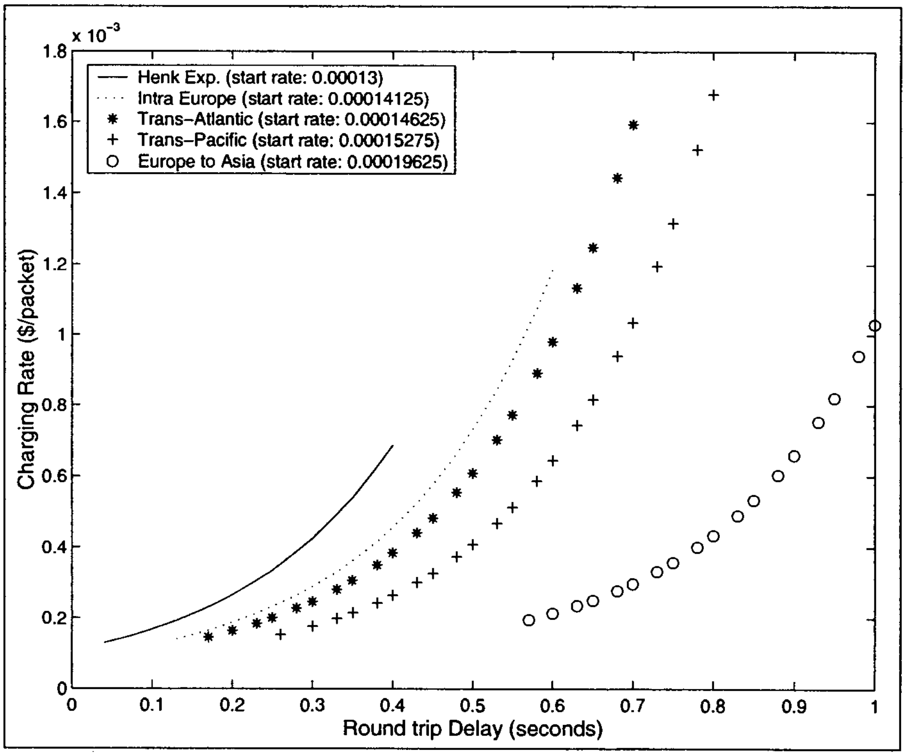

Figure 6.8. *Round Trip Delay versus Rate for Various Distances*

6.2.6 Distance versus Rate for Different Applications

In this section, how the charging rate behaves with varying distance for different applications is examined.

Table 6.8. *Application Types and Their Coefficients A2 and B*

Application	A2	B
Email	0.00009	0
Web Browser	0.00011	5
TELNET	0.00013	8
Voice over IP	0.00017	10
Video on Demand	0.00019	15

There are four coefficients involved in this study: A2 and B that associated with each application type as shown in Table 6.8, and propagation delay coefficient D which ranges from 0.04 to 0.57 seconds. As for the other distance coefficient A1, it is calculated using the formula below:

$$A1(D) = 0.00002 + 0.000005 * \frac{D-0.04}{0.04}$$

As the study is focusing on the effects of different applications and various distances on the charging rate, the congestion effect is neglected here. It is assumed that congestion status (C-D) is fixed at 0.05 seconds.

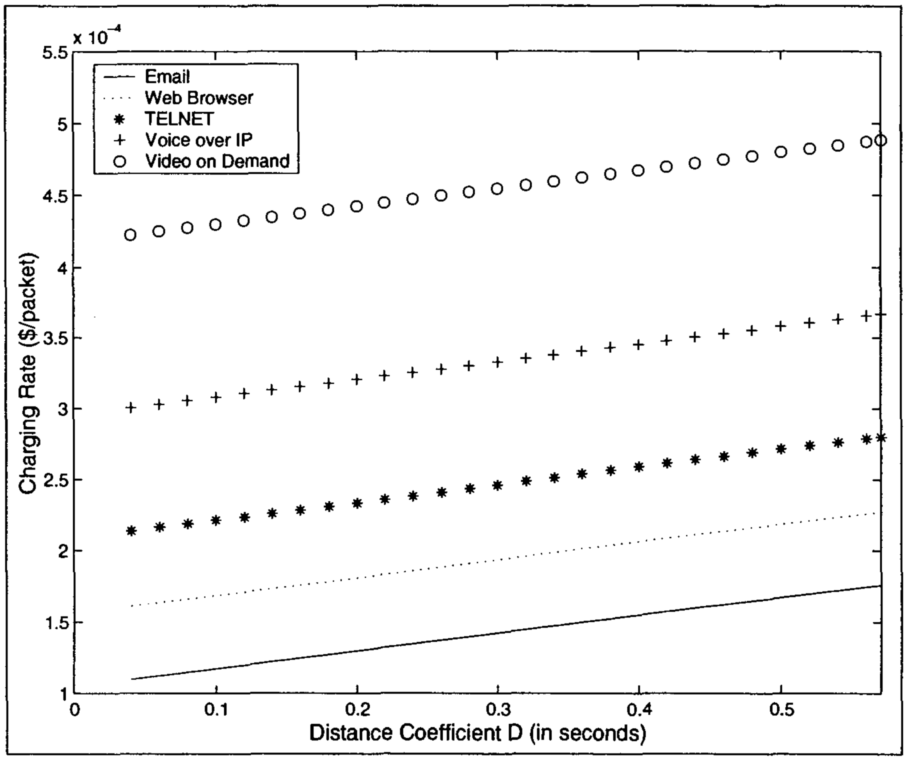

Figure 6.9. *Distance versus Rate for Different Applications*

The characteristics of application-based and distance-based charging are shown in this study:

- The distance factor is reflected as an upward sloping rate for all applications as shown in Figure 6.9. With the assumption that resources allocated for the same request increase linearly with the distance, the charging rate is designed also to increase linearly with the distance to offset the cost.

- The application-based characteristic is reflected by having different applications charge differently. Since applications with higher QoS requirements have higher resource usage requirements, the charging rate is designed such that applications with higher QoS requirements are charged higher than applications with lower QoS requirements.

6.2.7 A Summary of the Quantitative Analysis

In this quantitative analysis of the proposed Integrated Volume-Based Charging Scheme, the effects of each coefficient on the charging rate are examined. More complicated cases with multiple changing coefficients are investigated to get a deeper understanding and to verify the objectives of the scheme. The conclusions of the quantitative analysis with respect to our design objectives are:

- More resources are needed for longer distance traffic. The charging rate is designed to increase linearly with the distance coefficient A1 as shown in Figure 6.1.
- More resources are allocated for applications with higher QoS requirements, i.e., a higher A2 coefficient. These applications are charged higher than applications with lower QoS requirements as shown in Figure 6.2.
- The coefficient B is used to differentiate applications in a way that applications with higher QoS requirements will be assigned a higher B. With the exponential function in the charging formula, charging rates for those applications with higher QoS requirements increase more quickly when facing congestion along the transmission route as shown in Figure 6.3. This mechanism is utilized to control congestion as it is expected that more users will give up their service requests at higher charging rates. As a result, quality can be maintained for the remaining applications.
- Increasing the round trip delay coefficient C while keeping the other coefficients fixed means congestion along the route is getting worse. It is a design objective to have

the charging rate increase exponentially so that congestion control is in effect when it occurs. This result is shown in Figure 6.4.

- Increasing the propagation delay coefficient D while keeping the other coefficients fixed means there is less congestion along the route. The charging rate is expected to decrease as D increases as shown in Figure 6.5.
- From Figure 6.6, it is learned that among the five coefficients, round trip delay C has the most impact on the rate while distance coefficient A1 has the least impact. This agrees with our design objective that congestion control has the highest priority; while pure distance is at the lowest.
- The charging rate for an application should increase exponentially with the round trip delay C. Moreover, applications with higher QoS requirements consume more resources and therefore they cannot tolerate severe congestion along the route. As shown in Figure 6.7, these applications are charged higher (as indicated by A2) and their rates increase more quickly than lower QoS applications (as indicated by B).
- An increase of distance (as indicated by A1 and D) should result in a higher rate. In addition, the charging rate should increase exponentially as congestion (coefficient C) increases in order to achieve congestion control. These expectations are shown in Figure 6.8.
- It is clearly shown in Figure 6.9 that charging rate increases with distance, and applications with higher QoS requirements. These are desired objectives.

In conclusion, the proposed charging scheme is indeed distance-based, application-based, and congestion-based. The design objectives are met as expected.

6.3 Qualitative Analysis

After a thorough quantitative analysis of the IVB charging scheme, we confirmed that it has the desired features and characteristics of our design objectives. A qualitative analysis based on the evaluation dimensions will be provided in this section. Advantages and disadvantages will also be discussed.

6.3.1 Applying the Eight Evaluation Dimensions

The IVB scheme can be catalogued by applying the charging parameters proposed in Chapter 3.

Catalogue (IVB) = <Various Time Scale, User Input Choice, Volume, Dynamic; Distributed Computing, Decision by ISP; Best Effort and QoS Guaranteed, Multiple ISPs>

The performance of the IVB scheme is shown by applying the evaluation function introduced in Chapter 4.

Evaluate (IVB) = <Individual; Practicality; High Economic Efficiency; Application; High Admin-Network Efficiency; Feasible; No Theory; No Regulation>

The above evaluation results are explained in details by applying each evaluation dimension to the scheme:

- Individual Dimension: the proposed charging scheme is preferred by individual users. Users are informed of the charging rate in advance, and QoS guaranteed services are offered. Moreover, users have the flexibility to change service and price. For example, when the routes are congested, users can switch their Voice over IP service to Email service for a lower price.
- Practicality Dimension: the charging scheme takes practical implementation into considerations. The IVB scheme has a simple algorithm and an un-complex rate calculation, however, the probing process would need to be implemented. It would be best to have the probing done by the edge routers, so that when users request the service from their ingress routers, the process will be carried out before the session is admitted.
- Economics Dimension: economic efficiency is reached in the IVB scheme. Users are charged according to their usage, and are differentiated by their willingness to pay.

When the network gets congested, the marginal cost of sending one more packet is increased, consequently the charging rate is increased. Furthermore, since different applications' charging rates increase differently according to their sensitivity to the delay, users' utilities are maximized with constrained resources.

- Application Dimension: the proposed charging scheme takes each application's characteristics into account to determine the rate. Congestion for delay insensitive applications, such as email, is not considered, while delay sensitive, real time applications' charging rates increase quickly with congestion. Moreover, other congestion control techniques, such as CAC (Call Admission Control) could be activated when congestion (i.e., C-D) reaches some threshold for these applications.
- Admin-Network Dimension: the following objectives of the Admin-Network dimension, as discussed in Chapter 4, are reached in the charging scheme:
    - Utilization efficiency of the network resources: In the IVB scheme, users are charged according to their usage and the rates are adjusted according to the route congestion status. Unlike the flat rate pricing, only valuable traffic will be raised by the users, network resources are thus utilized efficiently. Moreover, as users could shift their traffic to less expensive (less congested) routes that can provide the same services, the traffic engineering objective is achieved.
    - Support for congestion control or traffic management: Since charging rate increases with congestion on the route, under the assumption that the number of users who will withdraw their requests increases with the increase of rate, congestion control or traffic management is achieved in the IVB scheme.
    - Informative on the network and user state: With the knowledge of changing charging rate and billing for each user, the IVB scheme provides information of the status about the network and the users. Also, an ISP can learn from the high rates regarding the parts of the network that are frequently congested. Expansion can be planned accordingly.

  - Maximizing the ISP's revenue: The IVB scheme adjusts the charging rates with congestion status. When the network has light load, the rate is lower which will attract more users to utilize the network. When the network is congested, the rate is then increased, and the ISP's revenue is thus maximized.

- Technology Dimension: The proposed charging scheme is compatible with ATM network since the application's characteristics are considered and applications could be classified in a way similar to the ATM's [30]. The probing process should be incorporated at the ISP side, and made transparent to the application or the user.
- Theory Dimension: The new charging scheme is still based on volume, except that the charging rate reflects the distance, application type and congestion. Though the scheme is innovative and heuristic, and it integrates features of complicated charging schemes into a simple usage-based charging scheme, no new theories are introduced.
- Regulation Dimension: Issues for regulation purpose are not considered in the charging scheme. However, the scheme is somewhat socially fair as poorer users can still receive those services which rates do not increase or increase slowly with congestion.

6.3.2 Other Considerations

Upon closer examination, one can identify similarities of the proposed charging scheme to other classes of Internet pricing schemes. It has the following benefits of a simple usage-based charging scheme:

- Simple and convenient
- Compliant with existing technologies
- Low billing measurement and cost of implementation
- User preferred as budget control is easy with fixed rate within a session

At the same time, the IVB charging scheme also has the following advantages as other congestion-based charging schemes:

- Facilitate congestion control: As congestion will drive the rate higher, users of these routes would probably shift their traffic from prime period to a less busy period, or shift from a higher price destination to a lower price destination that provides the same service.
- Encourage network efficiency: As users will be charged according their usage, less valuable traffic would be withdrawn by users from the congested routes. In addition, higher price on congested routes would make some users shift their required services to other destinations along less congested routes. This improves network efficiency as traffic is shifted from scarce to available resources.
- Encourage economic efficiency: Network resources are allocated to services which are appreciated (utility) more by users who are willing to pay more for these services when the network is more congested. Total utility is maximized.
- Revenue provides information for further expansion. A congested route not only provides the ISP with higher revenue due to higher charging rates, it also provides network information to ISP for future expansion consideration.

Obviously, there are some disadvantages in the proposed charging scheme. Management overheads, including the probing packets and pricing table lookup, would need to be implemented. However, these overheads would be offset if the session is relatively long.

The other potential problem would be the modification to existing technology in order to accommodate the charging algorithm. The code would be ideally be placed on the ISP side interfacing to users' requests. In addition, the practicality of having the ISPs at both ends of the route to agree on the probing protocol needs further investigation.

6.4 Chapter Conclusion

In this chapter, we have analyzed the proposed Integrated Volume-Based Charging Scheme from both quantitative and qualitative points of view. Quantitative results have shown that it is indeed a congestion-based, application-based and distance-based charging scheme. A big advantage of using the proposed scheme is that the Internet Service Provider could adjust the values of the coefficients in the charging formula to satisfy its own objectives.

The qualitative analysis is based on the eight evaluation dimensions introduced in Chapter 4, and the results are quite satisfying in most dimensions. It can be concluded that the proposed charging scheme can be adopted for real use with some modifications.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Many charging schemes have been proposed by economists and engineers in the past decade. With the further development of complex Internet applications, and their penetration into people's everyday life, Internet Pricing would have a more important role as a solution of congestion control, and for economic and network efficiency improvement.

In this thesis, a hierarchical structure is proposed to examine factors considered in existing charging schemes. Eight evaluation dimensions are introduced to give a better understanding of the design objectives as well as the advantages and disadvantages of each charging scheme. Classifications of existing charging schemes according to how well they perform in the economics and technology dimensions are shown.

In Chapter 5, a new charging scheme, the Integrated Volume-Based Charging Scheme, is proposed. The design objective is to combine the advantages of economic and network efficiency, usually provided by more complicated charging schemes, and the advantages of simplicity from Volume-Based charging. The proposed scheme was evaluated numerically using MATLAB. The quantitative results indeed meet design expectations. Moreover, a qualitative analysis made using the eight evaluation dimensions has shown that the proposed scheme is feasible and useful in practical situation.

7.2 Future Work

Possible future work includes a refinement of the evaluation cube, and the classification of additional charging schemes not covered in this work.

As mentioned in Chapter 3, charging schemes which exploit the advantages of allowing applications to be the decision maker could be another interesting and meaningful research topic.

Furthermore, with the eventual arrival of 3G and WiFi (Wireless Fidelity) technology, wireless Internet applications will be the next technology focus. Extending our study of

Internet Pricing to wireless applications area would be a worthwhile project.

Bibliography

[1] Qwest Global Internet Service Level Agreement. [Online]. Available: http://www.qwest.com/legal/Global_SLA_05_17_02.pdf

[2] The Tragedy of the Commons. [Online]. Available: http://www.econlib.org/library/Enc/TragedyoftheCommons.html

[3] M. Baglietto, R. Bolla, F. Davoli, M. Marchese, A. Mainero, and M. Mongelli, "A Unified Model for a Pricing Scheme in a Heterogeneous Environment of QoS-Controlled and Best Effort Connections," *Courmayeur 2002*, Januray 2002.

[4] R. Bohn, H.-W. Braun, K. Claffy, and S. Wolff, "Mitigating the Coming Internet Crunch: Multiple Service Levels via Precedence," *High Speed Networks*, vol. 3, no. 4, pp. 335–349, 1994.

[5] D. Clark, "Adding Service Discrimination to the Internet," *Telecommunications Policy*, no. 20, pp. 169–181, 1996.

[6] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Examples," *IEEE/ACM Transactions on Networking*, 1993.

[7] A. Dasilva, "Pricing for QoS-Enabled Networks: A Survey," *IEEE Communications Surveys, Second Quarter*, 2000.

[8] M. Falkner, M. Devetsikiotis, and I. L. Badaris, "An Overview of Pricing Concepts for Broadband IP Networks," *IEEE Communications Surveys, Second Quarter*, 2000.

[9] P. C. Fishburn, A. M. Odlyzko, and S. R. C., "Fixed Fee Versus Pricing for Information Goods: Competition, Equilibrium, and Price Wars," *First Monday*, vol. 2, no. 7, July 1997.

[10] T. Henderson, J. Crowcroft, and S. Bhatti, "Congestion Pricing," *IEEE Internet Computing*, pp. 85–89, September-October 2000.

[11] A. Jorn and K. Chu, "How to Charge for Network Services - Flat-Rate or Usage-Based," *Computer Networks*, no. 36, pp. 519–531, 2001.

[12] F. P. Kelly, "Tariffs and Effective Bandwidths in Multiservice Networks," in *International Teletraffic Conference ITC'14*, pp. 401–410.

[13] F. P. Kelly, A. Maulloo, and D. Tan, "Rate Control for Communication Networks:

Shadow Prices, Proportional Fairness and Stability," *Operational Research Society*, 1998.

[14] S. Khan, K. Li, E. Manning, R. Watson, and G. Shoja, "Optimal Quality of Service Routing and Admission Control Using the Utility Model," *Future Generation Computer System*, vol. 19, no. 7, pp. 1063–1073, Oct. 2003.

[15] C.-C. Liu, S.-C. Chang, and H.-H. Cheng, "Pricing and Fee Sharing for Point to Multipoint and Quality Guaranteed Multicast Services," in *Seventh International Conference on Parallel and Distributed Systems (ICPADS'00)*, Iwate, Japan, July 2000.

[16] J. K. Mackie-Mason and H. R. Varian, "Pricing the Internet," ' in Public Access to the Internet,B. Kahin and J. Keller, eds., Prentice-Hall, Englewood Cliffs, NJ, 1995.

[17] J. K. Mackie-Mason and H. R. Varian. (1994, September) Some FAQs about Usage-Based Pricing. [Online]. Available: http://citeseer.nj.nec.com/mackie-mason94some.html

[18] A. Mas-Colell, M. Whinston, and J. Green, *Microeconomic Theory*, 6th ed. Oxford, 1995.

[19] J. Misic and S. Chanson, "Charging Schemes for ATM Networks Based on Virtual Effective Bandwidth," *INFOCOM*, 1997.

[20] A. Odlyzko. (1997, September) A Modest Proposal for Preventing Internet Congestion. [Online]. Available: http://www.research.att.com/amoi

[21] A. Odlyzko. (1998, July) The Economics of the Internet: Utility, Utilization, Pricing, and Quality of Service. [Online]. Available: http://www.research.att.com/ amo/doc/complete.html

[22] A. Odlyzko, "Should Flat-Rate Internet Pricing Continue?" *IT Pro.*, pp. 48–51, September/October 2000.

[23] A. Odlyzko, "Internet Pricing and the History of Communications," *Computer Networks*, no. 36, pp. 493–517, 2001.

[24] P. Reichl, S. Burkhard, and T. Ziegler, "Charging Multi-dimensional QoS with the Cumulus Pricing Scheme," in *Proceedings of ITCom 2001*, Denver, Colorado, August 2001.

[25] B. Stiller, P. Reichl, and S. Leinen, "Pricing and Cost Recovery for Internet Services: Practical Review, Classification, and Application of Relevant Models," *Netnomics - Economic Research and Electronic Networking*, vol. 2, no. 1, January 2000.

[26] TechBuddha. (2000, August). [Online]. Available: http://www.techbuddha.com/nextlevel/wappt3imode.html

[27] H. Uijterwaal, C. Bovy, H. Mertodimedjo, G. Hooghiemstra, and

P. Mieghem. (2002, January 18th) Analysis of End-to-End Delay Measurements in the Internet. Amsterdam. [Online]. Available: http://www.ripe.net/test-traffic/Talks/0201_RIPE41_TUD/sld001.html

[28] H. Varian. (2000, August) Estimating the Demand for Bandwidth. [Online]. Available: http://www.INDEX.Berkeley.EDU/public/index.html

[29] J. Wang and K. Li, "Understanding Internet Pricing: An Objective-Oriented Classification," in *the IEEE Canadian Conference on Electrical and Computer Engineering*, Montreal, Canada, May 4th-7th 2003.

[30] S. William, *Data & Computer Communications*, 6th ed. Prentice-Hall, Inc., 2000.

Appendix A

Quantitative Analysis Code

A.1 Effects of the Coefficients on the Charging Rate

A.1.1 Effect of Distance Coefficient A1 on Charging Rate

```
A1=[0.00002,0.000025,0.00003,0.000035,0.00004,0.000045,0.00005,
    0.000055,0.00006,0.000065,0.00007,0.000075,0.00008,0.000085,0.00008625];
    A2=0.00011;
    B=5;
    C=0.2;
    D=0.04;
    R = A1 + A2 * exp (B * (C - D));
    plot (A1, R, ’k-’);
    xlabel(’Distance Coefficient A1 ($/packet)’,’Fontsize’,12)
    ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.1.2 Effect of Application Coefficient A2 on Charging Rate

```
A2=[0.00009,0.00011,0.00013,0.00017,0.00019];
    A1=0.00002;
    B=5;
    C=0.2;
    D=0.04;
    R = A1 + A2 * exp (B * (C - D));
    bar(A2, R);
    xlabel(’Application Coefficient A2 ($/packet)’,’Fontsize’,12)
```

```
    ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.1.3 Effect of Application Coefficient B on Charging Rate

```
B=[0,5,8,10,15];
    A1=0.00002;
    A2=0.00011;
    C=0.2;
    D=0.04;
    R = A1 + A2 * exp (B * (C - D));
    bar(B, R);
    xlabel(’Application Coeffi cient B’,’Fontsize’,12)
    ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.1.4 Effect of Round Trip Delay C on Charging Rate

```
C=[0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.15,0.2,0.25,0.3,0.35,0.4];
    A1=0.00002;
    A2=0.00011;
    B=5;
    D=0.04;
    R = A1 + A2 * exp (B * (C - D));
    plot (C, R, ’k-’);
    xlabel(’Round Trip Delay C (seconds)’,’Fontsize’,12)
    ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.1.5 Effect of Propagation Delay D on Charging Rate

```
D=[0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.11,0.12,0.13,0.14,0.15,0.16,0.17,0.18,0.19,0.2];
    A1=0.00002;
    A2=0.00011;
    B=5;
    C=0.4;
    R = A1 + A2 * exp (B * (C - D));
    plot (D, R, ’k-’);
    xlabel(’Propagation Delay D (seconds)’,’Fontsize’,12)
```

```
    ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.1.6 A Rate Sensitivity Study

```
A1=0.00002;
    A2=0.00011;
    B=5;
    C=0.2;
    D=0.04;
    R = A1 + A2 * exp (B * (C - D));
    Range=[-50,-45,-40,-35,-30,-25,-20,-15,-10,-5,0,5,10,15,20,25,30,35,40,45,50];
    R1 = A1*( 1 + Range/100) + A2 * exp (B * (C - D));
    R1 = 100*(R1-R)/R;
    R2 = A1 + A2*(1 + Range/100)* exp (B * (C - D));
    R2 = 100*(R2-R)/R;
    R3 = A1 + A2 * exp (B * (1 + Range/100)*(C - D));
    R3 = 100*(R3-R)/R;
    R4 = A1 + A2 * exp (B * (C * (1 + Range/100) - D));
    R4 = 100*(R4-R)/R;
    R5 = A1 + A2 * exp (B * (C - D * (1 + Range/100)));
    R5 = 100*(R5-R)/R;
    plot (Range, R1, ’k-’, Range, R2,’k:’, Range, R3,’k*’, Range, R4,’k+’, Range, R5,’k^’);
    xlabel(’Coeffi cient Percentage Change %’,’Fontsize’,12)
    ylabel(’Charging Rate Percentage Change %’,’Fontsize’,12)
```

A.2 Round Trip Delay versus Rate for Different Applications

```
C=[0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.11,0.12,0.13,0.14,0.15,0.16,0.17,0.18,0.19,0.2];
    A1=0.00002;
    A21=0.00009;
    A22=0.00011;
    A23=0.00013;
    A24=0.00017;
    A25=0.00019;
```

```
B1=0;
B2=5;
B3=8;
B4=10;
B5=15;
D=0.04;
R1 = A1 + A21 * exp (B1 * (C - D));
R2 = A1+ A22 * exp (B2* (C - D));
R3 = A1+ A23 * exp (B3* (C - D));
R4 = A1+ A24 * exp (B4* (C - D));
R5 = A1+ A25 * exp (B5* (C - D));
plot (C, R1, ’k-’, C, R2,’k:’, C, R3,’k*’, C, R4,’k+’,C, R5,’ko’);
xlabel(’Round Trip Delay (seconds)’,’Fontsize’,12)
ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.3 Round Trip Delay Versus Rate for Various Distances

```
C1=[0.04,0.05,0.06,0.07,0.08,0.09,0.1,0.11,0.12,0.13,0.14,0.15,0.16,
    0.17,0.18,0.19,0.2,0.25,0.3,0.35,0.4];
    C2=[0.13,0.15,0.18,0.2,0.23,0.25,0.28,0.3,0.33,0.35,0.38,0.4,0.43,0.45,
    0.48,0.5,0.53,0.55,0.58,0.6];
    C3=[0.17,0.2,0.23,0.25,0.28,0.3,0.33,0.35,0.38,0.4,0.43,0.45,0.48,0.5,
    0.53,0.55,0.58,0.6,0.63,0.65,0.68,0.7];
    C4=[0.26,0.3,0.33,0.35,0.38,0.4,0.43,0.45,0.48,0.5,0.53,0.55,0.58,0.6,
    0.63,0.65,0.68,0.7,0.73,0.75,0.78,0.8];
    C5=[0.57,0.6,0.63,0.65,0.68,0.7,0.73,0.75,0.78,0.8,0.83,0.85,0.88,0.9,
    0.93,0.95,0.98,1.0];
    A2=0.00011;
    B=5;
    A11=0.00002;
    A12=0.00003125;
    A13=0.00003625;
    A14=0.00004275;
    A15=0.00008625;
```

```
D1=0.04;
D2=0.13;
D3=0.17;
D4=0.26;
D5=0.57;
R1 = A11 + A2 * exp (B* (C1 - D1));
R2 = A12 + A2 * exp (B* (C2 - D2));
R3 = A13 + A2 * exp (B* (C3 - D3));
R4 = A14 + A2 * exp (B* (C4 - D4));
R5 = A15 + A2 * exp (B* (C5 - D5));
plot (C1, R1, ’k-’, C2, R2,’k:’, C3, R3,’k*’, C4, R4,’k+’,C5, R5,’ko’);
xlabel(’Round trip Delay (seconds)’,’Fontsize’,12)
ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```

A.4 Distance versus Rate for Different Applications

```
D=[0.04,0.06,0.08,0.1,0.12,0.14,0.16,0.18,0.2,0.22,0.24,0.26,
    0.28,0.3,0.32,0.34,0.36,0.38,0.4,0.42,0.44,0.46,0.48,0.5,0.52,0.54,0.56,0.57];
    A1=0.00002;
    A21=0.00009;
    A22=0.00011;
    A23=0.00013;
    A24=0.00017;
    A25=0.00019;
    B1=0;
    B2=5;
    B3=8;
    B4=10;
    B5=15;
    R1 = A1 + 0.000005*(D-0.04)/(0.04) + A21 * exp (B1 * 0.05);
    R2 = A1 + 0.000005*(D-0.04)/(0.04) + A22 * exp (B2 * 0.05);
    R3 = A1 + 0.000005*(D-0.04)/(0.04) + A23 * exp (B3 * 0.05);
    R4 = A1 + 0.000005*(D-0.04)/(0.04) + A24 * exp (B4 * 0.05);
    R5 = A1 + 0.000005*(D-0.04)/(0.04) + A25 * exp (B5 * 0.05);
```

```
plot (D, R1, ’k-’, D, R2,’k:’, D, R3,’k*’, D, R4,’k+’, D, R5,’ko’);
xlabel(’Distance Coefficient D (in seconds)’,’Fontsize’,12)
ylabel(’Charging Rate ($/packet)’,’Fontsize’,12)
```