The University of Oklahoma
Graduate College

Pricing in Multi-Service Communication Networks: A Game-theoretic Approach

A dissertation
Submitted To The Graduate Faculty
in partial fulfillment of the requirements for the
Degree of
Doctor Of Philosophy

by

Fan Zhang
Norman, Oklahoma
2011

Pricing in Multi-Service Communication Networks: A Game-theoretic Approach

A Dissertation Approved For The
School Of School of Electrical and Computer Engineering

By

Pramode K. Verma, Chair

Yi Zhou

James J. Sluss

Sudarshan K. Dhall

Samuel Cheng

To my beloved husband.

Acknowledgments

I would like to express my gratitude to my advisor, Dr. Pramode Verma, whose patience, encouragement, and in-depth knowledge has provided invaluable help for my research work. This thesis would have not been possible without his help and support. I would like also to thank Dr. Samuel Cheng for his insightful comments and help during my Ph.D. study. Besides my advisors, I would like to thank the rest of my thesis committee: Dr. Yi Zhou, Dr. Sudarshan K. Dhall and Dr. James J. Sluss, for their comments and questions on my Ph.D. work.

Much gratitude to my parents, Bo Zhang and Mingjing Xu, who have offered me support without any reservation ever since I was born. To my husband, Xuelin Li, for his help, long-time support, and encouragement throughtout these years of study.

I also want to thank my friends in the department for their friendship. And thanks to all the people who helped me during my Ph.D. study.

Table of Contents

List of Figures

List of Tables

Abstract

The promise of multi-class communication networks is gradually becoming a reality. The term multi-class means that the network provides different classes of service that can support diverse application requirements and heterogeneous users demand. This dissertation focuses on establishing an equitable price for each class of service in multi-class networks, considering fairness among the classes and economic efficiency. We adopt a game-theoretic approach to the problem in order to take into account the interdependence among users' service choices.

We investigate subsidy-free prices for each class of service under two distinct service architectures: in multi-class priority-based networks, traffic from each class is assigned priority level in the queue; in multi-class DiffServ networks, network resource is allocated to each class. In both cases, classes of traffic having longer average waiting time receive monetary compensations from other classes and the subsidy-free price for each class of service is developed based on inter-class compensations. This work provides a framework to set subsidy-free price or sustainable price for each class of service which is assumed crucial to network providers if they are to survive the competition in the market place.

We further consider market-clearing prices for each class of service in a competitive market in which each user endowed with an initial budget will purchase bandwidth from each class of the network resource to maximize his or her utility function. A competitive equilibrium is reached when the total bandwidth is allocated, each user spends all his or her budget, and the utility functions are independently and simultaneously maximized. Our research shows that such equilibrium always exists and, under fixed bandwidth supply for each class of service, the equilibrium is also unique. Furthermore, we discuss how to adjust the initial endowment of each user to meet his or her individual bandwidth constraint, either from constraint on the access network or from the limitation of the user equipment. Under this bandwidth constraint condition, the proposed competitive equilibrium yields the price for each class of service, the budget redistribution and the bandwidth allocation among all users. We also develop an iterative algorithm for budget allocation to satisfy each user's bandwidth constraint. The presented competitive market model provides a solution for pricing a multi-class network and allocating network resource among users. And we find this solution achieves higher social utilization, better individual satisfaction and the QoS of each class.

Another advanced topic in communication networks is net neutrality, which has become the subject of fierce debate among the stakeholders of public telecommunication services. Broadband access providers argue that preservation of the integrity of the network services requires them to use discriminatory traffic management practices to slow down certain applications or to purge certain packets that would compromise the integrity of the network. We propose a solution based on the idea of inter-user compensations that could control network congestion and yet maintain fairness among heavy and light users without violating net neutrality. Users consuming less network resource will receive compensations from heavy users. Our research provides a method for broadband access providers to shape the traffic characteristics of users and thus controlling network congestion and maintaining network performance without inflicting discriminatory treatment on network traffic.

CHAPTER 1

Introduction: Pricing and Communication Networks

The tremendous surge in the use of the Internet for business and entertainment is among the most important social phenomena of the last several decades. The development of cost-effective optical networking technologies and the wide acceptance of Internet Protocols as the platform for communication, further enhanced by the imaginative edge software glue of World Wide Web 2.0 (or 3.0), have provided a converged network that supports a multitude of applications. This has raised users' expectation for the network's ability to provide applications with different quality of service (QoS) requirements. The demand for applications that deliver voice, video, image, and text, often in real-time mode, leads to network service differentiation and controllable QoS in the Next Generation Network [1].

Multi-service networks can support diverse application requirements, and there is a clear need for incentives to be offered to customers to encourage them to choose the service that is most appropriate for their needs, thereby discouraging over-allocation of resources. Pricing is commonly assumed to be an effective solution to this end. The relation between QoS and price of services in a communication network is a major theme of this dissertation. We focus on a study of establishing equitable prices for each class of service in multi-class networks, taking into account fairness among the classes and economic efficiency. While the network pricing schemes are often dominated by the network providers' policies and market forces, our study will provide an important input for actual pricing.

Communication network services have changed the way we engage in social life, business and politics. The pricing of these services, therefore, plays an important role. Although a tariff must be charged by the service provider to recover costs and remain solvent, pricing has other important roles such as shaping users' demand, controlling network congestion. In order to understand pricing's other roles, we need first consider the characteristics of communication networks and communication services. In this chapter, we begin with the network's externality characteristics and then present special features of communication services. Section 1.2 states the role of pricing in communication networks, and Section 1.3 lists recent research on pricing communication networks. Section 1.4 contains a statement of the problem and a summary of contributions of this dissertations. Finally, we finish the chapter with a description of the organization of this dissertation.

1.1 Characteristics of Communication Networks

1.1.1 Network Externalities

Network externality is the notion to describe that a network's value increases with its size. Metcalfe's Law [2], which is named after the inventor of Ethernet, describes that the value of a network increases by the square of the number of its users. The foundation of this law is the observation that in a communication network with n users, each can make $n - 1$ connections with other users. The total value of the network is proportional to $n(n - 1)$, that is, approximately, n^2. This makes a large user base a competitive advantage because each of the large network's users can communicate with a greater number of other users.

During the Internet bubble of the late 1990s, entrepreneurs, venture capitalists, and engineers believed in the steady commercial growth of the Internet; and Metcalfe's Law gave a quantitative explanation for the Internet boom's various reasons, like "first move advantage," "Internet time" and "network effects". At that time, many companies invested heavily in new fiber infrastructures not only at a backbone level but also at the metropolitan level. Dense Wavelength Division Multiplexing made it possible to transport up to 160 light waves on a single strand of fiber with a combined bit rate into the range of tera-bits per second [3]. Ethernet technologies and the Internet Protocol made the connectivity services to be supported on these fiber infrastructure very inexpensive. A fact that is likely responsible for an overinvestment in the telecommunication infrastructure during the 1990s [4].

One fundamental flaw underlying Metcalfe's law is the assignment of equal value to all connections. For example, in a large network such as the Internet, there are millions of potential connections between users. In general, connections are not all used with the same intensity and most of them are not used at all. As a result, assigning equal value to all connections is not justified and a revision of Metcalfe's law is proposed in [5], based on the assumption that not all connections are equally valuable to the users. The assignment of value to each connection is based on the ZIPF's Law [6]. It says, for example, that in a long English language text, the most popular word, "the," usually accounts for 7 percent of all word occurrences; the second-place word, "of," makes up 3.5 percent; the third-place word, "and," accounts for "2.8" percent. In other words, the sequence of word occurrence frequencies corresponds closely with the $\frac{1}{k}$ sequence, $(1, \frac{1}{2}, \frac{1}{3}...)$. This logic can be extended to a communication network with n users. For each user, the value of connections will be proportional to $1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n-1}$, which approaches roughly to $log(n)$. There are other $n - 1$ users who get similar value from the network and the value of the network is proportional to $nlog(n)$ in the revised metcalfe's law.

The revised metcalfe's law shows that the value of the network grows faster than its size in linear terms and has a form of $nlog(n)$. This growth, which is faster than the linear growth, helps explain the occasional Internet successes we have experienced. On the other hand, the $nlog(n)$ valuation, which is smaller than the Metcalfe's square growth (n^2), describes a slower growth in the value of dot-com companies and explains the Internet bubble from another angle.

In this section, we have reviewed the network externalities: the value of the network increases as $nlog(n)$, where n is as its user base. This is faster than the linear growth. Since the cost of the network is, at most, linear to its user base, a network provider has a great incentive to price its services attractively to expand the user base so as to increase the value of the network.

1.1.2 Communication Services

When a communication network is built, the construction cost is largely a fixed cost and the variable operating cost is extremely small compared to the fixed cost. This is somewhat similar to information goods that are costly to produce but cheap to reproduce. The first copy of a software product bears all the development cost, and it is a sunk cost. All additional copies can be produced at almost zero marginal cost. Similarly, once a network is built, the marginal cost of providing a unit of communication service can be almost zero, especially when there is no congestion. And it is well-known that a competitive market drives prices toward marginal costs. As a result, there is a danger for the communication industry that the prices of communication services can be driven close to zero.

Returning to the subject of communication services, it should be noted that they can sell at both low and high prices. For example, there are hundreds of web sites providing email services, and it seems they cannot charge users because there are many nearly equivalent sites providing similar service. Such a service is termed as *commodity*, which has little pricing flexibility. Providers of a commodity service would find it more profitable to concentrate on differentiating their services by providing value-added features like security for which some users may be willing to pay for this feature. Besides the fact that different applications require different QoS as discussed in the beginning of this chapter, avoiding becoming a commodity is another reason for the communication networks to support service differentiation. The reason is that if a good is not a commodity, it can sell at a price that reflects its value to users, rather than its production cost, that is, its marginal cost. From a regulatory perspective, it can be noted that as long as network operators offer equal treament to users' traffic within each class of transport service, the service differentiation across classes does not violate the net neutrality [7].

A special feature of communication services is their reliance on statistical multiplexing. This is because the data traffic is often bursty and sporadic, and the network does not need to reserve bandwidth for users equal to their maximum demand. Therefore, statistical multiplexing produces economy of scale, that is, the size of the user base increases more than proportionately to the raw quantity of the network resource for identical performance. This fact shows that the value of the network should increase much faster than the investment of building the network because of the economy of scale. Besides network externality, statistical multiplexing is also incentive for network provider to price its services attractively so as to expand its user base.

Another special feature of the communication services is that the performance obtained by any network user is not only determined by his own traffic and service choice but also by other users' traffic and service choices. While this is also true for Internet best effort (single grade QoS) networks, in multi-service networks this interdependence is much more complex because priority service traffic may cause performance variation to all others, even when the aggregate traffic load remains constant. This interdependence among users can be addressed through a game theoretic framework in which each user makes a service choice and traffic volume or throughput rate that maximizes his or her utility while taking into consideration all the other users' choices. This dissertation shows that such a delicate balance is achievable by applying the principles of game theory.

In this section, we have discussed features of telecommunication services: the marginal cost of the communication services is close to zero, and in order to recover the huge sunk cost and be profitable, service differentiation provides an avenue. Statistical multiplexing produces economy of scale for communication networks; the interdependence among users from a usage and pricing perspective can be effectively addressed using a game-theoretic framework.

1.2 The Role of Pricing in Communication Networks

One obvious role of pricing is for network providers to recover the capital investment plus extra revenue and become profitable in the market place. As discussed in Section 1.1.1, the network externalities effect prompts the network providers to increase the value of the network by reducing price to attract more demand. Network providers can also increase prices to control congestion and smooth bursty customer demand. Therefore, pricing can be viewed as a control mechanism to shape users' demand.

However, some commentators like David Isenberg believe that future networks will be overprovisioned [8]. By taking advantage of the reduced cost of new technologies, overprovisioning can solve the problem of congestion, and network providers don't need to use pricing to control traffic. Overprovisioning might be reasonable for the backbone of the network because it consists of a fairly small number of links. But there is substantially less fiber installed in the access part of a network, which connects users to the core network. The core network infrastructure is shared by all users, but the access network is used by much fewer users. Some experts believe that it would take 20 to 30 times as much time and expense to overprovision an access part of the network as it has taken to build the fiber infrastructure in the backbone [4]. Although AT&T and Verizon have the fiber to home services like Uverse and FIOS respectively [9], these services are only provided in a limited number of metropolitan cities.

In addition, it is always hard for any network operator to predict demand. Just as it was overestimated in the 90s, it may now be underestimated. There are increasing amounts of traffic on the network generated by programs and devices connected to the Internet, rather than by humans. These programs and devices can ultimately greatly outnumber human users, network traffic has the potential to grow extremely rapidly. The question is how fast will the demand outgrow the supply. In addition, capacity cannot be provided in arbitrarily small increments. Because the demand is not accurately predictable and capacity expansion cannot be provided in "real-time" in response to every increase in demand, pricing can serve an important role by increasing stability and reducing quality fluctuations. As the network transport service market will be constantly in a transient phase, we believe that pricing will always play an important role in safeguarding the network performance while equitably addressing users' needs. Without effective pricing schemes, a network provider won't have enough incentive to invest, and, therefore, overprovisioning may never happen in the market place. Without enough network capacity, internet innovations will be restricted and will adversely affect users' experience. Thus, in addition to being a control mechanism, pricing is also an important factor in keeping the information industry economically viable.

Pricing can produce the right incentives for users to choose levels of service in service differentiated networks most appropriate for them so as to help ensure that users do not waste important resource that they do not value. For example, one user with an email application should not require the same QoS transport service as another user with a Voice over IP application. Pricing can provide the right incentives by appropriately charging higher QoS transport service. While the proposed implementations vary in studies [10, 11, 12, 13, 14, 15], the basic idea is that the appropriate pricing policy will provide incentives for users to behave in ways that improve overall network performance.

Pricing can be further viewed as a signal from the network operator to user base that there are incentives to use the network efficiently. Pricing information that is signaled to the edges of the network can play a significant role in providing rational end-users with the appropriate incentive to control their traffic. This is very similar to the Transmission Control Protocol (TCP) in the Internet. When TCP receives a congestion signal from the network, it reduces the sending rate; otherwise it increases the sending rate. However, a user might cheat by rewriting the protocol to disobey the TCP and send at a greater rate than it should. This would not be an issue if the congesting traffic were to be charged more [16]. Use of this contrivance provides stability and robustness to the network.

In this section, we have talked about using pricing as a means of shaping users' demand and described its role in signaling. By increasing price, an operator can reduce demand, reduce congestion, and ensure that services are provided to the users most willing to pay.

1.3 Related Works on Pricing for Communication Networks

1.3.1 Pricing for Regulated Telecommunication Services

Providers of services like telephony offered in the Public Switched Telephone Network (PSTN) are considered common carriers. Pricing of such services is generally subject to regulation. Traditionally, there has been very little competition in such markets. With the advent of Internet services like VoIP, email, etc., the market place for these services has changed much; however, it is still interesting to review the traditional regulatory pricing scheme for network operators.

In a highly regulated, monopoly environment, telecommunication providers could maximize their profits by raising prices to inefficiently high levels at the expense of users' welfare and reduced demand for telecommunication services. Because of the traditionally high barriers to enter the telecommunication industry, the monopoly situation tends not to be a temporary phenomenon. In order to protect users from monopoly pricing, federal and state policymakers enforce rate regulation on incumbents.

Like any other public utility, an incumbent has made a regulatory pact with the government in which the company is given an opportunity to earn a "reasonable rate of return" on its overall regulated investment. Under a rate-of-return regime, federal and state regulation gives dominant local incumbents opportunities to charge retail rates sufficient to cover their anticipated expense plus a reasonable return on their net investment [17].

The rate-of-return regulation tends to give these incumbents incentives to "gold plate" their assets, that is, to spend more than is efficient or necessary simply to increase the rate, thus to increase profits. In the 1980s and 1990s, federal and state regulators sought to address these problems by adopting a price cap scheme for retail rate regulation of the incumbents. A price cap analysis starts with the retail rates calculated in a given year under the traditional rate-of-return regulation. In the succeeding years, however, retail rates were not determined by the rate-of-return process but by mathematical adjustments designed to reflect the following two factors. The first is driven by technological and other innovations resulting in industry-wide increases in efficiency; the second by fluctuations in inflation and other macroeconomic variables. This price cap approach, unlike the traditional rate-of-return regime, rewards the incumbents for efficiency over time by allowing them to keep much of the extra profit they generate as a result of cutting unnecessary costs. Right now, the retail

Table 1.1: Arguments for and against flat rates and usage-dependent pricing

	Advantages	Disadvantages
FLAT RATE	•Easy to implement	•Unfair to light users
	•Little billing overhead	•May lead to service overuse
USAGE-DEPENDENT	•Increased fairness	•Adverse response from users
	•Can be used for congestion control	•Billing complexity
		•Reduced usage

rates of most incumbents are under the regulation of price cap.

1.3.2 Pricing Internet Services

Internet access has always been categorized under U.S. law as an information service, not as a telecommunications service. Thus it has not been subject to common carrier regulations [18]. In this non-monopolized market, network providers compete against one another for users, and this competition theoretically keeps the price of service at reasonably efficient levels. The most widely used pricing schemes for the Internet services include access-rate dependent charges, usage-dependent charges, or a combination of both [19]. An access-rate dependent charge has the following two forms: unlimited use, or limited time of the connection and charging a per-minute fee for additional connection time. Similarly, the access and usage-dependent charging scheme allows a fixed access fee for a defined usage to be transmitted, and then imposes per-unit volume charge for additional use. A brief summary of the main advantages and disadvantages of flat rate and usage-dependent rate are presented in Table 1.1.

An access-rate dependent flat rate is the method used in the United States to charge for Internet use. Light users (e.g., email, occasional web browsing) may, therefore, subsidize the heavy users (e.g., multimedia applications, frequently downloading of large files) [20]. In addition, the unbridled consumption may lead to overuse of the network resources. However, because users have strong preference for flat rate pricing, despite the above disadvantages, service providers still stick to a flat rate model in order to avoid losing customers to a competitor [21].

Usage-dependent pricing can be a solution for the problem of fairness and service overuse. However, this policy makes it difficult for users to budget for a network expense, not only because it is hard for users to predict their own traffic statistics, but also because the Internet is an interactive experience and users are not fully in control of their usage. These evidences makes Internet users not react favorably to a usage-dependent pricing scheme [22]. Furthermore, for network providers, the additional costs in billing may be substantial and must be offset by the gains brought by usage-based pricing. In traditional telephony, more than half of what users pay for a call goes to cover the cost of providers' accounting system [23], and this is in a circuit-switched system in which there is no need to count how many packets traverse the network. Finally, usage-dependent pricing tends to discourage the use of the Internet which is in contrast to the network externality discussed in Section 1.1.1.

Well-known proposals for Internet pricing rely on a centralized optimization process to maximize the total users utilities [16, 24, 25, 26, 27]. Kelly [16] forms a distributed flow control algorithm using the gradient ascent method from optimization theory which continuously informs the selfish users prices according to the network condition. Selfish users, who seek to maximize their own net benefit, are given the prices that have the right incentives to globally optimize the social benefits. An Explicit Congestion Notification (ECN)-based marking has been proposed in [28] to convey congestion information back to the end points. The resulting system converges to an optimal system state as long as all utility functions are strictly concave. Instead of only marking the packets during periods of congestion, [29] has proposed assigning each packet a price to reflect the congestion of the network. However, it is not clear whether all these theoretical results hold in the presence of transmission delay at the scale of a large network. In addition, all these schemes assume network services are best-effort and rely on a pure market mechanism to maximize social benefits.

The Internet, a single-service or best-effort service network, cannot support the performance needs of heterogeneous applications unless it is extremely overprovisioned [30]. Moving from a single-service to multiple-service architecture adds new dimensions to the pricing issue. It is obvious that a flat rate would no longer provide adequate incentives for users' choice of services, and therefore service-class-dependent, congestion-sensitive approaches must also be investigated. The next subsection reviews some of the progress made in pricing of multi-service communication networks.

1.3.3 Recent Literature on Pricing Multi-service Communication Networks

A number of articles have been published on telecommunication engineering and economics investigating the subject of pricing for multi-service networks. We summarize some of these studies.

Pricing based on network resource consumption has been considered in [32, 33, 34, 35, 36, 37]. A study [32] has proposed a pricing algorithm in a DiffServ environment based on the cost of providing different levels of services and on long-term average user resource demand of a service class. The network service is dynamically priced based on the level of service, usage, and congestion-sensitive parameters. The study [36] has presented a mechanism that introduced a priority system with the objective of providing a higher and a lower quality of service to two customer groups. The non-priority traffic carries a lower price tag and a lower quality of service. An important characteristic of the proposed pricing schemes is that the overall revenue associated with the network would remain constant as long as the total demand is confined within a relatively large bound, termed the region of operation, for the network. Like the region of operation defined in [36], in order to make sure the prices for higher QoS are larger than prices for lower QoS, [32] also assumes the long-term demand for higher QoS traffic is lower than demand for lower QoS traffic. Reference [34] has proposed that users be charged a price per unit of effective bandwidth used. Assuming that the network knows its capacities and virtual path routing, as well as users' benefit function and traffic stream characterization, the paper has discussed the role of pricing in meeting users' needs, network resource allocation, and contract negotiation to form a complete connection provisioning process. Reference [35] has studied a network that offers its bandwidth and buffers for rent. The network periodically adjusts prices based on monitored user requests for resources with the objective of maximizing social welfare. Users reserve resources based on individual traffic parameters and delay requirements so as to maximize their utilities subject to budget constraints.

Microeconomic supply-demand principles have been applied to network traffic management problems. The studies in [38, 39, 40] rely on a centralized optimization process to maximize the total user utility. Kelly [38] has described a system in which users reveal how much they are prepared to pay per unit time. Then the network determines allocated rates so that the rate per unit charge are proportionally fair. The author has determined that the optimum system in this case is achieved when users' choices of charges and the network's choice of rates are in equilibrium. Reference [39] has studied the efficiency of using one bit to carry streams with differential QoS requirements in an attempt to maximize network revenue. In [41, 42], the resource is priced to reflect demand and supply. The method in [41] relies on well-defined source model and cannot adapt well to changing traffic demands; while the scheme in [42] also takes into account network dynamics (sessions join or leave) and source traffic characteristics, and allows different equilibrium prices over different time periods. An economic equilibrium model is proposed in [40] which describes utility maximization by users and revenue optimization by service providers. In the presence of competing providers, the equilibrium prices reduce to the marginal costs. Study [43] has borrowed the framework described in [29] and calculates a price for each packet based on its bandwidth consumption, service level and buffer occupancy. Reference [44] adjustes bandwidth and buffer allocations among classes to guarantee the target delay and loss.

Several studies have demonstrated through experiments or simulations that service-class sensitive pricing results in higher network performance. Reference [45] has proposed a Paris Metro Pricing (PMP) scheme which partitions the network into logically separated classes with different prices for each. It is expected that the higher-priced class will have less load and will provide better service. The behavior of PMP under equilibrium conditions is considered and compared with a uni-class pricing system in [46, 47]. Study [48] has analyzed the equilibrium of such a system using non-cooperative game theory. Reference [49] has considered a similar framework based on queuing theory and experiments. All of the above works consider the impact of differential pricing on the relative performance of the system as a result of user self selection process. References [50, 51] have used simulations to study the problem of customer decisions in a two-priority network, where a fixed per unit price is associated with each priority class. These studies have concluded that, through the use of class-dependent pricing, it is possible to set prices so that all users are more satisfied with the resultant cost/benefit provided by the network.

Several opportunity cost-based mechanisms have been studied. Reference [52] has addressed the impact of QoS on bandwidth requirements in IntServ networks and proposes a scheme in which a service provider can develop compensatory and fair prices for users with varying QoS. Since exclusive allocation of bandwidth to a specific flow has a performance penalty on delay and jitter to other flows, [52] has derived the additional capacity required to maintain the desired performance of other flows and has proposed a compensatory scheme that will fairly charge the specific flow requesting exclusive bandwidth. Reference [33] has developed a grade-of-service (GoS) based pricing scheme that results in efficient utilization of the network bandwidth and buffers. Essentially, each traffic is charged an amount of money based on the QoS degradation caused to other users sharing network resources. Price is, therefore, a function of the network utilization as well as individual utilities. Reference [53] has presented an approach based on the notion of cost in the context of providing services with differentiated levels of quality. In [53], the authors have investigated the impact of multiple traffic classes on the carrying capacity of a network with a prescribed threshold of blocking probability in a DWDM ring network architecture.

Auction-based mechanisms have been studied in [12, 54, 55]. The smart market model has been studied in [12], in which prior to transmission, users inform the network of how much they are willing to pay for the transmission of a packet; packets are admitted if their bids exceed the current cutoff amount, determined by the marginal congestion cost imposed by an additional packet. Users do not pay the price they actually bid, but rather the market-clearing price, which is always lower than the bids of all admitted packets. However, this mechanism only provides a priority relative to others and it does not promise quality of service. The Generalized Vickrey Action (GVA) model in [54] supports multiple levels of QoS guarantees. But the optimal solution requires substantial computation, which increases as polynomial-time with the number of users. The Progressive Second Price auction (PSP) scheme in [55] has extended the traditional, single, non-divisible object auction to the allocations of arbitrary shares of the total available resource with associated bids.

A set of game-theoretic analyses have been proposed for QoS provision and network pricing. In [56], packets are marked according to users' QoS requirements and the costs incurred to users are dependent on performance. The authors in [57] have investigated the dimensioning of network capacity for different service classes. References [58, 59, 60] have studied a static pricing scheme based on the priority classes. Reference [58] has described a method to predict each user's service choice in the game-theoretic framework given any price difference between services and an estimate of users' utility functions. Therefore, the service provider can determine the price ranges that encourage users to exhibit behavior that is beneficial to both users and providers. The work in [61] has generalized the idea in [12] to support auctions for different service levels.

In this dissertation, we analyze desired prices for networks considering current network infrastructure and multiple service classes under different contexts described in Section 1.4.1. In order to properly take into account the interdependence among users' service choices, we employ game-theoretic concepts.

1.4 Problem Statement and Summary of Contribution

1.4.1 Contexts for Deriving Prices

The work presented here is primarily motivated by the advent of networks that support more than one class of service. The need for a mechanism that makes profitable and efficient use of existing resources leads to the importance of reasonable pricing policy to be adopted by service providers.

In thinking about how price is determined, the first rationale is to set subsidy-free prices or sustainable prices. Imagine that an incumbent firm wishes to protect itself against competitors who might enter the market. If the incumbent is to be secure against new entrants seducing away some of its users, the prices it charges for different services must not involve any cross-subsidization. If the incumbent uses the revenue from selling one service to subsidize the cost of producing another, then the firm is in danger if a competitor produce only the first product and sell it for less. However, in a communication network, a large part of the total cost is common cost, and it is difficult to apportion that cost rationally among different services. Service providers must figure a way to set prices to be subsidy-free and sustainable if they are to survive the competition in the market place.

A second rationale for setting prices is driven by the objective to match supply and demand in the market place. Supply and demand at given prices depend upon the supplier's technological capacities, the costs of the supply, and how users value the services. If prices are set too low, there will be insufficient incentive to supply and there is likely to be unsatisfied demand. On the other hand, if prices are set too high, suppliers may over-supply the market and find there is insufficient demand at the higher price. The right price should be "market-clearing," that is, it should be the price at which demand exactly equals supply.

The two rationales discussed above for setting prices do not necessarily lead to the same prices. There is possibly no single recipe for setting prices that satisfies all possible requirements. Pricing may also depend on the context. For example, a monopoly supplier in a market sets price to maximize its profits. If this market is under regulation, a regulator may arrange prices to maximize the social welfare, which not only includes supplier's benefits but also users' welfare. This means that the task of pricing requires a careful balance between customers' need, their willingness to pay, the underlying principle of the technology as well as the regulatory environment.

In this dissertation, we model the communication system in a game-theoretic framework to reflect the interdependence of each class of service. We first study the desirable subsidy-free price for each class of service, and then investigate market-clearing price for each class of service under competitive market model.

1.4.2 Distinguishing Characteristics of Our Approach

A well-engineered packet-switched network will make use of statistical multiplexing gains to decrease capacity costs while still being capable of guaranteeing adequate QoS. Therefore, resources that are not directly allocated to any user are typically made available to all. The inclusion of this particular characteristic into our model adds interdependence among users' service choice and resource allocation among different classes. We also recognize the fact that different applications have different QoS objectives. We model such distinct objectives through a utility function which will be discussed in Chapter 3.

As shown in Section 1.3.3, the bulk of the literature in multi-service network pricing focuses on developing the desired price to maximize users' utility, network provider's revenue, and social welfare [33, 34, 35, 39, 40]. Its objective is to keep the system in equilibrium [58, 59, 60] or to maintain constant revenue [36, 37, 53, 52]. However, the subsidy-free pricing issue remains unexplored and part of the reason is that it is difficult to apportion cost rationally among different classes of service. In this dissertation, we try to investigate general guidelines for finding the price difference among different classes of service which can be an important input for actual pricing for service providers.

Our study of subsidy-free pricing for multi-service networks is largely inspired by the pioneering work of Maniquet [62], who studied the monetary transfer between agents in a queuing problem where each agent had a different unit waiting cost. We extend the model described in [62] to multi-class communication networks, and generalize it by adding the stochastic property of communication traffic. We also propose a fair and efficient way to get the price differences among classes of service.

Further, we investigate the market-clearing prices for each class of service in a multi-service network under the competitive economy model. The work of Ye [63] in spectrum management using competitive economy serves as a major source for our study. We use revenue as the utility function for the service provider and enhance

Kelly's utility function [16] by including a QoS parameter for users. Given the initial endowment for each user, we show that a competitive equilibrium (price for each class of service and bandwidth allocation among all users) for the competitive multi-class network resource market always exists. And under this equilibrium, both individual optimality and social economic efficiency are achieved in a way that all users' utilities are maximized simultaneously.

1.4.3 Summary of Contributions

The main contributions of this dissertation are the following:

- We investigate how, in multi-class priority-based networks, appropriate subsidy-free prices for each class of service are calculated in order to protect a service provider against entry by potential competitors.
- We investigate how, in multi-class DiffServ networks, appropriate subsidy-free prices are calculated in order to protect the service provider against entry by potential competitors.
- We show that, when the network utilization is high, the price difference between high QoS and low QoS services increases rapidly, which gives users the incentive to truthfully disclose their QoS requirements. This information is of great value to the traffic management task.
- We demonstrate that the market-clearing price always exists for a multi-service network and, with this price, both individual optimality and social economic efficiency are achieved simultaneously.
- We further discuss how to adjust the initial budget for each user to meet their bandwidth constraint (either from constraint on the access network or from limitation of the user equipment).
- We propose a solution for network providers to control network congestion and yet maintain fairness among heavy and light users without violating net neutrality.

1.5 Structure of This Document

This dissertation is structured into ten chapters:

1. **Introduction** - In this chapter, we motivate the research and provide related work on pricing for multi-service networks.

2. **Game Theory and its Applications to Communication Networks** - This chapter includes background information on Game Theory and its applications to communication networks.

3. **User, Network, and Pricing Model** - Chapter 3 formally states our model of user and network provider behavior. The model is made general enough to fit the different network architectures discussed in the following chapters.

4. **Subsidy-free Prices in Priority-based Networks** - Networks in which service differentiation is accomplished through the assignment of priorities are employed in this chapter. Rules for assigning positions for each class of traffic in the queue and corresponding subsidy-free prices are investigated.

5. **Subsidy-free Prices in Class-based Networks** - Networks in which QoS for each class is guaranteed through the allocation of resources are discussed in this chapter. Analysis of the subsidy-free price for each class of service is presented.

6. **Market-clearing Prices in Class-based Networks** - This chapter investigates the existence and uniqueness of market-clearing prices for multi-service class-based networks. We further discuss how to adjust the initial budget for each user to meet their bandwidth constraints.

7. **A User-friendly Constant Revenue Model for Net Neutrality** - Chapter 7 proposes a solution for broadband access providers that would control network congestion and yet maintain fairness among heavy and light users without violating net neutrality. The broadband service provider's revenue remains constant under this proposal.

8. **A Constant Revenue Model for Packet Switched Network** - Chapter 8 presents a mechanism that introduces a priority system with the objective of providing a higher and a lower quality of service to the two customer groups. The nonpriority traffic carries a lower price tag and a lower quality of service. An important characteristic of the proposed pricing schemes is that the overall revenue associated with the network remains constant as long as the total demand is confined within a relatively large bound, termed the region of operation, for the network.

9. **A Two-step Quality of Service Provisioning in Multi-class Networks** - This Chapter proposes a novel distributed QoS provisioning architecture for multi-class networks by two steps: inter-class resource allocation and intra-class flow control. In the proposed architecture, the service provider only supports limited number of classes. The inter-class network resource allocation is modeled as a centralized optimization problem to maximize the social welfare while maintaining the quality of service for each class as a whole. A distributed game theoretic framework is proposed to regulate flow behavior within each class.

10. **Summary and Future Work** - This chapter summaries the dissertation and gives the directions of future work.

CHAPTER 2

Game Theory and its Applications to Communication Networks

Game theory has been used for years as an economic analysis tool to understand and predict what will happen in economic contexts [65]. Because of the interdependence among users of communication networks, game theory also provides a useful framework for modeling users' decisions. In this chapter we discuss some of the game-theoretic terminology and basic concepts that are relevant to the present work and review some of their applications to communication networks.

2.1 A Brief Introduction to Game Theory

A game consists of a principal (e.g., the network service provider) and a finite set of players (e.g., network users) $N = \{1, 2, ..., n\}$. The network service provider supports $M = \{1, 2, ..., m\}$ different QoS classes. Each player will then choose a strategy $x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ with the objective of maximizing its payoff function u_i, where x_{ij} is the amount of traffic from service j that user i consumes. The following terminology applies to the classes of games we study:

- Player i's strategy, $i \in N$, is a M-dimensional vector x_i;
- A player's strategy space $X_i \subseteq R^M$ is the set of strategies available to user i, thus $x_i \in X_i$;
- A joint strategy x is the vector containing the strategies of all players: $x = \{x_1, x_2, ..., x_n\}$;
- The joint strategy space X is defined as the Cartesian product of the strategy spaces of all players: $X = \times_{i \in N} X_i$;
- Each player's payoff is a scalar-valued function of the joint strategy and we denote this function by $u_i(x) : X \rightarrow R$.

Games can be differentiated into non-cooperative games and cooperative games. In a non-cooperative game, each player chooses his or her strategy independently while in a cooperative game, players are able to form binding commitments and communications are always assumed to be allowed among players.

2.1.1 Non-cooperative Games

Game theory attempts to predict the outcome of such a game or, when this is not feasible, properties of the predicted outcome, such as its existence and uniqueness. This leads to the important definition of the Nash Equilibrium in a non-cooperative game, a joint strategy where no player can increase his or her payoff by unilaterally changing his or her strategy.

Definition 2.1.1 ***Nash Equilibrium**: Strategy $x \in X$ is a Nash equilibrium if $u_i(x) \geq u_i(x_i^*, x_{-i}), \forall x_i^* \in X_i, \forall i \in N$, where x_{-i} represents all components of vector x except its i^{th}component.*

The Nash equilibrium is considered to be a consistent predictor of the outcome of the game, in the sense that if all players predict that a Nash equilibrium will occur, then no player has an incentive to choose a different strategy [66]. In general, the uniqueness or even the existence of a Nash equilibrium is not guaranteed; neither is convergence to an equilibrium when the equilibrium exists.

A direct interpretation of Definition 2.1.1 is that the Nash equilibrium is a mutual best response from each player to the other players' strategies. In order to formally state this result, we first define the best reply mapping [67]:

Definition 2.1.2 ***Best Reply Mapping**: The best-reply mapping for player i is a point to set mapping that associates each joint strategy $x \in X$ with a subset of X_i according to the following rule: $\psi_i(x) = arg\ max_{x_i^* \in X_i} u_i(x_i^*, x_{-i})$. The best reply mapping for the game is then defined as $\psi(x) = \times_{i \in N} \psi_i(x)$.*

It is sometimes convenient to make use of the following alternate definition of Nash equilibrium [67].

Definition 2.1.3 ***Nash Equilibrium**: Strategy x is a Nash equilibrium if and only if $x \in \psi(x)$.*

We emphasize that the idea of the Nash equilibrium as a consistent predictor of the outcome of the game does not necessarily require perfect knowledge on the part of the players regarding other players' payoff functions. Even without this knowledge, in a quasi-static environment, players may converge to an equilibrium through a learning process.

A desirable property of an equilibrium is that it is efficient. We use the concept of Pareto optimality to determine the efficiency of a Nash equilibrium. In addition, if multiple equilibria exist but only one is Pareto optimal, then we consider it to be superior to others. An equilibrium is Pareto optimal if there is no other joint strategy which one or more players would prefer and to which all others would be indifferent.

Definition 2.1.4 ***Pareto Optimality***: *A strategy x is Pareto optimal if there does not exist $x' \in X$ such that:*

1. $u_i(x') \geq u_i(x), \forall i \in N$; *and*
2. $u_i(x') > u_i(x)$ *for at least one $i \in N$.*

2.1.2 Cooperative Games

In 1950, Melvin Dresher and Merrill Flood at the RAND Corporation devised a game (see Table 2.1) to illustrate that a non-cooperative game could have an equilibrium outcome which is unique, but fails to be Pareto optimal [68].

Table 2.1: The original Prisoner's Dilemma. A is "don't confess", B is "confess".

		Colin	
		A	B
Rose	A	(0,0)	(-2,1)
	B	(1,-2)	(-1,-1)

- If one of them confesses and the other does not, the confessor will get a reward (payoff +1) and his partner will get a heavy sentence (payoff -2);
- If both confess, each will get a light sentence (payoff -1);
- If neither confesses, both will go free (payoff 0).

In the years since 1950 this game has become known as the Prisoner's Dilemma. Strategy B is dominant for both players, leading to the unique equilibrium at BB. However, this equilibrium is non-Pareto-optimal, since both players would do better at AA.

Instead of choosing their strategies independently, when we assume that a player can communicate and form a coalition (e.g., both promise to play strategy A), this is a cooperative game. Unlike the game we describe in Table 2.1 with only two players, in our modern connected world, most economic, social and biology games involve more than two player. The questions are:

- Which coalitions should form?
- How should a coalition divide its winnings among its members?

Before looking into these questions, we first define characteristic functions.

Definition 2.1.5 ***Characteristic Function**: A game in characteristic function form is a set of N players, together with a function v which for any subset $S \subseteq N$ gives a number $v(S)$.*

The number $v(S)$ is the amount that the players in S could win if they formed a coalition and the function v is called the characteristic function of the game. To calculate $v(S)$, assume that the coalition S forms and then plays optimally against an opposing coalition $N - S$.

There is an important relation among the values of different coalitions which holds for games in characteristic function form: superadditive[1].

Definition 2.1.6 ***Superadditive**: A characteristic function form game (N, v) is called superadditive if $v(S \cup T) \geq v(S) + v(T)$ for any two disjoint coalitions S and T.*

If two coalitions S and T, with no common members, decide to join together to form $S \cup T$, Definition 2.1.6 says that they can always assure themselves of at least $v(S) + v(T)$ because they can simply continue to do what they would do if they hadn't joined. And, they may often be able to do better than this by coordinating their actions.

From Definition 2.1.6, we find that it is in all players' interest to form a coalition N and get $v(N)$. Instead of asking about the possible results of actual coalitional behavior, we consider one class of cooperative games and ask if there might be a single payoff $|N|$-dimensional vector which could represent a fair distribution of payoffs to all players. This payoff vector might not arise from the competitive behavior of coalitions, but it would be the payoff vector an outside arbitrator might impose, taking into account the relative strengths of the various coalitions. For example, in Neumann and Morgensterm's Divide the Dollar game, three players will be given a dollar if they

[1] This holds for all games in characteristic function form which rise from games in normal form.

can decide how to divide the dollar among themselves by majority vote. We can see that a likely outcome might be that one of the three two-person coalitions would form and divide the dollar equally between its two members. As a result, we get one of the three payoff vectors $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$ or $(0, \frac{1}{2}, \frac{1}{2})$. An outside arbitrator considers the symmetry of this game and decides the fair division is certainly $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This is the case in cost sharing of multi-service communication networks. We don't prefer one coalition to another, but would like a fair distribution of cost among different classes of service.

In 1953, Lloyd Shapley gave a general answer to this fair division question and it has come to be known as the Shapley Value of a cooperative game in characteristic function form. We first look at three definitions which capture a fair distribution of payoffs [68]. Here, we use payoff vector $\varphi = (\varphi_1, \varphi_2, ..., \varphi_n)$ to denote the fair payoff to each player.

Definition 2.1.7 ***Efficiency***: *The total gain is distributed:* $\sum_{i \in N} \varphi_i = v(N)$ *and therefore payoff allocation* φ *is Pareto optimal.*

Definition 2.1.8 ***Symmetry***: φ *should depend only on* v *and should respect any symmetries in* v*. That is, if plays* i *and* j *have symmetric roles in* v*, then* $\varphi_i = \varphi_j$*.*

Definition 2.1.9 ***Zero Player***: *If* $v(S) = v(S - i)$ *for all coalitions* $S \subseteq N$*, that is, if player* i *is a dummy who adds no value to any coalition, the* $\varphi_i = 0$*. Furthermore, adding a dummy player to a game does not change the value of* φ_j *for any other players* j *in the game.*

Definition 2.1.10 ***Additivity***: $\varphi[v + w] = \varphi[v] + \varphi[w]$

The above definition is about the sum of two games. Suppose that (N, v) and (N, w) are two games with the same player set N. Then we can define the game $v + w$ by defining $(v + w)(S) = v(S) + w(S)$ for all coalitions S. Now we have three games under consideration and use $\varphi[v], \varphi[w]$ and $\varphi[v + w]$ to denote payoff vector for each game. It means that if it is fair for player i to get $\varphi_i[v]$ in v and $\varphi_i[w]$ in w, it would seem fair to get the sum of these two payoffs in the game $v + w$. For example, the cost sharing of communication network with fixed cost and operation cost is the cost sharing of the fixed cost plus the cost sharing of the operation cost.

Shapley has proved that there is one and only one method of assigning payoff vector φ for a game (N, v) which satisfies all above definitions.

Shapley Value: The Shapley Value of each player i in the cooperative game has the following expression:

$$\varphi_i = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S) - v(S \cup \{i\})] \tag{2.1}$$

The meaning behind the Shapley value is that each player's payoff depends on the incremental cost for which he or she is responsible when provision of the services accumulates in random order. Let's illustrate it with an example.

Suppose there are three airplanes A, B, C sharing a runway. Airplane A requires 1km to land, Airplane B requires 2km and Airplane C requires 3km. When a runway of 3km is built, how much should each airplane pay? Before we applying Equation (2.1), we look at their incremental cost in the six possible orders in Table 2.2 (cost is measured in unites per kilometer):

Table 2.2: Cost sharing problem using the Shapley Value

	Incremental cost		
Order	A	B	C
A, B, C	$v(A) - \phi = 1$	$v(A, B) - v(A) = 1$	$v(A, B, C) - v(A, B) = 1$
A, C, B	$v(A) - \phi = 1$	$v(A, B, C) - v(A, C) = 0$	$v(A, C) - v(A) = 2$
B, A, C	$v(A.B) - v(B) = 0$	$v(B) - \phi = 2$	$v(A, B, C) - v(A, B) = 1$
B, C, A	$v(A, B, C) - v(A, B) = 0$	$v(B) - \phi = 2$	$v(B, C) - v(B) = 1$
C, A, B	$v(A, C) - v(C) = 0$	$v(A, B, C) - v(A, C) = 0$	$v(C) - \phi = 3$
C, B, A	$v(A, B, C) - v(B, C) = 0$	$v(B, C) - v(C) = 0$	$v(C) - \phi = 3$
Total	2	5	11

Therefore, based on the Shapley value Equation (2.1), each airplane should be responsible for $(\frac{2}{6}, \frac{5}{6}, \frac{11}{6})$, respectively. Let's calculate this problem by treating it as sum of three games. The first kilometer is shared by all airplanes and so its cost should be $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; the second kilometer is shared by airplane B and C and their cost should be $(0, \frac{1}{2}, \frac{1}{2})$; the last kilometer is used only by airplane C and the allocated cost should be $(0, 0, 1)$. Based on the Definition 2.1.10, the cost sharing of this runway is the sum of above vector and the result is $(\frac{2}{6}, \frac{5}{6}, \frac{11}{6})$.

2.2 Game Theory in Communication Networks

Although the original applications of game theory tended to be in the field of economics, over the years its usefulness has been recognized in other disciplines such as biology, political science and philosophy. Recently, game theory has been applied to numerous networking problems.

- **Spectrum Management**: In a communication system where multiple users share a common frequency bank such as cognitive radio, each user's performance, measured by a Shannon utility function, depends on not only the power allocation (across spectrum) of its own, but also those of other users in the system. This spectrum management problem can be formulated either as a non-cooperative game [63, 69, 70] or as a cooperative utility maximization problem [71, 72]. In [63], each user is given an initial budget to purchase its own transmit power spectra (taking others as given) in order to maximizing its Shannon utility or payoff function, which includes the effects of interference. It has proved that an equilibrium always exists for a discrete version of the problem, and, under a weak-interference condition or the Frequency Division Multiple Access (FDMA) policy, the equilibrium can be computed in polynomial time.

- **Flow Control**: Papers [73, 74] study the issue of flow control using a game-theoretic framework. Reference [73] defines each flow's utility function as its transmission rate and the QoS it receives and model this interdependence under a game-theoretic framework. Instead of using edge routers to shape incoming flows, this paper has assumed that flows in the network are responsible and has proved there exists an Pareto optimal equilibrium. In [73], each user's objective is to maximize the average throughput subject to an upper bound on average delay. Since users share a network of quasi-reversible queues, each user's strategy affects all other users' performance and the authors have determined the existence of an equilibrium for such a system.

- **Congestion Control**: Reference [75] considers the problem of whether a given switch service rule will lead to an operating point that is fair and efficient or not using game theory. In [75], users' payoff function is defined as the amount of traffic and received QoS provided by a switch and users are selfish to maximize their payoff by varying the level of traffic.

- **Routing**: References [76, 77] have investigated routing issues in communication networks using the game theory model. Reference [76] has studied the existence of routing strategies of the network manager that drives the system to an optimum operating point. Reference [77] has investigated how to partition the traffic from a number of users among a number of parallel links to maximize their payoff function, which is defined as a measure of performance and satisfaction.

- **Pricing**: Several multi-service network pricing schemes using game-theoretic models are described in Section 1.3.3.

Game theory provides a powerful tool for studying the performance and QoS issues in communication networks. In this dissertation, we study the pricing issue of multi-service networks unde the game-theoretic framework.

2.3 Chapter Summary

This chapter has introduced background information of game theory: The Nash Equilibrium in non-cooperative games and the Shapley Value in cooperative games. Further, it has reviewed its application in communication networks including spectrum management, flow control, congestion control and routing problems.

In the next chapter, we will introduce network and pricing models used in this dissertation.

CHAPTER 3

Users, Network, and Pricing Model

In multi-service networks, performance of each class of service and users' satisfaction will be directly influenced by all users' service choices. In best-effort (single QoS) networks, it is also true that individual performance is affected by the characteristics of aggregate traffic, however, this interdependence among users is even more complicated in multi-service networks. In this dissertation, traffic from each class of service has different sensitivity to delay (disutility to average delay). The network resource allocation among all classes is modeled as a cooperative game to minimize the total disutility.

In addition, since each user's satisfaction (utility) is influenced not only by the network provider's pricing policy but also by all other users' service choices, we model this interdependence among users' service choices through a non-cooperative game-theoretic framework. The service provider sets price for each class of service to maximize its revenue. Users purchase service from service provider to maximize their utility functions independently under their budget constraints. The operating point is determined by the equilibrium where the price for each class of service is market-clearing price.

In this chapter, our basic assumptions regarding users' behaviors, as well as their utility functions and service provider's objective, are discussed in Section 3.1. Section 3.2 contains assumptions made about the network, while in Section 3.3 we discuss a general pricing model. We close with a brief summary of this chapter in Section 3.4.

3.1 Users and Network Provider Models

The pricing game we study in this dissertation has one principal, the network provider, and a finite set of players, network users. The principal sets a price for each class of service; based on these prices, each user decides on service choices. Users do not cooperate with one another when deciding on an optimal strategy (hence characterizing a non-cooperative game), but rather each user acts individually striving to maximize their own payoff functions (utility functions).

3.1.1 Utility Functions for Users

Network users' preferences are modeled through utility functions, which describe how sensitive users are to changes in QoS and/or amount of network resource made available to them. In the context of this dissertation, it is useful to think of utility as users' willingness to pay for a certain resource available to them.

Through users' actions, we can sometimes assess their willingness to pay for certain improvements in quality. A complete characterization of actual utility function is unlikely in practice. However, it is generally reasonable to assume that user *i*'s utility function u_i possesses the following properties:

Assumption 3.1.1 ***Allocated bandwidth and QoS***: *u_i depends on the bandwidth allocated to user i and its received QoS.*

In data communication networks, a user's utility not only depends on the allocated bandwidth but also depends on the received network QoS. For example, a 2 Mbps bandwidth with the average delay 50 ms has better utility for most users than the same bandwidth with 500 ms average delay.

Assumption 3.1.2 ***Monotonicity***: *u_i is a monotonic function of its variables.*

This assumption is also quite intuitive. For instance, one would expect the utility to be monotonically increasing with bandwidth and monotonically decreasing with the QoS parameter, such as the average delay. In general, we don't assume strict monotonicity, since there may exist a point beyond which further increase in the QoS or bandwidth does not yield any additional benefit for the user. An example of this is the case of a constant bit rate application like VoIP – availability of bandwidth greater than that constant rate typically does not result in any improved performance.

Assumption 3.1.3 ***Concaveness***: *u_i is concave function of its variables.*

This assumption arises from a diminishing returns argument. We expect a user's marginal utility to decrease with the bandwidth and QoS. It means that the more bandwidth and better quality, the less the user is willing to pay for further improvement. This assumption is also consistent with [78].

Combinations of above assumptions are adopted in many pricing studies that employ the concept of utility functions, including [16, 11, 33, 32, 79].

The well-known utility function in data communication networks, proposed by Kelly [16], has the form $u_i = w_i \log(x_i)$, where w_i is user's willingness to pay and x_i is the allocated bandwidth. Although this utility function fits into our assumptions about utility Assumption 3.1.2 and 3.1.3, it does not take QoS as a parameter into consideration for the utility function. Here, we redefine user's utility as a function of the allocated bandwidth and the QoS of the network as follows:

$$u = \frac{\beta}{T_{now}} \log(\frac{x}{\tilde{x}}) \tag{3.1}$$

where $\beta > 0$ is the weighting factor and it describes the flow's relative sensitivity to the QoS parameter based on the fact that applications exhibit varying degree of sensitivity to QoS parameters (here we use delay as the QoS parameter for the network). For example, real-time voice and video are very sensitive to delay; packets that do not arrive within some delay bound cannot be used for playback and are in effect considered lost (although there are ways to make these applications less sensitive to such losses using coding or extra buffer). On the other hand, traditional data applications such as email service, file transfer are typically not very sensitive to delay. Thus, we use β to denote applications' QoS sensitivity characteristic.

In this dissertation, we consider both elastic and inelastic users as defined by Shenker [31]. Traditionally, real-time voice and video applications that employ constant bit rate coding with no tolerance to eventual packet losses require a fixed amount of bandwidth for adequate QoS. There are numerous ways in which real-time applications can be made tolerant of changes in available bandwidth through proper coding and interpolation of the received data; however, some minimum bandwidth is nevertheless often required. Although traditional data applications are elastic in nature and tend to be tolerant of variations in delay and can take advantage of even minimal amount of bandwidth, to guarantee users' network experience, we still assume a minimum bandwidth requirement. Here in Equation (3.1) we use $\tilde{x}$ to denote this minimum bandwidth requirement. Notice that we use the present average delay T_{now} to represent the QoS parameter of network. The reason for this choice is that users always keep record of the present network situation such as Round Trip Time, packet loss rate, etc. We can easily calculate T_{now} from these information.

3.1.2 Utility Functions for Network Provider

Our model treats the network provider as a monopolist, a common assumption employed by most recent pricing studies discussed in Section 1.3.3.

In order to deter future competitors, it is better for network provider to set up subsidy-free price for each class of service. Due to the statistical multiplexing characteristic of networks as discussed in Section 1.1.2, each class of traffic will incur a waiting cost c_i based on related delay. The utility function for the network is then defined as the sum of costs incurred by all classes and the network provider try to minimize this total waiting cost:

$$u_s = \sum_{i=1}^{n} c_i \tag{3.2}$$

Alternatively, the network provider may use revenue as its utility function [11, 38]. When we use a vector p to denote the price of each class of service and vector s denote the supply of each class of service, the utility function for the service provider in this situation is defined as:

$$u_s = p \times s \tag{3.3}$$

In this section, we have discussed the general forms of utility functions of both users and service provides.

3.2 Network Model

Here, we consider two different network models. The first model consists of a single source-destination pair common to all users. The study of a single queue is applicable to local and metropolitan area networks (e.g., Expedited Forwarding Per-Hop behavior traffic [80]), which are sometimes modeled as a single server with a queue that is distributed among all stations. This common source-destination model is also employed in [58, 76, 11].

Unlike in the first model, the network maintains a single priority-based queue for all classes of services; in the second DiffServ model, each class has a promised QoS and maintains a separate queue. The network resource is allocated among these classes using scheduling schemes like Weighted Round Robin (WRR), Class-based Queuing (CBQ) [81] and dynamic WRR [82]. This class-based network structure has been used in [32, 36, 73, 82].

Routing is assumed to be fixed and independent of the pricing policy and such simplifying assumptions free us from routing concerns. In a real commercial network, we find that routing will not play a fundamental role in the pricing issue since it is unlikely that the service provider will ask users to pay different prices depending on the actual route, given same QoS level.

A non-preemptive priority-based multi-class system (single queue) will be used in chapter 4 and the class-based DiffServ network architecture (seperate queue for each class) will be examined in chapters 5, 6 and 9.

3.3 Pricing Model

The amount charged for network services may be a function of the combination of several factors, most notably as follows.

- **Access cost**: A charge for accessing the network services.
- **Service type**: In networks with multiple service categories, each class of service will be priced differently to reflect the QoS it provides.
- **Usage charge**: the usage charge is determined by the level of service provided to user and actual amount of traffic consumed by user. Usage-based charge component can be used to discourage over-consumption and provide better network performance.
- **Time-of-day sensitive charge**: The time of consumption network is relevant when implementing time-of-day pricing. The main objective of time-of-day pricing is to produce the smoothing of traffic by encouraging users to shift their demand to times when the network is more lightly loaded.

Table 3.1: Pricing mechanisms employed in various studies

	[51]	[58]	[79]	[32]	[36]
Access cost		×		×	
Service type	×	×	×	×	
Usage charge		×	×	×	×
Time-of-day sensitive charge	×				

In Table 3.1, we list components of pricing scheme of some studies on network pricing as discussed in Chapter 1. We combine the factors discussed above into a general pricing policy model and the charge P_{ij} of class of service j to user i according to the following expression:

$$P_j = c_j + p_j(t) * x_{ij} \tag{3.4}$$

In Equation (3.4), c_j is a fixed access charge assigned to service j and x_{ij} is the amount of service j consumed by user i. The unit price of service j, p_j, is dependent on time t and this makes the model general enough to encompass time-of-day pricing.

In this dissertation, we attempt to find the desired price vector $p = (p_1, p_2, ..., p_j, ..., p_m)$ for service provider that supports m classes of services under pricing contexts described in Section 1.4.1.

3.4 Chapter Summary

A user's preferences are modeled through a utility function which is expressed in terms of the QoS parameter such as delay and the allocated bandwidth. These utility functions are assumed to be monotonic and concave. Each user will independently choose a strategy with the objective of maximizing his or her own utility function under his or her budget constraint.

The network service provider is assumed to be a monopolist with the aim of either minimizing the total waiting cost or maximizing its revenue. We also discussed two different network models used in this dissertation and a general form of pricing policy.

In the next two chapters, we discuss the subsidy-free price for each class of service under different network models as discussed in Section 3.2.

CHAPTER 4

Subsidy-free Prices in Priority-based Networks

Some of the simplest types of multi-service networks are those in which the distinction among service classes is accomplished exclusively through the assignment of different service priorities on a per-packet basis. In this chapter, we study the desirable subsidy-free price of each class of service on such priority-based networks taking into account the fairness among classes and economic efficiency requirements (minimizing the waiting cost associated with all classes of service) [83].

This chapter considers price differences among different classes as a cooperative queuing problem. Each class of service has a different waiting cost per unit of time (waiting cost factor). A cooperating queue is organized to minimize the total waiting cost of all classes while monetary compensations are set up for those classes which have to wait longer time. It is reasonable to assume this queuing problem as a transfer utility game and solve it by applying the Shapley Value. We consider the total cost of a coalition as the total waiting cost its members (classes) would incur if they had the power to be served first. The waiting cost associated with each class corresponds to the Shapley Value of the queuing game. In this chapter, we also find the solutions associated with the Shapley Value which satisfy many fairness properties like any classes which are served before another class are responsible to compensate the latter for their waiting cost; the sum of all compensation transfers is equal to zero.

Using Shapley Value to deal with the cost sharing queuing problem has been considered in [62, 84, 85, 86, 87]. In [62], the author has studied monetary transfer between agents based on a model where each agent has a different unit waiting cost. For such a model, this paper has characterized the Shapley Value rule using classical fairness axioms. Reference [84] has interpreted the worth of a coalition of agents in a different manner for the same model as in [62], and derived a different rule. It has also characterized this different rule using similar fairness axioms. In [85, 86], the queuing problem is studied from a strategic point of view under the assumption that all agents have identical unit waiting cost. The study in [87] is also based on the same model described in [62], and has considered cost sharing when both unit waiting cost and processing time of agents are present.

In this chapter, we extend the model described in [62] to multi-class communication networks and add the stochastic property of the communication network to it. We also propose a fair and efficient way to get the price difference among different classes of service.

The rest of this chapter is organized as follows. The model adopted for priority service is discussed in Section 4.1. The queuing games and the Shapley Value are studied in Section 4.2. In Section 4.3, we define the fairness axioms and check our results with these axioms. An illustrative example is given in Section 4.4, and Section 4.5 captures our conclusions.

4.1 Problem Statement and the Model

The network model proposed in this chapter is depicted in Fig. 4.1. The packet switched network is represented by a single communication server with a defined capacity c and a non-preemptive priority scheme is assumed. There are n different classes of service in the network. The packet length for all classes is considered to have the same statistics, and an exponential distribution with the average packet length equal to $\frac{1}{\mu}$ is assumed.

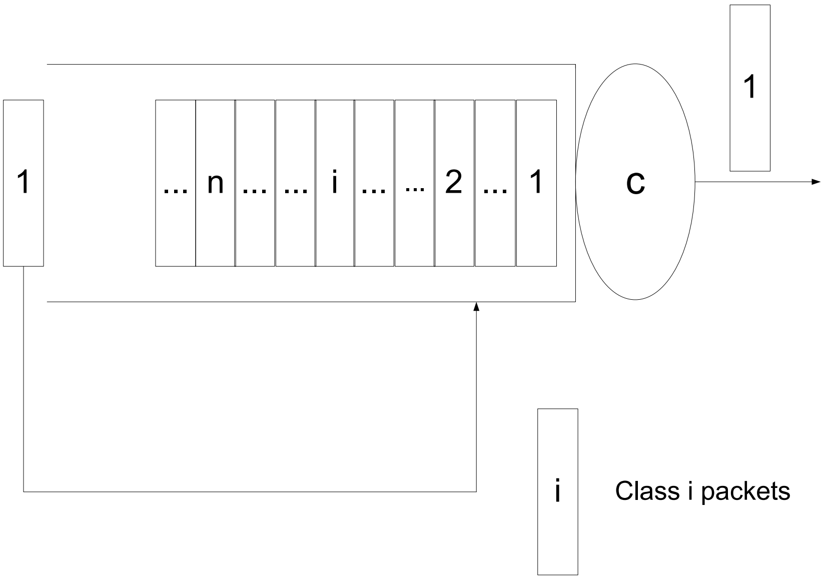


Figure 4.1: FIFO non-preemptive priority schedule

The set of classes are denoted as $N = \{1, ..., n\}$. $\sigma : N \to N$ is priority ordering of n classes of service and σ_i denotes the priority of class i. Each class i is distributed with Poisson arrivals and identified by two parameters: (λ_i, θ_i). λ_i is the average arrival rate and θ_i is the waiting cost factor for class i. A queuing problem is defined by a list $q = (N, \lambda, \theta) \in Q$, where Q is the set of all possible lists. Given an ordering σ, the waiting cost incurred by class i is given by:

$$v_i(\sigma) = \lambda_i w_i \theta_i \tag{4.1}$$

where w_i is the average waiting time of class i packets given the ordering σ.

The total waiting cost incurred by packets from all classes given an ordering σ can be written as:

$$v(N, \sigma) = \sum_{i=1}^{n} v_i(\sigma) = \sum_{i=1}^{n} \lambda_i w_i \theta_i \tag{4.2}$$

Kleinrock [88] presents closed-form results for the waiting time in a non-preemptive priority discipline for a single M/G/1 queue, which we can apply here. We use $\bar{x}_i$, $\bar{x}_i^2$ to denote the first two moments of service time for class i packets. We have already assumed that all packets have identical statistical distribution with the average packet length $\frac{1}{\mu}$, therefore, $\bar{x}_i = \frac{1}{\mu c}, \bar{x}_i^2 = \frac{2}{(\mu c)^2}, \forall i = 1, ..., n$.

First, we define:

$$w_0 = \sum_{i=1}^{N} \frac{\lambda_i \bar{x_i^2}}{2} = \frac{\sum_{i=1}^{n} \lambda_i}{(\mu c)^2} \tag{4.3}$$

We also define $\lambda^i = \bar{x}_i \sum_{j=1}^{i} \lambda_i = \frac{\sum_{j=1}^{i} \lambda_i}{\mu c}$, and $\lambda^0 = 1$. The waiting time for packets in class i is [88]:

$$w_i = \frac{w_0}{(1 - \lambda^{i-1})(1 - \lambda^i)} \tag{4.4}$$

Here we define an allocation for a queuing problem $q = (N, \lambda, \theta) \in Q$ as $\psi(\sigma, t)$, where σ is an ordering, and t_i is the transfer related to class i packets. Given an ordering σ and a transfer t_i, the waiting cost share for class i packets is defined as,

$$u_i = v_i(\sigma) - t_i = \lambda_i \theta_i w_i - t_i \tag{4.5}$$

Here we define an efficient allocation as follows. *An allocation $\psi(\sigma, t)$ is efficient* for queuing problem $q = (N, \lambda, \theta) \in Q$ whenever it minimizes the total cost of waiting minimize $v(N, \sigma)$ and the algebraic sum of transfer is equal to 0, $\sum_{i=1}^{n} t_i = 0$.

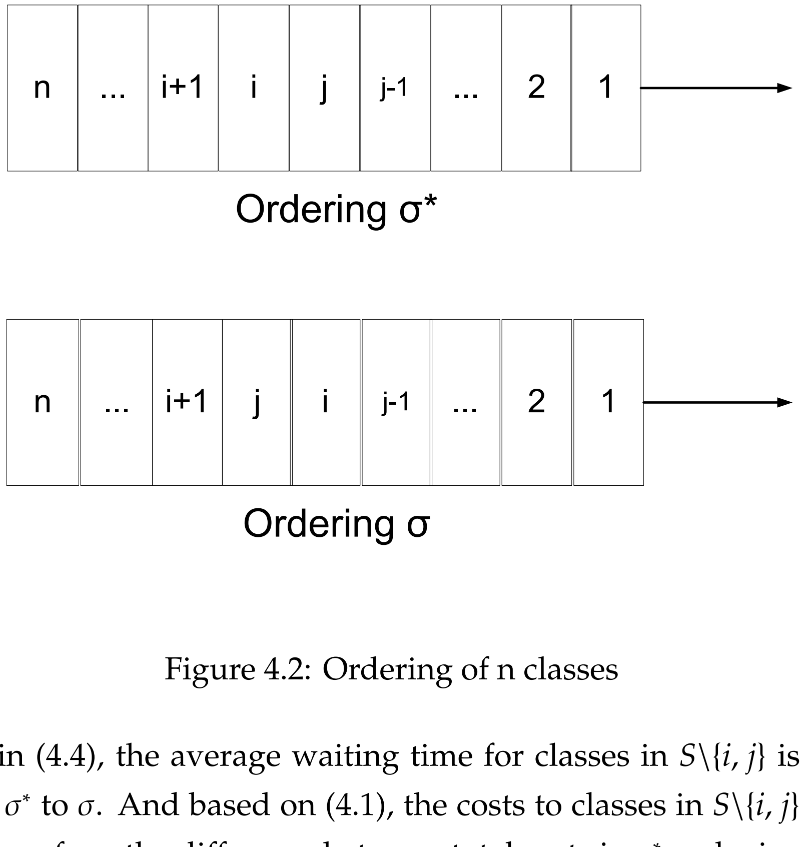
An efficient ordering σ^* for the queuing problem $q = (N, \lambda, \theta) \in Q$ is the one which minimizes the total waiting cost $v(N, \sigma^*)$ incurred by packets from all classes. It means that $v(N, \sigma^*) \leq v(N, \sigma), \forall \sigma$. For notational simplicity, we will write the total waiting cost in the efficient ordering of classes from N as $v(N)$, whenever it is not confusing. In some cases, we will deal with only a subset classes $S \subseteq N$. The ordering σ will then be defined on the classes in S only, and we will write the total waiting cost from an efficient ordering of classes in S as $v(S)$.

The following lemma shows that classes are ordered in decreasing θ is an efficient ordering.

Lemma 4.1.1 *For any $S \subseteq N$, let σ^* be an efficient ordering of classes in S. For every $i \neq j$, $i, j \in S$, if $\sigma^*_i > \sigma^*_j$, then $\theta_i < \theta_j$.*

proof:

Assume that the statement of the lemma is not true. This means that we can find two consecutive classes $i, j \in S(\sigma_i = \sigma_j + 1)$ such that $\theta_i > \theta_j$. We can then define a new ordering σ by interchanging i and j in σ^* as shown in Fig. 4.2.



Ordering σ*

Ordering σ

Figure 4.2: Ordering of n classes

As shown in (4.4), the average waiting time for classes in $S \backslash \{i, j\}$ is not changed from ordering σ^* to σ. And based on (4.1), the costs to classes in $S \backslash \{i, j\}$ also remains unchanged. Therefore, the difference between total costs in σ^* and σ is given by,

$$v(S, \sigma^*) - v(S, \sigma) = \lambda_i w_i^{\sigma^*} \theta_i + \lambda_j w_j^{\sigma^*} \theta_j - (\lambda_j w_j^{\sigma} \theta_j + \lambda_i w_i^{\sigma} \theta_i) \qquad (4.6)$$

As shown in Fig. 4.2, we can get the average waiting time for class i, j packets in ordering σ^* and σ as follows:

$$w_i^{\sigma^*} = \frac{w_0}{(1 - \lambda^{j-1} - \frac{\lambda_j}{\mu c})(1 - \lambda^{j-1} - \frac{\lambda_j}{\mu c} - \frac{\lambda_i}{\mu c})}$$

$$w_j^{\sigma^*} = \frac{w_0}{(1 - \lambda^{j-1})(1 - \lambda^{j-1} - \frac{\lambda_j}{\mu c})}$$

$$w_i^{\sigma} = \frac{w_0}{(1 - \lambda^{j-1})(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c})}$$

$$w_j^{\sigma} = \frac{w_0}{(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c})(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c} - \frac{\lambda_j}{\mu c})}$$

Now, we take the above equations into (4.6), we get:

$$v(S, \sigma^*) - v(S, \sigma) = \frac{\lambda_i \lambda_j (\theta_i - \theta_j)(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c} + 1 - \lambda^{j-1} - \frac{\lambda_j}{\mu c})}{(1 - \lambda^{j-1})(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c})(1 - \lambda^{j-1} - \frac{\lambda_j}{\mu c})(1 - \lambda^{j-1} - \frac{\lambda_i}{\mu c} - \frac{\lambda_j}{\mu c})}$$

We have already assumed that σ^* is an efficient ordering, and we get $v(S, \sigma^*) - v(S, \sigma) \leq 0$, this give us $\theta_i \leq \theta_j$, which is a contradiction.

Thus, we prove Lemma 4.1.1.

Notice that the efficient queuing problem is independent of the transfer and is unique when all classes have different unit waiting cost. And we can rewrite that an allocation $\psi(\sigma, t_i)$ is efficient for the queuing problem $q = (N, \lambda, \theta) \in Q$ whenever σ is an efficient ordering and $\sum_{i=1}^{n} t_i = 0$.

Until now, in (4.5), we can calculate the actual waiting cost for each class $v_i(\sigma)$ based on efficient ordering as described in Lemma 4.1.1. In the next section, we will consider the waiting cost share problem as a cooperative game and set up the waiting cost share for each class u_i using the Shapley value. For each class, if we know the actual waiting cost $v_i(\sigma)$ and waiting cost share u_i, based on (4.5), we can find the transfer t_i for each class.

The inequality $t_i > 0$ shows that class i will receive compensation and $t_i < 0$ shows it will compensate other classes. After getting the transfer t_i between different classes, we are able to define the price difference Δp_{ij} between class i and j as follows.

$$\Delta p_{ij} = \frac{t_i}{\lambda_i} - \frac{t_j}{\lambda_j} \tag{4.7}$$

4.2 Waiting Cost Sharing Using the Shapley Value

As discussed in Section 4.1, we solve the waiting cost share queuing problem by treating it as a cooperative game. In this section, we first define the coalitional cost of this game and then analyze the solution based on Shapley Value of the corresponding game.

Given a queue $q \in Q$, the waiting cost of a coalition of $S \subseteq N$ classes in the queue is defined as the cost incurred by the classes in S if they have the power to be served first in the queue and use an efficient ordering in S.

And, the cost of a coalition $S \subseteq N$ is as:

$$v(S) = v(S, \sigma) = \sum_{i \in S} \lambda_i w_i \theta_i \tag{4.8}$$

where $\sigma = \sigma(S)$ is an efficient ordering considering classes in S only. In [62], the author also studied another equivalent way to define the worth of a coalition using the dual function of the cost function. Other ways to define the worth of a coalition is addressed in [84] which assumed that a coalition of classes are served after the classes not in the coalition.

The marginal contribution of class $i \in N$ to a coalition S in $v(S)$, $i \notin S$ is a sum of the costs associated with each member of S. Indeed, those classes having a higher unit waiting cost θ than class i impose a waiting cost on it, and those having a lower unit waiting cost θ than class i have to wait additional units of time. That is the marginal contribution is composed of the cost of waiting of class i itself, and the cost its existence imposes on those classes who follow it in the new queue. Formally, for $q = (N, \lambda, \theta) \in Q, S \subset N, i \in N \backslash S$, the marginal contribution of class i is :

$$v(S \cup \{i\}) - v(S) = v(S \cup \{i\}, \sigma') - v(s, \sigma) = \sum_{i \in S \cup \{i\}} \lambda_i w_i^{\sigma'} \theta_i - \sum_{i \in S} \lambda_i w_i^{\sigma} \theta_i \tag{4.9}$$

where $\sigma' = \sigma(S \cup \{i\}), \sigma = \sigma(S)$ are efficient orderings considering classes in $S \cup \{i\}$ and S respectively.

The Shapley Value (waiting cost share) of class i is defined as a weighted sum of the class's marginal contribution to coalitions. Based on the definition of the Shapley Value as described in Section 2.1.2, for all $q = (N, \lambda, \theta) \in Q, i \in N$, the payoff assigned to class i is given by:

$$u_i = SV_i = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[v(S \cup \{i\}) - v(S)] \tag{4.10}$$

The Shapley Value allocation rule says that classes are ordered using an efficient ordering and transfers (compensations) are assigned to classes such that the cost share of each class is equal to its Shapley Value. Based on the efficiency property of Shapley Value described in Section 2.1.2, the total gain is distributed among N players, that is $\sum_{i \in N} SV_i$ = minimum $v(N, \lambda, \theta)$.

Another way to write the Shapley Value formula is as follows [89],

$$SV_i = \sum_{S \subseteq N : i \in S} \frac{\Delta(S)}{|S|} \tag{4.11}$$

where $\Delta(S) = v(S)$ if $|S| = 1$ and $\Delta(S) = v(S) - \sum_{T \subset S} \Delta(T)$. This gives $\Delta(\{i\}) = v(\{i\}) = \lambda_i w_i^{\{i\}} \theta_i$, $\forall i \in N$, $w_i^{\{i\}}$ is the waiting time for class i when class i packets have the power to be served first in the queue. For any $i, j \in N$, that is $|S| = 2$ we have,

$$\Delta(S) = \Delta\{i, j\} = v(\{i, j\}) - v(\{i\}) - v(\{j\}) = \lambda_i w_i^{\{i,j\}} \theta_j + \lambda_j w_j^{\{i,j\}} \theta_j - \lambda_i w_i^{\{i\}} \theta_i - \lambda_j w_j^{\{j\}} \theta_j$$

We assume $\theta_i \geq \theta_j$ and consider (4.4), we have,

$$\Delta(S) = \Delta\{i, j\} = v(\{i, j\}) - v(\{i\}) - v(\{j\}) = \lambda_j \theta_j (w_j^{\{i,j\}} - w_j^{\{j\}})$$

If $|S| = 3$, say $S = \{i, j, k\}$ then,

$$\Delta(S) = \Delta\{i, j, k\} = v(\{i, j, k\}) - \Delta(\{i, j\}) - \Delta(\{j, k\}) - \Delta(\{i, k\}) - \Delta(\{i\}) - \Delta(\{j\}) - \Delta(\{k\})$$

If we further assume $\theta_i \geq \theta_j \geq \theta_k$, we have,

$$\Delta(S) = \Delta\{i, j, k\} = \lambda_k \theta_k (w_k^{\{i,j,k\}} - w_k^{\{i,k\}} - w_k^{\{j,k\}} + w_k^{\{k\}})$$

It is easy to use induction to show that when $S = \{1, 2, ..., n\}$ with $\theta_1 \geq \theta_2 \geq ... \geq \theta_n$, $\sigma_1 = 1, \sigma_2 = 2, ..., \sigma_n = n$ (we will use i to denote the position of each class i in the queue instead of σ_i for simplicity) $\Delta(S)$ is:

$$\Delta(S) = \Delta\{1, 2, ..., n\} = \lambda_n \theta_n w_n^{\Delta(S)} \tag{4.12}$$

where $w_n^{\Delta(S)} = \sum_{T \subseteq S, n \in T} (-1)^{|T|} w_n^T$, when $|S|$ is even, and $w_n^{\Delta(S)} = \sum_{T \subseteq S, n \in T} (-1)^{|T+1|} w_n^T$, when $|S|$ is odd. We can see that the $\Delta(S)$ only depends on the lowest priority class, that is $\lambda_n \theta_n w_n^{\Delta(S)}$.

Now, we are ready to consider $SV_i = \sum_{S \subseteq N: i \in S} \frac{\Delta(S)}{|S|}$ in more detail. Given a length of a set $|S|, S \subseteq N, i \in S$, there are $\binom{i-1}{|S|-1}$ situations where class i is the lowest priority class in set S. Denote $\mathcal{A}_i$ as the set of S satisfies the aforementioned situations. Similarly, for class j which has the position $j > i$, there are $\binom{j-2}{|S|-2}$ situations where class j is the lowest priority class in S. Denote $\mathcal{B}_j$ as the set of S satisfies the situations.Therefore, we can rewrite (4.11) as:

$$SV_i = \sum_{S \subseteq N, i \in S} \frac{\sum_{S' \in \mathcal{A}_i} \lambda_i \theta_i w_i^{\Delta(S')} + \sum_{j=i+1}^{n} \sum_{S' \in \mathcal{B}_j} \lambda_j \theta_j w_j^{\Delta(S')}}{|S|} \tag{4.13}$$

Using (4.13), we can also show the efficiency property of the Shapley Value, i.e., $\sum_{i=1}^{n} SV_i = \sum_{i=1}^{n} \lambda_i \theta_i w_i^{\{1,\dots,n\}} = \sum_{i=1}^{n} v_i(\sigma) = v(N, \sigma)$.

After we get SV_i, the transfer t_i for each class can be calculated as follows:

$$t_i = v_i(\sigma) - SV_i = \lambda_i \theta_i w_i^{\{1,\dots,n\}} - SV_i \tag{4.14}$$

Lemma 4.2.1 *Using the Shapley Value as the waiting cost share for each class, the allocation $\psi(\sigma, t)$ is efficient.*

proof:

We already stated the efficient allocation definition in Section 4.1: if an allocation $\psi(\sigma, t)$ for queuing problem $q = (N, \lambda, \theta) \in Q$ minimizes the total waiting cost $v(N, \sigma)$ and no transfer is lost ($\sum_{i=1}^{n} t_i = 0$), then this allocation is efficient.

First, from the efficiency property of the Shapley value, we have $\sum_{i=1}^{n} SV_i$ = minimize $v(N, \sigma)$.

From (4.14), we know that: $\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} v_i(\sigma) - \sum_{i=1}^{n} SV_i$.

The σ in (4.14) is an efficient ordering and from Lemma 4.1.1, $\sum_{i=1}^{n} v_i(\sigma)$ is the minimum system cost, minimum $v(N, \sigma)$.

Therefore, we have, $\sum_{i=1}^{n} t_i$ =minimum $v(N, \sigma)$– minimum $v(N, \sigma) = 0$.

Thus, we prove Lemma 4.2.1.

Taking (4.14) into (4.7), we get the complete formation of price difference between class i and class j as follows:

$$\Delta p_{ij} = \frac{\lambda_i \theta_i w_i^{\{1,\dots,n\}} - SV_i}{\lambda_i} - \frac{\lambda_j \theta_j w_j^{\{1,\dots,n\}} - SV_j}{\lambda_j} \tag{4.15}$$

We have thus developed the subsidy-free price difference between classes based on the inter-class compensations. Since a network generally maintains a limited number of classes of service, the calculation of waiting cost share SV_i and actual waiting cost c_i for each class does not suffer from the scalability problem.

4.3 Axiomatic Characterization of the Shapley Value

As shown in Section 4.2, the price difference between classes directly depends on the waiting cost share rule in the network, specifically, using the Shapley Value as the waiting cost share for each class. In this section, we define several axioms on fairness and characterize the Shapley Value using them.

Definition 4.3.1 *The waiting cost sharing rule satisfies the efficiency rule if and only if for all $q = (N, \lambda, \theta)$, $\psi(\sigma, t)$ is efficient.*

As shown in Lemma 4.2.1, when we use Shapley Value as the waiting cost sharing for each class, $\psi(\sigma, t)$ is efficient.

The next definition is as in literature. For example, two similar classes should be compensated such that their cost shares are equal (equal treatment of equals).

Definition 4.3.2 *The waiting cost sharing rule satisfies equal treatment of equals if and only if for all $q = (N, \lambda, \theta) \in Q, \psi(\sigma, t), i, j \in N$, then $\lambda_i = \lambda_j, \theta_i = \theta_j \Longrightarrow u_i = u_j$.*

Using Shapley Value as the waiting cost share for each class obviously satisfies equal treatment of equals axiom from (4.10).

Assume that the impatience of class i increases. In this case, the total cost of waiting may increase. The following axiom, called independence of preceding classes impatience (IPAI), proposes that classes being served after class i be not affected by the increase. Since those classes are served after class i, they do not impose any cost of waiting to class i. Only the classes being served before class i impose a cost of waiting to class i, so that they should bear the consequences of an increase in that waiting cost.

Definition 4.3.3 *The waiting cost sharing rule satisfies independence of preceding classes impatience (IPCI) if and only if for all $q = (N, \lambda, \theta), q' = (N, \lambda, \theta') \in Q, \psi(\sigma, t), \psi(\sigma', t')$, and for all $i \in N, \lambda_i = \lambda'_i, i \in N \backslash k : \theta_i = \theta'_i$ and $\theta_k < \theta'_k$, then for all $j \in N$ such that $\sigma_j > \sigma_k : u_j = u'_j$.*

The proof using Shapley Value as the waiting cost share for each class which satisfies IPCI is given here.

Given that $q = (N, \lambda, \theta), q' = (N, \lambda, \theta') \in Q, \psi(\sigma, t), \psi(\sigma', t'), k \in N$, and for all $i \in N \backslash k : \theta_i = \theta'_i$ and $\theta_k < \theta'_k$, we get:

$\sigma_k \geq \sigma'_k$ and for any $j \in N \backslash k, \sigma_j > \sigma_k, \theta_j = \theta'_j$, and have the same ordering.

From (4.13), given a length of a set $|S|, S \subseteq N, j \in S$, there are $\binom{j-1}{|S|-1}$ situations where class j is the lowest priority class in set S. Denote $\mathcal{A}_j$ as the set of S satisfies the aforementioned situations. For class i which has the position $i > j$, there are $\binom{i-2}{|S|-2}$ situations where class i is the lowest priority class in S. Denote $\mathcal{B}_j$ as the set of S satisfies the situations. Therefore, we have:

$$SV_j = \sum_{S \subseteq N, j \in S} \frac{\sum_{S' \in \mathcal{A}_j} \lambda_j \theta_j w_j^{\Delta(S')} + \sum_{i=j+1}^{n} \sum_{S' \in \mathcal{B}_i} \lambda_i \theta_i w_i^{\Delta(S')}}{|S|}$$

Similarly, given a length of a set $|S|, S \subseteq N, j \in S$, there are $\binom{j'-1}{|S|-1}$ situations where class j' is the lowest priority class in set S. Denote $\mathcal{A}'_j$ as the set of S satisfies the aforementioned situations. For class i' which has the position $i' > j'$, there are $\binom{i'-2}{|S|-2}$ situations where class i' is the lowest priority class in S. Denote $\mathcal{B}_j$ as the set of S satisfies the situations. Therefore, we have:

$$SV'_j = \sum_{S \subseteq N, j' \in S} \frac{\sum_{S' \in \mathcal{A}'_j} \lambda'_j \theta'_j w_j'^{\Delta(S')} + \sum_{i'=j'+1}^{n} \sum_{S' \in \mathcal{B}'_i} \lambda'_i \theta'_i w_i'^{\Delta(S')}}{|S|}$$

We already have for all $i \in N, \lambda_i = \lambda'_i$. For any $j \in N$, when $\sigma_j > \sigma_k$, we have $\sigma_j = \sigma'_j$. From (4.4), we have $w_j = w'_j$. In addition, when we look at (4.12) and we can also find that $w_j^{\Delta(S)} = w_j'^{\Delta(S)}, S \subseteq N$. Thus, we conclude that $SV_j = SV'_j$ when $\sigma_j > \sigma_k$ and using Shapley value as the waiting cost share for each class satisfies IPCI.

The last definition is consistent with the idea that if two classes are served before a third one (all three classes have identical traffic load λ), then the former are equally responsible for the waiting cost incurred by the latter. To capture this idea, we consider that the network no longer charges the last class (the unit waiting cost θ is 0). Then it is not necessary to change the queue. On the other hand, the transfer that had to be allocated to the last agent has to be redistributed among the remaining agents. Equal responsibility requires that it be redistributed equally among them all.

Definition 4.3.4 *Waiting cost sharing rule satisfies equal responsibility (ER) if and only if for all $q = (N, \lambda, \theta), \in Q, \psi(\sigma, t)$, if for all $i \in N, \lambda_i = \lambda$, $q' = (N, \lambda', \theta') \in Q$ such that for all $i \in N$: $\lambda'_i = \lambda$, for all $i \in N \backslash n, \theta'_i = \theta_i, \theta'_n = 0$, there exists $\psi(\sigma', t')$, such that for all $i \in M$:*

$\sigma'_i = \sigma_i$, and

$t'_i = t_i + \frac{t_n}{n-1}$.

Using Shapley Value as the waiting cost share for each class satisfies ER. The proof is given here.

Under Definition 4.3.4 , for all $i \in N, \lambda_i = \lambda'_i = \lambda$, we have for $|S| = i, i = 1, ..., n$, $w^{\Delta(S)}$ is a constant and it is not related to the elements in S. For example, $|S_1| = |S_2| = 3$, and $S_1 = \{i, j, k\}, S_2 = \{a, b, c\}$ with $\theta_i \geq \theta_j \geq \theta_k, \theta_c \geq \theta_b \geq \theta_c$, From the $w^{\Delta(S)}$ described in (4.12), we have,

$$w_k^{\Delta(S_1)} = w_k^{\{i,j,k\}} - w_k^{\{i,k\}} - w_k^{\{j,k\}} + w_k^{\{k\}}$$

$$w_c^{\Delta(S_2)} = w_c^{\{a,b,c\}} - w_c^{\{a,c\}} - w_c^{\{b,c\}} + w_c^{\{c\}}$$

For all $i \in N, \lambda_i = \lambda$, from (4.4) we get,

$$w_k^{\Delta(S_1)} = w_c^{\Delta(S_2)} = \frac{w_0}{(1-\frac{2\lambda}{\mu c})(1-\frac{3\lambda}{\mu c})} - \frac{w_0}{(1-\frac{\lambda}{\mu c})(1-\frac{2\lambda}{\mu c})} - \frac{w_0}{(1-\frac{\lambda}{\mu c})(1-\frac{2\lambda}{\mu c})} + \frac{w_0}{1-\frac{\lambda}{\mu c}}$$

For simplicity, in the following equation, we use $w^{|S|}$ to denote $w^{\Delta(S)}$ without differentiating among S as long they have same size.

From (4.13), we have,

$$SV_i = \sum_{|S|=1}^{n} \frac{\lambda w^{|S|}((\binom{i-1}{|S|-1})\theta_i + \sum_{j=i+1}^{n} \binom{j-2}{|S|-2}\theta_j)}{i}$$

In this case, the only difference between $q = (N, \lambda, \theta), \psi(\sigma, t)$ and $q = (N, \lambda', \theta'), \psi(\sigma', t')$ is $\theta'_n = 0$. Therefore the efficient ordering σ' of q' is same as the efficient ordering σ of q ($\sigma'_i = \sigma_i$, for $i \in N$).

Since $\theta'_n = 0$, the Shapley Value for SV'_i is,

$$SV'_i = \sum_{|S|=1}^{n-1} \frac{\lambda w^{|S|}((\binom{i-1}{|S|-1})\theta_i + \sum_{j=i+1}^{n-1} \binom{j-2}{|S|-2}\theta_j)}{i}$$

From (4.5), we have,

$$t_i = \lambda\theta_i w_i - SV_i$$

$$t'_i = \lambda\theta'_i w'_i - SV'_i$$

Since for $i \in N, \lambda_i = \lambda'_i = \lambda$, we find $w_i = w'_i = \frac{w_0}{(1-\frac{(i-1)\lambda}{\mu c})(1-\frac{i\lambda}{\mu c})}$.

Therefore,

$$t'_i - t_i = SV_i - SV'_i = \sum_{|S|=1}^{n-1} \frac{\lambda w^{|S|}\binom{n-2}{|S|-2}\theta_n}{|S|}$$

From above equation we find that for $i = 1, ..., n-1$, $t'_i - t_i$ is equal to the identical value $\sum_{|S|=1}^{n-1} \frac{\lambda w^{|S|}\binom{n-2}{|S|-2}\theta_n}{|S|}$. In order to find out what this value stands for, we assumes that:

$$t'_i = t_i + x, i = 1, ..., n-1. \tag{4.16}$$

Since allocations $\psi(\sigma, t), \psi(\sigma', t')$ are efficient, $\sum_{i=1}^{n} t_i = 0, \sum_{i=1}^{n-1} t'_i = 0$.

Summation of both side of (4.16),we get,

$$\sum_{i=1}^{n-1} t'_i = \sum_{i=1}^{n-1} t_i + (n-1)x$$

Therefore, we get $x = \frac{t_n}{n-1}$, and $t'_i = t_i + \frac{t_n}{n-1}$. Note that we also find $\frac{t_n}{n-1} = \sum_{|S|=1}^{n-1} \frac{\lambda w^{|S|}\binom{n-2}{|S|-2}\theta_n}{|S|}$.

Thus, we have proved that using Shapley Value as the waiting cost share for each class satisfies ER axiom.

4.4 An Illustrative Example

This section presents a numerical example of the pricing scheme for the multi-class priority-based network proposed in this chapter.

To emphasize the methodology, we simply assume that the network has two different classes with the non-preemptive priority scheme. We use λ to denote the total traffic in the network and k the percentage of traffic choosing class one service which has θ_1 as the waiting cost factor. Here, we assume $\theta_1 > \theta_2$. It means that class one packets have higher priority based on Lemma 4.1.1, the efficient ordering is $\sigma = \{1, 2\}$. Using (4.13), the waiting cost share for each class is as follows,

$$SV_1 = k\lambda\theta_1 w_1^{\{1\}} + \frac{(1-k)\lambda\theta_2(w_2^{\{1,2\}} - w_2^{\{2\}})}{2}$$

$$SV_2 = (1-k)\lambda\theta_2 w_2^{\{2\}} + \frac{(1-k)\lambda\theta_2(w_2^{\{1,2\}} - w_2^{\{2\}})}{2}$$

As described in Section 4.2, $w_i^S, i \in S$ is the average waiting cost of class i in an efficient ordering of S assuming that S has the power to be served first. Using (4.4), the average waiting time for each class in an efficient ordering can be calculated as:

$$w_1^{\{1\}} = w_1^{\{1,2\}} = \frac{\lambda}{\mu c(\mu c - k\lambda)}$$

$$w_1^{\{2\}} = \frac{\lambda}{\mu c(\mu c - (1-k)\lambda)}$$

$$w_2^{\{1,2\}} = \frac{\lambda}{(\mu c - k\lambda)(\mu c - \lambda)}$$

As defined in (4.5), we can calculate the transfer for each class as follows,

$$t_1 = k\lambda\theta_1 w_1^{\{1,2\}} - SV_1 = -\frac{(1-k)\lambda\theta_2(w_2^{\{1,2\}} - w_2^{\{2\}})}{2}$$

$$t_2 = (1-k)\lambda\theta_2 w_2^{\{1,2\}} - SV_2 = \frac{(1-k)\lambda\theta_2(w_2^{\{1,2\}} - w_2^{\{2\}})}{2}$$

Using the definition in (4.7), the price difference between class one service and class two service is,

$$\Delta p_{12} = \frac{t_2}{(1-k)\lambda} - \frac{t_1}{k\lambda} = \frac{\theta_2(w_2^{\{1,2\}} - w_2^{\{2\}})}{2k}$$

For simplicity, we assume the network capacity c is equal to 100 and the average length of packets in the network $\frac{1}{\mu} = 1$. The unit waiting cost for each class as $\theta_1 = 1.5$ and $\theta_2 = 1$. Fig. 4.3 describes $w_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ profiles against change λ total arrival rate in the network with fixed percentage of traffic choosing class one service k. In Fig. 4.3(a), average waiting time for both class increases as the total average arrival rate increases while class one packets experiences much lower average delay. Fig. 4.3(b) shows the waiting cost share (Shapley value) for each class against the increasing total traffic arrival rate. In Fig. 4.3(c), it describes the transfer of each class, to be specifically, the amount class one packets should compensate the class two packets. As shown in Section 4.1, total transfer is equal to 0, $t_1 + t_2 = 0$. Fig. 4.3(d) shows that the price difference between the two classes increases as the total traffic in

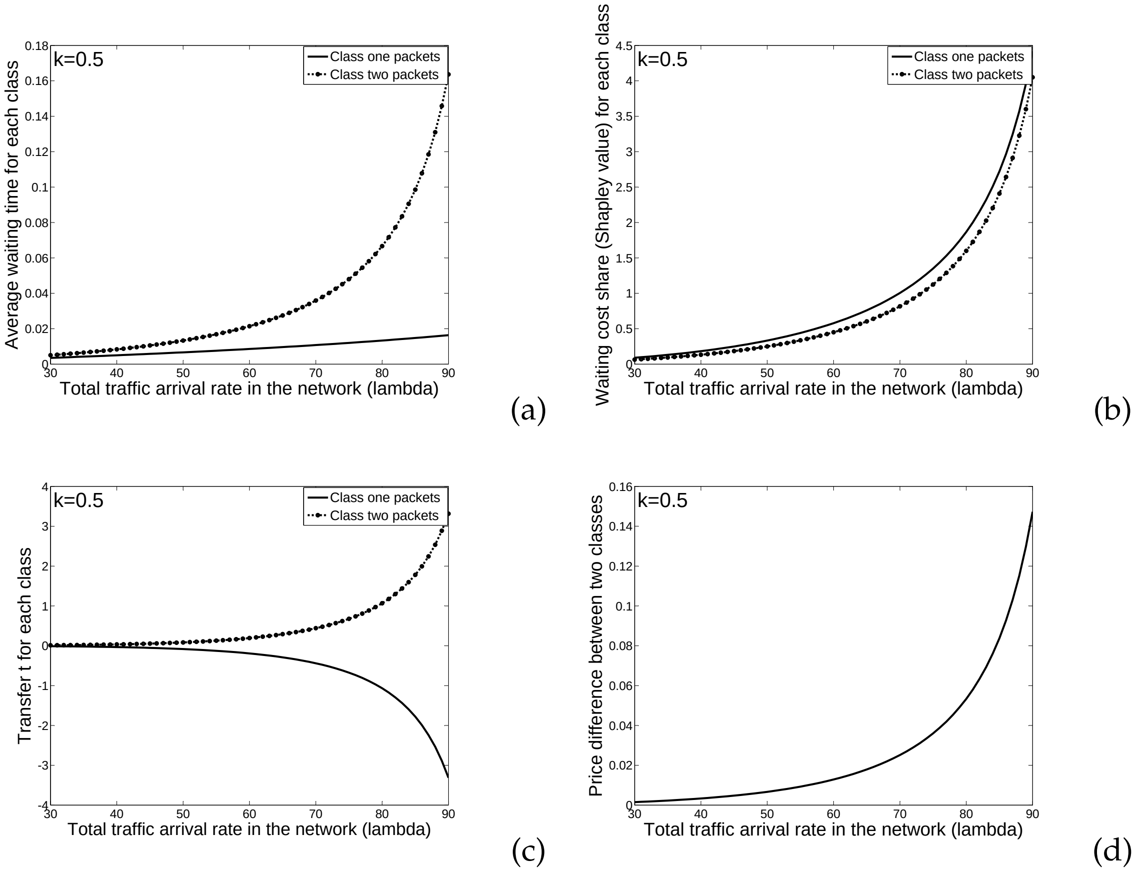


Figure 4.3: Profiles $w_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ against changing λ (total arrival rate in the network) with fixed k (percentage of traffic choosing class one service). (a) Average waiting time $w_i, (i = 1, 2)$ for each class against λ. (b) Waiting cost share (Shapley value) $SV_i, (i = 1, 2)$ for each class against λ. (c) Transfer $t_i, (i = 1, 2)$ for each class against λ. (d) Price difference between the two classes Δp_{12} against λ.

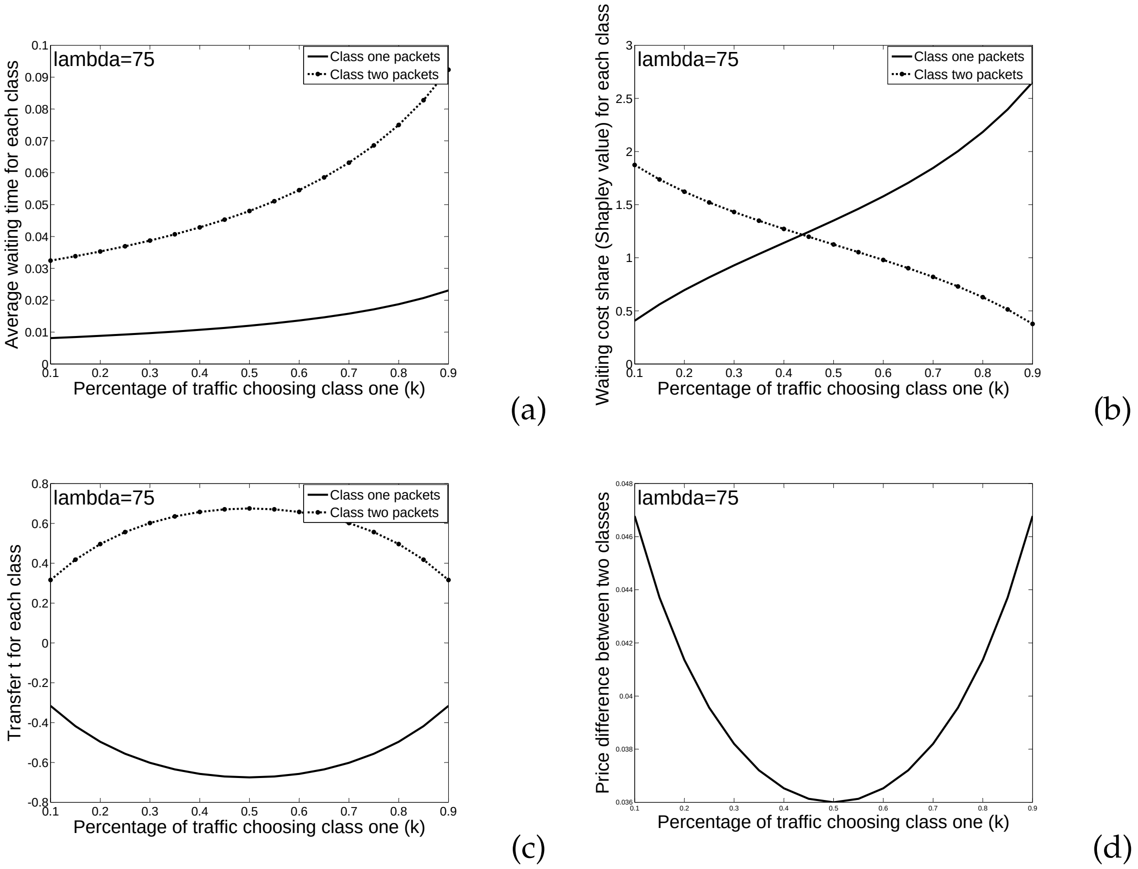


Figure 4.4: Profiles $w_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ against changing k (percentage of traffic choosing class one service) with fixed λ (total arrival rate in the network). (a) Average waiting time $w_i, (i = 1, 2)$ for each class against k. (b) Waiting cost share (Shapley value) $SV_i, (i = 1, 2)$ for each class against k. (c) Transfer $t_i, (i = 1, 2)$ for each class against k. (d) Price difference between the two classes Δp_{12} against k.

the network increases since class one packets needs to pay higher price for the better service (lower average waiting time) when the network resource is in short.

Figures shown in Fig. 4.3 in under the situations with the fixed percentage of traffic choosing class one service k. And Fig. 4.4 shows all the $w_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ profiles against changing k with a given total arrival rate in the network (λ). Fig. 4.4(a) shows the average waiting time for both classes increase as the percentage of packets choosing class one service (k). In Fig. 4.4(b), the waiting cost share (Shapley Value) for each class increase as the percentage of traffic choosing that class increases. In other words, the waiting cost share for class one is increasing with percentage of traffic choosing class one (k) and the waiting cost share for class two is increasing with percentage of traffic choosing class two ($1 - k$). Fig. 4.4(c) shows the amount class one packets should compensate class two packets as a function of k with the property $t_1 + t_2 = 0$. In Fig. 4.4(d), the price difference between the two classes is minimized when each of the class has identical amount of traffic, i.e., $\lambda_1 = \lambda_2$. In makes sense in a way that when the percentage of traffic choosing class one (k) is small, the price difference between the two classes is higher since class one packets enjoy little economic scale. And when the percentage of traffic choosing class one (k) is large, the class one packets need to pay more to compensate class two packets since class two packets experiencing much worse network quality (large average waiting time).

4.5 Chapter Summary

In this chapter, we have investigated the problem of pricing multi-class priority-based network services. Compensation is transfered from higher priority classes to lower priority classes which experience longer average waiting time. In the model proposed in this chapter, each class has Poisson arrivals and exponentially distributed packet length with identical mean, but a different waiting cost factor. We present an efficient ordering scheme by assigning each class its position in the queue, and then calculate compensation for each class based on its waiting cost share which is its Shapley Value in a cooperative game. We have also characterized the Shapley Value using different intuitive fairness axioms. A numerical example illustrates how the analytical results presented in the chapter can be used in a practical situation.

CHAPTER 5

Subsidy-free Prices in Class-based Networks

While priority networks are especially amenable to analysis thanks to existing results from queuing theory, in practice most multi-service networks do not discriminate among service classes solely through the tagging of packets with different priority levels. Rather, these networks may attempt to meet QoS of each class of service through resource allocation techniques. In this chapter, we will investigate subsidy-free prices for each class of service in class-based DiffServ networks.

The network model adopted in this chapter assumes that the network maintains separate queues for each class of service, and the network resource is allocated among these classes in order to maximize some social utility or minimize network disutility. Further, we establish the subsidy-free price for each class, based on the associated network resource consumption [90]. The remainder of the chapter is organized as follows. A discussion of the model used can be found in Section 5.1. The cooperative game and the Shapley Value are studied in Section 5.2. In Section 5.3, we investigate the pricing scheme. An illustrative example is given in Section 5.4 and Section 5.5 captures our conclusions.

5.1 Problem Statements and the Model

We consider a packet switched network with n classes of service as shown in Fig. 5.1. The network is modeled as a queuing network with First in First out (FIFO) discipline. The network resource c is allocated among the n classes and we use c_i to denote the resource allocated to class i. The packet length of all classes is considered to have an exponential distribution with the average packet length equal to $\frac{1}{\mu}$.

The set of classes are denoted as $N = \{1, ..., n\}$. The traffic in each class i is distributed with Poisson arrivals and identified by two parameters: (λ_i, θ_i). λ_i is the average arrival rate and θ_i is the impatience per unit time for class i. Then, given a resource allocation $\hat{c} = (c_1, c_2, ..., c_n)$, the disutility incurred by class i is given by:

$$v_i(c_i) = \lambda_i \theta_i d_i \tag{5.1}$$

where d_i is the average delay of class i packets.

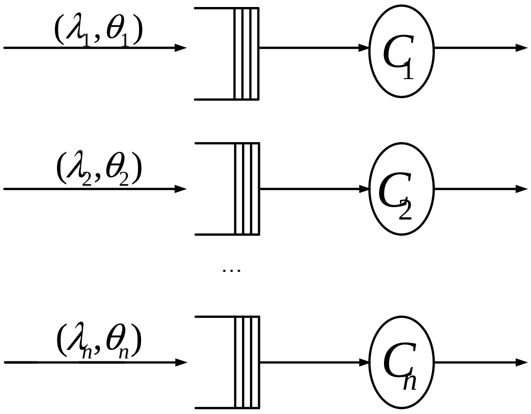


Figure 5.1: Network model in chapter 5

The total disutility incurred by packets from all classes due to an resource allocation $\hat{c} = (c_1, c_2, ..., c_n)$ can be written as:

$$v(N) = \sum_{i=1}^{n} d_i(c_i) = \sum_{i=1}^{n} \lambda_i \theta_i d_i \tag{5.2}$$

We have already assumed Poisson arrivals in each class with identical mean packet length and exponential distribution. Therefore, the average delay for packets in class i is [88]:

$$d_i = \frac{1}{\mu c_i - \lambda_i} \tag{5.3}$$

Here we define an allocation as $\psi(\hat{c}, t)$, where $\hat{c} = (c_1, c_2, ..., c_n)$ is the network resource allocation to each class and $t = (t_1, t_2, ..., t_n)$ is the monetary transfer related to each class. Given a network resource allocation $\hat{c}$ and a transfer t, the disutility share for class i, φ_i is defined as,

$$\varphi_i = v_i(c_i) + t_i = \lambda_i \theta_i d_i + t_i \tag{5.4}$$

An allocation $\psi(\hat{c}, t)$ is efficient whenever it minimizes the total disutility incurred by packets from all classes (i.e. minimum $v(N)$) and no transfer is lost (i.e. $\sum_{i=1}^{n} t_i = 0$).

An efficient network resource allocation $\hat{c}$ is the one which minimizes the total disutility incurred by packets from all classes $v(N)$.

Taking (5.3) into (5.2), the total disutility minimization problem becomes:

$$\text{minimize}_{\hat{c}}\ v(N) = \sum_{i=1}^{n} \lambda_i \theta_i d_i = \sum_{i=1}^{n} \frac{\lambda_i \theta_i}{\mu c_i - \lambda_i} \tag{5.5}$$

subject to the following constraints:

$$\sum_{i=1}^{n} c_i \leq c \tag{5.6}$$

$$\lambda_i \leq \mu c_i, i = 1, ...n \tag{5.7}$$

It can be observed that the $-v(N)$ is a strictly concave function over a closed and bounded set defined by (5.6) and (5.7). Therefore, a unique minimum always exists. We now use Lagrangian multipliers to append constraints to the objective. Thus, we can rewrite this minimization problem as:

$$\text{minimize}_{\hat{c}}\ \sum_{i=1}^{n} \frac{\lambda_i \theta_i}{\mu c_i - \lambda_i} - \gamma_0 (\sum_{i=1}^{n} c_i - c) + \sum_{i=1}^{n} \gamma_i (\mu c_i - \lambda_i) \tag{5.8}$$

The necessary and sufficient Karush-Kuhn-Tucker (KKT) Conditions [91] applicable to (5.8) are given by:

$$\frac{-\lambda_i \theta_i \mu}{(\mu c_i - \lambda_i)^2} - \gamma_0 + \gamma_i \mu = 0 \tag{5.9}$$

$$\gamma_0 (\sum_{i=1}^{n} c_i - c) = 0 \tag{5.10}$$

$$\gamma_i (\mu c_i - \lambda) = 0, i = 1, ..., n \tag{5.11}$$

Since a network generally maintains a limited number of classes of service, the calculation of efficient network resource allocation $\hat{c}$ does not suffer from the scalability problem. So far, we have calculated $v_i(c_i)$ in (5.4) based on efficient network resource allocation $\hat{c}$. In the next section, we will consider the disutility share problem as a cooperative game and set up the disutility share for each class φ_i using the Shapley value.

5.2 Disutility Share for Each Class

Another way of solving the disutlity share φ_i for each class is by viewing the process as a cooperative game. As stated in Section 5.1, each class tries to get more network resource to reduce its disutility as expressed in (5.1). This can be modeled as a cooperative game and one can use the Shapley Value as the payoff (disutility share) for each class. We first define the worth of a coalition and then compute the Shapley value.

We define the worth of a coalition $v(S)$, $S \subseteq N$ as the sum of its members' disutility assuming they have the privilege to be allocated more network resource in an efficient network resource allocation. The privilege will allow them to increase their network resource allocation by increasing their impatience θ. Let's look at the minimization problem as described in (5.5); the larger the θ_i, the more the network resource that will be allocated to class i. When there is a set $N = \{1, 2, ..., n\}$ with $\theta_1 > \theta_2 > ... > \theta_n$, a coalition $S \subseteq N$ means that its members will be given the $|S|$ highest impatience, that is, $\{\theta_1, \theta_2, ..., \theta_{|S|}\}$, by its original impatience order. At the same time, members in $N \backslash S$ will have the $|N| - |S|$ lowest impatience $\{\theta_{|S|+1}, ..., \theta_n\}$ by its original impatience order. Therefore, we establish a new impatience $\theta' = (\theta'_1, ..., \theta'_n)$. when there is a coalition S.

As an example, we assume that there are four different classes of service $N = \{1, 2, 3, 4\}$ with $\theta = \{\theta_1 = 5, \theta_2 = 4, \theta_3 = 2, \theta_4 = 1\}$. When we consider a coalition $S = \{2, 3\}$, its members class 2 and 3 will have the two highest impatience, that is $\{5, 4\}$. By its original impatience order, $\theta_2 > \theta_3$, we then have $\theta'_2 = 5, \theta'_3 = 4$. Meanwhile, members in the set $N \backslash S = \{1, 4\}$ will have the two lowest impatience $\{1, 2\}$ and by their original impatience order, $\theta_1 > \theta_4$, we thus have $\theta'_1 = 2, \theta'_4 = 1$. Thus, we establish a new impatience $\theta' = \{\theta'_1 = 2, \theta'_2 = 5, \theta'_3 = 4, \theta'_4 = 1\}$ when there is a coalition $S = \{2, 3\}$.

After establishing the new impatience θ' when there is a coalition S, each class will be allocated network resource in an efficient network resource allocation based on their new impatience θ' and the worth of this coalition S is defined as sum of its members' disutility as follows:

$$v(S) = v(S, \hat{c}') = \sum_{i \in S} \lambda_i \theta_i d'_i = \frac{\lambda_i \theta_i}{\mu c'_i - \lambda_i} \tag{5.12}$$

where $\hat{c}'$ is the efficient network resource allocation based on their new impatience θ'.

The marginal contribution of a class $i \in N$ to a coalition S in v, $i \notin S$ is:

$$\begin{aligned} v(S \cup \{i\}) - v(S) &= v(S \cup \{i\}, \hat{c}'') - v(S, \hat{c}') \\ &= \sum_{i \in S \cup \{i\}} \lambda_i \theta_i d''_i - \sum_{i \in S} \lambda_i \theta_i d'_i \\ &= \sum_{i \in S \cup \{i\}} \frac{\lambda_i \theta_i}{\mu c''_i - \lambda_i} - \sum_{i \in S} \frac{\lambda_i \theta_i}{\mu c'_i - \lambda_i} \end{aligned}$$

where $\hat{c}''$ is the efficient network resource allocation when there is a coalition $S \cup \{i\}$, while $\hat{c}'$ is the efficient network resource allocation when there is a coalition S.

The Shapley Value is defined as a weighted sum of the classes' marginal contribution to coalitions. Let us recall the definition of the Shapley Value in Section 2.1.1. For all $i \in N$, the payoff (disutility share) to class i is given by:

$$SV_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{5.13}$$

We have thus developed the disutility share for each class by using the corresponding Shapley value.

5.3 Pricing Scheme

Given an efficient network resource allocation $\hat{c}$ and a transfer t_i, the disutility share φ_i for class i is equal to $v_i(\hat{c}) + t_i$ as shown in (5.4) in Section 5.1. And in Section 5.2, we calculated the disutility share for each class in another way-by modeling this problem as a cooperative game and using the corresponding Shapley Value as the disutility share for each class. This means $\varphi_i = SV_i$ and we can then rewrite (5.4) as follows:

$$SV_i = v_i(\hat{c}) + t_i \tag{5.14}$$

Now we can calculate the monetary transfer t_i for each class i from Equation (5.14) as $SV_i - v_i(\hat{c})$. Equation (5.14) also shows that when the disutility share SV_i of class i is larger than its actual disutility $v_i(\hat{c})$, class i packets will compensate others $SV_i - v_i(\hat{c})$ because it has been allocated more network resource and experiences less average delay. On the other hand, if the disutility share SV_i of class i is smaller than its actual disutility $v_i(\hat{c})$, class i packets will receive compensations $SV_i - v_i(\hat{c})$ from others since it has been allocated less network resource and experiences longer average delay.

Let's now consider $\sum_{i=1}^{n} t_i$, from (5.14) we have,

$$\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} SV_i - \sum_{i=1}^{n} v_i(\hat{c})$$

From the Shapley value's efficiency property, we know $\sum_{i=1}^{n} SV_i$ = minimum $v(N)$. We already stated that $\hat{c}$ is an efficient network resource allocation and it means that $\sum_{i=1}^{n} v_i(\hat{c})$ = minimum $v(N)$. From above equations, we observe that $\sum_{i=1}^{n} t_i = 0$.

We can now state that the allocation $\psi(\hat{c}, t)$ is efficient when $\hat{c}$ is an efficient network resource allocation and t_i is calculated using (5.14).

After getting the transfer t_i for each class based on (5.14), we are able to define the price difference Δp_{ij} between class i and j as follows:

$$\Delta p_{ij} = \frac{t_i}{\lambda_i} - \frac{t_j}{\lambda_j} \tag{5.15}$$

Let's now take (5.1)(5.3)(5.14) into (5.15), and get:

$$\Delta p_{ij} = \frac{SV_i - \frac{\lambda_i \theta_i}{\mu c_i - \lambda_i}}{\lambda_i} - \frac{SV_j - \frac{\lambda_j \theta_j}{\mu c_j - \lambda_j}}{\lambda_j} \tag{5.16}$$

where $\hat{c}$ is an efficient network resource allocation and SV_i is the Shapley Value corresponding to the cooperative game.

We have thus developed the price difference between classes based on the inter-class compensations.

5.4 An Illustrative Example

This section presents a numerical example of the pricing scheme for the class-based network proposed in this chapter.

To emphasize the methodology, we simply assume that the network supports two different classes. We use λ to denote the total traffic in the network and k the percentage of traffic choosing class one service which has θ_1 impatience per unit time. Here, we assume $\theta_1 > \theta_2$. Using Equation (5.13), the waiting cost share for each class is as follows,

$$SV_1 = \frac{v\{1,2\}}{2} + \frac{v\{1\}}{2} - \frac{v\{2\}}{2}$$
$$SV_2 = \frac{v\{1,2\}}{2} + \frac{v\{2\}}{2} - \frac{v\{1\}}{2}$$

As described in Section 5.2, $v(S)$ is the sum of its members' disutility in an efficient network resource allocation assuming that its members have the $|S|$ largest impatience. Therefore, from (5.12) we get:

$$v\{1,2\} = k\lambda\theta_1 d'_1 + (1-k)\lambda\theta_2 d'_2 = \frac{k\lambda\theta_1}{\mu c_1 - k\lambda} + \frac{(1-k)\lambda\theta_2}{\mu c_2 - (1-k)\lambda}$$

where $\hat{c} = (c_1, c_2)$ is the efficient network resource allocation when $\theta'_1 = \theta_1, \theta'_2 = \theta_2$. Further,

$$v\{1\} = k\lambda\theta_1 d'_1 = \frac{k\lambda\theta_1}{\mu c_1 - k\lambda}$$

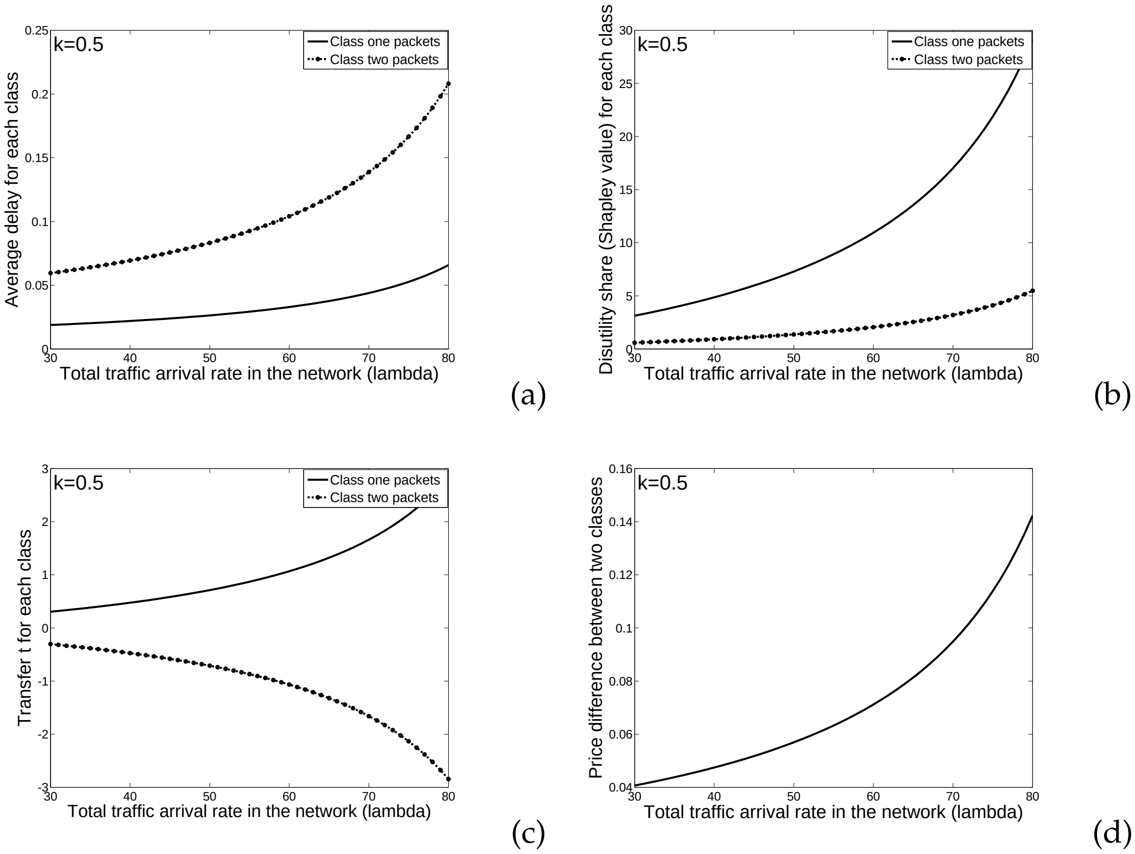


Figure 5.2: Profiles $d_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ against changing λ(total arrival rate in the network) with fixed k(percentage of traffic choosing class one service). (a) Average delay $d_i, (i = 1, 2)$ for each class against λ. (b) Disutility share (Shapley value) $SV_i, (i = 1, 2)$ for each class against λ. (c) Transfer $t_i, (i = 1, 2)$ for each class against λ. (d) Price difference between the two classes Δp_{12} against λ.

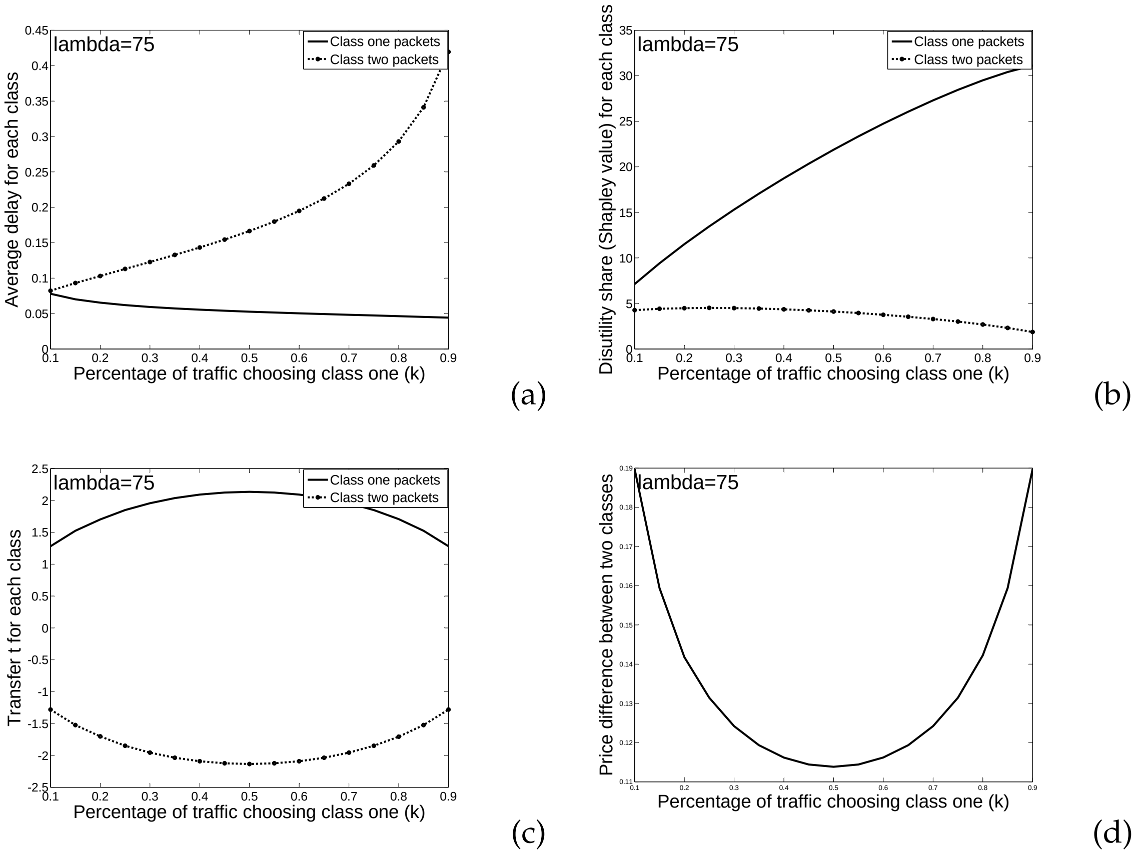


Figure 5.3: Profiles $d_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ against changing k (percentage of traffic choosing class one service) with fixed λ (total arrival rate in the network). (a) Average delay $w_i, (i = 1, 2)$ for each class against k. (b) Disutility share (Shapley value) $SV_i, (i = 1, 2)$ for each class against k. (c) Transfer $t_i, (i = 1, 2)$ for each class against k. (d) Price difference between the two classes Δp_{12} against k.

where $\hat{c} = (c_1, c_2)$ is the efficient network resource allocation when $\theta'_1 = \theta_1, \theta'_2 = \theta_2$. And,

$$v\{2\} = (1-k)\lambda\theta_2 d'_2 = \frac{(1-k)\lambda\theta_2}{\mu c_2 - (1-k)\lambda}$$

where $\hat{c} = (c_1, c_2)$ is the efficient network resource allocation when $\theta'_1 = \theta_2, \theta'_2 = \theta_1$.

As defined in (5.14), we can now calculate the transfer for each class as follows,

$$t_1 = \frac{(1-k)\lambda\theta_2}{2(\mu c_2^{\{\theta'_1=\theta_1,\theta'_2=\theta_2\}} - (1-k)\lambda)} - \frac{(1-k)\lambda\theta_2}{2(\mu c_2^{\{\theta'_1=\theta_2,\theta'_2=\theta_1\}} - (1-k)\lambda))}$$

$$t_2 = \frac{(1-k)\lambda\theta_2}{2(\mu c_2^{\{\theta'_1=\theta_2,\theta'_2=\theta_1\}} - (1-k)\lambda)} - \frac{(1-k)\lambda\theta_2}{2(\mu c_2^{\{\theta'_1=\theta_1,\theta'_2=\theta_2\}} - (1-k)\lambda))}$$

where $c_2^{\{\theta'_1=\theta_1,\theta'_2=\theta_2\}}$ is network resource allocated to class 2 in the efficient network resource allocation when $\theta'_1 = \theta_1, \theta'_2 = \theta_2$; $c_2^{\{\theta'_1=\theta_2,\theta'_2=\theta_1\}}$ is network resource allocated to class 2 in the efficient network resource allocation when $\theta'_1 = \theta_2, \theta'_2 = \theta_1$.

Using the definition in (5.15), the price difference between class one service and class two service is,

$$\Delta p_{12} = \frac{\theta_2}{2k(\mu c_2^{\{\theta'_1=\theta_1,\theta'_2=\theta_2\}} - (1-k)\lambda)} - \frac{\theta_2}{2k(\mu c_2^{\{\theta'_1=\theta_2,\theta'_2=\theta_1\}} - (1-k)\lambda)}$$

For the numerical example, we assume that the network capacity c is equal to 100 and the average length of packets in the network $\frac{1}{\mu} = 1$. The unit waiting cost for each class is $\theta_1 = 10$ and $\theta_2 = 1$. Fig. 5.2 depicts $d_i, SV_i, t_i, \Delta p_{12}, (i = 1, 2)$ profiles against λ with a fixed $k = 0.5$. In Fig. 5.2(a), the average delay for both classes increases as the total arrival rate increases while class one packets experience much lower average delay. Fig. 5.2(b) shows the disutlity share (Shapley value) for each class against λ. Fig. 5.2(c) describes the transfer for each class, specifically, the amount class one packets should compensate the class two packets. As shown in Section 5.3, the total transfer is equal to 0, $t_1 + t_2 = 0$. Fig. 5.2(d) shows that the price difference between the two classes increases as the total traffic in the network increases since class one packets need to pay higher price for the better service (lower average delay) when the network resource is short.

Fig. 5.2 has addressed a fixed percentage of traffic choosing class one service, i.e., $k = 0.5$. Fig. 5.3 shows all the corresponding profiles against k with a given $\lambda = 75$. Fig. 5.3(a) shows that the average delay for class two increase as the percentage of packets choosing class one service (k) increases. Due to economy of scale, the average delay for class one packets slightly decreases as the percentage of traffic choosing class one service increases. In Fig. 5.3(b), the disutility share (Shapley value) for each class increases as the percentage of traffic choosing that class increases. In other words, the waiting cost share for class one increases with percentage of traffic choosing class one and the waiting cost share for class two increases with percentage of traffic choosing class two. Fig. 5.3(c) shows the amount class one packets should compensate class two packets as a function of k. In Fig. 5.3(d), the price difference between the two classes is minimized when each of the classes has identical amount of traffic, i.e., $\lambda_1 = \lambda_2$. This is also intuitive because that when the percentage of traffic choosing class one is small, the price difference between the two classes is higher since class one packets don't have economy of scale. On the other hand, even when the class one traffic is very high, the price difference is still large because despite the economy of scale, class one packets enjoy a significantly higher Quality of Service (QoS).

5.5 Chapter Summary

In this chapter, we have established the subsidy-free prices for each class of service in a class-based network using inter-class compensations. A lower QoS class will receive compensations from other classes for experiencing longer average delay. In the model proposed in this chapter, each class has a different level of impatience, arrival rate, but exponentially distributed packet lengths with identical mean. We have developed an efficient network resource allocation for each class and then computed compensations among the classes based on the disutility share of each class. The disutility share of a class is its Shapley Value in a cooperative game. We have then proved that the allocation is efficient and presented a pricing scheme based on inter-class compensations. The pricing scheme proposed in this chapter has nice fairness property: the higher QoS service price is always higher than the lower QoS service even when the higher QoS service enjoys a large economy of scale. Further, for any given ratio of higher vs. lower QoS traffic, the price differential between the higher and lower QoS services increases as the total network demand increases. The price difference is minimized when the higher QoS service demand is equal to the lower QoS service demand for any given level of total traffic.

CHAPTER 6

Market-clearing Prices in Class-based Networks

Class-based network architectures such as Diffserv [64] have become the most viable solution for providing Quality-of-Service (QoS) over IP networks. In multi-class networks, incentives are needed to be offered to users in order to encourage them to choose the amount of bandwidth from each class that is most appropriate for their needs, thereby discouraging over-allocation and maximizing the social welfare and individual utility. In this chapter, we are particularly concerned about the market-clearing price for each class of service under competitive market model [92].

The study of competitive economy equilibrium was first started by Walras [93] over a hundred years ago. In this problem, each user participating in the market initially has an endowment of some amount of each of n goods $w_i = (w_{i1}, ..., w_{in})$. Every agent sells the entire initial endowment and buys a bundle of goods $x_i = (x_{i1}, ..., x_{in})$ to maximize his utility function $u_i(x_i)$. Subject to the following constraint, $x_i p^T \leq w_i p^T$, Arrow and Debreu [94] have showed that there exists equilibrium prices $p = (p_1, ..., p_n)$ for the n goods such that the market is cleared (the demand of each of the goods equals the supply) if the utility function of each agent u_i were concave. Reference [95] has provided an algorithm to compute this competitive economy equilibrium.

In modern networks, terminals/users no longer use one type of service. Instead, they consume multiple services at the same time. For example, when a user is playing an on-line game, he may also need to chat with another party or use text messaging. It is natural for us to model a multi-class network as a competitive market where agents are divided into two sets: users and the service provider. The service provider supports n types of services and each service has a different QoS guarantee. Users spend money to buy a bundle of n types of services and maximize their individual utility. An equilibrium is a set of prices for n services so that the market is cleared and each user's utility is maximized. This model is a special case of Walras' model when money is also considered as a good.

For general concave and homogeneous utility functions, the equilibrium problem is reduced to a social utility maximization problem over a convex set defined by the supply-demand linear constraints and the equilibrium prices derived from the Lagrangian multipliers of these constraints [96, 97, 98]. References [63, 99] have investigated the competitive market equilibrium with non-homogeneous utility functions which include goods purchased by other users. It means that each user's utility function not only depends on his purchase action but also related to other users' purchase choices. This is also the case for communication networks in that the performance obtained by any given network user is determined by all users' traffic and service choices.

In this chapter, we consider pricing, resource allocation and QoS provisioning in a multi-class network under the competitive economy model. We use revenue as the utility function for the service provider and enhance Kelly's utility function by including a QoS parameter for users. Given the initial endowment for each user, we show that a competitive equilibrium (price for each class of service and bandwidth allocation among all users) for the competitive multi-class network resource market always exists. And under this equilibrium, both individual optimality and social economic efficiency are achieved in a way such that all users' utilities are maximized simultaneously. In other words, this equilibrium is Pareto optimal. We further show that under the fixed supply condition for each class of the service, this equilibrium is unique which can be computed in polynomial time.

In addition, this chapter also discusses how to adjust the initial endowment for each user to meet their bandwidth constraint (either from constraint on the access network or from limitation of the user equipment). Under this constraint, the competitive equilibrium is the price for each class of service, the budget redistribution and the bandwidth allocation among all users. The equilibrium conditions are analyzed and the existence of the equilibrium is also proved. A procedure to recompute the equilibrium is proposed.

The rest of the chapter is organized as follows. Section 6.1 lists the mathematical notations used in this chapter. In Section 6.2, we introduce our model of competitive multi-class network resource market. In this section, both the service provider's and users' utility functions are presented and investigated. Section 6.3 proves the existence and uniqueness of the competitive equilibrium of this market. In Section 6.4, we discuss how to adjust the users' initial budget under each user's bandwidth constraint. Section 6.5 gives a numerical example resulting in the competitive equilibrium in pricing and resource allocation in multi-class networks, and Section 6.6 captures the conclusions of this chapter.

6.1 Problem Statements and Mathematical Notations

Mathematical notations used in this chapter are described in this section. We use R^n to denote the n-dimensional Euclidean space and R^n_+ to denote the subset of R^n where each coordinate is non-negative. We also use R and R_+ to represent the set of real numbers and the set of non-negative real numbers. Throughout this chapter, for a vector x, $x > 0$ means that every component of x is larger than 0. Other comparative notations $<, \geq, \leq$ are extended in a similar manner.

We assume the network provides n services. Each class of service has a different QoS guarantee and is suitable for different applications. There are m users sharing the network resource. We use $X \in R^{mn}_+$ to denote the set of ordered m-tuples $X = (x_1, ..., x_m)$ and use $\bar{X} \in R^{(m-1)n}_+$ to denote the set of ordered $(m-1)$-tuples $\bar{X} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_m)$, where $x_i = (x_{i1}, ..., x_{in}) \in X_i \in R^n_+$ for $i = 1, ..., m$. For each user i, we use $u_i(x_i, \bar{x}_i)$ to denote his utility function where $x_i \in X_i$ and $\bar{x}_i \in \bar{X}_i$. It means that user i's utility depends on his own action x_i, as well as actions made by all other users $\bar{x}_i$.

The following definitions [100] will be used in the proof of the existence and uniqueness of competitive equilibrium.

Definition 6.1.1 *For a twice-differentiable function u_i: $R^n_+ \to R_+$, if $\nabla_{x_i} u_i$ is positive and $\nabla^2_{x_i} u_i$ is negative, we have u_i as a monotonically increasing function and concave with respect to x_i.*

Definition 6.1.2 *A function u_i: $R^n_+ \to R_+$ is concave if for any $x, y \in R^n_+$ and any $0 \leq a \leq 1$, we have $u_i(ax + (1-a)y) \geq au_i(x) + (1-a)u_i(y)$. And it is strictly concave, if $u_i(ax + (1-a)y) > au_i(x) + (1-a)u_i(y)$ for $0 < a < 1$.*

6.2 Competitive Multi-class Network Resource Market

In the competitive multi-class network resource market, as stated in Section 6.1, the network supports n types of services and there are m users in the network. There are three types of entities in the market: users, service provider and market.

Each user i is endowed a monetary budget $w_i (> 0)$ and uses it to purchase some amount of each of n services $x_i = (x_{i1}, ..., x_{in})$ from an open market so as to maximize its own utility $u_i(x_i, \bar{x}_i)$, where $\bar{x}_i$ represents the amount of services obtained by all other users. The budget w_i for each user can represent the budgets for users to pay for services as in [32]. As noted in Section 6.1, in this chapter x_{ij} is used to denote the amount of service j that user i purchases; $x_i = (x_{i1}, ..., x_{in})$ is a n dimensional vector which shows some amount of each of n services user i purchase; $\bar{x}_i = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_m)$ is an $(m-1) \times n$ vector and represents the amount of each of n services obtained by all other users except user i.

We now consider the network service provider. It sets up the network and allocates limited network resource to n types of services from a convex and compact set C to maximize its utility.

The third entity, the market, sets unit prices $p = (p_1, ..., p_n)$ for n types of services. Each type of service has a different QoS guarantee and p_j can be interpreted as the unit price for service j. For example, $p_1 = 1$ and $p_2 = 2$ simply means that users can use one unit service 2 to trade for two units of service 1.

Then, user i's $(i = 1, ..., m)$ individual utility maximization problem is as follows.

$$\text{maximize}_{x_i} u_i(x_i, \bar{x}_i) \tag{6.1}$$

subject to the following constraints:

$$\begin{aligned} x_i p^T &\leq w_i \\ x_i &\geq 0 \end{aligned} \tag{6.2}$$

Equation (6.2) shows that the total payment for the purchased n types of services should not exceed his or her endowed budget w_i.

As discussed in Section 3.1.1, we redefine user's utility as a function of the allocated bandwidth and the QoS of the network as follows:

$$u = \frac{\beta}{T_{now}} \log(\frac{x}{\tilde{x}}) \tag{6.3}$$

where $\beta > 0$ is the weighting factor and it describes the flow's relative sensitivity to the QoS parameter. We use the present average delay T_{now} to represent the QoS parameter of network. $\tilde{x}$ is the minimum bandwidth requirement. Here we use the present average delay since users always keep record of the present network situation like Round Trip Time, packet loss rate, etc.

Then user i's utility in a multi-class network is the sum of utilities from each class of service as follows:

$$u_i(x_i, \bar{x}_i) = \sum_{j=1}^{n} \frac{\beta_{ij}}{T_{j_{now}}} \log(\frac{x_{ij}}{\tilde{x_{ij}}}) \tag{6.4}$$

where β_{ij} is a weighting factor and shows user i's relative sensitivity to bandwidth and QoS in service class j; $\tilde{x_{ij}}$ is user i's minimum bandwidth requirement for class j

service. It is assumed that the network has an overall resource that exceeds the sum of the minimum bandwidth requirements.

As we stated before, the network supports n classes of service. Within each class j, we assume that the traffic follows the Poisson distribution and this poissonian arrival discipline is generally considered to be a good model for the aggregate traffic from a large number of independent users [88]. For simplicity, we further assume that the length of all packets in the network is exponentially distributed with average length equal to $\frac{1}{\mu}$. The network is modeled as a queuing network with First in First out (FIFO) discipline. Therefore, the actual QoS parameter delay for each class j, $T_{j_{now}}$ is calculated using M/M/1 queuing model as follows:

$$T_{j_{now}} = \frac{1}{\mu c_j - \sum_{k=1,k\neq i}^{m} x_{kj}} \tag{6.5}$$

where c_j is the allocated resource for class j.

Together with (6.4) and (6.5), we can derive a more explicit description of each user i's utility as follows.

$$u_i(x_i, \bar{x_i}) = \sum_{j=1}^{n} \beta_{ij}(\mu c_j - \sum_{k=1,k\neq i}^{m} x_{kj}) \log(\frac{x_{ij}}{\tilde{x_{ij}}}) \tag{6.6}$$

In order to keep different QoS for each class, we have the maximum possible arrival rate s_j for each class as follows:

$$T_{j_{SLA}} = \frac{1}{\mu c_j - s_j} \tag{6.7}$$

where $T_{j_{SLA}}$ is the QoS agreed to by service provider and users in Service Level Agreement for class j. Vector $s = (s_1, ..., s_n)$ represents maximum possible arrival rate for each class in the network. For simplicity, in the following analysis, we further assume the average message lengths are equal to 1 ($\frac{1}{\mu} = 1$). Therefore, numerically, the message arrival rate is the same as bandwidth.

The service provider's individual utility maximization problem is represented as:

$$\text{maximize}_s u_s(s, p) = sp^T \tag{6.8}$$

subject to the constraint:

$$s \in S \tag{6.9}$$

where s is a feasible set of available bandwidth supply in each class.

A competitive market equilibrium is a point which we denoted as $[p, s, x_1, ..., x_m]$ where $s = (s_1, ..., s_n)$ and s_j is the total amount of bandwidth available for class j, and $p = (p_1, ..., p_n) \in R^n_+$ and p_j is the unit price for class j set by the market; such that:

1 (User optimality) x_i is a maximizer of (6.4) given $\bar{x}_i$ and p for every i.

2 (Service provider optimality) s is a maximizer of (6.8) given p.

3 (market efficiency) $p \geq 0$, $\sum_{i=1}^m x_{ij} \leq s_j$, $p_j(\sum_{i=1}^m x_{ij} - s_j) = 0$ for all j.

The last condition implies that for class j, when the supply is larger than the demand, the equilibrium price for class j is equal to 0.

6.3 Equilibrium Characteristics

In Section 6.2, we have given out assumptions about user's utility in data communication networks and our proposed utility function in a multi-class network. In this section, we will investigate the existence and uniqueness of the competitive equilibrium of the multi-class network resource market.

Theorem 6.3.1 *Using the utility functions defined in (6.6) and (6.8), the multi-class network resource market has a competitive equilibrium.*

proof:

From the proof given in [94], if the utility function $u_i(x_i, \bar{x}_i)$ of each agent is continuous and concave in $x_i \in R^n_+$ for every $\bar{x}_i \in R^{(m-1)n}_+$, and each agent's strategy set $X_i \in R^n_+$ is a closed convex, the existence of a competitive equilibrium is guaranteed.

First, we check the monotonicity of the utility function $u_i(x_i, \bar{x}_i)$. The partial derivative of $u_i(x_i, \bar{x}_i)$ is:

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = \frac{\beta_{ij}}{x_{ij}}(\mu c_j - \sum_{k=1, k\neq i}^{m} x_{kj}) > 0 \tag{6.10}$$

Then, we check the concavity of the utility function $u_i(x_i, \bar{x}_i)$ using the second partial derivative of $u_i(x_i, \bar{x}_i)$, we have,

$$(\nabla^2_{x_i} u_i(x_i, \bar{x}_i))_j = -\frac{\beta_{ij}}{x^2_{ij}}(\mu c_j - \sum_{k=1, k\neq i}^{m} x_{kj}) < 0 \tag{6.11}$$

From (6.10) and (6.11), we find that each user's utility function is monotonically increasing and is concave with respect to each variable x_{ij}, given $\bar{x}_i$ based on Definition 6.1.1.

Now, we will check the monotonicity and concavity of service provider's utility function u_s as described in (6.8) as follows.

$$\nabla_{s_j} u_s = p_j \tag{6.12}$$

$$\nabla^2_{s_j} u_s = 0 \tag{6.13}$$

From (6.12) and (6.13), we also find that the service provider's utility function is also monotonically increasing and concave with respect to each variable s_j (although u_s is not strictly concave with respect to the variable s_j).

We note that x_i is bounded under the linear constraint (6.2). It is a closed and convex set. Since the network has fixed limited network resource, the supply for each class of service s is also a closed and convex set.

Until now, we have proved that all agents'(users and service provider) utility functions are continuous and concave with respect to its variables (either x_i or s_j). $x_i \in R^n_+$ and $s \in R^n_+$ are both closed and convex sets. We claim that this competitive multi-class network resource market has an equilibrium. Thus, we prove Theorem 6.3.1.

We note that if p and s are fixed and the users are the only agents in the game, the equilibrium problem reduces to a Nash Equilibrium problem. By allowing p and s to change in the game, we can potentially achieve a more efficient equilibrium point. And each competitive equilibrium is Pareto optimal [94].

Now considering the optimality conditions of (6.6) and (6.8), we can find the following conditions using the Lagrangian multiplier. In other words,

$$\text{maximize}_{x_i} u_i(x_i, \bar{x}_i) - \lambda(x_i p^T - w_i) + \gamma x_i^T \tag{6.14}$$

where $\lambda \geq 0$ and $\gamma \geq 0$ are Lagrangian multipliers. Note that λ is a scalar and γ is a vector.

From the KKT [91] condition, we have:

$$\nabla_{x_i} u_i(x_i, \bar{x}_i) - \lambda p^T + \gamma^T = 0 \tag{6.15}$$

$$\lambda(x_i p^T - w_i) = 0 \tag{6.16}$$

$$\gamma x_i = 0 \tag{6.17}$$

Since the Lagrangian multiplier $\gamma \geq 0$, take it into (6.15), we have:

$$\nabla_{x_i} u_i(x_i, \bar{x}_i) \leq \lambda p^T \tag{6.18}$$

From (6.16) and (6.18), we get the following inequality:

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p \geq w_i \nabla_{x_i} u_i(x_i, \bar{x}_i) \tag{6.19}$$

Together with the constraints in (6.6) and (6.8), we have the complete necessary and sufficient conditions for a competitive equilibrium as follows:

$$\begin{aligned} (\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p \geq w_i \nabla_{x_i} u_i(x_i, \bar{x}_i) \\ x_i p^T \leq w_i \\ \sum_{i=1}^{m} x_{ij} \leq s_j, \forall j \\ s p^T \leq \sum_{i=1}^{m} w_i \\ x_i, p, s_j \geq 0, \forall i, j \end{aligned} \tag{6.20}$$

Now, multiplying $x_i \geq 0$ to both sides of (6.19), we have $x_i p^T \geq w_i$ for all i, and together with (6.20), we have:

$$\sum_{i=1}^{m} w_i \geq s p^T = \sum_{j=1}^{n} s_j p_j \geq \sum_{j=1}^{n} \sum_{i=1}^{m} x_{ij} p_j = \sum_{i=1}^{m} x_i p^T \geq \sum_{i=1}^{m} w_i,$$

This means that every inequality in this sequence must be equal. Thus, we have the following characterization of a competitive equilibrium.

Theorem 6.3.2 *Every competitive Equilibrium in multi-class network resource market has the following properties:*
1 (Supply is equal to demand), $\sum_i^m x_{ij} = s_j, \forall j$;
2 (All users' budgets go to provider), $s p^T = \sum_i^m w_i$;
3 (Every user only purchases the most valuable class of service resource), if $x_{ij} > 0$, *then:*
$(\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p_j - w_i (\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = 0$.

proof:

We have already showed the properties 1 and 2 above. We will only prove property 3 here.

From the KKT condition (6.17), if $x_{ij} > 0$, we have, $\gamma = 0$.

Now take this $\gamma = 0$ into (6.15), we have, $\nabla_{x_i} u_i(x_i, \bar{x}_i) - \lambda p^T = 0$.

Together with (6.16), we get the property 3 as follows, $(\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p_j - w_i (\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = 0$.

Thus, we prove Theorem 6.3.2.

We notice that the necessary and sufficient equilibrium conditions (6.20) are all linear, except (6.19):

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p \geq w_i \nabla_{x_i} u_i(x_i, \bar{x}_i)$$

We further assume the multi-class network has fixed bandwidth supply for each class, that is $S = \{s\}$ is unique. We can now prove the uniqueness of the competitive equilibrium as follows.

We already have the partial derivative of $u_i(x_i, \bar{x}_i)$ from (6.10) as,

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = \frac{\beta_{ij}}{x_{ij}} (\mu c_j - \sum_{k=1, k \neq i}^{m} x_{kj}), \forall j$$

Multiply x_i to both sides of above equation, we find:

$$\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i = \sum_{j=1}^{n} \beta_{ij} (\mu c_j - \sum_{k=1, k \neq i}^{m} x_{kj}) \tag{6.21}$$

From the equilibrium property 1 of the Theorem 6.3.2, we have, $\sum_{i=1}^{m} x_{ij} = s_j, \forall j$. Take it into (6.10) and (6.21), we get the following equations:

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = \frac{\beta_{ij}}{x_{ij}} (\mu c_j - s_j + x_{ij}), \forall j$$

$$\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i = \sum_{j=1}^{n} \beta_{ij} (\mu c_j - s_j + x_{ij})$$

Now, we take the above two equations into (6.19) and write this nonlinear inequality using the logarithmic transformation as:

$$\log(w_i) \leq \log(p_j) + \log(\frac{x_{ij}}{\beta_{ij}(\mu c_j - s_j + x_{ij})}) + \log(\sum_{j=1}^{n} \beta_{ij} (\mu c_j - s_j + x_{ij})) \tag{6.22}$$

We have already assumed that s is unique and, in Section 6.2, we defined the relationship between s_j and c_j in (6.7), therefore, the vector c is also fixed. We also find that the function on the right side of (6.22) is a strictly concave function in x_{ij} and p_j. The left hand side of (6.22) is a constant, therefore, (6.22) is a convex inequality. Based on the property of convex inequality in [101], we can now state the following theorem.

Theorem 6.3.3 *In a multi-class network, with a fixed network resource for each class, the competitive equilibrium set is convex and the equilibrium can be computed in polynomial time.*

To show the uniqueness of this solution, we proceed as follows.

From property 3 of Theorem 6.3.2, when $x_{ij} > 0$, we have:

$$\log(w_i) = \log(p_j) + \log(\frac{x_{ij}}{\beta_{ij}(\mu c_j - s_j + x_{ij})}) + \log(\sum_{j=1}^{n} \beta_{ij}(\mu c_j - s_j + x_{ij}))$$

Let $[x^1, p^1]$ and $[x^2, p^2]$ be two distinct competitive equilibriums. Since the equilibrium set is convex, the point $[0.5x^1 + 0.5x^2, 0.5p^1 + 0.5p^2]$ is also an equilibrium, so that:

$$\log(w_i) = \log(0.5p_j^1 + 0.5p_j^2) + \log(\frac{0.5x_{ij}^1 + 0.5x_{ij}^2}{\beta_{ij}(\mu c_j - s_j + 0.5x_{ij}^1 + 0.5x_{ij}^2)}) + \log(\sum_{j=1}^{n} \beta_{ij}(\mu c_j - s_j + 0.5x_{ij}^1 + 0.5x_{ij}^2))$$

$\forall x_{ij}^1, x_{ij}^2$, satisfy either $x_{ij}^1 > 0$ or $x_{ij}^2 > 0$,

As showed before, the right side of above equation is strictly concave in p and x. From definition 6.1.2, we have,

$$\log(0.5p_j^1 + 0.5p_j^2) + \log(\frac{0.5x_{ij}^1 + 0.5x_{ij}^2}{\beta_{ij}(\mu c_j - s_j + 0.5x_{ij}^1 + 0.5x_{ij}^2)}) + \log(\sum_{j=1}^{n} \beta_{ij}(\mu c_j - s_j + 0.5x_{ij}^1 + 0.5x_{ij}^2)) >$$

$$0.5(\log(p_j^1) + \log(\frac{x_{ij}^1}{\beta_{ij}(\mu c_j - s_j + x_{ij}^1)}) + \log(\sum_{j=1}^{n} \beta_{ij}(\mu c_j - s_j + x_{ij}^1))) +$$

$$0.5(\log(p_j^2) + \log(\frac{x_{ij}^2}{\beta_{ij}(\mu c_j - s_j + x_{ij}^2)}) + \log(\sum_{j=1}^{n} \beta_{ij}(\mu c_j - s_j + x_{ij}^2))) \geq$$

$$0.5\log(w_i) + 0.5\log(w_i) = \log(w_i)$$

$\forall x_{ij}^1, x_{ij}^2$, satisfy either $x_{ij}^1 > 0$ or $x_{ij}^2 > 0$.

Thus, we must have $p^1 = p^2$, and $x^1 = x^2$, which imply that the equilibrium point is unique. We can now state,

Theorem 6.3.4 *In a multi-class network, with fixed network resource for each class, the competitive price equilibrium $p = (p_1, ..., p_n)$ and resource allocation $x = (x_1, ..., x_m)$ is unique.*

6.4 Budget Allocation in Competitive Network Resource Market

In data communication networks, the bandwidth constraint exists either because of the limitation of the user equipment or the access speed of the network. In almost all cases, either the speed of the access network or the speed limitation of the user's equipment is predefined. Therefore, the bandwidth constraint for each user is fixed. In this section, we consider how to adjust the initial budget to satisfy each user's bandwidth constraint.

Assume there is a budget agent who adjusts budget for each user to satisfy their access bandwidth constraint b_i and we assume that $\sum_i b_i \geq \sum_j s_j$. That means that the total users' bandwidth constraint is higher or equal to the total available network bandwidth supply. A competitive market equilibrium $[w, p, x_1, ..., x_m]$ must satisfy:

1 (User optimality) x_i is a maximizer of (6.4) given $\bar{x}_i$, w and p for every i.

2 (Service provider optimality) s is a maximizer of (6.8) given p.

3 (Market efficiency) $p \geq 0$, $\sum_{i=1}^m x_{ij} \leq s_j$, $p_j(\sum_{i=1}^m x_{ij} - s_j) = 0$ for all j.

4 (Budget adjustment) given x, w is a minimizer of:

$$\text{minimize}_{w_i} \sum_i (max\{0, \sum_j x_{ij} - b_i\}) w_i; s.t. \sum_i w_i = m, w \geq 0 \qquad (6.23)$$

Equation (6.23) says that if user i's access bandwidth constraint is broken, that is $\sum_j x_{ij} - b_i \geq 0$, then the budget agent will allocate less budget to user i. And any budget allocation is optimal if $\sum_j x_{ij} \leq b_i$ for all i. That means every user's access bandwidth constraint is met.

Since the budget adjustment problem is a bounded linear optimization problem and all other maximizations are identical as described in Section 6.3, we can state the following theorem.

Theorem 6.4.1 *The multi-class network resource market has a competitive equilibrium which satisfies access bandwidth constraint and each competitive equilibrium has the following properties.*
1 (Supply is equal to demand), $\sum_i^m x_{ij} = s_j, \forall j$;
2 (All users' budgets go to provider), $sp^T = \sum_i^m w_i$;
3 (Each user's bandwidth constraint is met), $\sum_j x_{ij} \leq b_i, \forall i$;
4 (Every user only purchases the most valuable class resource), if $x_{ij} > 0$, *then:*

$$(\nabla_{x_i} u_i(x_i, \bar{x}_i)^T x_i) p_j - w_i (\nabla_{x_i} u_i(x_i, \bar{x}_i))_j = 0.$$

As already shown in Section 6.3 that in a multi-class network, with a fixed network resource for each class, the competitive price equilibrium $[p, x_1, ..., x_m]$ is unique and can be calculated in polynomial time. We use the following iterative algorithm for budget allocation to satisfy the bandwidth constraint:

```
Algorithm 6.4.2 Initialize budget assigned to each user $w_i = 1, i = 1, ..., m$;
repeat
Compute competitive economy equilibrium $[x_1, ..., x_m, p]$ under $s = (s_1, ..., s_n)$ and $w = (w_1, ..., w_m)$.
Obtain the allocated bandwidth to each user i, $\sum_j x_{ij}$;
Calculate average bandwidth surplus, $avg_s = \frac{\sum_i (b_i - \sum_j x_{ij})}{m}$;
Update $w_i = w_i + \frac{(b_i - \sum_j x_{ij}) - avg_s}{k}, i = 1, ..., m$.
Until $b_i - \sum_j x_{ij} \geq$ error tolerance, $i = 1, ..., m$.
```

In each iteration, the unique competitive equilibrium is derived given network resource for each class s_j and budget for each user w_i. The user budget is reassigned according to the bandwidth shortage of each user in the equilibrium solution. The idea of comparing the user's bandwidth surplus with average surplus means less budget allocation to the users with lower bandwidth surplus while keeping the total budget unchanged. Here, k is a scalar parameter. We can adjust k to get to the competitive equilibrium with bandwidth constraint with different pace.

6.5 An Illustrative Example

This section presents a numerical example about the pricing and resource allocation in a multi-class network resource market as described in this chapter. We will show the competitive equilibrium properties of this market.

We assume the network supports three different service classes ($n = 3$) with the capacity c=100. The service class 1 supports real time gaming and the QoS parameter average delay defined in SLA as $T_{1_{SLA}} = 0.04s$; the service class 2 and service class 3 are designed to carry interactive streaming service and non-interactive streaming service respectively with $T_{2_{SLA}} = 0.1$ and $T_{3_{SLA}} = 0.2$. To emphasize the method, we assume there are three users ($m = 3$) competing for the network resource. Based on the utility function proposed in (6.4), the utility functions for the users are,

$$u_1 = (\mu c_1 - x_{21} - x_{31}) \log(\frac{x_{11}}{4}) + (\mu c_2 - x_{22} - x_{32}) \log(\frac{x_{12}}{2}) + (\mu c_3 - x_{23} - x_{33}) \log(\frac{x_{13}}{5}) \tag{6.24}$$

Class 1	Class 2	Class 3
$T_{1SLA} = 0.04s$	$T_{2SLA} = 0.1s$	$T_{3SLA} = 0.2s$
$c_1 = 45$	$c_2 = 30$	$c_3 = 25$
$s_1 = 20$	$s_2 = 20$	$s_3 = 20$

Table 6.1: Resource allocation among classes

$$u_2 = 1.1(\mu c_1 - x_{11} - x_{31}) \log(\frac{x_{21}}{4}) + (\mu c_2 - x_{12} - x_{32}) \log(\frac{x_{22}}{2}) + (\mu c_3 - x_{13} - x_{33}) \log(\frac{x_{23}}{5}) \quad (6.25)$$

$$u_3 = 1.2(\mu c_1 - x_{11} - x_{21}) \log(\frac{x_{31}}{4}) + (\mu c_2 - x_{12} - x_{22}) \log(\frac{x_{32}}{2}) + (\mu c_3 - x_{13} - x_{23}) \log(\frac{x_{33}}{5}) \quad (6.26)$$

As shown in (6.24) (6.25) (6.26), all three users have the minimum bandwidth requirement for class 1 service as $\tilde{x_{11}} = \tilde{x_{21}} = \tilde{x_{31}} = 4$, class 2 as $\tilde{x_{12}} = \tilde{x_{22}} = \tilde{x_{32}} = 2$ and class 3 as $\tilde{x_{13}} = \tilde{x_{23}} = \tilde{x_{33}} = 5$. The weighting factors of the first user for different class service as $\beta_{11} = \beta_{12} = \beta_{13} = 1$. The second user and the third user have higher weighting factors for class one service and identical weighting factors for class two and three services as, $\beta_{21} = 1.1, \beta_{31} = 1.2, \beta_{22} = \beta_{23} = \beta_{32} = \beta_{33} = 1$.

As proven in Section 6.3, when the network resource allocation among classes is fixed, the market has an unique competitive equilibrium. We further assume the network allocate 45% of network resource to class 1 service, 30% to class 2 service and 25% to class 3 service. To satisfy the QoS agreement in each class's SLA, based on the Equation (6.7), the available network resource for each class is 20 as shown in Table 9.1. We also assume the average length of packets in the network $1/\mu = 1$ and the initial endowments for each users as $w_1 = 8; w_2 = 10; w_3 = 12$. Then the competitive solution is as follows.

$$p_1 = 0.8294253; p_2 = 0.3939162; p_3 = 0.2766585;$$
$$x_{11} = 5.1842; x_{12} = 5.6658; x_{13} = 5.3071;$$
$$x_{21} = 6.6393; x_{22} = 6.7135; x_{23} = 6.6822;$$
$$x_{31} = 8.1765; x_{32} = 7.6207; x_{33} = 8.0107;$$

The above results show that under competitive equilibrium:

1 Each user spent all the budget: $w_i = \sum_{j=1}^{3} p_j x_{ij}, \forall i = 1, 2, 3;$

2 All users budget goes to service provider: $\sum_{i=1}^{3} w_i = \sum_{j=1}^{3} p_j s_j;$

3 Demand is equal to supply for each class bandwidth: $s_j = \sum_{i=1}^{3} x_{ij}, \forall j = 1, 2, 3;$

4 Each class's QoS is well-maintained: $D_j = \frac{1}{\mu c_j - \sum_{i=1}^{3} x_{ij}} = T_{j_{SLA}}, \forall j = 1, 2, 3.$

From the solution we also get the total bandwidth for the first user is $x_1 = \sum_{j=1}^{3} x_{1j} = 16.1571$, the second user is $x_2 = \sum_{j=1}^{3} x_{2j} = 20.0349$ and the third user is $x_3 = \sum_{j=1}^{3} x_{3j} = 23.8080$. we can also calculate the utility for the first user as $u_1 = 35.7132$, for the second user as $u_2 = 59.5291$ and for the third user as $u_3 = 83.9198$; and therefore the social utility has the value $u_1 + u_2 + u_3 = 179.1621$.

Now consider each of them has a bandwidth constraint $b_1 = 15, b_2 = 20, b_3 = 30$. From the above example, if the budget agent sets $w_1 = 8; w_2 = 10; w_3 = 12$, the bandwidth allocation can not satisfy the bandwidth constraint. By the algorithm proposed in Section 6.4, we can adjust the initial budget endowments and get the competitive results as follows. In this example, we set the error tolerance as 0.01.

$$w_1 = 7.474; w_2 = 9.984; w_3 = 12.542;$$

$$p_1 = 0.829419; p_2 = 0.393402; p_3 = 0.277179;$$

$$x_{11} = 4.8997; x_{12} = 5.2783; x_{13} = 4.8115;$$

$$x_{21} = 6.6319; x_{22} = 6.7182; x_{23} = 6.6398;$$

$$x_{31} = 8.4684; x_{32} = 8.0035; x_{33} = 8.5487;$$

As shown above, when we adjust the initial budget endowments to $w_1 = 7.474, w_2 = 9.984$ and $w_3 = 12.542$, on the competitive equilibrium, the bandwidth allocation for all three users allocated bandwidth satisfy their bandwidth constraints: $x_1 = \sum_{j=1}^{3} x_{1j} = 14.9894 \leq b_1 = 15, x_2 = \sum_{j=1}^{3} x_{2j} = 19.99 \leq b_2 = 20$ and $x_3 = \sum_{j=1}^{3} x_{3j} = 25.0206 \leq b_3 = 30$.

6.6 Chapter Summary

This chapter has considered multi-class network resource in a competitive market where each user endowed with an initial budget will purchase bandwidth from each class of the network resource to maximize its utility function. After defining utility functions for users in a multi-class network, we have proved that there exists a unique competitive equilibrium $[p, x_1, ..., x_m]$ when we fix the available bandwidth for each class. This competitive equilibrium can be calculated in polynomial time. For bandwidth constraint due to users' equipment or the speed of the access network, the proposed algorithm adjusts the initial budget for each user to satisfy their respective constraints. Further, we have proved that a competitive equilibrium $[w, p, x_1, ..., x_m]$, under bandwidth constraints, always exists for the multi-class resource network as long as the sum of the total bandwidth constraints is equal to or exceeds the total bandwidth supply. Since the competitive equilibrium is Pareto optimal, the proposed solution achieves both higher social utilization and better individual satisfaction than the Nash equilibrium.

CHAPTER 7

A User-friendly Constant Revenue Model for Net Neutrality

The recent lawsuit between Comcast and BitTorrent has brought widespread attention and the net neutrality debate has polarized the major stakeholders, hardening their respective stands. On August 17, 2007, Comcast was reported to prevent BitTorrent users from seeding files [102]. Later, Comcast's limiting of Bit Torrent applications was further confirmed in a study conducted by the Electronic Frontier Foundation [103]. At the same time, Comcast argued that it considered choking BitTorrent traffic as a way to let the network traffic remain available for everyone. In January 2008, FCC Chairman Kevin Martin stated that the FCC was going to investigate complaints that Comcast "actively interferes with Internet traffic as its subscribers try to share files online" [104]. On August 21, 2008, the FCC issued an order which stated that Comcast's network management was unreasonable and that Comcast must terminate the use of its discriminatory network management by the end of the year. In December 2009 Comcast admitted no wrongdoing in its proposed settlement of up to 16 million dollars and decided to appeal the FCC's ruling with the US DC Court of Appeals to protect its rights, claiming that the FCC's decision was not based on any existing legal standards. The Court of Appeals sided with Comcast in its April 7, 2010 decision, pointing out that FCC has never been given by the Congress the authority to control the Internet network management, and agreed that there were no legal grounds for making Comcast stop its practice [105]. The court's decision means that broadband service providers are now free to manage their networks as they wish and this could prompt more discrimination on the Internet.

Net neutrality was first proposed by Columbia Law School professor Tim Wu, and is used to signify the idea that "maximally useful public information network aspires to treat all content, sites and platforms equally" [7]. While a formal process for the implementation of the principle does not exist, net neutrality usually means that broadband service providers charge consumers only once for the Internet access and do not favor one content provider over another, and do not charge content providers for sending varying amounts of information over broadband lines to end users [106]. Simply put, the network neutrality principle is that all Internet traffic should be treated equally, a philosophy which network neutrality proponents (online content providers like Google, Microsoft and others) claim would preserve the principles on which the Internet was founded. Tim Berners-Lee, the founder of the World Wide

Web, also favors keeping net neutrality in place, since “the Internet is the basis of a fair competitive market economy” [107].

However, Broadband service providers like at&t, Verizon, and Comcast, among others, view net neutrality as being unfair to: (a) broadband service providers themselves and (b) light network users (compared to heavy users with the same access charge). Opponents also believe that prioritization of bandwidth is necessary for future innovation on the Internet [108]. The broadband service providers argue it is the service providers who have put their resources which they have to maintain and upgrade for their customers. They also argue that heavy-duty users and popular content providers (like Google, Skype) get a relatively “free ride” on their network which costs billions of dollars to build [109]. Lack of additional sources of revenue might act as a disincentive for broadband service providers to upgrade their infrastructure which, in turn, will affect the service providers’ plans of increasing capacities. Further, it is estimated that 80% of Internet traffic is caused by 5% of the user population. This 5% is causing all the traffic using P2P applications such as BitTorrent which is optimized, in many cases, to hog bandwidth. In order to keep network traffic flowing for all consumers, broadband service providers argue it is reasonable for them to use network traffic management practices to slow down P2P application performances.

Broadband service providers claim that, under net neutrality, the incentive to expand the capacity and the capabilities of the existing infrastructure for the next generation of broadband services is much lower compared to the case when they are allowed to charge the online content providers for preferential treatment. Study [110], however, shows that the incentive for the broadband service provider to expand under net neutrality is unambiguously higher than under the no net neutrality regime. This is against the assertion of the broadband service providers that under net neutrality, they have limited incentive to expand. Results presented in [111] also indicate that non net neutrality networks are not always more favorable in terms of social welfare compared to net neutrality networks.

Although it seems reasonable for broadband service providers to choke certain applications which overwhelm the network in order for traffic flowing for everyone, people are afraid that broadband service providers will slow down or even block services and applications, they consider undesirable, freely. Especially under the situation that broadband and content providers are merging and turning to digital content distribution (e.g., Comcast bought NBC Universal on Jan, 2010), without net neutrality, as Lawrence Lessig and Robert W. McChesney said, the Internet would start to look like cable TV: A handful of massive companies would controlling access and distribution of content, deciding what you get to see and how much it would cost [112]. In this chapter, we propose a solution for broadband service providers to control network congestion and maintain fairness among all consumers [113].

We propose the concept of inter-user compensations among users based on their usage of network resource. Users consuming less network resource will receive compensations from other users. One notable characteristic of this scheme is that the algebraic sum of all inter-user compensations is equal to zero which means that no inter-user compensations are lost. While compensations are among the users, broadband service providers' revenue are not affected by this scheme. In other words, broadband service providers' revenue remains constant under this model. We view this characteristic as being important since the network users would see such a mechanism positively in as much as the network provider does not take advantage of them by increasing its own profitability.

In this chapter, we assume that all users are responsible for the cost of the network and we model this cost sharing problem as a cooperative game. The cost share associated with each user corresponds to its Shapley Value of the cooperative game. We consider that the total cost of a coalition as the network resource required to maintain the desired QoS for the traffic in the coalition. The inter-user compensations are established based on the difference between their cost share and the actual price they pay by way of access fees to the broadband service provider. The broadband service provider can use a scale parameter to these inter-user compensations. During the peak period, the broadband service providers can increase these inter-user compensations to regulate the heavy users so as to control network congestion; otherwise, the broadband service provider can reduce these inter-user compensations to keep the network load at a desirable level. Another characteristic of this scheme is that the broadband service providers control network congestion and maintain fairness among all users without discriminatory treatment of any traffic flowing on the network. In other words, the net neutrality is well-maintained.

The rest of chapter is organized as follows. In Section 7.1, we present the model. The cooperative game and cost share are also studied in this section. Section 7.2 investigates the inter-user compensations scheme. We present an illustrative example in Section 7.3. We also discuss the application of this mechanism. Section 7.4 captures our conclusions.

7.1 Problem Statement and the Model

The network is modeled as a queuing network with First in First out (FIFO) discipline. Users in the network are denoted by a set as $N = \{1, ..., n\}$, and we use $|N|$ to denote the length of the set or the number of users in the network. The most prevalent pricing scheme in communication networks is access-rate dependent flat-rate charge. For example, Verizon offers DSL Internet service which includes a starter package ($19.99 per month for download speeds up to 1.0 Mbps); power package ($29.99 per month for download speeds up to 3.0 Mbps) and turbo package ($39.99 per month for download speeds up to 10.0 Mbps). To emphasize our inter-user compensations mechanism, we further assume that all users choose the same package. Our mechanism can be easily extended to multiple packets of service when we consider inter-user compensations within the subscribers of each package.

Although all users pay the same amount to access the Internet, the traffic generated by each user is different. We use the average arrival rate λ_i to denote user i's traffic, and all users' traffic is assumed to follow Poisson distribution. The average delay D is used as the predefined QoS objective which is agreed to by network users and the network provider in their service level agreement (SLA). We also assume that the packet lengths of all packets in the network are exponentially distributed with the average packet length of $\frac{1}{\mu}$.

Based on [88], the aggregate traffic from a number of independent users with Possion distribution still follows the Poissonian arrival discipline. We use λ to denote the aggregate traffic arrival rate. To maintain the required QoS D as described in the SLA, the required network resource c can be calculated using the M/M/1 queuing model as follows:

$$D = \frac{1}{\mu c - \lambda} \tag{7.1}$$

We can rewrite (7.1) as:

$$c = \frac{1}{D\mu} + \frac{\lambda}{\mu} \tag{7.2}$$

From (7.2), we can find the total resource required to support the QoS promised by the network provider for the aggregate traffic λ. This problem of resource allocation among users generating varying levels of traffic can be modeled as a joint cost allocation problem and we can use the Shapley Value to solve it.

The Shapley Value has been used as a method of joint-cost allocation instead of the traditional accounting allocation bases since the 1970s [114]. The Shapley Value was introduced by Shapley in 1953 as a method for each player to assess the benefits he would expect from playing a game. It consistently produces a unique allocation that virtually all researchers consider fair and equitable. This method distributes the total gain of cooperation based upon the assumption that the cost of a participant in a coalition is determined by the incremental cost attributed to that participant in the coalition. Since the order in which participants join a coalition affects the incremental payoff produced, the Shapley Value considers all orderings equally likely and weights them equally. This generates an allocation solution that impartial observers would consider fair and desirable [62].

To show its application to the problem of equitably assigning joint cost among individual users, the total network resource c as described in (7.2) among all users, we first define the worth of a coalition function $c(S)$. After that, we will compute the Shapley value. The coalition function $c(S)$ describes the required network resource to maintain the QoS for users in the coalition $S, S \subseteq N$ when the users in S share the network resource together. By (7.2), we find:

$$c(S) = \frac{1}{D\mu} + \frac{\sum_{i \in S} \lambda_i}{\mu} \tag{7.3}$$

The incremental network resource used by a user $i \in N$ to the coalition $S, i \notin S$ is:

$$c(S \cup \{i\}) - c(S) = \frac{\lambda_i}{\mu} \tag{7.4}$$

From (7.2) and (7.4) together, we can see the economy of scale in sharing the network resource: To maintain the average delay D, the incremental network resource required $\frac{\lambda_i}{\mu}$ for user i if users in $S \cup \{i\}$ sharing the resource together is smaller than the required network resource $\frac{1}{D\mu} + \frac{\lambda_i}{\mu}$ when user i was to be allocated the network resource individually.

The Shapley Value is defined as a weighted sum of the user's marginal contribution to all possible coalitions [84]. For all $i \in N$, the Shapley Value to user i is given by:

$$SV_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [c(S \cup \{i\}) - c(S)] \tag{7.5}$$

As shown before, the term $c(S \cup \{i\}) - c(S)$ calculates the incremental network resource attributed to user i in the coalition S. This incremental resource allocation occurs for exactly those orderings in which the participant i is preceded by $|S|$ other players in $S \cup \{i\}$ and followed by $|N| - |S| - 1$ players not in $|S|$. This means there are exactly $|S|!(|N| - |S| - 1)!$ orderings of interest. $|N|!$ determines the total number of coalition permutations that can be created from the participants. Taken together, the expression $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ is the weighting factor that assigns equal share of the marginal contribution generated to each coalition of interest. The network resource consumed by the user i is thus weighted and summed for all possible coalitions where i appears in S. Thus, each user i is allocated a value equal to its expected incremental contribution across all possible coalitions. We have thus developed the desirable network resource share for each class by using the corresponding Shapley value.

Take (7.4) into (7.5), we get a more clear expression of network resource share for each user:

$$SV_i = \frac{1}{|N|D\mu} + \frac{\lambda_i}{\mu} \tag{7.6}$$

From (7.6), the network resource attributed to each user i depends on their usage λ_i: the larger the λ_i, the more the network resource consumed by user i. In addition, (7.6) also shows the economy of scale in the network resource consumption. For example, if user j's traffic is equal to $\lambda_j = 2 * \lambda_i$, the network resource share for user j, is smaller than $2 * SV_i$; that is $SV_j = \frac{1}{|N|D\mu} + \frac{\lambda_j}{\mu} < \frac{2}{|N|D\mu} + \frac{2*\lambda_i}{\mu}$.

7.2 Inter-user Compensations Scheme

From Section 7.1, we find out how much network resource is consumed by each user using the Shapley Value in a joint-cost allocation circumstance. In the beginning of Section 7.1, we also stated that all users pay the same amount to access the network. In order to solve the cross-subsidization between light and heavy users, we propose an Inter-user compensations scheme. Besides the fairness between light users and heavy users, broadband service providers can also use the inter-user compensations scheme to control the behavior of heavy users to further control congestion in the network, flatten the peak hours of usage, and thus maintain the QoS.

The rational behind the flat rate access is that all users are supposed to consume the same amount of network resource. From (7.2), the total network resource required to maintain average delay D for all users is:

$$c(N) = \frac{1}{D\mu} + \frac{\sum_{i=1}^{n} \lambda_i}{\mu} \tag{7.7}$$

Therefore, each user is assumed to consume a network resource equal to $\frac{c(N)}{|N|}$, that is,

$$\hat{c} = \frac{c(N)}{|N|} = \frac{1}{|N|D\mu} + \frac{\sum_{i=1}^{n} \lambda_i}{|N|\mu} \tag{7.8}$$

Further, we define the inter-user compensation t_i as the difference between actual network resource it consumed, SV_i and its allocated resource consumption $\hat{c}$:

$$t_i = k * (SV_i - \hat{c}) = k * \frac{|N| * \lambda_i - \sum_{j=1}^{n} \lambda_j}{|N|\mu} \tag{7.9}$$

where k is a scale parameter discussed below.

Equation (7.9) shows that when $SV_i \geq \hat{c}$, user i consumed more network resource than it paid for the service and therefore it will compensate others by an amount equal to t_i. And when $SV_i \leq \hat{c}$, user i consumed less network resource than it paid for the service and therefore it will receive a compensation equal to t_i from others. The scale parameter k is used to to control the inter-compensation dynamically to further control network congestion. For example, when total traffic on the network is heavy, the broadband service provider can increase k to increase the compensation heavy users give to light users in order to control heavy users' usage. This mechanism maintains fairness among users and solves the network congestion problem on an equitable basis.

We can check the algebraic sum of inter-user compensations as follows:

$$\sum_{i=1}^{n} = k * \sum_{i=1}^{n} (SV_i - \hat{c}) = k * \sum_{i=1}^{n} \frac{|N| * \lambda_i - \sum_{j=1}^{n} \lambda_j}{|N|\mu} = 0 \tag{7.10}$$

Equation (7.10) shows that sum of inter-user compensations is equal to zero and no inter-user compensations are lost. Thus the inter-user compensation mechanism is strictly among users. Broadband service providers use a scale parameter k to control the actual amount by which the heavy users compensate the light users. The circulation of the compensation is among the users only and the broadband service provider is a neutral party in this mechanism. The proposed mechanism thus guarantees the neutrality of the network. The broadband service provider's revenue is the same as before the proposed inter-user compensation mechanism was instituted. The flat rate access fee remained unchanged.

When the required network resource $c(N)$ to maintain the QoS is smaller than a threshold $c_{threshold}$, the scale parameter k is set to 0. The threshold can be assumed as a trigger for invoking the inter-user compensations mechanism. Broadband service providers can set this trigger point by observing the network traffic pattern. For example, the threshold could be set at the point corresponding to the network utilization of 0.5. When the total network load is low, the relatively heavy users should not be punished for using the network resource, thus encouraging them to shift their load to low usage periods.

When the required network resource $c(N)$ to maintain the QoS exceeds the threshold $c_{threshold}$, it signals the broadband service provider to control the heavy users' usage by setting the scale parameter k appropriately. We can assume k to be a function of $c(N)$, the larger the $c(N)$, the larger the value of k. In considering setting the inter-user compensation among users, the broadband service provider also needs to ensure that the compensations light users receive should always be smaller than the price of access.

7.3 An Example and Further Discussion

In this section, we first present a numerical example to show the inter-user compensation mechanism proposed in this chapter. Further, we discuss the application of this mechanism in broadband networks.

We assume there are three users in the network and the QoS parameter D agreed in the SLA is equal to $0.05s$. The average length of packets in the network is equal to 1, that is $\frac{1}{\mu} = 1$. The average arrival for each user is: $\lambda_1 = 5; \lambda_2 = 15; \lambda_3 = 20$. Based on (7.7), the total network resource required to maintain average delay $0.05s$ is equal to:

$$c(1,2,3) = \frac{1}{D\mu} + \frac{\lambda_1 + \lambda_2 + \lambda_3}{\mu} = 60$$

The network resource share for each user can be calculated based on (7.6):

$$SV_1 = \frac{1}{3 * D\mu} + \frac{\lambda_1}{\mu} = 11.67$$

$$SV_2 = \frac{1}{3 * D\mu} + \frac{\lambda_2}{\mu} = 21.67$$

$$SV_3 = \frac{1}{3 * D\mu} + \frac{\lambda_3}{\mu} = 26.66$$

Therefore, using (7.9) we can find the inter-user compensation as follows:

$$t_1 = k * (SV_1 - \frac{c(1,2,3)}{3}) = -8.33 * k$$

$$t_2 = k * (SV_2 - \frac{c(1,2,3)}{3}) = 1.67 * k$$
$$t_3 = k * (SV_3 - \frac{c(1,2,3)}{3}) = 6.66 * k$$

From the above, we find that user one will receive a compensation of $8.33 * k$; while users two and three will compensate user one by $1.67 * k$ and $6.66 * k$ respectively. The total inter-user compensation is equal to 0.

When the total available network resource is large, let's say the network capacity is equal to 100, then network utilization at this moment is:

$$\rho = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\mu * 100} = 0.4$$

If the broadband network provider assume this network utilization is relatively low, then they can set $k = 0$ without deterring more traffic from entering the network and penalizing heavy users when such users don't cause congestion.

However, if the network capacity is equal to 50, the network utilization becomes:

$$\rho = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\mu * 50} = 0.8$$

In this case, the broadband service provider will chose k (which is no longer zero) appropriately in order to restrain the usage of users two and three so as to reduce the network utilization to a desirable level.

One reason that Comcast argued in its appeal to FCC's ruling with the US DC Court of Appeals is that Comcast does not target traffic from specific applications. Instead it begins to throttle traffic from heavy users. However, this also violates the philosophy of net neutrality since all Internet traffic should be treated equally by the neutral network. The proposed inter-user compensation scheme proposes an acceptable economic concept to manage the network without trying to target any specific entity within the network.

Since, in the proposed scheme, the broadband service provider is neutral and the compensations are circulated among users, this mechanism should be much more acceptable by the users as the traffic management scheme. It will also appeal to broadband service providers. The service provider can set up the threshold point and also the actual amount that the heavy users will pay light users. The service provider can be guided by the traffic pattern on the network and its own strategy to manage the network. For example, the service provider can trigger the inter-user compensations early but the compensations can be relatively small, or trigger the inter-user compensations late with relatively larger amounts.

Implementation of this mechanism will inevitably add more work to the network management system. However, from (7.9), the computational complexity of finding the inter-user compensation is linear and it should be acceptable by the broadband service provider.

7.4 Chapter Summary

In this chapter, we have proposed an inter-user compensations mechanism for a broadband service provider to manage the network traffic while maintaining the philosophy of net neutrality. Users consuming less network resource will receive compensation from heavy users. All these compensations are among users only and the broadband service provider keeps neutral in the process. The computational complexity of calculating the inter-user compensations is linear and this mechanism should be acceptable for the broadband service provider's management system. The broadband service provider's revenue remains constant. The proposed mechanism discourages heavy users during peak periods thus flattening the network usage.

CHAPTER 8

A Constant Revenue Model for Packet Switched Network

The Quality of Service (QoS) offered by a Packet switched networks is characterized by delay experienced by a packet as it transitions through the network. Such characterization is always on a statistical basis that typifies the macro behavior of the network to incident traffic. As is well known, unlike circuit switched networks where all blocked calls are lost (and therefore characterized by the probability of blocking), packet switched networks are characterized by their delay behaviour. The delay rises rapidly as the level of average traffic load increases, making the network unstable as the utilization of the network increases beyond, say, 80%. A wide variety of techniques including Integrated Services, Differentiated Services and Multi-Protocol Label Switching has been proposed and/or implemented in network to provide quality of service (QoS).

This chapter proposes another approach to QoS by implementing a differentiated pricing scheme by classifying customers into priority or non-priority groups. The priority customers continue to be offered a higher QoS than the nonpriority customers. Priority customers would pay a higher price for this privilege. Two pricing schemes are considered. In both the cases considered, the higher priority services receive a statistically guaranteed QoS while the lower priced services receive varying levels of service. The pricing for the higher priority services varies depending upon the total incident traffic and the fraction of customers choosing the higher priority service. The two cases are differentiated depending on whether the pricing for the lower priced service is kept fixed, or there is variable pricing for both the higher and lower priced services.

A fundamental assumption in developing the pricing scheme for both the pricing schemes is that the gross revenue of the network is kept constant while the network is operating within a predefined range. As described in Chapter 7, we offer two reasons for making this assumption. First, the clients of the network would view such a pricing mechanism positively in as much as the network does not take advantage of a high demand by increasing its own profitability. Second the cost associated with a telecommunications network is largely fixed and independent of the level of traffic requesting service. (This characteristic is quite unlike other distribution networks such as the power grid where the largest cost component, namely fuel, can be adjusted depending on demand.) With a constant level of cost, it makes sense to keep the revenue constant, targeting the profit to a fixed value consistent with realities of a competitive marketplace.

As discussed, the approach we propose in this chapter aims to keep the revenue for the network provider constant. Reference [115] has presented an approach similar to ours but directed to a circuit switched environment. The approach taken in this chapter addresses the packet switched environment. The remainder of the chapter is organized as follows. Section 8.1 presents a network model on which the proposed two pricing schemes are based. Section 8.2 presents details of the pricing strategy proposed. Section 8.3 presents results of the analysis and an illustrative example using the proposed model. Section 8.4 captures out conclusion.

8.1 Problem Statements and the Model

The network model proposed in this chapter is depicted in Fig. 8.1 The packet-switched network is represented by a single communication server with a defined capacity. An M/M/1 traffic model is assumed. The average packet arrival rate is denoted by λ. We assume the capacity of the communications server to be C and the average length of the packets $\frac{1}{\mu}$.

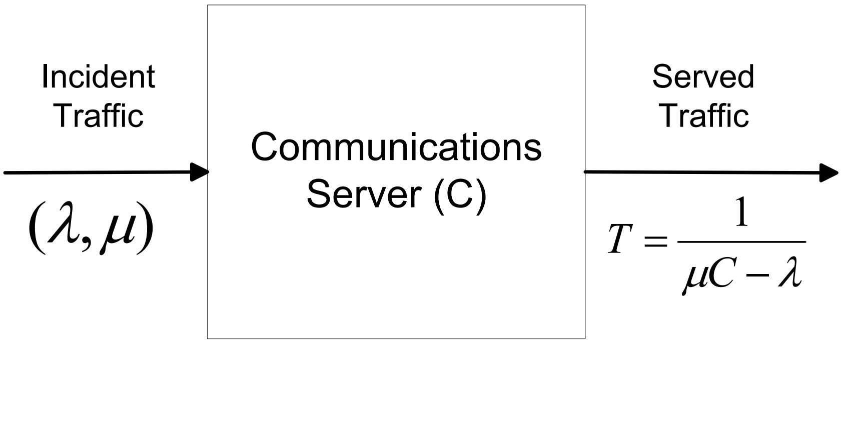


Figure 8.1: Network model in chapter 8

Based on above network model, Fig. 8.2 shows the relationship between the mean delay and the incident traffic as a function of the utilization factor. Fig. 8.2 assumes the network parameters to be: $C = 250, \mu = 1$.

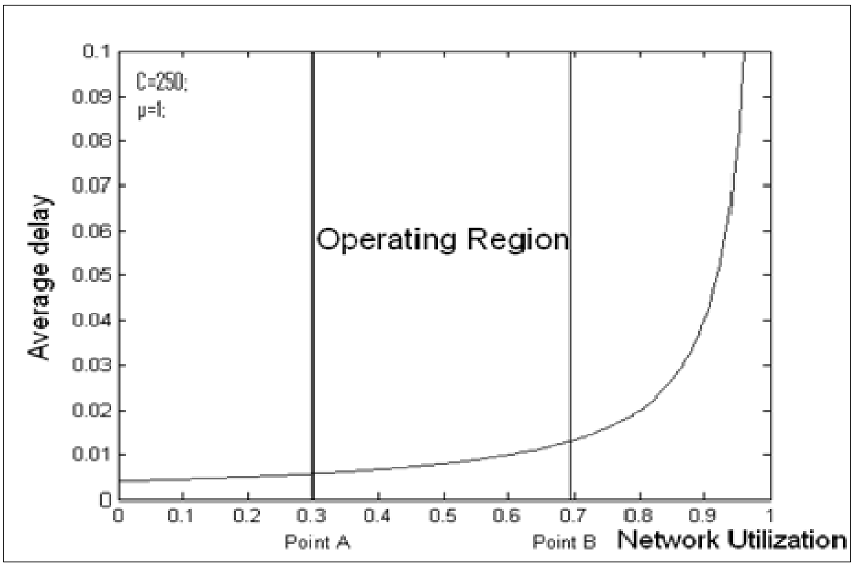


Figure 8.2: Average delay vs. network utilization

As we can see from Fig. 8.2, the average delay increases rapidly when the network utilization exceeds its design limit. For most applications in practice, the network utilization is rarely allowed to exceed 0.8 because of the risk of indefinite queue build up and the resultant instability. Since no common user network can maintain a predefined QoS and avoid the negative impact of congestion as the traffic increases indefinitely, its important to institute and manage a process that will control traffic. There are multiple ways of accomplishing it. One way to avoid congestion is to simply not accept new traffic while the network is in or tending toward a state of congestion. The other way is to institute a differentiated pricing mechanism such that users are motivated to control the originating traffic on their own based on the needs of the application and their willingness to pay.

We adopt the latter approach in this chapter and, except under a very low or high level traffic, have the network offer two tiers of service. The point at which the differentiated services are introduced is indicated as point A in Fig. 8.2. A differentiated pricing scheme introduced prior to the network reaching a traffic level corresponding to point A will not accomplish its objective because the delay experienced by a packet will be only fractionally above the minimum possible delay. A mature network, in general, will be operating well past point A. Fig. 8.2 assumes, somewhat arbitrarily, that point A corresponds to a level of utilization equal to 0.3. The two-tier pricing scheme is introduced at this point.

Admission control is introduced at point B when the network cannot provide the guaranteed QoS to higher priority service if all the customers chose this service and have the willingness to pay the higher price. At this point, the average delay for the non-priority traffic would become indefinitely large. Fig. 8.2 assumes that the average delay guaranteed for the priority traffic within the operating range of the network is 0.012 seconds which we assume to be a fair value based on ITU recommendation for interactive stream applications [116].

Beyond the point A, customers have the option to choose priority service with guaranteed end to end average delay, or non-priority service on a best effort basis. We use q as the parameter representing the fraction of traffic that chooses priority. Fig. 8.3 illustrates that when the level of traffic reaches point A, the network introduces the priority and nonpriority services as mentioned earlier. As shown in this figure, $q * \lambda$ traffic chooses priority service and $(1 - q) * \lambda$ traffic chooses non-priority service.

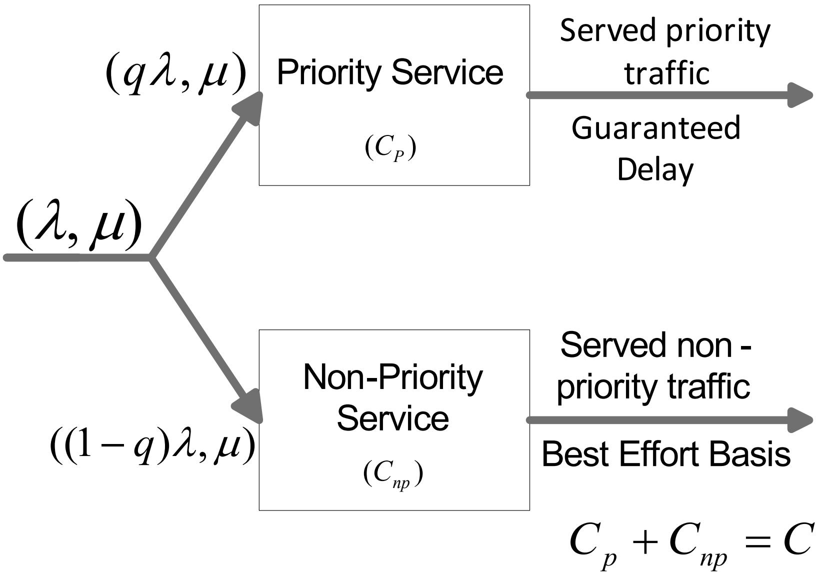


Figure 8.3: Priority and non-priority services in the operating region

8.2 Pricing Strategies

As stated in the beginning of this chapter, two pricing strategies are introduced in this chapter. The common factor associated with each of the strategies is that the overall revenue of the network remains constant.

8.2.1 Delay Analysis

In order to provide priority and non-priority services in the network, the network capacity C is divided into two groups, C_p serving the priority traffic and C_{np} serving the non priority traffic. Obviously, we have:

$$C = C_p + C_{np} \tag{8.1}$$

We already know that q is the fraction of traffic choosing the priority service. In order to keep the average delay for priority traffic as D, C_p can be evaluated from,

$$D = \frac{1}{\mu C_p - q\lambda} \tag{8.2}$$

The average delay for non-priority traffic T is:

$$T = \frac{1}{\mu C_{np} - (1-q)\lambda} \tag{8.3}$$

Using Equation (8.1), (8.2) and (8.3), we get:

$$T = \frac{1}{\mu C - \lambda - 1/D} \tag{8.4}$$

An interesting observation from the Equation 8.4 is that the average delay T of non-priority traffic is independent of q.

To keep the total revenue constant at R, different prices need to be charged to priority and non-priority traffic, respectively. In following sections, we use P_{1p} and P_{1np} to denote prices charged for priority traffic and non-priority traffic in pricing strategy 1, while we use P_{2p} and P_{2np} to indicate prices charged for priority traffic and non-priority traffic in pricing strategy 2.

8.2.2 Pricing Strategy 1: Pricing Based on Resource Consumption by the Respective Service Categories

In this section, prices for priority traffic (P_{1p}) and non-priority traffic (P_{1np}) are calculated dynamically based on their relative consumption the network resource. We, therefore, have,

$$P_{1p} * q\lambda = \frac{R * C_p}{C} \tag{8.5}$$

And we can rewrite it as:

$$P_{1p} = \frac{R * C_p}{C * q\lambda} \tag{8.6}$$

Similarly, for non-priority group, we have:

$$P_{1np} * (1 - q)\lambda = \frac{R * C_{np}}{C} \tag{8.7}$$

And we can rewrite it as:

$$P_{1np} = \frac{R * C_{np}}{C * (1 - q)\lambda} \tag{8.8}$$

As can be seen from Equations (8.6) and (8.8), the total revenue generated by the two service classes is constant at R, while the contribution made by each of the two service classes is equal to their respective traffic volumes multiplied by the respective prices. Additionally, for fairness in pricing, we must have the following inequality satisfied,

$$P_{1p} > P_{1np} \tag{8.9}$$

within the operating region of the network.

8.2.3 Pricing Strategy 2: Constant Non-priority Service Price

In the second pricing strategy, price for the non-priority service is kept constant at a certain level (P_{2np}), while the price for the priority service (P_{2p}) is calculated based on P_{2np}, the parameter q and the overall revenue for network provider R.

The non-priority service price (P_{2np}) is computed at point B since this is the point at which the network is operating most efficiently. We have earlier defined the point A somewhat arbitrarily. Point B at which the network must institute an admission control mechanism to prevent the network from going into congestion can be defined with a higher degree of specificity. We define it as the point beyond which the network can no longer provide the guaranteed QoS to the higher priority service if all the customers chose this service and were willing to pay the higher price associated with the service.

We use λ_B to denote the arrival rate at point B, and q_B to represent the percentage of traffic choosing priority service at point B. We can then calculate the price for priority service P_{1pB} and non-priority service P_{1npB} based on pricing strategy 1, developed in Section 8.2.2. We use this P_{1npB} as constant price for non-priority service. We thus have,

$$P_{2np} = P_{1npB} \tag{8.10}$$

$P_{2p} * q * \lambda + P_{2np} * (1 - q) * \lambda = R$; and we can rewrite it as:

$$P_{2p} = \frac{R - P_{2np} * (1 - q) * \lambda}{q * \lambda} \tag{8.11}$$

Just as in strategy 1, since the price for non-priority service cannot exceed that of the priority service, we must have,

$$P_{2p} > P_{2np} \tag{8.12}$$

8.3 An Illustrative Example

This section presents a numerical example based on pricing schemes proposed in Section 8.2. We will develop pricing and delay profiles against the traffic load λ and the fraction of priority traffic q in operating region.

We choose the following parameter for the network: $C = 250$, $\mu = 1$, the guaranteed delay for priority traffic D as 0.012 seconds [116]. We further assume that the revenue R for the network provider is 275 for a 10 percent profit to make it competitive in telecommunications industry.

Fig. 8.4 presents the resulting delay for both priority and non-priority traffic against traffic load with different q.

It can be observed from Fig. 8.4 that the priority traffic, as required by our assumption, always has a constant end-to-end average delay regardless of the network load and percentage of traffic choosing priority service within the operating region. For

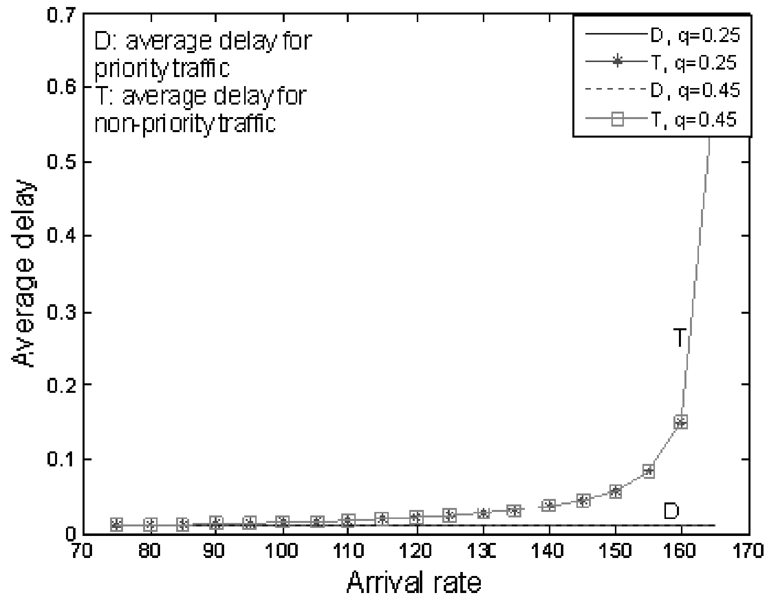


Figure 8.4: Delay against arrival rate with different percentages priority traffic

the non-priority traffic, the average delay increases as the arrival rate increases. As shown in Equation (8.4), it can be observed that the average delay for non-priority traffic is identical regardless of q for any given network load.

8.3.1 Results for Pricing Strategy 1

Fig. 8.5 presents the resulting prices for both priority and non-priority traffic based on their network resource consumption while keeping the overall revenue constant. The percentage of traffic choosing priority (q) is parameterized for two values, $q = 0.25$ and $q = 0.45$.

It can be seen from Fig. 8.5 that prices for both priority and non-priority traffic decrease as the network traffic load increase. For a given arrival rate λ, the price for the priority traffic decreases as q increases while the opposite is the case for non-priority traffic. This phenomenon reflects the economy of scale with increasing traffic for the priority and non-priority service categories. Fig. 8.4 and Fig. 8.5 indicate that a service category is more attractive when more people choose it since the price for that category will be lower with the same latency.

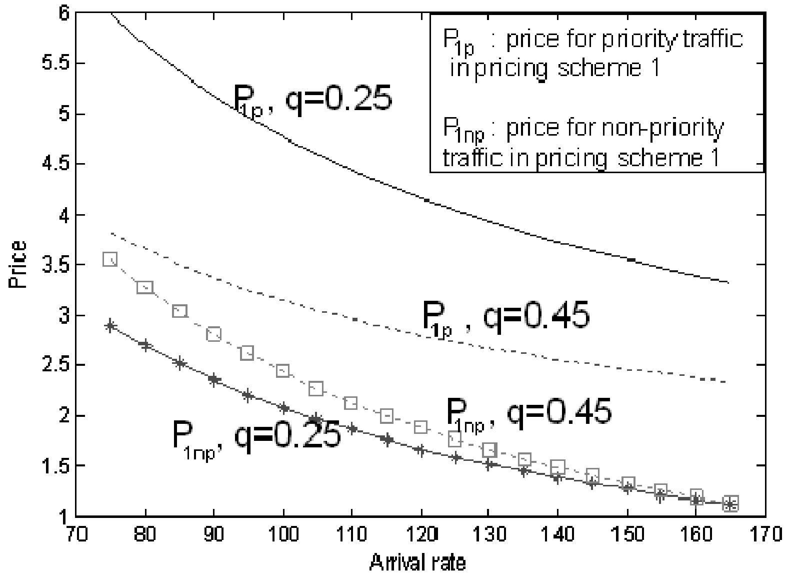


Figure 8.5: Prices for priority and non-priority traffic based on pricing strategy 1

8.3.2 Results for Pricing Strategy 2

Fig. 8.6 presents the resulting prices for both priority and non-priority traffic based on the constant price for non-priority traffic while keeping the overall revenue constant. As stated in Section 8.2.3, the constant non-priority service price is computed at point B and we further assume the percentage of traffic choosing priority service at point B, $q_B = 0.1$. Using Equation (8.8) we can calculate this constant price for non-priority traffic (P_{2np}) and the price for priority service (P_{2p}) can be computed from Equation (8.11). The percentage of traffic choosing priority (q) is parameterized as 0.25 and 0.45 as in the previous case.

As can be seen in Fig. 8.6, the price for the non-priority traffic is kept constant in the operating region. The price for priority traffic decreases as the traffic increases. At any given arrival rate, prices for the priority service are lower with higher q.

8.3.3 Discussion of Results

In pricing strategy 1, priority traffic has a constant guaranteed delay and varying prices, while both delay and price for non-priority traffic are changing based on relative resource consumption. In pricing strategy 2, priority traffic has fixed delay and varied prices, while non-priority traffic has fixed price and varied average delay.

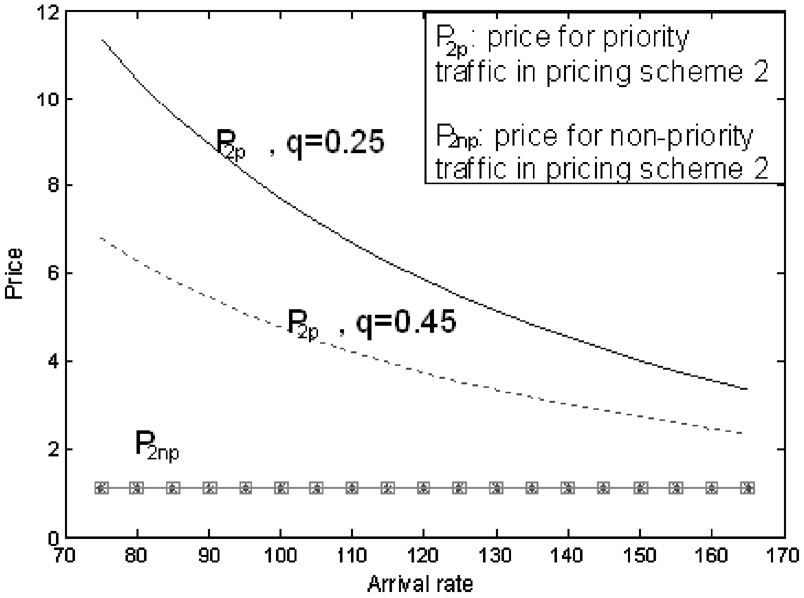


Figure 8.6: Prices for priority and non-priority traffic based on pricing strategy 2

We found that the price for priority servicer is relative higher in pricing scheme 2 than in pricing scheme 1 given identical λ and q. For example, when $\lambda = 90$ and $q = 0.25$, the price for priority service is nearly 5 in pricing scheme 1 and 8 in pricing scheme 2. It also can be seen that pricing scheme 1 shows relatively little variation in priority prices both as a function of varying λ and q. The reason for above phenomenon is that in pricing scheme 2, constant price for non-priority traffic calculation is at the most efficient network operating point B. As a result that non-priority traffic is taking advantage of priority traffic in pricing scheme 2 in operating region.

8.4 Chapter Summary

This chapter has presented a two-tier pricing schemes for instituting priority and non-priority services in a packet switched network. The pricing schemes are such that the network generates constant revenue even with varying load and varying fraction of customers choosing priority and non-priority services. The first pricing scheme is based on the two service classes being priced on the basis of their respective resource consumption. The other scheme keeps the price for the non-priority service fixed in the operating region.

As expected, prices for both priority and non-priority service classes decrease with increasing overall traffic, validating the economy of scale principle. Prices for priority traffic decrease with increasing q and the opposite is the case for non-priority traffic except for scheme 2 when the price for non-priority traffic is kept constant. We assume pricing scheme 1 is a fairer method than pricing scheme 2 since the prices calculated in scheme 1 depend on the network resource consumption, namely cost. However, since there is strong public preference for a constant pricing scheme for telecommunications services [117], the constant price for non-priority traffic is still very attractive to customers.

CHAPTER 9

A Two-step Quality of Service Provisioning in Multi-class Networks

In this chapter, we investigate the resource allocation issue in a multi-class DiffServ network. The scalability and efficient network resource consumption of class-based networks compared to flow based network architectures like IntServ [119] has been clearly established. In a class-based network, IP flows are classified and aggregated into different forwarding classes which provide different QoS services according to the Service-Level-Agreement (SLA) between the network provider and the end user class.

In a common user network, the overall network resource is generally not dedicated to a single class or a certain user but shared by multiple classes and all users. Uncontrolled consumption of network resources might result in lower profit for the service provider and reduced satisfaction to users since it will lead to deteriorated throughput and unacceptable level of delay. In order to maximize the network utilization while guaranteeing service levels for different classes as described in SLAs, bandwidth allocation and flow control should be enforced.

Since class-based network architecture is stateless from a per flow perspective, flow control should be enforced on the edge of network and the network resource should be allocated on a per-class basis in the network core. Reference [82] has considered an edge router for metering, policing and shaping the incoming flows before aggregation into a limited number of classes. Recently, there are increasing numbers of responsible applications which are able to adjust their transmission rate according to the network condition. Unlike [82], we assume all flows are responsible flows in this chapter. Responsible flows have also been considered in [32, 120, 16]. The well-known resource scheduling schemes used in the core network include priority queuing (PQ), Weighted Round- robin (WRR) and Class-Based Queuing (CBQ) [121, 81]. However, these schedulers all keep a static service weight regardless of the actual number of aggregations in each class. Although dynamic schedulers like Fair WRR [82] have been proposed, the change of service weight is done somewhat arbitrarily without explicit network performance objective.

In this chapter, we model the network resource allocation among different classes (inter-class) as a centralized optimization problem to maximize the social welfare which is defined as the sum of all user utilities. For flow control in each class (intra-class), we develop a distributed game theoretic framework to regulate the individual flow behavior, where each flow competes for resources within the class to maximize its own performance.

Modeling the network resource allocation as a centralized maximization social benefit problem on the basis of the knowledge of user utility functions has been considered in [16, 38, 122]. The maximization approach used in [38] solves the resource allocation problem in the context of network providing best-effort service. This chapter considers the resource allocation problem in a multi-class network environment where each class has an explicit QoS guarantee. In addition, unlike the utility function used in [38] which is only related to the transmission rate, we take not only the transmission rate but also the QoS parameter into the consideration for defining the utility function. In order to solve the scalability problem, when the number of sources becomes large, [38] forms a distributed flow control algorithm using gradient ascent algorithm from optimization theory. Our approach to the resource allocation mechanism does not suffer from the same scalability problem since the network provider only supports a limited number of classes. And the intra-class flow control mechanism is enforced separately using the Nash arbitration/bargaining framework.

The remainder of this chapter is organized as follows. Section 9.1 describes the network structure considered in this chapter. In Section 9.2, we introduce a game theoretic framework to control the flow behavior. For each class, Nash Arbitration Scheme (NAS) is computed and the characteristics of flows are explored. Section 9.3 shows dynamic network resources allocation among different classes to maximize social welfare while keeping the QoS in each class at a predefined value. Admission control conditions are also discussed. Section 9.4 gives a numerical example of the proposed flow control and resource allocation architecture and Section 9.5 captures the conclusions of this chapter.

9.1 Problem Statements and the Model

We consider a packet switched network with m classes as shown in Fig. 9.1. The network is modeled as a queuing network with First in First out (FIFO) discipline. The network resource C is allocated among the m classes in order to maximize social benefits which will be discussed in Section 9.3. Within each class, n_i users (or flows) share the allocated C_i and compete to maximize their individual performance objectives (defined later).

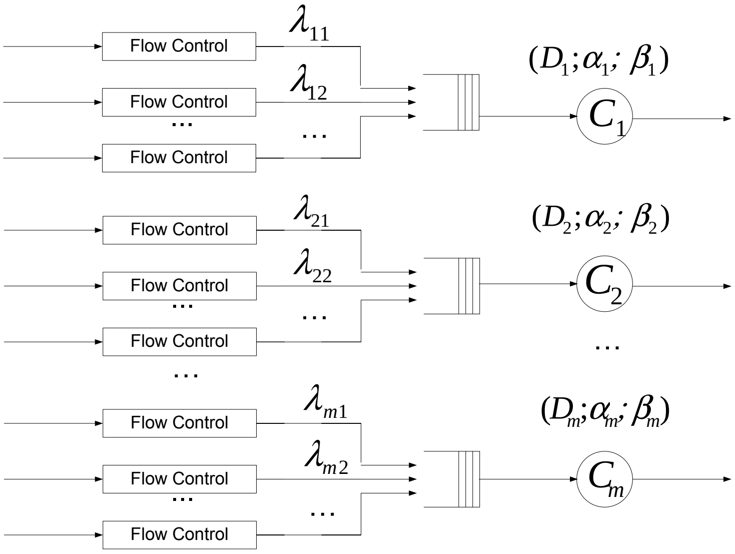


Figure 9.1: Network model in chapter 9

D_i is used as the predefined QoS objective for class i which is agreed to by class i users and the network provider in their SLA. T_i is the actual QoS experienced by class i flows. We denote a flow j in class i as $flow_{ij}$: $\lambda_{ij} \geq \lambda_{ijm}, T_i \leq D_i$, where λ_{ij} describes this flow's transmission rate and λ_{ijm} is the minimum rate requirement. We also assume the packet length of all flows is exponentially distributed with average length equal to $1/\mu$. Obviously, we have the actual average delay T_i for class i as:

$$T_i = \frac{1}{\mu C_i - \lambda_i} \tag{9.1}$$

where $\lambda_i = \sum_{j=1}^{n} \lambda_{ij}$ as the total traffic in class i.

Each flow is trying to maximize its performance objective, which is expressed as power given by Kleinrock in [123]: The weighted throughput of the system divided by the corresponding average delay in the system. We extend Kleinrock's definition to a multi-class situation. As a result, the power for flow j in class i is as follows:

$$P_{ij} = \frac{\lambda_{ij}^{\alpha_{ij}}}{T_i^{\beta_{ij}}} \tag{9.2}$$

where α_{ij}, β_{ij} are the sensitivity parameters for flow j in class i. α_{ij} describes the flow's sensitivity to the transmission rate and β_{ij} denotes the flow's tolerance to delay. We can see that if $\alpha_{ij} \geq \beta_{ij}$, the flow is more sensitive to transmission rate, or in another word, insensitive to delay, and vice versa. Without loss of generality, we assume

$0 < \alpha_{ij} \leq 1, 0 < \beta_{ij} \leq 1$.

Normally, flows within the same class have the same QoS and minimum bandwidth requirement and, in our following analysis, we further assume flows belonging to the same class have identical sensitivity parameters and minimum transmission rate, that is: $\alpha_i = \alpha_{ij}, \beta_i = \beta_{ij}, \lambda_{im} = \lambda_{ijm}, i = 1, ..., m; j = 1, ..., n_i$.

9.2 Intra-class Flow Control

Within each class i, every flow competes for the limited allocated network resource C_i and tries to maximize its power. From (9.1) and (9.2), we can derive a more explicit description of power as follows,

$$P_{ij}(\lambda_i) = \lambda_{ij}^{\alpha_{ij}} * (\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij})^{\beta_{ij}} \tag{9.3}$$

As stated before, n_i is used to denote the number of flows in class i and vector $\lambda_i = (\lambda_{i1}, \lambda_{i2}, ..., \lambda_{in_i})$ is the transmission rate for each flow in class i. We can see from (9.3) that the power of $flow_{ij}$ (P_{ij}) not only depends on its own transmission rate λ_{ij} but also on transmission rates of other flows in that class. Therefore, it is natural for us to model this problem as a n-party game.

The Pareto optimal point is a globally efficient solution and has better or at least equal payoff for each player than the Nash equilibrium point. We further assume that this game is cooperative and the Pareto optimality can be found. Let a vector $\lambda_i^* = (\lambda_{i1}^*, \lambda_{i2}^*, ..., \lambda_{in_i}^*)$ be the transmission rate for each flow in class i under Pareto optimality. By the definition of Pareto optimality [124], the condition,

$$P_{ij}(\lambda_i^*) \leq P_{ij}(\lambda_i^* + \Delta), j = 1, ..., n_i \tag{9.4}$$

can not be met (Δ is a non-zero vector). It means that the Pareto optimality represents global maximization and it is impossible to find another point which leads to better payoff for at least one player without degrading the payoff to others.

In general, in a game with n players, the Pareto optimal points form an $n - 1$ dimensional hypersurface and it means that there are an infinite number of points which are Pareto optimal. As we said before, an optimal network operation point should be a Pareto optimal point. The question here is: Which of these infinite Pareto optimal points should we choose to operate the network?

One way to find suitable Pareto-optimal points for operation is by introducing further criteria. When we consider network resource sharing, one of the natural criteria is the notion of fairness. The notion of fairness is not well defined. There are many different ways to express it like proportional fairness [38], max-min fairness [125], etc. In this chapter, we use the axioms of the fairness from game theory as the fairness criteria [126]. Nash arbitration scheme (NAS) which encapsulates the requirements of yielding Pareto optimality as well as satisfying the axioms of fairness is proposed in this section to find the suitable Pareto optimal point for each class. Stefanescu et al [127] characterize the NAS as follows.

Let $f_j : X \to R, j = 1, 2, ..., n$ be concave upper-bounded functions defined on X which is a convex and compact subset of R^n and $f(x) = (f_1(x), ..., f_n(x))$.

Let $U = \{u \in R^n : \exists\, x \in X, \text{s.t. } f(x) \geq u\}$ and $X(u) = \{x \in X : f(x) \geq u\}$, $X_0 = X(u_0)$ the subset of strategies that enable the users to achieve at least their initial performance.

Then the NAS is given by the point which maximizes the following unique function,

$$\text{maximize}_x \prod_{j=1,...,n} (f_j(x) - u_j^0) \tag{9.5}$$

From (9.3), $P_{ij}(\lambda_i)$ is defined on $(\lambda_{i1}, \lambda_{i2}, ..., \lambda_{in_i})$ and the transmission rate of each flow j in class i should be smaller than the service rate of each class i, μC_i, therefore, we have $0 \leq \lambda_{ij} \leq \mu C_i$, for $j = 1, ..., n_i$. Thus, $P_{ij}(\lambda_i)$ is defined on a convex and compact subset of R^{n_i}.

Now, we take the partial derivation and second partial derivation of $P_{ij}(\lambda_i)$ to check its concavity and upper-bound characteristics.

$$\nabla_{\lambda_{ij}}(P_{ij}(\lambda_i)) =$$

$$\lambda_{ij}^{\alpha_{ij}-1}(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij})^{\beta_{ij}-1}[\alpha_{ij}(\mu C_i - \sum_{k=1,k\neq j}^{n_i} \lambda_{ik}) - (\alpha_{ij} + \beta_{ij})\lambda_{ij}]$$

From above equation, we find that when $\lambda_{ij} \leq \frac{\alpha_{ij}(\mu C_i - \sum_{k=1,k\neq j}^{n_i} \lambda_{ik})}{\alpha_{ij}+\beta_{ij}}$, P_{ij} is monotonically increasing with respect to λ_{ij}, and when $\lambda_{ij} \geq \frac{\alpha_{ij}(\mu C_i - \sum_{k=1,k\neq j}^{n_i} \lambda_{ik})}{\alpha_{ij}+\beta_{ij}}$, P_{ij} is monotonically decreasing with respect to λ_{ij}. Therefore, we say that P_{ij} is upper-bounded at $\lambda_{ij} = \frac{\alpha_{ij}(\mu C_i - \sum_{k=1,k\neq j}^{n_i} \lambda_{ik})}{\alpha_{ij}+\beta_{ij}}$ given the $n_i - 1$ vector $(\lambda_{i1}, ..., \lambda_{i(k-1)}, \lambda_{i(k+1)}, ..., \lambda_{in_i})$.

Now, let's check the second partial derivation of $P_{ij}(\lambda_i)$.

$$\nabla^2_{\lambda_{ij}}(P_{ij}(\lambda_i)) = \lambda_{ij}^{\alpha_{ij}-2}(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij})^{\beta_{ij}-2}[\alpha_{ij}(\alpha_{ij}-1)(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij}) +$$

$$\beta_{ij}(\beta_{ij}-1)(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij}) - \alpha_{ij}\beta_{ij}\lambda_{ij}(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij})]$$

From (9.1), we know $(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij}) \geq 0$ since delay can not be negative. And we already assumed in Section 9.1 that the sensitive parameters α and β are defined as, $0 < \alpha \leq 1$ and $0 < \beta \leq 1$. From above equation, we conclude that $\nabla^2_{\lambda_{ij}} P_{ij}(\lambda_i) < 0$ and P_{ij} is a concave function on λ_{ij}.

We also assume that each user has an initial arrival rate 0 and now we are ready to calculate the suitable Pareto optimal by solving the following maximization problem,

$$\text{maximize}_{\lambda_i} \prod_{j=1,\ldots,n_i} (P_{ij}(\lambda_i)) \tag{9.6}$$

which leads to:

$$\nabla_{\lambda_{ij}}(\prod_{j=1,\ldots,n_i} (P_{ij}(\lambda_i))) = 0, j = 1, \ldots, n_i \tag{9.7}$$

And (9.7) is equivalent to:

$$\alpha_{ij}(\mu C_i - \sum_{j=1}^{n_i} \lambda_{ij}) - \lambda_{ij}(\sum_{j=1}^{n_i} \beta_{ij}) = 0, j = 1, \ldots, n_i \tag{9.8}$$

From (9.8), we obtain a linear system of equations with a unique solution as follows,

$$\lambda_{ij}^* = \frac{\mu C_i \alpha_{ij}}{\sum_{j=1}^{n_i}(\alpha_{ij} + \beta_{ij})} \tag{9.9}$$

Given the assumption in Section 9.1 that all flows within the same class have identical QoS requirement, the transmission rate for each flow in class i under NAS therefore is:

$$\lambda_{i1}^* = \lambda_{i2}^* = \ldots = \lambda_{in_i}^* = \frac{\mu C_i}{n_i(1 + \frac{\beta_i}{\alpha_i})} \tag{9.10}$$

The actual system delay experienced by each flow can be calculated using (9.1):

$$T_i = \frac{1 + \frac{\alpha_i}{\beta_i}}{\mu C_i} \tag{9.11}$$

Equation (9.10) shows that under NAS, each flow has identical transmission rate. In the definition of Power in Section 9.1, α_i and β_i are related to the QoS requirements for each class. Equation (9.11) further exemplifies this statement in a way that the parameter T_i is dependent on $\frac{\alpha_i}{\beta_i}$ with a given class resource C_i. From (9.11), we also find that the equilibrium T_i does not degrade by the increasing number of flows in the class. Therefore, we can use $\frac{\beta_i}{\alpha_i}$ as a QoS indicator of class i. The bigger $\frac{\beta_i}{\alpha_i}$ is, the smaller T_i becomes and the better QoS class i gets.

Under the game theoretic flow control framework described in this section, the QoS of each class can be maintained no matter how many flows reside in them. With this property, a fair and efficient inter-class resource allocation mechanism among different classes is proposed in Section 9.3.

9.3 Inter-class Resource Allocation

The network resource is allocated dynamically among classes based on the network situation to maximize the social benefit. Social benefit is the sum of the utility functions of each user. The most well-known utility function as proposed by Kelly [38] has the form $U_{ij} = w_{ij} log \lambda_{ij}$, where w_{ij} is user's willingness to pay and λ_{ij} is the transmission rate.

In this chapter, we define a new utility function as follows:

$$U_{ij} = \frac{\beta_i}{\alpha_i} log \lambda_{ij} \tag{9.12}$$

Equation (9.12) describes that the utility's relation to the QoS indicator $\frac{\beta_i}{\alpha_i}$ and the transmission rate λ_{ij}. This is consistent with the notion that the user's utility is proportional to the QoS and the logarithm of the transmission rate since the better the QoS, the more utility the user gets. This utility function also fits into Kelly's utility function in a way that the better the QoS the more willingness to pay for the service. The reason we use a logarithmic function lies in the marginal utility as a function of transmission rate diminishing as the rate increases. This is consistent with [78].

In Section 9.2, we have shown that under NAS in each class, each flow converges to the identical flow rate. As a result, the sum of all flows' utility function in each class i can be given as:

$$U_i = n_i\frac{\beta_i}{\alpha_i}log\lambda_{ij} = n_i\frac{\beta_i}{\alpha_i}log\frac{\mu C_i}{n_i(1+\frac{\beta_i}{\alpha_i})} \tag{9.13}$$

As assumed in Section 9.1, there are m classes in the network, therefore the social benefits maximization problem becomes:

$$\text{maximize}_{C_i}\sum_{i=1}^{m} n_i\frac{\beta_i}{\alpha_i}log\lambda_{ij} = \sum_{i=1}^{m} n_i\frac{\beta_i}{\alpha_i}log\frac{\mu C_i}{n_i(1+\frac{\beta_i}{\alpha_i})} \tag{9.14}$$

subject to the following constraints:

$$\sum_{i=1}^{m} C_i \leq C \tag{9.15}$$

$$T_i = \frac{1+\frac{\alpha_i}{\beta_i}}{\mu C_i} \leq D_i, i = 1, ..., m \tag{9.16}$$

$$\lambda_i = \frac{\mu C_i}{1+\frac{\beta_i}{\alpha_i}} \geq n_i\lambda_{im}, i = 1, ..., m \tag{9.17}$$

Inequality (9.15) states that the resource allocated among all m classes subject to the total available network resource C. Inequality (9.16) holds that for each class, the average delay experienced should be smaller than the promised parameter in SLA. It has been shown in Section 9.2 that, under NAS, each flow within the same class has the same transmission rate and (9.17) is used to maintain minimum bandwidth for each flow.

We simplify the constraints (9.16) and (9.17), then get:

$$C_i \geq \frac{1+\frac{\alpha_i}{\beta_i}}{\mu D_i}, i = 1, ..., m$$

and

$$C_i \geq \frac{n_i(1+\frac{\beta_i}{\alpha_i})\lambda_{im}}{\mu}, i = 1, ..., m$$

Thus, constraints (9.16) and (9.17) can be simplified as a single constraint as follows,

$$C_i \geq C_{im}, i = 1, ..., m \tag{9.18}$$

where $C_{im} = \max(\frac{1+\frac{\alpha_i}{\beta_i}}{\mu D_i}, \frac{n_i(1+\frac{\beta_i}{\alpha_i})\lambda_{im}}{\mu})$.

When $\sum_{i=1}^{m} C_{im} \geq C$, the network is oversubscribed and by no means can provide the promised QoS and minimum bandwidth requirement for each flow as defined in SLA, and therefore admission control has to be introduced. As a result, for the admission control, we have:

$$\sum_{i=1}^{m} C_{im} \leq C \tag{9.19}$$

It can be observed that the expression in (9.14) is a strictly concave function over a closed and bounded set defined by (9.15) and (9.18). Therefore a unique maximum always exists. We now use Lagrangian multipliers to append constraints to the objective. Thus, we can rewrite it as:

$$\text{maximize}_{C_i} \sum_{i=1}^{m} n_i \frac{\beta_i}{\alpha_i} log \frac{\mu C_i}{n_i(1+\frac{\beta_i}{\alpha_i})} - \gamma_0(\sum_{i=1}^{m} C_i - C) + \sum_{i=1}^{m} \gamma_i (C_i - C_{im}) \tag{9.20}$$

The necessary and sufficient Karush-Kuhn-Tucker (KKT) [91] conditions applicable to (9.20) are given by:

$$\frac{n_i \beta_i}{C_i \alpha_i} - \gamma_0 + \gamma_i = 0 \tag{9.21}$$

$$\gamma_0(\sum_{i=1}^{m} C_i - C) = 0 \tag{9.22}$$

$$\gamma_i(C_i - C_{im}) = 0, i = 1, ..., m \tag{9.23}$$

We denote the optimum network resource allocation among m classes as $(C_1^*, C_2^*, ..., C_m^*)$.

This centralized network resource allocation mechanism doesn't suffer from the scalability problem for the reason that the network provider only needs to maintain a limited number of classes of service and our resource allocation is made among classes. Furthermore, since the utility function we used in this section is logarithmic, the solution obtained has the proportionally fair property as shown by Kelly in [38].

Thus, the resource allocation state in this chapter is an efficient and fair mechanism from both intra-class and inter-class perspectives. The intra-class resource allocation is modeled as a multi-party game. Each flow tries to maximize its power in a distributed way. On the NAS, each flow has identical fraction of network resources and the QoS is not impacted by the number of flows in the class. The inter-class resource allocation is based on maximizing social benefits while allowing each class to maintain its QoS and minimum bandwidth requirement.

We can now make some observations regarding the admission control in each class.

In class i, when the allocated resource C_i^* satisfies the following equation:

$$\frac{\mu C_i^*}{1 + \frac{\beta_i}{\alpha_i}} = n_i \lambda_{im} \tag{9.24}$$

the admission control is considered to begin in this class. Equation (9.24) suggests that using the KKT condition (9.23), $C_i^* - C_{im} = 0$. And together with (9.18), we have:

$$T_i^* = \frac{1 + \frac{\alpha_i}{\beta_i}}{\mu C_i^*} \leq D_i \tag{9.25}$$

From (9.11) and (9.24), the average delay for class i can be rewritten as:

$$T_i^* = \frac{\alpha_i}{\beta_i n_i \lambda_{im}} \tag{9.26}$$

Equations (9.25) and (9.26) show that in class i, the flows will receive the lower delay T_i^* than the SLA agreed delay D_i and T_i^* decreases with increasing number of flows in this class for the scale efficiency. Admission control is considered when (9.24) is first satisfied because the network load is becoming heavy when (9.24) is met. Another reason is that since T_i^* is already lower than D_i, decreasing T_i^* is not as valuable as the increasing the transmission rate for flows in other classes. Therefore, this is the point at which admission control for that class needs to be introduced.

9.4 An Illustrative Example

This section presents a numerical example about the network resource allocation and flow control mechanisms proposed in this chapter. We will develop the QoS of each class and allocated bandwidth of each flow profile using the proposed mechanisms.

For simplicity, we assume the network has three different service classes ($m = 3$) with the capacity $C = 100$. The service class 1 is supposed to support real time gaming and the QoS parameter average delay defined in SLA as $D_1 = 0.04s$; while each flow within this class has the identical bandwidth to delay weighting factor $\frac{\alpha_1}{\beta_1} = 0.9$ and minimum transmission rate requirement $\lambda_{1m} = 1.8$. Interactive streaming service will be carried in class 2; the delay and transmission rate requirement is

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
Number of flows in Class 1	$n_1 = 6$	$n_1 = 7$	$n_1 = 8$	$n_1 = 9$	$n_1 = 10$	$n_1 = 11$	$n_1 = 12$	$n_1 = 13$	$n_1 = 14$
Number of flows in Class 2	$n_2 = 5$	$n_2 = 6$	$n_2 = 7$	$n_2 = 8$	$n_2 = 9$	$n_2 = 10$	$n_2 = 11$	$n_2 = 12$	$n_2 = 13$
Number of flows in Class 3	$n_3 = 4$	$n_3 = 5$	$n_2 = 6$	$n_3 = 7$	$n_3 = 8$	$n_3 = 9$	$n_3 = 10$	$n_3 = 11$	$n_3 = 12$
Allocated resource for class 1	$C_1 = 48.0$	$C_1 = 47.5$	$C_1 = 47.5$	$C_1 = 47.5$	$C_1 = 47.5$	$C_1 = 47.5$	$C_1 = 47.5$	$C_1 = 49.4$	$C_1 = 53.2$
Allocated resource for class 2	$C_2 = 32.8$	$C_2 = 32.7$	$C_2 = 32.2$	$C_2 = 32.0$	$C_2 = 31.8$	$C_2 = 31.6$	$C_2 = 31.5$	$C_2 = 30.3$	$C_2 = 29.8$
Allocated resource for class 3	$C_3 = 19.2$	$C_3 = 19.8$	$C_3 = 20.3$	$C_3 = 20.5$	$C_3 = 20.7$	$C_3 = 20.9$	$C_3 = 21.0$	$C_3 = 20.3$	$C_3 = 17.0$

Table 9.1: Network resource allocation under different network scenarios

as $D_2 = 0.1s, \lambda_{2m} = 1.2$, the bandwidth to delay sensitive parameter as $\frac{\alpha_2}{\beta_2} = 1.1$. The service class 3 is designed to support non-interactive streaming service and has following parameters: $D_3 = 0.3s, \lambda_{3m} = 0.8, \frac{\alpha_3}{\beta_3} = 1.5$. We also assume the average length of packets in the network $\frac{1}{\mu} = 1$.

Table 9.1 shows the network resource allocation among the three classes using the mechanisms proposed in this chapter under different network scenarios.

For example, in scenario 1 when there are 6 flows in class 1, 5 flows in class 2 and 4 flows in class 3, class 1 will be allocated with 48.0%, class 2 will be allocate with 32.8% and class 3 with 19.2% of the network resource.

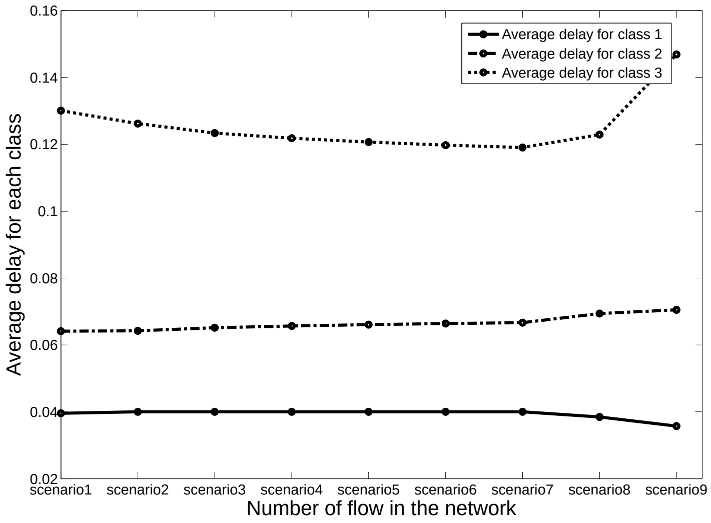


Figure 9.2: Delay experienced by each class under different network scenarios

Under the network resource allocation shown in Table 9.1, using the flow control mechanism described in Section 9.2, Fig. 9.2 describes the average delay curve of each class under different network scenarios shown in Table 9.1 and Fig. 9.3 presents how the flow transmission rate in each class changes with the increasing number of flows in the network.

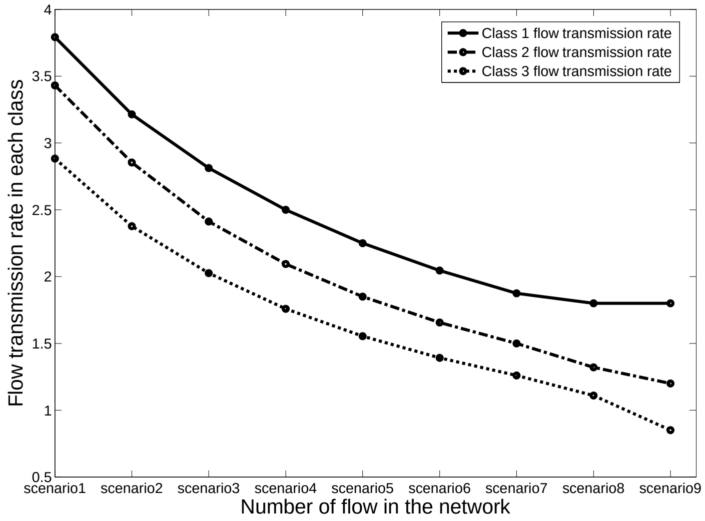


Figure 9.3: Flow rate in each class under different network scenarios

As can be seen from Fig. 9.2, even as the average delay of each class changes, they all satisfy the SLA QoS agreement; the average delay of class 1 is less than 0.04s; the average delay of class 2 is less than 0.1s; the average delay for class 3 is less than 0.3s. In Fig. 9.3, it can be observed that although the flow rate of each class decreases as the number of flows increases, they all satisfy the minimum transmission rate requirement in each class. The minimum flow rate in class 1 is 1.8; minimum flow rate in class 2 is 1.2 and minimum flow rate in class 3 is 0.8.

In Fig. 9.3, the flow rate for class 1 is kept at the minimum rate 1.8 in scenarios 8 and 9. In Fig. 9.2, we find that the average delay for class 1 in scenario 8 and 9 is decreasing. This gives a signal to consider admission control in class 1 as described in Section 9.3.

Fig. 9.3 also shows the proportionally fair transmission rate of the resource allocation mechanisms described in this chapter. The network resources allocation mechanism doesn't favor any special class but allocates the resources in a way to maximize the social benefits. For example, in scenario 1, the flow rate in class 1 is 3.8, flow rate in class 2 is 3.4 and flow rate is class 3 is 2.8. All these flow rates are above their minimum requirement by different percentages and the network utility is maximized.

9.5 Chapter Summary

In this chapter, we have presented a two-step mechanism to provide QoS in the multi-class network environment. The inter-class resource allocation problem is modeled as a centralized optimization problem to maximize the sum of all users′ utilities where we define the utility as a function of each user′s transmission rate and the QoS it received. This centralized optimization doesn′t suffer from the scalability problem since the network only needs to maintain a limited number of classes. We use a game theoretic framework to control the optimal rate within each class leading to the NAS expected in a distributed manner. As shown in Section 9.4, these mechanisms assure the QoS for each flow as described in the SLA. Further, the resource allocation to each flow results in maximizing the social benefits of the network.

CHAPTER 10

Summary and Future Work

In this dissertation, we have discussed pricing issues in multi-class communication networks. Our work can be summarized as follows:

Firstly, we have discussed desirable subsidy-free prices for each class of service in multi-class, priority-based networks. We have presented rules for assigning positions for each class in the queue and classes of packets with lower priorities and longer average waiting times receiving compensation from others. Price differences among various classes are developed based on inter-class compensations.

Secondly, we have discussed desirable subsidy-free prices for each class of service in class-based DiffServ networks. We have presented rules for allocating network resources to each class. Classes of packets that allocated less network resource and longer average delays receive monetary compensations from others. The inter-class price differential is determined by the inter-class compensations.

Thirdly, we have considered the market-clearing prices in multi-class DiffServ networks. We have proved that the market-clearing price always exists and both individual optimality and socio-economic efficiency are achieved simultaneously under this pricing scheme. Further, we have discussed how to adjust each user's initial budget to meet his or her bandwidth constraint.

Fourthly, we have provided a solution for network providers to manage the network without violating net neutrality based on inter-user compensations. Users consuming less network resource receive compensation from heavy users of network resources. Broadband access providers can use such inter-user compensation to shape the traffic characteristics of users without inflicting discriminatory treatment on network traffic.

All the above methods are proved to be economically efficient, and we believe these pricing methods can help network providers manage network traffic more effectively.

10.1 Directions for Future Work

Our work can be extended in several directions.

1. **Extensions to multi-provider market.** In this dissertation, we assume that

there exist only one service provider in the market. We would like to relax this assumption by considering multi-provider market scenarios. By doing this we will use game theory to simulate the competition among different network providers.

2. **Advanced market clearing prices for each class of service.** In Chapter 6, we prove that there exist a unique equilibrium where demand of network resource is equal to supply. One assumption in this chapter is that the network provider already knows which classes of service each user requests. Such assumption may not be practicable in reality beacuase it may be difficult for some network providers to obtain such data. Nevertheless, these network providers may still be able to estimate - from history record or survey - the probability distribution of the user classes of service requests. We plan to develop a stochastic version on the market clearing prices for each class of service.

3. **In depth experimental investigation with various utility and cost functions.** Our methods in this dissertation apply to any kinds of utility and cost functions. However, the effects of our methods may vary with the definition of the utility or cost functions. We would like to look for alternative utility and cost functions, and perform in-depth study on the impact of these functions using the methods in the disseration.

References

[1] "NGN Release 1 - Release definition," in *ETSI TR 180 001, v.1.1.1*, 2006.

[2] C. Shapiro and H. R. Varian, *Information Rules, pp. 187*. Harvard Business Press, 1999.

[3] NTT(2010-03-25), "World Record 69-Terabit Capacity for Optical Transmission over a Single Optical Fiber," in *Press release*, Retrieved Jun. 2010.

[4] C. Courcoubetis and R. Weber, *Pricing communication networks : economics, technology, and modelling*. West Sussex, England ; Hoboken, NJ : Wiley, 2003.

[5] B. Briscoe, A. Odlyzko, and B. Tilly, "Metcalfe's law is wrong - communications networks increase in value as they add members-but by how much?" *IEEE Spectrum*, vol. 43, no. 7, pp. 34 – 39, July, 2006.

[6] C. D. Manning and H. Schtze, *Foundations of Statistical Natural Language Processing, pp. 24*. MIT Press, 1999.

[7] T. Wu, "Network Neutrality, Broadband Discrimination," in *Journal of Telecommunications and High Technology Law 2: 141. doi:10.2139/ssrn.388863. SSRN 388863.*, 2003.

[8] D. Isenberg, "Research on Costs of Net Neutrality(Jul. 2007)," in *http://isen.com/blog/2007/07/research-on-costs-of-net-neutrality.html*, Retrieved Jun. 2010.

[9] "Who Wins: Verizon FiOS vs AT&T U-Verse, Aug. 2008," in *http://gigaom.com/2008/08/19/who-wins-verizon-fios-vs-att-u-verse/*, Retrieved Jun. 2010.

[10] A. Gupta, B. Jukic, M. Parameswaran, D. O. Stahl, and A. B. Whinston, "Streamlining the Digital Economy: How to Avert a Tragedy of the Commons," in *IEEE Internet Computing, I(6): 38-47*, Nov. 1997.

[11] M. L. Honig and K. Steiglitz, "Usage-based Pricing of Packet Data Generated by a Heterogeneous User Population," in *Proc. of IEEE INFOCOM, vol.2, pp. 867-874, Boston, MA*, Apr. 1995.

[12] J. K. MacKie-Mason and H. R. Varian, "Pricing the Internet," in *Public Access to the Internet, JFK School of Government*, May 1993.

[13] ——, "Pricing Congestible Network Resources," in *IEEE JOURNAL on Selected Areas in Communications, 13(7): 1141-1148*, Sep. 1995.

[14] C. Parris, S. Keshav, and D. Ferrari, "A Framework for the Study of Pricing in Integrated Networks," in *Technical Report, International Computer Science Institute, Berkeley, CA*, 1992.

[15] L. Murphy, J. Murphy, and J. K. MacKie-Mason, "Feedback And Efficiency In ATM Networks," in *Proc. of the International Conference on Communications (ICC'96), pp. 1045-1049, Dallas, TX*, Jun. 1996.

[16] F. Kelly and D. T. A. Maulloo, "Rate control in communication networks: shadow prices, proportional fairness and stability," in *J. Oper.Res. Soc., vol. 49, pp. 237-252*, 1998.

[17] J. Nuechterlein and P. Weiser, *Digital Crossroads: American Telecommunications Policy in the Internet Age*. The MIT Press; Cambridge, Massachusetts; London, England, 2004.

[18] "Federal Communications Commission, New Principles Preserve and Promote the Open and Interconnected Nature of Public Internet, Aug. 2005," in *http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-260435A1.pdf*, Retrieved Jun. 2010.

[19] P. Reichl, S. Leinen, and B. Stiller, "A Practical Review of Pricing and Cost Recovery for Internet Services," in *Proceedings of the 2nd Internet Economics Workshop Berlin (IEW '99), (Berlin, Germany)*, May 1999.

[20] R. Edell and P. Varaiya, "Providing Internet Access: What We Learn From INDEX," in *IEEE Network, vol. 13, no. 5, pp. 18-25*, Sep. 1999.

[21] J. Bailey and L. Mcknight, "Internet Economics: What Happens When Constituencies Collide," in *INET'95, pp. 659-666, Honolulu, HI*, Jun. 1995.

[22] L. Mcknight and J. Bailey, "Internet Economics: When Constituencies Collide in Cyberspace," in *IEEE Internet Computing, I(6):30-37*, Dec. 1997.

[23] H. Brody, "Internet@crossroads," in *Technology Review*, 1995.

[24] C. Courcoubetis and V. Siris, "Managing and Pricing Service Level Agreements for Differentiated Services," in *Proc. IEEE/IFIP IWQoS'99, London, U.K.*, Jun. 1999.

[25] P. Reichl, S. Leinen, and B. Stiller, "A Practical Review of Pricing and Cost Recovery for Internet Services," in *Proc. 2nd Internet Economics Workshop Berlin (IEW'99), Berlin, Germany*, May 1999.

[26] J. Altmann and K. Chu, "A Proposal for a Flexible Service Plan that is Attractive to Users and Internet Service Providers," in *Proc. IEEE INFOCOM, Anchorage, AK*, Apr. 2001.

[27] S. Floyd and V. Jacobson, "Link-sharing and Resource Management Models for Packet networks," in *IEEE/ACM Transaction on Networiing, vol.3 no. 4, pp, 365-386*, Aug. 1995.

[28] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," in *Automatica, vol. 35, pp. 1969-1985*, 1999.

[29] R. S. A. Ganesh, K. Laevens, "Congestion pricing and user adaptation," in *Proc. IEEE INFOCOM, Anchorage, AK, pp. 959-965*, Apr. 2001.

[30] S. J. Shenker, "Service Models and Pricing Policies for an Integrated Services Internet," in *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.6269*, 1993.

[31] ——, "Fundamental Design Issues for the Future Internet," in *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, 1995.

[32] X. Wang and H. Schulzrinne, "Pricing Network Resources for Adaptive Applications," in *IEEE/ACM Transaction on Networking, Vol.14, NO.3*, Jun. 2006.

[33] H. Ji, J. Y. Hui, and E. Karasan, “GoS-based pricing and resource allocation for multimedia broadband networks,” in *Proc. IEEE INFOCOM, San Francisco, CA, pp. 1020-1027*, Mar. 1996.

[34] H. Jiang and S. Jordan, “A pricing model for high speed networks with guaranteed quality of service,” in *Proc. IEEE INFOCOM, San Francisco, CA, pp. 888-895*, Mar. 1996.

[35] S. H. Low and P. P. Varaiya, “A New Approach to Service Provisioning in ATM Networks,” in *IEEE/ACM Transactions on Networing, 1(5):547-553*, Oct. 1993.

[36] F. Zhang and P. Verma, “A Constant Revenue Model for Packet Switched Network,” in *IEEE GIIS 09, Hammamet, Tunisia*, Jun. 2009.

[37] H. R. Sukasdadi and P. K. Verma, “A Constant Revenue Model for Telecommunication Networks,” in *International Conference on Systems and International Conference on Mobile Communications and Learning Technologies*, 2006.

[38] F. P. Kelly, “Charging and Rate Control for Elastic Traffic,” in *European Transactions on Communications, 8:33-37*, 1997.

[39] S. Ramesh, C. Rosenberg, and A. Kumar, “Revenue maximization in ATM networks using the CLP capability and buffer priority management,” in *IEEE/ACM Transactions on Networking, 4(6):941-950*, Dec. 1996.

[40] K. Kumaran, M. Mandjes, D. Mitra, and I. Saniee, “Resource Usage and Charging in a Multi-Service Multi-QoS Packet Network,” in *Proc. MIT Workshop on Internet Service Quality Economics*, Dec. 1999.

[41] N. Anerousis and A. A. Lazar, “A framework for pricing virtual circuit and virtual path services in atm networks,” in *Proc. ITC-15*, Dec. 1997.

[42] E. W. Fulp and D. S. Reeves, “Distributed network flow control based on dynamic competive markets,” in *Proc. ICNP98*, Oct. 1998.

[43] A. J. O’Donnell and H. Sethu, “A Novel, Practical Pricing Strategy for Congestion Control and Differentiated Services,” in *Proc. ICC, pp.986-990*, Apr. 2002.

[44] S. Jordan, “Pricing of Buffer and Bandwidth in a Reservation-based QoS Architecture,” in *Proc. ICC, pp. 1521-1525*, May. 2003.

[45] A. Odlyzko, “Paris Metro Pricing: the Minimalist Differentiated Services Solution,” in *Proc. NOSSDAV’99, Basking Ridge, NJ*, Jun. 1999.

[46] R. Gibbens, R. Mason, and R. Steinberg, “Internet service classes under competition.”

[47] R. Jain, T. Mullen, and R. Hausman, “Analysis of Paris Metro Pricing Strategy for QoS with a Single Service Provider,” in *Proc. IWQoS, pp. 44-48*, Jun. 2001.

[48] P. Marbach, “Pricing Differentiated Services Networks: Bursty Traffic,” in *Proc. IEEE INFOCOM, pp. 650-658, Anchorage, AK*, Apr. 2001.

[49] J. Altmann, H. Daanen, H. Oliver, and A. S.-B. Surez, “How to market-manage a QoS network ,” in *Proc. IEEE INFOCOM, vol. 1, pp. 284-293*, Nov. 2002.

[50] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, "A study of priority pricing in multiple service class networks," in *ACM SIGCOMM Computer Communication Review, vol. 21, issue 4, pp. 123-130*, Sep. 1991.

[51] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer Networks: Motivation, Formulation, and Example," in *IEEE/ACM Transactions on Networking, 1(6): 614-627*, Dec. 1993.

[52] M. H. Dahshan and P. K. Verma, "Pricing for quality of service in high speed packet switched networks," in *High Performance Switching and Routing workshop, Poznan*, Oct. 2006.

[53] Y. Qu and P. Verma, "Notion of cost and quality in telecommunication networks: an abstract approach," in *IEE Proc., Commun., vol. 152, Issue 2, pp.167171*, Apr. 2005.

[54] J. F. MacKie-Mason, "A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network," in *Univ. of Michigan, Tech. Rep*, Sep. 1997.

[55] N. Semret and A. Lazar, "The progressive second price auction mechanism for network resource sharing," in *Proc. 8th Int. Symp. Dynamic Games, The Netherlands*, Jul. 1998.

[56] S. Chen and K. I. Park, "An Architecture for Noncooperative QoS Provision in Many Switch Systems," in *Proc. IEEE/INFOCOM, pp. 865-872*, Mar. 1999.

[57] P. Fuzesi and A. Vidacs, "Game Theoretic Analysis of Network Dimensioning Strategies in Differentiated Services networks," in *Proc. ICC, pp. 1069-1073*, Apr. 2002.

[58] L. A. Dasilva, D. W. Petr, and N. Akar, "Equilibrium Pricing in Multi-service Priority-based Networks," in *Proc. IEEE/Global Telecommunications Conference (GLOBECOM) vol. 3, pp. 1373-1377*, Nov. 1997.

[59] M. Mandjes, "Pricing Strategies under Heterogeneous Service Requirements," in *Proc. IEEE/INFOCOM, pp. 1210-1220*, Apr. 2003.

[60] P. Marbach, "Analysis of a Static Pricing Scheme for Priority Services," in *IEEE/ACM Transaction on Networing, vol.12, no. 2, pp, 312-325*, Apr. 2004.

[61] J. Shu and P. Varaiya, "Pricing Network Services," in *Proc. IEEE/INFOCOM pp. 1221-1230*, Apr. 2003.

[62] F. Maniquet, "A Characterization of the Shapley Value in Queueing Problems," in *Journal of Economic Theory, 109: 90–103*, 2003.

[63] Y. Ye, "Competitive Communication Spectrum Economy and Equilibrium," in *Working Paper*, 2007.

[64] S. Blake, D. Black, M. Carhn, E. Daviq, Z. Wang, and W. Weiss, "An Architecture for Differentiated Servicss," in *RFC2475*, Dec. 1998.

[65] D. M. Kreps, *Game Theory and Economic Modelling*. Clarendon Press, 1990.

[66] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1993.

[67] J. W. Friedman, *Game Theory with Applications to Economics*. Oxford University Press, 2nd edition, 1990.

[68] P. D. Straffin, *Game Theory and Strategy*. The Mathematical Association of America, 1993.

[69] Z. Q. Luo and J. Pang, "Analysis of Iterative Water-Filling Algorithm For Multi-User Power Control In Digital Subscriber Lines," in *Special issue of EURASIP Journal on Applied Signal Processing on Advanced Signal Processing Techniques for Digital Subscriber Lines, Vol.2006, Article ID 24012*, 2006.

[70] N. Yamashita and Z. Q. Luo, "A Nonlinear Complementarity Approach To Multi-User Power Control for Digital Subscriber Lines," in *Optimization Methods and Software, Vol. 19, pp. 633652*, 2004.

[71] R. Cendrillon, W. Yu, M. Moonen, J. Verliden, and T. Bostoen, "Optimal Multi-User Spectrum Management for Digital Subscriber Lines," in *IEEE Transactions on Communications*, 2007.

[72] W. Yu, R. Lui, and R. Cendrillon, "Dual Optimization Methods For Multi-User Orthogonal Frequency Division Multiplex Systems," in *newblock IEEE Global Communications Conference (Globecom), Vol. 1, pp. 225229, Dallas, USA*, 2004.

[73] F. Zhang and P. K. Verma, "A Two-step Quality of Service Provisioning in Multi-class Networks," in *18th International Conference on Telecommunications (ICT 2011), Ayia Napa, Cyprus*, May 8-11, 2011.

[74] Y. A. Korilis and A. A. Lazar, "On the existence of Equilibria om Mpm-cooperative Optimal Flow Control," in *Journal of the Association for Computing Machinery, 42(3):584-613*, May 1995.

[75] S. Shenker, "Making Greed Work in Networks: A Game Theoretic Analysis of Switch Service Disciplines," in *IEEE/ACM Transactions on Networking, 3(6):819-831*, Dec. 1995.

[76] Y. A. korilis, A. Lazar, and A. Orda, "Achieving Network Optima using Stakelberg Routing strategies," in *IEEE/ACM Transactions on Networking, 1(5):161-172*, Feb. 1997.

[77] A. Orda, R. Rom, and N. Shimkin, "Competitive Routing in Multiuser Communication Networks," in *IEEE/ACM Transactions on Networking, 1(5):510-521*, Oct. 1993.

[78] A. Watson and M. Sasse, "Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems," in *Interacting with Computers, vol.8, no. 3, pp. 255-275*, 1996.

[79] H. Yaiche, R. Mazumdar, and C. Rosenberg, "A Game Theoretic Framework for Bandwidth. Allocation and Pricing in Broadband Networks," in *IEEE/ACM Transactions on Networking (TON), vol. 8 , no. 5*, Oct. 2000.

[80] B. Davie and A. Charny, "RFC 3246, An Expedited Forwarding PHB (Per-Hop Behavior)," Mar. 2002.

[81] G. Mamais, M. Markaki, G. Politis, and I. S. Venieris, "Efficient buffer management and scheduling in a combined IntServ and DiffServ architecture: a performance study," in *Proc. ICATM, pp. 236-242*, 1999.

[82] S. Yi, X. Deng, G. Kesidis, and C. Das, "Providing fairness in DiffServ architecture," in *Global Telecommunications Conference, pp. 1435- 1439 vol.2*, 2002.

[83] F. Zhang, P. Verma, and S. Cheng, "Pricing Multi-class Priority-based Network Services for Stochastic Traffic Using the Shapley Value," in *submitted to the Proceeding of Communication, IET*, Oct, 2010.

[84] Y. Chun, "A Note on Maniquet's Characterization of the Shapley Value in Queueing Problems," in *Working Paper, Rochester University*, 2004.

[85] H. Moulin, "On Scheduling Fees to Prevent Merging, Splitting and Transferring of Jobs," in *Working Paper, Rice University*, 2004.

[86] ——, "Split-proof Probabilistic Scheduling," in *Working Paper, Rice University*, 2004.

[87] D. Mishra and B. Rangarajan, "Cost sharing in a Job Scheduling Problem Using the Shapley Value," in *Proceedings of the 6th ACM conference on Electronic commerce, Vancouver, BC, Canada, pp: 232 - 239*, 2005.

[88] L. Kleinrock, *Queuing Systems*. vol. 2: Computer Applications. John Wiley and Sons, 1976.

[89] J. C. Harsanyi, "Contributions to Theory of Fames IV, Chapter A Bargaining Model for Cooperative n-person Games," in *Princeton University Press*, 1995.

[90] F. Zhang and P. Verma, "pricing Class-based Networks Using the Shapley Value," in *submitted to Journal of NETNOMICS*, Jan. 2011.

[91] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[92] F. Zhang, P. Verma, and S. Cheng, "Pricing, Resource Allocation and QoS in Multi-class Networks with Competitive Market Model," in *IET Commun., Volume 5, Issue 1, p.5160, doi:10.1049/iet-com.2009.0694*, Jan. 2011.

[93] L. Walras, "Elements of Pure Economics; Or the Theory of Social Wealth," in *Lausanne, Paris*, 1874.

[94] K. J. Arrow and G. Debreu, "Existence of an Equilibrium for a Competitive Economy," in *Econo-metrica 22 (3), pp. 265-290*, 1954.

[95] W. C. Brainard and H. Scarf, "How to Compute Equilibrium Prices in 1891," in *Cowles Foundation Discussion Paper 1270*, 2000.

[96] E. Eisenberg and D. Gale, "Consensus of Subjective Probabilities: The Pari-Mutuel Method," in *Annals of Mathematical Statistics 30, 165-168*, 1959.

[97] E. Eisenberg, "Aggregation of Utility Functions," in *Management Sciences 7(4), 337-350*, 1961.

[98] D. Gale, "The Theory of Linear Economic Models," in *McGraw Hill, N.Y.*, 1960.

[99] M. Ling, J. Tsai, and Y. Ye, "Budget allocation in a competitve communication spectrum economy," in *EURASIP Journal on Advances in Signal Processing*, Apr. 2009.

[100] W. Rudin, *Principles of Mathematical Analysis*. p.101, 3rd ed, New York: McGraw-Hill, 1976.

[101] L. Chen, Y. Ye, and J. Zhang, "A Note on Equilibrium Pricing as Convex Optimization," in *Working Paper*, 2007.

[102] "TorrentFreak, Comcast Throttles BitTorrent Traffic, Seeding Impossible, Aug, 2007,," in *http://torrentfreak.com/comcast-throttles-bittorrent-traffic-seeding-impossible/,*, Retrieved Nov, 2010.

[103] F. L. P. Eckersley and S. Schoen, "Packet Forgery By ISPs: A Report On The Comcast Affair, Nov, 2007, ," in *http://www.eff.org/files/eff_comcast_report.pdf*, Retrieved Nov, 2010.

[104] P. Svensson, "FCC to probe Comcast data discrimination, Jan, 2008,," in *http://www.msnbc.msn.com/id/22560491/ns/technology_and_science-internet,*, Retrieved Nov, 2010.

[105] "Comcast Wins the BitTorrent Throttling Case,," in *http://extratorrent.com/article/404/comcast+wins+the+bittorrent+throttling+case.html*, Retrieved Nov, 2010.

[106] R. Hahn and S. Wallsten, "The Economics of Net Neutrality," in *The Berkeley Economic Press Economists' Voice*, 2006.

[107] T. Berner-Lee, "Neutrality on the Net," in *http://dig.csail.mit.edu/breadcrumbs/node/132*, Retrieved Nov, 2010.

[108] H. D. Jonathan, "Internet Law," in *BNA Books. p. 750. ISBN 1570186839, 9781570186837*, 2007.

[109] P. Waldmeir, "The net neutrality dogfight that is shaking up cyberspace," in *Financial Times, New York,*, Mar, 2006.

[110] S. B. H. K. Cheng and G. Hong, "Debate on Net Neutrality: A Policy Perspective," in *Information Systems Research*, Jun, 2008.

[111] G. S. J. Musacchio and J. Walrand, "A Two-Sided Market Analysis of Provider Investment Incentives With an Application to the Net-Neutrality Issue," in *Review of Network Economics, Vol. 8, Issue 1,*, Mar, 2009.

[112] L. Lessig and R. W. McChesney, "No Tolls on The Internet," in *Columns (Washington Post), from http://www.washingtonpost.com/wp-dyn/content/article/2006/06/07/AR2006060702108.html*, Retrieved Nov, 2010.

[113] F. Zhang and P. Verma, "A User-friendly Constant Revenue Model for Net Neutrality," in *submitted to the Journal of Telecommunications Management*, Jan. 2011.

[114] A. E. Roth and R. E. Verrecchia, "The Shapley value as Applied to cost Allocation: A Reinterpretation," in *Journal of Accounting Research, Vol. 17 No. 1*, Spring 1979.

[115] H. R. Sukasdadi and P. K. Verma, "A Constant Revenue Model for Telecommunication Networks," in *International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, ICN/ICONS/MCL*, 2006.

[116] ITU-T, “One-way transmission time,” in *Recommendation G.114*, 1996.

[117] A. Odlyzko, “Internet Pricing and history of communications,” in *AT&T Labs-Research*, Feb. 8, 2001.

[118] “3GPP TS. 23.223, “IP Multimedia Subsystem(IMS),” in *v.8.6.0.*, 2008.

[119] R. Braden, D. Clark, and S. Shenker, “Integrated Services in the Internet Architecture: an Overview,” in *RFC 1633*, June 1994.

[120] S. Shenker, “Fundamental design issues for the future Internet,” in *IEEE J Selected Areas Commun,13: 1176-1188*, 1995.

[121] V. Jacobson, K. Nichols, and K. Poduri, “An expedited forwarding PHB group,” in *IETF RFC 2598*, Jun. 1999.

[122] A. Ganesh, K. Laevens, and R. Steinberg, “Congestion pricing and user adaptation,” in *Proc. IEEE INFOCOM, Anchorage, AK, pp. 959-965*, Apr. 2001.

[123] L. Kleinrock, “Power and deterministic rules of thumb for probabilistic problems in computer communications,” in *International Conference on Communications, Boston, Mass., pp. 43.1.1-43.1.10*, June 1979.

[124] G. O. T. Basar, “Dynamic Noncooperative Game Theory,” in *Academic Press*, 1982.

[125] D. Bertsekas and R. Gallager, “Data networks,” in *Prentice-Hall, Inc., Upper Saddle River, NJ*, 1987.

[126] J. Nash, “The bargaining problem,” in *Econometrica, vol. 18, pp. 155-162*, 1950.

[127] A. Stefanescu and M. W. Stefanescu, “The arbitrated solution for multiobjective convex programming,” in *Rev. Roum. Math. Pure Applicat., vol. 29, pp. 593-598*, 1984.