國立政治大學應用數學系

博士學位論文

Analysis of Bandwidth Allocation on End-to-End QoS Networks under Budget Control

在預算限制下分配隨機數位網路最佳頻寬之研究

博士班學生：王 嘉 宏
指導教授： 陸 行 博士

中 華 民 國 一 百 年 六 月

Abstract

This thesis considers the problem of bandwidth allocation on communication networks with multiple traffic classes, where bandwidth is determined under the budget constraint. Due to the limited budget, there exists a risk that the network service providers can not assert a 100% guaranteed availability for the stochastic traffic demand at all times. We derive the blocking probabilities of connections as a function of bandwidth, traffic demand and the available number of virtual end-to-end paths for all service classes. Under general assumptions, we prove that the blocking probability is directionally (i) decreasing in bandwidth, (ii) convex in bandwidth for specific regions, (iii) increasing in traffic demand, and (iv) decreasing in the number of virtual paths. We also demonstrate the monotone and convex relations among those model parameters and the expected path occupancy. As the number of virtual paths is huge, we derive a heavy-traffic queueing model, and provide a diffusion approximation and its asymptotic analysis for the blocking probability, where the traffic intensity increases to one from below.

Taking the blocking probability into account, two revenue management schemes are introduced to allocate bandwidth under budget control. The revenue/profit functions are studied in this thesis through the monotonicity and convexity of the blocking probability and expected path occupancy. Optimality conditions are derived to obtain an optimal bandwidth allocation for two revenue management schemes, and a solution algorithm is developed to allocate limited budget among competing traffic classes. In addition, we present three elasticities of the blocking probability to study the effect of changing model parameters on the average revenue in analysis of economic models. The sensitivity analysis and economic elasticity notions are proposed to investigate the marginal revenue for a given traffic class by changing bandwidth, traffic demand and the number of virtual paths, respectively.

The main contribution of the present work is to prove the relationship between the blocking probability and allocated bandwidth under the budget constraint. Those results are also verified with numerical examples interpreting the blocking probability, utilization level, average revenue, etc. The relationship between blocking probability and bandwidth allocation can be applied in the design and provision of broadband communication networks by optimally choosing model parameters under budget control for sharing bandwidth in terms of blocking/congestion costs.

中文摘要

本論文針對隨機數位網路提出一套可行的計算機制，以提供網路管理者進行資源分配與壅塞管理的分析工具。我們研究兩種利潤最佳化模型，探討在預算控制下的頻寬分配方式。因為資源有限，網路管理者無法隨時提供足夠頻寬以滿足隨機的網路需求，而量測網路連結成功與否的阻塞機率 (Blocking Probability) 為評估此風險之一種指標。我們利用頻寬分配、網路需求量和虛擬端對端路徑的數量等變數，推導阻塞機率函數，並證明阻塞機率的單調性 (Monotonicity) 和凸性 (Convexity) 等數學性質。在不失一般性之假設下，我們驗證阻塞機率是 (1) 隨頻寬增加而變小；(2) 在特定的頻寬分配區間內呈凸性；(3) 隨網路需求量增加而變大；(4) 隨虛擬路徑的數量增加而變小。

本研究探討頻寬分配與阻塞機率之關係，藉由推導單調性和凸性等性質，提供此兩種利潤模型解的最適條件與求解演算法。同時，我們引用經濟學的彈性概念，提出三種模型參數對阻塞機率變化量的彈性定義，並分別進行頻寬分配、網路需求量和虛擬路徑數量對邊際利潤函數的敏感度分析。當網路上的虛擬路徑數量非常大時，阻塞機率的計算將變得複雜難解，因此我們利用高負荷極限理論 (Heavy-Traffic Limit Theorem) 提供阻塞機率的估計式，並分析其漸近行為 (Asymptotic Behavior)。本論文的主要貢獻是分析頻寬分配與阻塞機率之間的關係及其數學性質。網路管理者可應用本研究提出的分析工具，在總預算限制下規劃寬頻網路的資源分配，並根據阻塞機率進行網路參數的調控。

Acknowledgment

I am greatly indebted to my advisor, Professor Hsing Paul Luh, for his encouragement and support during my formative years. As an advisor, Professor Luh has always maintained high standards for us and has been doing a lot to help his students grow. Without his instructions and advices, this Ph.D. dissertation would not have been completed.

Besides, I would like to sincerely thank the committee members at my Ph.D. thesis defense, which include Professor Jeh-Nan Pan, Professor Kuo-Hsiung Wang, Professor Chih Chang, Professor Chien-Chih Yu, Professor Chi-Chun Lo and Professor Ming-Long Liu, for their insightful comments and suggestions on this dissertation.

Special thanks go to Professor Zhe George Zhang in Department of Decision Sciences, Western Washington University, USA, for teaching me how to do a good research and letting me know what hard working is. I also would like to thank Professor Yew-Sing Thomas Lee in Department of Information and Decision Sciences, University of Illinois at Chicago, USA, for advising me during my visiting at University of Illinois at Chicago. Additionally, many thanks to the National Science Council, Taiwan for financially supporting me to pursue advanced research in USA from Year 2007 to Year 2008 under grant No. NSC 96-2917-I-004-015.

I have benefited enormously from all the friendly members of the Department of Mathematical Sciences at National ChengChi University, which provides a helpful research environment. I wish to express my sincere gratitude to all my friends for their kindly help and valuable advices. Last but not least, I want to appreciate my dear family for their support and encouragement in pursuing my Ph.D. degree.

Chia-Hung Wang

June 24, 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The telecommunication industry is moving toward a converged network because of the rapid growth of Internet traffic, aggressive deployment of broadband fiber optic network, advance of Voice over IP technology, and the global standardization of IP technology [41, 42]. The converged network uses a single global IP-based network carrying all types of network traffics to replace the traditional separated packet switching and circuit switching networks [84]. A core network is a part of the global IP-based network under a single organization or Internet Service Provider (ISP) [84]. Applications expected to produce the bulk of traffic in the future multi-service Internet can be broadly categorized as streaming or elastic according to the nature of the connections they produce [33, 76]. Connections using the core network are typically generated by a very large population of users independently communicating with an equivalently large population of servers and correspondents for a variety of applications [91]. According to traffic demand and network management settings, it requires suitable bandwidth allocation to individual connection to achieve guaranteed Quality of Service (QoS) level [3, 4, 14, 16].

Every core network has a centralized network manager, known as the Bandwidth Broker (BB), which is aware of the network topology and status, using the underlying routing protocols [73, 130]. First the client (maybe a single user, a corporate network or an aggregated sub-network) negotiates its QoS requirements, known as Service Level Agreement (SLA) negotiation [71], with the BB in its core network. For example, real-time multimedia applications require stringent QoS guarantees, including hard bounds on bandwidth and packet loss probability [59, 130]. Then the BB in the source network negotiates resource allocation with the BBs in intermediate and destination networks. The users accessing the network via the ISP are guaranteed an amount of bandwidth under the total available budget for their connections [28, 33, 42].

Bandwidth provisioning while increasing the revenues of ISPs is still one of the major issues in communication systems [4, 28, 44, 74]. We deal with the problem of dimensioning bandwidth on core networks [118, 120], and the objective is to determine the amount of bandwidth to serve all traffic demands under the budget constraint [123]. This thesis introduces two revenue management schemes that maximizes ISPs' revenue subject to the budget constraint, and solution analysis is studied through theoretical derivations and numerical illustrations. Network managers' economic profit consists of all revenue gained by providing bandwidth to connections [121, 130], and the opportunity cost (in terms of revenue loss) is determined through calculating the risk that is caused by blocking connections [133]. Traffic flows in communication networks are traditionally modelled by applying the theory of stochastic processes [17, 24, 97, 105]. For example, classical telecommunication networks were well studied under assumptions of Markovian processes and its variants [24, 105]. However, it becomes difficult or even impossible to compute blocking probability in modern communication networks due to a huge number of connections [23, 35, 38].

1.2 Bandwidth Allocation with QoS

Network design today often considers the problem of designing networks that carry elastic traffic [60, 65]. The network design problem becomes difficult if the network is also used for other types of QoS guaranteed connections [129, 101]. Resource allocation models may be used to solve network design problems [67, 89, 93]. Resource allocation decisions are concerned with the allocation of limited bandwidth so as to achieve the best system performances [33, 74, 89]. Network dimensioning with elastic traffic may be thought of as a search for such network flows that will maximize the network throughput while staying within a budget constraint for the costs of link bandwidth [3, 16]. However, such a problem of formulation would lead to the starvation of bandwidth between certain network nodes [27, 29]. The volume of bandwidth assigned to one connection is strongly dependent on the bandwidths assigned to other connections [4, 14].

Telecommunication networks are facing the increasing demand for Internet services. The idea of a single shared physical network that will support multiple heterogeneous applications with different traffic characteristics and different QoS requirements, is widely regarded as the way to meet the telecommunication challenges of the future. Packet-switched networks have been proposed to offer the QoS guarantees in integrated-services networks because in networks individual packets may exhibit a significant variation in network service quality.

Packet switched networks suffer three major quality problems in offering time-sensitive services: long delay time, jitter, packet loss. The Universal Mobile Telecommunications System (UMTS) [108] has specified different traffic classes according to their QoS requirements for different applications. UMTS network services have proposed four different QoS classes indicating traffic demands:

i. Conversational class (voice, video telephony, video gaming);

ii. Streaming class (multimedia, video on demand, webcast);

iii. Interactive class (web browsing, network gaming, database access);

iv. Background class (email, SMS, downloading).

Key issues in the design of broadband communication networks include providing suitable bandwidth to meet connections' QoS requirements [7, 109], and achieving that goal in an efficient manner [9, 90, 114]. The ability to provide end-to-end guarantees depends to a large extent on the scheduling policy and service discipline employed in the nodes [60, 109]. Typical schedulers map delay guarantees into rate guarantees, and have nodes advertise the residual rate they have available [41, 65]. Data packets of incoming connection of a certain traffic class can be transmitted from the source to the destination along a specified virtual end-to-end path [28, 52, 72]. In practice, some data-carrying mechanisms such as Multi-Protocol Label Switching (MPLS) [25, 71] can create virtual paths from the source to the destination on telecommunications networks. The maximum throughput of those virtual paths are limited by either the total budget or bottleneck links which lie on those virtual paths [74, 72]. The blocking of connections occurs for a certain traffic class if the dynamic traffic demand exceeds the maximum throughput [8, 122].

In the long run, a network operator always requires economic profit from the administrated network [4, 28, 66]. For the possibility of this to be achieved, there are some obvious properties that a backbone must possess [17, 80]. Usually, vital parameters of the operator's income are proportional to the network's total throughput [78, 99, 130]. The simplest way of defining a network allocation rule, and also the most profitable way from the network operator viewpoint, is to maximize the total throughput of the network, which is the so-called throughput maximization [76]. To maximize throughput, it is obvious that all flows should be routed on their shortest paths [18, 76]. The solution to this optimization problem is not unique [12, 16, 82]. Hence, the maximum throughput principle does not give a unique bandwidth allocation for a fixed set of ongoing connections.

For a communication network providing performance guarantees, it has to reserve resources and exercise call admission control [26, 81, 130]. Network users are mainly interested in getting good quality of connections whenever they place requests [101, 115, 117]. It is the network providers' mission to have a virtual end-to-end path with suitable bandwidth [3, 28]. Clearly, it is too costly for the network providers to assert a 100% guaranteed availability for all connections under the budget constraint at any time [42]. This is also not necessary since demand for connections or bandwidth capacity varies over time.

One of the key challenges to be addressed in the next generation networks is the distribution of the available path capacity amongst the multiple traffic with different bandwidth requirements [84]. The call blocking probability is one such QoS parameter for the next generation network [42, 43], and it is desirable to bring out an analytic performance model [26, 37]. Estimating blocking probabilities is a fundamental ingredient in network design and engineering [8, 40]. To compute it, the classical approach of Erlang provided a very well tried solution [24, 105]. However, computing blocking probabilities becomes much harder in today's complex networks that carry highly heterogeneous traffic [22, 55, 62].

Roberts [95] provided a survey of recent results on the performance of a network handling elastic traffic under the assumption that flows are generated as a random process. Faragó [40] estimated the blocking probability and link utilization for general multi-rate, heterogeneous traffic, where the individual bandwidth demands may aggregate in complex, nonlinear ways. Atkinson [11] developed two heuristic approaches for the general, multi-server queueing loss system with an emphasis on the calculation of steady-state loss probabilities. When the packet arrival process is a Markov-modulated Poisson process, an algorithm was presented in [96] to compute the blocking probability in a Markovian bufferless queuing system with a finite number of servers. There are very few analytical results available for the throughput performance of bandwidth allocation under stochastic traffic [4, 27, 51]. This is mainly because the performance of these networks is not insensitive and depends significantly on detailed traffic characteristics [20, 5].

Traffic flow fluctuates with time, and connections do not last forever but occur at random times and vanish in the network once the corresponding digital document has been transferred completely [122, 123, 131]. This results in a random, dynamic set of active connections. Moreover, the bandwidth assigned to each connection would determine how long that connection will stay active and thus impact the evolution of the set of active connections [122]. The network provider chooses an optimal sharing scheme for the different users under the total budget to fulfill requirements of connections [31, 130]. In addition, the risk/probability of rejecting connection requests due to lack of resources is supposedly kept below a negotiated level [26, 71].

Bandwidth sharing efficiency in overload would be improved if it were possible to perform pro-active admission control rather than relying on user impatience to stabilize the system [26, 81, 95]. Admission control consists in rejecting a new connection on its arrival in order to preserve the performance of connections already in progress. Cho et al. [33] investigated the optimal partitioning of the end-to-end QoS budget to quantify the advantage of having a non-uniform allocation of the budget over the links in a path. Jin and Jordan [59] studied the sensitivity of resource allocation and the resulting QoS to resource prices in a reservation-based QoS architecture. Güven et al. [46] formulated an optimization problem of load balancing the traffic, where multiple paths are provided between a source and a destination using application-layer overlay. Bruni et al. [26] designed a connection admission control procedure for resource management in a telecommunication network. However, the relationship between performance, traffic demand and capacity has not been well investigated [21, 54, 55].

In this thesis, we aim to analyze the relationship between blocking probability, bandwidth, traffic demand and the available number of virtual paths on communication networks with service from ISPs, where requests for connections represent customers arriving at the system. As soon as requests are accepted by the system, the service begins. The installed bandwidth allocation is used to maintain a guaranteed connection availability where the blocking probability is kept below a certain negotiated level. Our intent is to analyze the sensitivity of the blocking probability

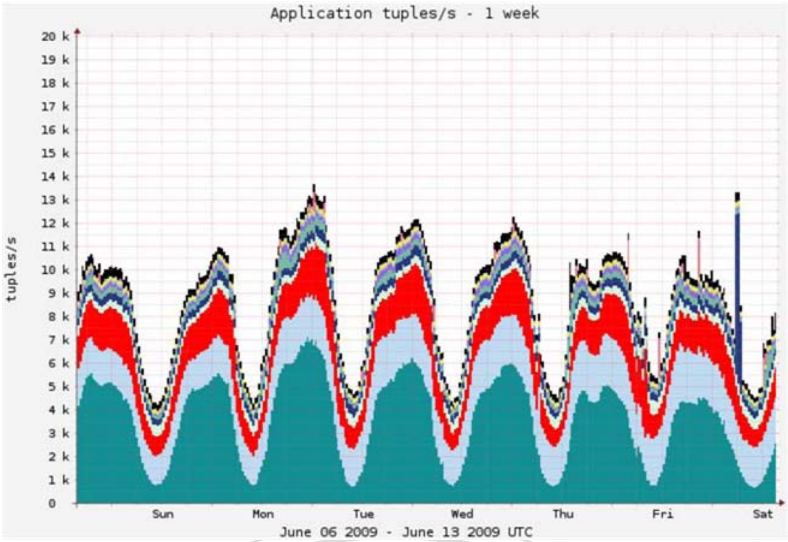

Figure 1.1: Traffic flow observed during June 6-13, 2009 from the network monitor framework at CAIDA [106], where data includes HTTP, TCP, UDP, SMTP, POP, FTP, etc. [Courtesy of CAIDA]

by system parameters, including directional monotone and convex properties. Then we study how these changes affect the long-run average revenue and find structural results. The relationship between the revenue function, the blocking probability and allocated bandwidth are to be investigated under the budget constraint, which has received relatively little attention in the literature [2, 28, 40, 130].

1.3 Motivations

Traffic flow fluctuates with time and it seemingly appears periodical patterns. For example, Fig. 1.1 shows real traffic flow observed from the network monitor framework at the Cooperative Association for Internet Data Analysis (CAIDA) [106]. In this thesis, we study the blocking probability of connections with allocated bandwidth, which is an important performance measure of network systems.

As soon as requests are accepted by the system, the service begins. Yacoubi et al. [130] and Smith [103] assume that the user population is infinite and connections join the core network according to a Poisson process. The sojourn time during which they occupy end-to-end paths has an arbitrary distribution. Nain [87] provided a solution to the classical Erlang blocking model on circuit-switched network, where the author obtained monotone and concave properties for loss probabilities, throughput and channel occupancy in terms of traffic intensity. Recently, Dutta and Chaubey [37] studied the performance analysis of all-optical network with wavelength converter using Erlang traffic model. In these literatures, researchers studied the network systems under the assumptions of Poisson arrivals, general distributions for sojourn times.

On the other hand, some researchers consider the network systems with independent general distributions for inter-arrival time and exponential service time distributions, see for example Choi et al. [34], Kim and Choi [63], Abramov [1], Halfin and Whitt [126] and references therein. The assumptions of renewal arrival process, exponential service times, finite servers and limited buffer size are commonly used in queueing systems [1, 11, 34, 48]. Therefore, we are going to study the network systems under $M/G/K/K$ and $GI/M/K/K$ assumptions individually in this thesis.

Bonald et al. [22] provided a queueing analysis of three usual bandwidth allocations, namely max-min fairness, proportional fairness and balanced fairness in a communication network. Antunes et al. [8] provided an analysis of loss networks with different classes of requests which move according to some routing policies. Faragó [40] gave an estimated blocking probability and link utilization for general multi-rate, heterogeneous traffic, where the individual bandwidth demands may aggregate in complex ways. Maglaras and Zeevi [79] studied the equivalent behavior of communication systems in a single-class Markovian model under revenue and social optimization objectives. The approach in this thesis is different from prior works because we focus on obtaining blocking probabilities with multiple classes while allocating resources under the budget constraint. Note that there are bandwidth allocation and virtual (end-to-end) path selection issues in the network management scheme, which is based on the knowledge of the connection's requirements and

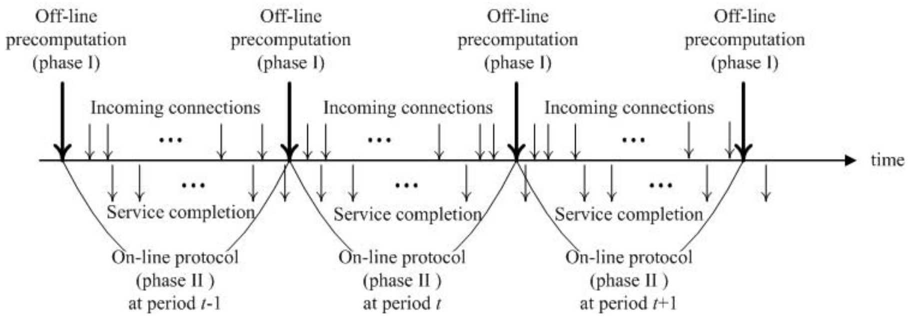

Figure 1.2: A two-phase network management scheme.

information about the availability of resources in the core network.

Time horizons for consideration of this QoS network can be disunited apart for certain periods according to its periodic demands of traffic. Wang and Luh [114] proposed a network management scheme by means of a two-phase procedure which is shown in Fig. 1.2. In the first phase, network managers determine bandwidth allocation and end-to-end paths for all traffic classes under budget and network constraints. The first phase is executed in advance as the off-line process, and its purpose is to predetermine solutions, which are summarized in a routing table for later use at the second phase. Bandwidth allocation and end-to-end paths can be predetermined and listed in a routing table at the beginning of the second phase during a certain planning period in this two-phase procedure. The second phase is activated as an online process, and its purpose is to select an adequate solution efficiently from the routing table when connections arrive. Network managers update the routing table periodically and distribute it to the ingress routers in order to make an accurate path selection.

The management scheme is carried out by means of a two-phase procedure [114]. In other words, the first phase prepares a routing table which enables to identify a suitable end-to-end path in order for requests of each connection through the second phase. By taking this precomputation-based network optimization scheme in advance, we can calculate the available bandwidth of the possible end-to-end paths, so that network managers are able to offer connections suitable end-to-end paths to satisfy their QoS requirements. After those end-to-end paths with bandwidth allocation are listed in the routing table, a simple procedure can be executed to assign each connection an optimal end-to-end path at the second phase. The key idea of this two-phase procedure is taken to effectively reduce the time needed in computing bandwidth allocation and determining end-to-end QoS routing paths when connections arrive.

We analyze the relationship between the blocking probability, allocated bandwidth, traffic demand and the available number of virtual paths on communication networks, where requests for connections represent customers arriving at the system. As soon as requests are accepted by the system, the service begins. The blocking of connections occurs due to the failure of meeting the demand for virtual paths of each traffic class, where the traffic demand is considered and written as a function of the product of the occurrence rate of connections and the average connection volume. The explicit expressions of the blocking probability and the expected path occupancy are to be derived in terms of model parameters under the budget constraint.

The main contribution of the present work is to prove the relationship between the blocking probability and allocated bandwidth under the budget constraint. Monotone and convex properties of the blocking probability are shown in both theoretical construction and numerical examples. The originality of our work lies in the derivation of bandwidth elasticity and demand elasticity of blocking in economic models by analyzing properties of the blocking probability with respect to allocated bandwidth, traffic demand and the number of virtual paths. The relationship between the blocking probability and bandwidth allocation derived in this thesis can be applied in the design of broadband communication networks under budget control.

1.4 Thesis Organization

The thesis is organized as follows. Chapter 1 is the introduction to the background and motivations of our research work, and provides a summary of this thesis. Problem definitions and two revenue management schemes with consideration of blocking are introduced in Chapter 2. Chapter 2 also provides an overview of related works on bandwidth sharing policies in literatures. The revenue functions are to be analyzed through deriving monotonicity and convexity of the blocking probability under the budget constraint.

In Chapter 3, we derive monotone and convex properties of the blocking probability and expected path occupancy with allocated bandwidth under assumptions of $M/G/K/K$. The risk of blocking connections is analyzed according to an Erlang loss model under assumptions of Poisson arrivals, general sojourn time, preset number of virtual paths with identical bandwidth, and no waiting space. The blocking probabilities of connections are to be formulated as a function of bandwidth, traffic demand and the available number of virtual paths for all service classes.

Chapter 4 provides an estimation of blocking probability under assumptions of $GI/M/K/K$. The embedded Markov chain technique is applied to derive the distributions of the number of connections in the system at arbitrary and pre-arrival epoches. Three illustrative examples are introduced to demonstrate the steady state probabilities under common distributions for inter-arrival times of exponential, deterministic and Erlang-r distributions. As the number of virtual paths is too huge to compute numerically, we derive a heavy-traffic queueing model and provide its approximation for the blocking probability in Chapter 4, where the traffic intensity approaches to one from below. Three illustrative examples are also provided to investigate the asymptotic behavior of the blocking probability under exponential, deterministic and Erlang-r inter-arrival time distributions, individually.

Chapter 5 presents a sensitivity analysis of two revenue management schemes given in Chapter 2 through monotonicity and convexity of the blocking probability.

The optimality conditions and monotonicity/convex properties of these two revenue functions are studied under budget constraints. A solution algorithm is also presented to allocate limited budget among competing traffic classes. As applications, three elasticities of blocking in economic models are proposed by analyzing properties of the blocking probability with respect to allocated bandwidth, traffic demand and the number of virtual paths, individually.

Sensitivity analysis with numerical illustrations are conducted in Chapter 6. Those numerical results interpret monotone and convex properties of the blocking probability, revenue functions, utilization level, and elasticities, etc. Concluding remarks and suggestions for future works are provided in Chapter 7. For the class of problems studied, it is concluded that the approximation is adequate for practical purposes.

Chapter 2

Bandwidth Allocation under Budget Constraints

2.1 Problem Definitions

Consider a communication (core) network $\mathbb{G}$, where the traffic occurs at the source node randomly, and it will be connected through the core network $\mathbb{G}$ if there exist available virtual paths. We presume the existence of a reservation-based QoS architecture using virtual paths for real-time applications and using scheduling policies that are capable of assigning bandwidth to aggregates of flows. Suppose there are m different traffic classes in the network $\mathbb{G}$, and denote $\mathbb{M} = \{1, 2, \ldots, m\}$ as an index set consisting of m traffic classes.

For each class $i \in \mathbb{M}$, there exist K_i available virtual (end-to-end) paths for connections of class i, where K_i is a positive integer. For any incoming connection, say j, of class $i \in \mathbb{M}$, when allowed to enter the core network $\mathbb{G}$, it will be routed through one path $p_{i,j}$ from those K_i virtual paths with allocated bandwidth x_i. In other words, data packets of incoming connection j of class $i \in \mathbb{M}$ can be transmitted along a virtual (end-to-end) path $p_{i,j}$. In general, every virtual path of class $i \in \mathbb{M}$ is allocated by the same amount of bandwidth x_i. Every virtual path of class $i \in \mathbb{M}$

has to meet the same minimum bandwidth requirement $b_i^{\min} \geq 0$, namely,

$$x_i \geq b_i^{\min}, \ \forall i \in \mathbb{M}. \tag{2.1}$$

Given limited budget B, network managers would like to determine the bandwidth allocation $\mathbf{x} = (x_1, \ldots, x_m)$ under the available number of virtual paths $\vec{K} = (K_1, \ldots, K_m)$. Due to the limited budget B on network planning, there exists the budget constraint

$$\sum_{i \in \mathbb{M}} K_i c_i x_i \leq B, \tag{2.2}$$

where $c_i > 0$ is the average cost of one unit bandwidth through virtual paths for class $i \in \mathbb{M}$. The goal is to determine the bandwidth allocation under the budget constraint so that the revenue earned by the network access providers is maximized. From the budget constraint (2.2) and the minimum bandwidth requirement (2.1), we can determine the possible range of available bandwidth for each class $i \in \mathbb{M}$, e.g.,

$$b_i^{\min} \leq x_i \leq \frac{B - \sum_{j \neq i} K_j c_j b_j^{\min}}{K_i c_i}, \ \forall i \in \mathbb{M}. \tag{2.3}$$

The maximum throughput of those virtual paths are limited by either the total budget B or bottleneck links which lie on those virtual paths. The blocking of connections occurs if the dynamic traffic demand exceeds the maximum throughput [113].

Definition 2.1 *The* ***maximum throughput*** *of class $i \in \mathbb{M}$ is defined as the maximum connection volume (in packets) that can be transmitted through those virtual paths in a unit of time [66, 119]. Namely, the maximum throughput Θ_i of K_i virtual paths for class $i \in \mathbb{M}$, is the product*

$$\Theta_i \triangleq K_i x_i. \tag{2.4}$$

Assume that connections of class $i \in \mathbb{M}$ occur at the source node with mean arrival rate $\lambda_i > 0$, and the average connection volume in terms of packets to be transmitted is $\sigma_i > 0$, $i \in \mathbb{M}$. For each class $i \in \mathbb{M}$, suppose the mean sojourn time of connections on virtual paths is $1/\mu_i$, which corresponds to the packet transmission time, and it is equal to average connection volume divided by bandwidth, i.e.,

$$\frac{1}{\mu_i} \triangleq \frac{\sigma_i}{x_i}. \tag{2.5}$$

Suppose that connections occupy the virtual paths in the order they occur and that sojourn times are identically distributed and mutually independent.

Definition 2.2 *The* ***traffic intensity*** *of virtual paths is defined as the fraction of the time in which virtual paths are occupied [49, 87, 103]. Namely, the traffic intensity of K_i virtual paths, $i \in \mathbb{M}$, is*

$$\rho_i \triangleq \frac{\lambda_i}{K_i\mu_i} = \frac{\lambda_i\sigma_i}{K_i x_i}, \tag{2.6}$$

which is the average occupancy of those K_i virtual paths.

Definition 2.3 *The* ***traffic demand*** *is defined as the product of the mean occurrence rate and the average connection volume [21, 119]. Equivalently, the traffic demand y_i for class $i \in \mathbb{M}$, is the product*

$$y_i \triangleq \lambda_i\sigma_i. \tag{2.7}$$

The principal quantity of interest is the blocking probability that all K_i virtual paths of class $i \in \mathbb{M}$ are occupied. In the situation here, the blocking is due to the failure of setting up the available number of virtual paths under limited budget B. For each class $i \in \mathbb{M}$, a connection gets dropped off upon its arrival at the source node when all K_i virtual paths are occupied. Otherwise, it will be routed through a virtual path $p_{i,j}$ with allocated bandwidth x_i. Our objective is to determine blocking probabilities in terms of bandwidth x_i.

Definition 2.4 *The* ***average throughput*** *of class $i \in \mathbb{M}$ is defined as the average connection volume transmitted through those virtual paths in a unit of time [22, 113]. That is, the average throughput $\overline{\Theta}_i$ of K_i virtual paths for class $i \in \mathbb{M}$, is*

$$\overline{\Theta}_i \triangleq \sum_{n=1}^{K_i} n x_i P_n, \tag{2.8}$$

where P_n is the steady-state probabilities that there are n virtual paths occupied by connections, $1 \leq n \leq K_i$.

Remark 2.1 *For each class $i \in \mathbb{M}$, let $\mathcal{L}(x_i, K_i, y_i)$ be the expected virtual path occupancy in the steady state. Then we have $\overline{\Theta}_i = x_i\mathcal{L}(x_i, K_i, y_i)$. It is worth noting that $\mathcal{L}(x_i, K_i, y_i)$ is a function of bandwidth x_i, the preset number of end-to-end paths K_i in the off-line optimization, but the traffic demand $y_i = \lambda_i\sigma_i$ from online traffic flow. The explicit expression of $\mathcal{L}(x_i, K_i, y_i)$ will be given in the next chapter.*

Definition 2.5 *The* ***utilization level*** *of virtual paths is defined as the percentage of maximum throughput [40]. That is, the utilization level of K_i virtual paths is*

$$U_i \triangleq \frac{\overline{\Theta}_i}{\Theta_i} = \frac{\mathcal{L}(x_i, K_i, y_i)}{K_i}. \tag{2.9}$$

2.2 Related Works on Bandwidth Sharing Policies

Kelly et al. [65, 66] show that an allocation policy can be expressed in terms of a utility function of bandwidth/rate, in the sense that the desired bandwidth allocation maximizes aggregate utility subject to constraints. The classical notion of fairness is the so-called max-min fairness [22]. In its original form, the max-min fairness rule of allocating bandwidth disregards network throughput, and instead emphasizes the importance of meeting the demands of the poor users. Massoulié and Roberts [81, 82] proposed an allocation policy called minimum potential delay with utility functions. Kelly et al. [66] advocated proportional fairness with concave logarithmic utility functions. Network managers can design rate control mechanisms that converge to the associated bandwidth allocation [66, 68, 69]. This method of allocating bandwidth yields a compromise between high total bandwidth allocation (total network throughput) and fairness between users.

Mo and Walrand [86] characterized the class of (w, α)-proportionally fair bandwidth allocation, for given weight w and any number $0 < \alpha \neq 1$. Due to the strict concavity of the function to be maximized, it defines a unique allocation referred to as α-bandwidth allocation [86]. The (w, α)-proportionally fair bandwidth allocation corresponds to the maximum throughput criterion when $\alpha \to 0$, to proportional fairness when $\alpha \to 1$, to the potential delay criterion when $\alpha \to 2$, and to max-min fairness when $\alpha \to \infty$ [86]. Wang and Luh [111, 116] proposed the proportional fair bandwidth allocation by using the achievement function of bandwidth, which is strictly increasing and concave. Depending on the specified aspiration and reservation levels of bandwidth, the utility/achievement functions transformed different QoS measurements onto a normalized scale [111]. The use of the logarithmic function prevents the possibility of assigning zero bandwidth to any user [112], and makes it non-profitable to assign too much bandwidth to the users [116].

In equilibrium, connections that share the same links do not necessarily equally share the available bandwidth [36, 58]. Rather, their shares reflect how they value the bandwidth as expressed by their utility functions and how their use of the bandwidth implies a cost on others [4, 66]. This could be a basis to provide differentiated services in terms of different bandwidth allocations [22, 77]. An equilibrium bandwidth allocation is usually characterized in terms of its fairness to users [21, 82]. Thus given a fixed number of users and fixed network capacities, one can typically arrange through an appropriate control mechanism to achieve an equilibrium which represents, according to some criterion, an equitable bandwidth allocation among users, e.g., [22, 69, 75, 86, 134].

2.3 Two Revenue Management Schemes

This thesis considers the scenario in which network users access the core network via ISP, where users belong to m different classes and the bandwidth received by each class is determined by revenue management schemes. To determine the amount of bandwidth required by connections for each traffic class, we introduce two queueing-based revenue management schemes that maximizes ISPs' revenue subject to the budget constraint.

2.3.1 Revenue Management Scheme I

Establishing a pricing scheme that charges the network users can regulate an overwhelming number of connections during congestion times [130]. Network users will be charged by ISPs to recover the operating cost based on the duration, the volume and the distance of a connection, which is incurred for setting and maintaining virtual paths [99]. Other than recovering costs, the ISP managers may handle congestion control by charging the users for regulating the mean occurrence rate λ_i of the connections during peak periods [6]. Network users are willing to pay a price because an excessive congestion would result in the inability of providing service to critical applications.

The operating costs can be determined by the types of traffic transmitted and the QoS guaranteed for such transfer, e.g., bandwidth allocation and blocking probability [130]. As far as QoS is concerned, bandwidth allocation $\mathbf{x}$ and blocking probability $\mathcal{P}(x_i, K_i, y_i)$ are the key elements of the pricing scheme, where the explicit expression of blocking probability $\mathcal{P}(x_i, K_i, y_i)$ will be given in the next chapter. For example, the long-run average revenue per unit time $F_i(x_i, K_i, y_i)$ for a given traffic class $i \in \mathbb{M}$ can be expressed as follows [130]:

$$F_i(x_i, K_i, y_i) = c_i^t \mathcal{L}(x_i, K_i, y_i) + c_i^b \lambda_i x_i (1 - \mathcal{P}(x_i, K_i, y_i)), \tag{2.10}$$

where $c_i^b > 0$ is the cost charged for using per unit of bandwidth and $c_i^t > 0$ is the cost per unit of time for the sojourn time $1/\mu_i = \sigma_i/x_i$ on those virtual paths. The total long-run average revenue is obtained by summing over (2.10) for all traffic classes. In this thesis, we illustrate the effect of changing x_i, y_i and K_i on the average revenue (2.10) by investigating properties of the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ and expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$. Those phenomena will also be illustrated in the numerical results.

The network managers may want to maximize the long-run average revenue in their bandwidth sharing policies [2, 92, 130, 133]. Here, we consider the long-run average revenue in (2.10) as the objective function for each traffic class. To maximize the average revenue under the budget constraint, we propose Revenue Management Scheme I as follows:

$$
\begin{aligned}
F = \quad & \max \quad \sum_{i \in \mathbb{M}} w_i F_i(x_i, K_i, y_i) \\
& \text{s.t.} \quad \text{budget constraint (2.2)},
\end{aligned}
$$

where the weight $0 < w_i < 1$ is assigned to traffic class $i \in \mathbb{M}$ by network managers.

2.3.2 Revenue Management Scheme II

In this subsection, we provide a congestion-based pricing scheme to allocate bandwidth while taking network users' utility into account. Bandwidth sharing in a network is frequently evaluated in terms of a utility function [6, 53], etc. For example, Kelly et al. [66] proposed an optimization framework in which the objective is to maximize the total utility of all network users over their transmission rates. The utility $f_i(x_i)$ of a connection of class $i \in \mathbb{M}$ is assumed to be an increasing function of its bandwidth x_i, as introduced by Kelly et al. [66]. The shape of the utility function $f_i(x_i)$ depends on the network user's behavior. For example, utility functions of risk-averse users are different from those of risk-seeking users. Examples of possible utility functions are $f_i(x_i) = \log x_i$ for class $i \in \mathbb{M}$, leading to so-called proportional fairness in Kelly et al. [66], and $f_i(x_i) = x_i^{1-\alpha}/(1-\alpha)$, $0 < \alpha < \infty$, for α-proportional fairness defined more generally by Mo and Walrand [86]. Max-Min fairness arises in the limit $\alpha \to \infty$ while proportional fairness corresponds to $\alpha \to 1$. In the limit $\alpha \to 0$, the objective is to maximize overall throughput by the detriment of fairness. More general notions of weighted fairness can be defined by multiplying the utility function with a class-dependent weight.

As introduced in [6, 22, 66], etc., network managers may consider the utility function $f_i(x_i) : [b_i^{\min}, b_i^{\max}] \to [0,1]$ for each traffic class $i \in \mathbb{M}$, where $b_i^{\min}$ is the minimum bandwidth requirement, and $b_i^{\max}$ is the upper bound of bandwidth determined from (2.3). The utility function $f_i(x_i)$ is assumed to be continuous, increasing, and concave [22, 66, 116]. For example, Wang and Luh [116] introduced the utility function

$$f_i(x_i) = \log_{\frac{a_i}{r_i}} \frac{x_i}{r_i}, \tag{2.11}$$

where a_i and r_i are the aspiration level and reservation level of bandwidth for class-i users, respectively, and they have $b_i^{\min} < r_i < a_i$. The logarithmic function is intimately associated with the concept of proportional fairness [66]. Depending on the specified reference levels, a_i and r_i, this utility function can be interpreted as a measure of the user's satisfaction with the value of the i-th criteria [116]. It is a strictly increasing function of bandwidth x_i, having value 1 if $x_i = a_i$, and value 0 if $x_i = r_i$. The utility function can map the different bandwidth requirement of traffic classes onto a normalized scale of the user's satisfaction.

Network managers' economic profit consists of all revenue gained by providing bandwidth x_i to class i and the opportunity cost (in terms of revenue loss) through calculating the risk of blocking connections/users. Suppose that, for each traffic class $i \in \mathbb{M}$, network managers gain the payoff $p_i > 0$ for achieving the utility $f_i(x_i)$ by providing bandwidth x_i. Meanwhile, we introduce the opportunity cost of blocking connections/users. Let $q_i > 0$ be the opportunity cost of increasing blocking probability $\mathcal{P}_i(x_i, K_i, y_i)$ for traffic class $i \in \mathbb{M}$. A higher blocking probability will lead to a higher opportunity loss in the network manager's revenue. Those two criteria can be combined into single objective function with payoff p_i and opportunity cost q_i for all classes $i \in \mathbb{M}$, where payoff p_i and opportunity cost q_i can be applied in designing network pricing mechanisms. By economic definition, let

$$G_i(x_i, K_i, y_i) = p_i K_i f_i(x_i) - q_i \mathcal{P}_i(x_i, K_i, y_i) \tag{2.12}$$

be the managers' economic profit gained from class $i \in \mathbb{M}$, which represents its payoff minus the opportunity cost. Network managers determine the optimal bandwidth allocation under the budget constraint when taking both users' utility and blocking probability into account. Thus, through the weighted sum of all traffic classes, the optimal bandwidth allocation can be determined by the following optimization model, Revenue Management Scheme II:

$$\begin{aligned} G = \quad & \max \quad \sum_{i\in\mathbb{M}} w_i G_i(x_i, K_i, y_i) \\ & \text{s.t.} \quad \text{budget constraint (2.2)}, \end{aligned}$$

where the weight $0 < w_i < 1$ is assigned to traffic class $i \in \mathbb{M}$ by network managers.

Remark 2.2 *If the budget B is finite and (2.3) holds, the feasible set is bounded since the bandwidth allocated to each class i has a upper bound, $\forall i \in \mathbb{M}$. Moreover, the feasible set shrinks to an empty set if $||\vec{K}||_2 = (\sum_{i=1}^{m} K_i^2)^{1/2}$ increases to a sufficiently large number, where $||\cdot||_2$ denotes the well-known Euclidean norm on the vector space $\mathbb{R}^m$.*

Next, we present a budget ratio to investigate the budget allocation among different classes in the revenue management schemes. Using the maximum throughput Θ_i defined in (2.4) for class $i \in \mathbb{M}$, the budget constraint (2.2) can be represented as

$$\sum_{i\in\mathbb{M}} \Theta_i + \tau = \Omega(\vec{K}, B), \tag{2.13}$$

where $\tau \geq 0$ is the reserved bandwidth, and $\Omega(\vec{K}, B)$ is the total available bandwidth purchased with limited budget B for preset numbers of virtual paths $\vec{K} = (K_1, \ldots, K_m)$ on the core network $\mathbb{G}$. Then, the budget ratio for each traffic class $i \in \mathbb{M}$ is given below.

Definition 2.6 *A **budget ratio** allocated to class i is defined as the fraction of total available bandwidth $\Omega(\vec{K}, B) < \infty$ allocated for class $i \in \mathbb{M}$. That is, the budget ratio allocated for class $i \in \mathbb{M}$ is*

$$B_i \triangleq \frac{\Theta_i}{\Omega(\vec{K}, B)}, \tag{2.14}$$

where the maximum throughput $\Theta_i = K_i x_i$.

If there exists sufficient bandwidth, the reserved bandwidth $\tau > 0$ may be shared among all the ongoing connections with budget ratio B_i for each class $i \in \mathbb{M}$ at the online process. That is, any remaining bandwidth is shared according to the budget ratio B_i. In case there is no bandwidth reserved ($\tau = 0$), the allocated bandwidth x_i will decrease if K_i increases when those preset numbers of virtual paths, K_j, $j \neq i$, of other classes are fixed.

Remark 2.3 *If the maximum throughput Θ_i increases in (2.13), the value $\sum_{j \neq i} \Theta_j$ will decrease. Since those numbers K_j are fixed for all $j \neq i$, the bandwidth x_j will decrease. That is, for certain class $j \neq i$, the bandwidth x_j may be snatched by class i as the maximum throughput Θ_i of class i becomes larger.*

Chapter 3

Blocking Probabilities of Connections under $M/G/K/K$

In this chapter, we derive monotone and convex properties of the blocking probability and expected path occupancy of connections. Those two revenue functions (2.10) and (2.12) can be analyzed through the monotonicity and convexity of the blocking probability with respect to bandwidth, traffic demand and the number of virtual paths, respectively. Under the limited budget, the blocking probability of connections is determined due to the failure of meeting the traffic demand of class $i \in \mathbb{M}$ for those K_i virtual paths.

3.1 Monotone and Convex Properties

We assume the user population is infinite, and connections join the core network according to a Poisson process [2, 37, 103, 130]. Connections of class i occur at the source node in accordance with independent Poisson processes at rate $\lambda_i > 0$, but the connection volume in terms of packets to be transmitted has an arbitrary distribution with mean $\sigma_i > 0$, $i \in \mathbb{M}$. The sojourn time during which they occupy virtual paths has an arbitrary distribution.

The risk of blocking possible connections is analyzed according to an Erlang loss model under assumptions of Poisson arrivals, general sojourn time, preset K_i virtual paths with identical bandwidth x_i, and no waiting space [37, 87]. Assume that, for each traffic class $i \in \mathbb{M}$, there exists the steady-state occupancy probabilities of n $(0 \leq n \leq K_i)$ connections, P_n. The unique steady-state probability of $M/G/K/K$ is given by

$$P_n = P_0 \frac{(\lambda_i/\mu_i)^n}{n!} = \frac{P_0}{n!}\left(\frac{\lambda_i \sigma_i}{x_i}\right)^n, \; n = 1, 2, \ldots, K_i. \tag{3.1}$$

Solving for P_0 in the equation $\sum_{n=0}^{K_i} P_n = 1$, we obtain that

$$P_0 = \left[\sum_{n=0}^{K_i} \frac{1}{n!}\left(\frac{\lambda_i}{\mu_i}\right)^n\right]^{-1} = \left[\sum_{n=0}^{K_i} \frac{1}{n!}\left(\frac{\lambda_i \sigma_i}{x_i}\right)^n\right]^{-1} \tag{3.2}$$

and then

$$P_n = \frac{1}{n!}\left(\frac{\lambda_i \sigma_i}{x_i}\right)^n \left[\sum_{j=0}^{K_i} \frac{1}{j!}\left(\frac{\lambda_i \sigma_i}{x_i}\right)^j\right]^{-1}, \text{ for } n = 1, 2, \ldots, K_i. \tag{3.3}$$

Thus, according to the Erlang loss model [24, 105], the blocking probability of incoming connections is formulated as

$$\mathcal{P}(x_i, K_i, y_i) = P_{K_i} = \frac{1}{K_i!}\left(\frac{y_i}{x_i}\right)^{K_i}\left[\sum_{n=0}^{K_i} \frac{1}{n!}(\frac{y_i}{x_i})^n\right]^{-1}, \tag{3.4}$$

where y_i is the traffic demand and is defined by (2.7). Moreover, the expected (virtual) path occupancy in the steady state can be formulated as

$$\begin{aligned}
\mathcal{L}(x_i, K_i, y_i) &= \sum_{n=1}^{K} \frac{1}{(n-1)!}\left(\frac{y_i}{x_i}\right)^n \left[\sum_{j=0}^{K} \frac{1}{j!}\left(\frac{y_i}{x_i}\right)^j\right]^{-1} \\
&= \frac{y_i}{x_i}\left(\sum_{n=0}^{K-1} P_n\right) \\
&= \frac{y_i}{x_i}\left(1 - \mathcal{P}(x_i, K_i, y_i)\right)
\end{aligned} \tag{3.5}$$

from the definition of $\mathcal{L}(x_i, K_i, y_i) = \sum_{n=0}^{K_i} nP_n$. We assume the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is twice differentiable with respect to bandwidth x_i and traffic demand y_i, respectively.

Proposition 3.1 shows that the blocking probability is decreasing in bandwidth. The monotonicity can be proved by deriving the first partial derivative of the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ in (3.4).

Proposition 3.1 *The blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is a decreasing function of bandwidth x_i, given $K_i \geq 1$ and $y_i > 0$ fixed.*

Proof. From (3.4), the first derivative of $\mathcal{P}(x_i, K_i, y_i)$ with respect to x_i is

$$\begin{aligned}\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} &= \frac{e^{-2y_i/x_i}\left\{-x_i + e^{y_i/x_i}(y_i - K_i x_i)(\frac{y_i}{x_i})^{-1-K_i}\Gamma[K_i+1, \frac{y_i}{x_i}]\right\}}{y_i x_i \left\{(\frac{y_i}{x_i})^{-1-K_i}\Gamma[K_i+1, \frac{y_i}{x_i}]\right\}^2} \\ &= \frac{-x_i + (y_i - K_i x_i)e^{y_i/x_i}(\frac{y_i}{x_i})^{-1-K_i}\Gamma[K_i+1, \frac{y_i}{x_i}]}{y_i x_i \left\{e^{y_i/x_i}(\frac{y_i}{x_i})^{-1-K_i}\Gamma[K_i+1, \frac{y_i}{x_i}]\right\}^2}, \qquad (3.6)\end{aligned}$$

where $\Gamma[a, z]$ is the upper incomplete gamma function which satisfies $\Gamma[a, z] = \int_z^\infty t^{a-1}e^{-t}dt$. It must be noted that if the arrival stream is infinite, it is sufficient to assume that $\mathcal{P}(x_i, K_i, y_i) > 0$ in any finite system. For positive integer K_i, we have

$$\Gamma[K_i+1, \frac{y_i}{x_i}] = K_i! e^{-y_i/x_i}\sum_{n=0}^{K_i}\frac{(y_i/x_i)^n}{n!},$$

which implies that

$$e^{y_i/x_i}(\frac{y_i}{x_i})^{-1-K_i}\Gamma[K_i+1, \frac{y_i}{x_i}] = \frac{x_i}{y_i \mathcal{P}(x_i, K_i, y_i)}.$$

Hence, given $K_i \geq 1$ and $y_i > 0$, we obtain that

$$\begin{aligned}\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} &= \frac{-x_i + (y_i - K_i x_i)\frac{x_i}{y_i\mathcal{P}(x_i,K_i,y_i)}}{y_i x_i\left(\frac{x_i}{y_i\mathcal{P}(x_i,K_i,y_i)}\right)^2} \\ &= \frac{-\mathcal{P}(x_i, K_i, y_i)}{x_i^2}\left[K_i x_i - y_i(1 - \mathcal{P}(x_i, K_i, y_i))\right] \\ &= \frac{-K_i\mathcal{P}(x_i, K_i, y_i)}{x_i}\left[1 - \frac{y_i}{K_i x_i} + \frac{y_i}{K_i x_i}\mathcal{P}(x_i, K_i, y_i)\right] \\ &= \frac{-K_i\mathcal{P}(x_i, K_i, y_i)}{x_i}\left[\frac{\sum_{n=0}^{K_i}(1-\frac{n}{K_i})\frac{(y_i/x_i)^n}{n!}}{\sum_{n=0}^{K_i}\frac{(y_i/x_i)^n}{n!}}\right] \\ &< 0 \qquad (3.7)\end{aligned}$$

for all $x_i > 0$. So, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is decreasing in bandwidth x_i. □

Furthermore, by deriving the second partial derivative of (3.4), it shows that the blocking probability is convex in bandwidth for a specific region.

Proposition 3.2 *For each $K_i \geq 1$ and $y_i > 0$, there exists a subset (or region) $\mathbb{S}_i$ of positive real numbers such that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex (concave) in bandwidth x_i for all $x_i \in (\notin)\ \mathbb{S}_i$.*

Proof. From the first derivative (3.6), we can derive the second derivative of $\mathcal{P}(x_i, K_i, y_i)$ with respect to x_i as follows. Considering the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ given in (3.4), we have

$$
\begin{aligned}
&\frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial x_i^2} \\
= \;& \frac{e^{-3y_i/x_i}}{y_i x_i^3 \left\{ (\frac{y_i}{x_i})^{-1-K_i} \Gamma[K_i+1, \frac{y_i}{x_i}] \right\}^3} \cdot \Bigg\{ 2x_i^2 + e^{y_i/x_i} (\frac{y_i}{x_i})^{-1-K_i} \\
& \cdot \Gamma[K_i+1, \frac{y_i}{x_i}] \cdot \left[x_i(-3y_i + (2+3K_i)x_i) + y_i^2 - 2y_i(1+K_i)x_i + K_i(1+K_i)x_i^2 \right] \Bigg\} \\
= \;& \frac{1}{y_i x_i^3 \left\{ \frac{x_i}{y_i \mathcal{P}(x_i, K_i, y_i)} \right\}^3} \cdot \left\{ 2x_i^2 + x_i/(y_i \mathcal{P}(x_i, K_i, y_i)) \right. \\
& \left. \cdot \left[x_i(-3y_i + (2+3K_i)x_i) + y_i^2 - 2y_i(1+K_i)x_i + K_i(1+K_i)x_i^2 \right] \right\} \\
= \;& \frac{y_i \mathcal{P}^2(x_i, K_i, y_i)}{x_i^5} \\
& \cdot \left\{ (K_i^2 + 4K_i + 2)x_i^2 - y_i(2K_i + 5 - 2\mathcal{P}(x_i, K_i, y_i))x_i + y_i^2 \right\},
\end{aligned}
\tag{3.8}
$$

given $K_i \geq 1$ and $y_i > 0$. To determine the sign of (3.8), it needs to consider $\mathcal{S}(P) = y_i^2(2K_i + 5 - 2P)^2 - 4y_i^2(K_i^2 + 4K_i + 2)$ for $P \in [0, 1]$. It can be easily checked that $\mathcal{S}(0) = y_i^2(4K_i + 17) > 0$, $\mathcal{S}(1) = y_i^2(1 - 4K_i) < 0$ and $d\mathcal{S}(P)/dP = -4y_i^2(2K_i + 5 - 2P) < 0$ for $P \in (0, 1)$. We obtain that $\mathcal{S}(P) = 0$ if $P = K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2}$. Meanwhile, we have $\mathcal{S}(P) < 0$ for $P \in (K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2}, 1]$ and $\mathcal{S}(P) > 0$ for $P \in [0, K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2})$. It should be noted that, as $K_i \to \infty$, the limit of the sequence $\{K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2} \mid K_i \in \mathbb{N}\}$ is 0.5, where $\mathbb{N}$ is the set of positive integers. Hence, we conclude that $\partial^2 \mathcal{P}(x_i, K_i, y_i)/\partial x_i^2 \geq 0$ for all x_i satisfying

$$K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2} \leq \mathcal{P}(x_i, K_i, y_i) \leq 1,$$

whose inverse image is

$$0 < x_i \leq \mathcal{P}^{-1}(K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2}).$$

Notice that the inverse image (or region) $\mathcal{P}^{-1}(\cdot)$ can be easily determined because $\mathcal{P}(x_i, K_i, y_i)$ is a decreasing function of bandwidth x_i by Proposition 3.1. In the case of bandwidth x_i satisfying

$$0 \leq \mathcal{P}(x_i, K_i, y_i) < K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2},$$

we have $\mathcal{S}(P) > 0$. Therefore, we conclude that there exist two inflection points x_i^* and x_i^{**} in its inverse image, where $x_i^* < x_i^{**}$, such that

$$\frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial x_i^2} < 0, \text{ for all } x_i^* < x_i < x_i^{**},$$

and

$$\frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial x_i^2} > 0, \text{ for all } 0 < x_i < x_i^* \text{ or } x_i > x_i^{**}.$$

So, we can find the region $\mathbb{S}_i$ of positive real numbers such that $\mathcal{P}(x_i, K_i, y_i)$ is convex in bandwidth x_i for all $x_i \in \mathbb{S}_i$. Otherwise, $\mathcal{P}(x_i, K_i, y_i)$ is concave in bandwidth x_i for all $x_i \notin \mathbb{S}_i$. □

Remark 3.1 *It should be noted that, as $K_i \to \infty$, the limit of the sequence $\{K_i + \frac{5}{2} - \sqrt{K_i^2 + 4K_i + 2} \mid K_i \in \mathbb{N}\}$ is 0.5, where $\mathbb{N}$ is the set of positive integers. As the number of virtual paths, K_i, is huge in real-world communication systems, Proposition 3.2 implies that $\mathcal{P}(x_i, K_i, y_i)$ is convex in bandwidth x_i if we have $0.5 < \mathcal{P}(x_i, K_i, y_i) \leq 1$. Otherwise, there exist two inflection points x_i^* and x_i^{**} when $0 \leq \mathcal{P}(x_i, K_i, y_i) < 0.5$.*

Remark 3.2 *For fixed $y_i > 0$ and $K_i \geq 1$, the expected path occupancy $L(x_i, K, y_i)$ is a decreasing function of bandwidth x_i. In addition, there exists an inflection point $\widetilde{x}_i$ such that for all $x_i \leq (\geq)\widetilde{x}_i$ the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is concave (convex) in bandwidth x_i. given $y_i > 0$ and $K_i \geq 1$ fixed.*

Remark 3.3 *It can also be observed that the utilization level U_i defined in (2.9) is a decreasing function of bandwidth x_i for given $y_i > 0$ and $K_i \geq 1$. This is because the utilization level U_i equals the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ divided by K_i. Meanwhile, there exists the same inflection point $\widetilde{x}_i$ as in $\mathcal{L}(x_i, K_i, y_i)$ such that for all $x_i \leq (\geq)\widetilde{x}_i$ the utilization level U_i is concave (convex) in bandwidth x_i.*

Next, we prove the monotone property of the blocking probability with respect to the traffic demand. The monotonicity can be observed by deriving those partial derivatives of the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ in (3.4) with respect to y_i.

Proposition 3.3 *The blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is increasing in traffic demand y_i, given $x_i > 0$ and $K_i \geq 1$ fixed.*

Proof. From (3.4), the first derivative of $\mathcal{P}(x_i, K_i, y_i)$ with respect to y_i is

$$\begin{aligned}
&\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial y_i} \\
&= \frac{y_i^{K_i-1}}{K_i! x_i^{K_i}} \left[\sum_{n=0}^{K_i} \frac{y_i^n}{n! x_i^n} \right]^{-2} \left\{ K_i \left(\sum_{n=0}^{K_i} \frac{y_i^n}{n! x_i^n} \right) - y_i \left(\sum_{n=1}^{K_i} \frac{n y_i^{n-1}}{n! x_i^n} \right) \right\} \\
&= \frac{y_i^{K_i-1}}{K_i! x_i^{K_i}} \left[\sum_{n=0}^{K_i} \frac{y_i^n}{n! x_i^n} \right]^{-2} \left\{ K_i + \sum_{n=1}^{K_i} \frac{(K_i - n) y_i^n}{n! x_i^n} \right\}, \qquad (3.9)
\end{aligned}$$

for all $y_i > 0$. Given $x_i > 0$ and $K_i \geq 1$, the first derivative (3.9) is always positive. Hence, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is increasing in traffic demand y_i. $\square$

When traffic demand y_i exceeds the maximum throughput Θ_i, for class i, i.e., $\lambda_i \sigma_i > K_i x_i$, the number of blocked connections increases indefinitely. It is shown that per-flow QoS depends critically on whether the traffic demand y_i is less than or greater than the maximum throughput Θ_i.

Remark 3.4 *For fixed $x_i > 0$ and $K_i \geq 1$, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex (concave) in traffic demand y_i if we have $\Theta_i \geq (\leq) y_i$. Furthermore, the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is an increasing and concave function of the traffic demand y_i, and the upper bound of $\mathcal{L}(x_i, K_i, y_i)$ is K_i no matter what y_i increases. The utilization level U_i in (2.9) is also an increasing and concave function of the traffic demand y_i, and the upper bound of U_i is 1 no matter what y_i increases.*

The following monotone and convex properties of the blocking probability with respect to K_i are consistent with those results of the Erlang-B function proved by Messerli [83], Jagers and Van Doorn [56], and Esteves et al. [39]. For detailed proofs, interested readers may refer to [83, 56, 39].

Proposition 3.4 *The blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is decreasing with respect to the number of virtual paths K_i, given $x_i > 0$ and $y_i > 0$.*

Proof. For all $K_i \geq 1$, the decrement of blocking probability $\mathcal{P}(x_i, K_i, y_i)$ with respect to K_i is determined by

$$\begin{aligned}
&\mathcal{P}(x_i, K_i + 1, y_i) - \mathcal{P}(x_i, K_i, y_i) \\
= \; & \frac{1}{(K_i+1)!}(\frac{y_i}{x_i})^{K_i+1}\left[\sum_{n=0}^{K_i+1}\frac{1}{n!}(\frac{y_i}{x_i})^n\right]^{-1} - \frac{1}{K_i!}(\frac{y_i}{x_i})^{K_i}\left[\sum_{n=0}^{K_i}\frac{1}{n!}(\frac{y_i}{x_i})^n\right]^{-1} \\
= \; & \frac{1}{K_i!}(\frac{y_i}{x_i})^{K_i}\left[\frac{\frac{y_i}{(K_i+1)x_i}}{\sum_{n=0}^{K_i+1}\frac{1}{n!}(\frac{y_i}{x_i})^n} - \frac{1}{\sum_{n=0}^{K_i}\frac{1}{n!}(\frac{y_i}{x_i})^n}\right]. \qquad (3.10)
\end{aligned}$$

Given $x_i > 0$ and $y_i > 0$, (3.10) is always negative. Hence, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is decreasing in K_i. □

Remark 3.5 *For fixed $x_i > 0$ and $y_i > 0$, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex in K_i, which is considered specifically in relaxation for continuous functions. Moreover, the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is an increasing function of K_i given x_i and y_i are fixed. It can also be observed that the utilization level U_i is an increasing function of K_i because the increment of $\mathcal{L}(x_i, K_i, y_i)$ is larger than the increment of K_i when both is increased with K_i by one unit.*

3.2 Blocking Probability with Respect to Traffic Intensity

Next,we present monotone and convex properties of the blocking probability with respect to traffic intensity ρ_i defined in (2.6). In (3.4), if we replace $y_i/(K_i x_i)$ by traffic intensity ρ_i, the equivalent expression of (3.4) is

$$\mathcal{P}(\rho_i, K_i) = \frac{(K_i\rho_i)^{K_i}}{K_i!}\left[\sum_{n=0}^{K_i}\frac{(K_i\rho_i)^n}{n!}\right]^{-1}. \qquad (3.11)$$

Meanwhile, the expected path occupancy (3.5) can be rewritten as

$$\mathcal{L}(\rho_i, K_i) = K_i \rho_i \left(1 - \mathcal{P}(\rho_i, K_i)\right). \tag{3.12}$$

After a little algebra, we derive the first and second derivatives of (3.11) with respect to ρ_i as follows:

$$\frac{\partial \mathcal{P}(\rho_i, K_i)}{\partial \rho_i} = \frac{K_i \mathcal{P}(\rho_i, K_i)}{\rho_i} \left[1 - \rho_i + \rho_i \mathcal{P}(\rho_i, K_i)\right] \tag{3.13}$$

and

$$\frac{\partial^2 \mathcal{P}(\rho_i, K_i)}{\partial \rho_i^2} = \frac{K_i \mathcal{P}(\rho_i, K_i)}{\rho_i^2} \left[1 - \rho_i + \rho_i \mathcal{P}(\rho_i, K_i)\right]^2 \mathcal{H}(\rho_i), \tag{3.14}$$

where

$$\mathcal{H}(\rho_i) = K_i + \frac{K_i \rho_i \mathcal{P}(\rho_i, K_i)}{1 - \rho_i + \rho_i \mathcal{P}(\rho_i, K_i)} - \frac{1}{\left[1 - \rho_i + \rho_i \mathcal{P}(\rho_i, K_i)\right]^2}. \tag{3.15}$$

Interested readers may refer to Bakry [15] and Harel [49] for detailed derivations.

The monotone and convex properties of the blocking probability in (3.11) are listed below, which are consistent with those properties shown by Harel [49]. The blocking probability in (3.11) is convex in the traffic intensity ρ_i if ρ_i is below certain inflection point ρ_i^* and concave if ρ_i is greater than ρ_i^*. For detailed proofs, interested readers may refer to Harel [49] and Bakry [15].

Theorem 3.1 *For each $K_i \geq 1$, the blocking probability $\mathcal{P}(\rho_i, K_i)$ is an increasing function of traffic intensity $\rho_i > 0$.*

Proof. Given $K_i \geq 1$, the first derivative of $\mathcal{P}(\rho_i, K_i)$ with respect to ρ_i is

$$\begin{aligned} \frac{\partial \mathcal{P}(\rho_i, K_i)}{\partial \rho_i} &= \frac{K_i \mathcal{P}(\rho_i, K_i)}{\rho_i} \left[1 - \rho_i + \rho_i \mathcal{P}(\rho_i, K_i)\right] \\ &= \frac{K_i \mathcal{P}(\rho_i, K_i)}{\rho_i} \left[\frac{\sum_{n=0}^{K_i} (K_i - n) K_i^{n-1} \rho_i^n / n!}{\sum_{n=0}^{K_i} (K_i \rho_i)^n / n!}\right] > 0, \end{aligned} \tag{3.16}$$

for all $\rho_i > 0$. So, the blocking probability $\mathcal{P}(\rho_i, K_i)$ is increasing in ρ_i. □

Theorem 3.2 *For each $K_i \geq 1$, there exists an inflection point ρ_i^* such that for all $\rho_i < (>) \rho_i^*$, the blocking probability $\mathcal{P}(\rho_i, K_i)$ is convex (concave) in traffic intensity $\rho_i > 0$.*

Proof. To show that such an inflection point ρ_i^* exists, we examine the function $\mathcal{H}(\rho_i)$ defined in (3.15). For $K_i = 1$, we have

$$\mathcal{P}(\rho_i, K_i) = \frac{\rho_i}{1+\rho_i},$$

so it implies

$$\mathcal{H}(\rho_i) = -2\rho_i < 0, \ \forall \rho_i > 0. \tag{3.17}$$

For $K_i = 2$, we obtain

$$\mathcal{P}(\rho_i, K_i) = \frac{2\rho_i^2}{1+2\rho_i+2\rho_i^2},$$

and

$$1-\rho_i+\rho_i\mathcal{P}(\rho_i, K_i) = \frac{1+\rho_i}{1+2\rho_i+2\rho_i^2},$$

thus it gives

$$\mathcal{H}(\rho_i) = \frac{1-6\rho_i^2-4\rho_i^3}{(1+\rho_i)^2}.$$

Hence, we find that $\rho_i^* = -0.5+\sqrt{3/4}$ and $\mathcal{H}(\rho_i) > (<)0$ for all $\rho_i < (>)\rho_i^*$. For $K_i \geq 3$, we use (3.11) and (3.15) to obtain

$$\begin{aligned} &\left[\sum_{n=0}^{K_i-1} \frac{(K_i-n)K_i^{n-1}\rho_i^n}{n!}\right]^2 \mathcal{H}(\rho_i) \\ = &\sum_{n=0}^{K_i-1} \frac{(K_i\rho_i)^n}{K_i}\left[\sum_{j=0}^{n} \frac{K_i(K_i-n-1)+j(n-j)}{(n-j)!j!}\right] \\ &+\sum_{n=0}^{K_i-1} \frac{(K_i\rho_i)^{K_i+n}\mathcal{T}(n)}{K_i(K_i+n)!}, \end{aligned} \tag{3.18}$$

where $\mathcal{T}(n)$ is defined as

$$\mathcal{T}(n) = \binom{K_i+n}{n} n(K_i+1-n) - \sum_{j=n}^{K_i} \binom{K_i+n}{j}\left[K_i-(K_i-j)(j-n)\right], \tag{3.19}$$

for $0 \leq n \leq K_i - 1$. It must further be noted that for every K_i, there exists $n^* < K_i$ such that $\mathcal{T}(n^*) \geq 0$, $\mathcal{T}(n) > 0$ for all $n < n^*$ and $\mathcal{T}(n) < 0$ for all $n > n^*$. Observe that (3.18) is a polynomial of ρ_i. It can be shown that the coefficients of (3.18) change sign only once. Thus, equating (3.18) to zero yields one positive real root, namely ρ_i^*. □

Remark 3.6 *For each $K_i \geq 1$, the expected path occupancy $\mathcal{L}(\rho_i, K_i)$ is increasing and concave in traffic intensity ρ_i, and the upper bound of $\mathcal{L}(\rho_i, K_i)$ is K_i. Moreover, the utilization level U_i is also increasing and concave in traffic intensity ρ_i, and the upper bound is 1.*

3.3 Blocking Probability with Huge Number of Connections

In real-world cases, the number of connections on networks is always huge, i.e., $K_i \gg 1$. Next, we investigate the blocking probability defined in (3.4) in the case of $K_i \gg 1$. If the traffic intensity $\rho_i = y_i/(K_i x_i) < 1$ holds, equation (3.4) can be rewritten as

$$\begin{aligned} \mathcal{P}(x_i, K_i, y_i) &= \frac{(y_i/x_i)^{K_i}}{K_i!}\left[\sum_{n=0}^{K_i} \frac{(y_i/x_i)^n}{n!}\right]^{-1} \\ &= \frac{(y_i/x_i)^{K_i}}{K_i!}\left[e^{y_i/x_i} - \mathcal{R}(K_i)\right]^{-1} \\ &\approx \frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i}}, \text{ as } K_i \gg 1, \end{aligned} \tag{3.20}$$

where $\mathcal{R}(K_i)$ is the K_ith-degree Taylor remainder term of e^{y_i/x_i}. The remainder term $\mathcal{R}(K_i) \approx 0$ as $K_i \gg 1$. Moreover, we can conclude that

$$\mathcal{L}(x_i, K_i, y_i) \approx \frac{y_i}{x_i}\left(1 - \frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i}}\right), \text{ as } K_i \gg 1. \tag{3.21}$$

From (3.20) and (3.21), it implies that

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} = (\frac{y_i}{x_i^2} - \frac{K_i}{x_i}) \cdot \mathcal{P}(x_i, K_i, y_i), \tag{3.22}$$

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial y_i} = (\frac{K_i}{y_i} - \frac{1}{x_i}) \cdot \mathcal{P}(x_i, K_i, y_i), \tag{3.23}$$

$$\begin{aligned} \frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial x_i \partial y_i} &= \frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial y_i \partial x_i} \\ &= \frac{1}{x_i}(1 + 2K_i - \frac{y_i}{x_i} - \frac{K_i^2 x_i}{y_i}) \cdot \mathcal{P}(x_i, K_i, y_i); \end{aligned} \tag{3.24}$$

$$\frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial x_i} = -\frac{y_i}{x_i^2}\left[1 - \mathcal{P}(x_i, K_i, y_i)\cdot(1 + K_i - \frac{y_i}{x_i})\right], \tag{3.25}$$

$$\frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial y_i} = \frac{1}{x_i}\left[1 - \mathcal{P}(x_i, K_i, y_i)\cdot(1 + K_i - \frac{y_i}{x_i})\right], \tag{3.26}$$

and

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(x_i, K_i, y_i)}{\partial x_i \partial y_i} &= \frac{\partial^2 \mathcal{L}(x_i, K_i, y_i)}{\partial y_i \partial x_i} \\ &= -\frac{1}{x_i^2}\left\{1 - \mathcal{P}(x_i, K_i, y_i)\cdot\left[(1 + K_i - \frac{y_i}{x_i})^2 - \frac{y_i}{x_i}\right]\right\} \end{aligned} \tag{3.27}$$

in the case of large $K_i \gg 1$ and traffic intensity $\rho_i < 1$.

In the case of large $K_i \gg 1$, Proposition 3.1 can be restated as follows.

Corollary 3.1 *If $K_i \gg 1$ and $\rho_i < 1$ holds, the first derivative of blocking probability $\mathcal{P}(x_i, K_i, y_i)$ with respect to bandwidth x_i is always negative for all $x_i > 0$, i.e.,*

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} = (\frac{y_i}{x_i} - K_i)\frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i+1}} < 0. \tag{3.28}$$

In the following result, we demonstrate the monotone property of the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ with respect to allocated bandwidth x_i, i.e., $\partial \mathcal{L}(x_i, K_i, y_i)/\partial x_i < 0$, for all traffic class $i \in \mathbb{M}$.

Proposition 3.5 *If the traffic intensity $\rho_i = y_i/K_i x_i > 1$ holds in the case of large $K_i \gg 1$, the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is a decreasing function of bandwidth x_i, given $y_i > 0$ fixed.*

Proof. Given $K_i \gg 1$ and $y_i > 0$, the first derivative of expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ with respect to bandwidth x_i can be determined from (3.21). By using (3.20), we have

$$\begin{aligned}\frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial x_i} &= -\frac{y_i}{x_i^2}(1 - \mathcal{P}(x_i, K_i, y_i)) - \frac{y_i}{x_i}\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} \\ &= -\frac{y_i}{x_i^2}\left[1 - \frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i}}(1 + K_i - \frac{y_i}{x_i})\right]. \end{aligned} \tag{3.29}$$

Note that $\rho_i = y_i/K_i x_i > 1$ implies $1 + K_i - \frac{y_i}{x_i} < 1$. Meanwhile, we have

$$\frac{y_i^{K_i} e^{-y_i/x_i}}{(K_i! x_i^{K_i})} = \mathcal{P}(x_i, K_i, y_i) < 1.$$

Thus, we can obtain

$$\frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial x_i} < 0.$$

□

Chapter 4

Blocking Probabilities of Connections under $GI/M/K/K$

In this chapter, our objective is to estimate the blocking probability that all K_i virtual paths of class $i \in \mathbb{M}$ are occupied under assumptions of $GI/M/K/K$. For every traffic class $i \in \mathbb{M}$, the mathematical derivation is conducted in general format. Hence, we skip the notation i in the following derivation to simplify the notation.

4.1 Embedded Markov Chain Technique

The assumptions of renewal arrival process, exponential service times, finite servers and limited buffer size are commonly used in queueing systems [1, 11, 34, 48]. Assume connections occur at the source node in accordance with independent general probability distribution, and we assume that the inter-arrival times of connections are independent and identically distributed (i.i.d.) random variables with cumulative distribution function (c.d.f.) $A(t)$, probability density function $a(t)$ for $t > 0$, Laplace-Stiltjes' transform $A^*(z)$, and mean $1/\lambda$. We also assume that the sojourn times of connections are i.i.d. random variables following exponential distribution with mean $1/\mu = \sigma/x$, where σ is the average connection volume and x is the allocated bandwidth. Suppose that the inter-arrival time and sojourn time are mutually independent. Connections occupy those K virtual paths in the order they occur, that is, the service discipline is First Come First Served.

We analyze the $GI/M/K/K$ queue through a combination of the supplementary variable and the embedded Markov chain techniques. We use the former technique to derive closed form relations between pre-arrival and arbitrary epoch probabilities and the latter one to obtain pre-arrival epoch probabilities. From the result of Laxmi and Gupta [110], we obtain the steady state probabilities of n $(0 \leq n \leq K)$ connections in the system at pre-arrival epoches, P_n^- by using the embedded Markov chain technique. First, we develop the relation between the distributions of the number of connections in the system at arbitrary and pre-arrival epoches, i.e., P_n and P_n^-. After some similar manipulation in [110], we obtain

$$P_{n+1} = \frac{y}{x(n+1)} P_n^-, \; n = 0, 1, \ldots, K-1, \tag{4.1}$$

where $y = \lambda\sigma$ is the traffic demand defined in (2.7). Once the P_n^- $(0 \leq n \leq K)$ are known, one can get P_n $(1 \leq n \leq K)$ from (4.1). Finally, P_0 is obtained by using $\sum_{n=0}^{K} P_n = 1$.

The steady state probabilities at pre-arrival epoches, P_n^-'s $(0 \leq n \leq K)$, can be determined by solving the system of linear equations

$$P_n^- = \sum_{m=0}^{K} P_m^- P_{m,n}, \; 0 \leq n \leq K, \tag{4.2}$$

where $P_{m,n}$'s are the one-step transition probabilities. The expression for $P_{m,n}$ is

$$P_{m,n} = \begin{cases} \int_0^\infty \binom{m+1}{n} e^{-\frac{xnt}{\sigma}} (1 - e^{-\frac{xt}{\sigma}})^{m+1-n} dA(t), & 0 \leq n \leq m < K, \\ \int_0^\infty \binom{K}{n} e^{-\frac{xnt}{\sigma}} (1 - e^{-\frac{xt}{\sigma}})^{K-n} dA(t), & 0 \leq n \leq m = K, \\ \int_0^\infty e^{-\frac{xnt}{\sigma}} dA(t), & 1 \leq m+1 = n \leq K, \\ 0, & 1 \leq m+1 < n \leq K. \end{cases} \tag{4.3}$$

Let $\mathbf{T} = [P_{m,n}]_{(K+1)\times(K+1)}$ denote the one-step transition probability matrix. Then the system of linear equations (4.2) can be rewritten as

$$\mathbf{P}^- = \mathbf{P}^- \mathbf{T}, \tag{4.4}$$

where $\mathbf{P}^- = (P_0^-, P_1^-, \ldots, P_K^-)'$ denotes the row vector of steady state probabilities at pre-arrival epoches. Hence, the state probability P_{K-1}^- can be computed as

$$P_{K-1}^- = \frac{1 - \int_0^\infty e^{-\frac{Kx}{\sigma}t} dA(t)}{\int_0^\infty e^{-\frac{Kx}{\sigma}t} dA(t)} P_K^-. \tag{4.5}$$

For simplicity, we denote

$$\mathcal{H}_K = \frac{1 - \int_0^\infty e^{-\frac{Kx}{\sigma}t} dA(t)}{\int_0^\infty e^{-\frac{Kx}{\sigma}t} dA(t)} \tag{4.6}$$

and $\mathcal{H}_n =$

$$\frac{\mathcal{H}_{n+1} - \sum_{j=n+1}^{K} \mathcal{H}_j \binom{j}{n} \int_0^\infty e^{-\frac{nx}{\sigma}t}(1 - e^{-\frac{x}{\sigma}t})^{j-n} dA(t) - \binom{K}{n} \int_0^\infty e^{-\frac{nx}{\sigma}t}(1 - e^{-\frac{x}{\sigma}t})^{K-n} dA(t)}{\int_0^\infty e^{-\frac{nx}{\sigma}t} dA(t)},$$

for all $n = K-1, K-2, \ldots, 1$. Then it gives $P_{K-1}^- = \mathcal{H}_K P_K^-$. Similarly, we can obtain the following iterative relations for other steady state probabilities

$$P_n^- = \mathcal{H}_{n+1} P_K^-, \ \forall n = K-1, K-2, \ldots, 1, 0. \tag{4.7}$$

The fact that the total sum of all steady state probabilities is equal to 1 implies

$$\sum_{n=0}^{K} P_n^- = (1 + \sum_{n=1}^{K} \mathcal{H}_n) P_K^- = 1.$$

Therefore, the steady state probability that all virtual paths are occupied at pre-arrival epoches can be computed as

$$P_K^- = \frac{1}{1 + \sum_{n=1}^{K} \mathcal{H}_n}. \tag{4.8}$$

We can obtain the steady state probabilities at pre-arrival epoches, P_n^-'s, by solving the system of linear equations (4.2). Then, we know all P_n, $0 \le n \le K$, from (4.1) and $\sum_{n=0}^{K} P_n = 1$. Hence, we can compute the blocking probability of connections in each QoS class, $\mathcal{P}(x, K, y)$. Therefore, we can determine the expected path occupancy in the steady state

$$\mathcal{L}(x, K, y) = \sum_{n=0}^{K} n P_n, \tag{4.9}$$

which represents the average number of connections occupying those K virtual paths.

To illustrate the above embedded Markov chain technique, we determine the steady state probabilities P_n^- under three common distributions for inter-arrival time: exponential, deterministic and Erlang-r distributions in the following examples.

Example 4.1 *(Exponential distribution) Assume the inter-arrival time is exponentially distributed with parameter λ. Then*

$$P_{m,n} = \begin{cases} \frac{y}{x}\frac{(m+1)!\Gamma(n+\frac{y}{x})}{n!\Gamma(m+2+\frac{y}{x})}, & 0 \le n \le m < K, \\ \frac{y}{x}\frac{K!\Gamma(n+\frac{y}{x})}{n!\Gamma(K+1+\frac{y}{x})}, & 0 \le n \le m = K, \\ \frac{y}{x}\frac{\Gamma(n+\frac{y}{x})}{\Gamma(n+1+\frac{y}{x})}, & 1 \le m+1 = n \le K, \\ 0, & 1 \le m+1 < n \le K. \end{cases} \tag{4.10}$$

From the above $P_{m,n}$'s, we can construct a K by K matrix as (4.39), where we can easily observe that

$$P_{K-1}^- = \frac{Kx}{y}P_K^-. \tag{4.11}$$

From iteratively backward computation, we can find

$$P_n^- = \mathcal{H}_{n+1}P_K^-, \ \forall n = K-1, K-2, \ldots, 1, 0,$$

where

$$\mathcal{H}_K = \frac{Kx}{y} \tag{4.12}$$

and

$$\mathcal{H}_n = \left(\frac{nx}{y}+1\right)\left\{\mathcal{H}_{n+1} - \left[\sum_{j=n+1}^{K}\frac{yj!\Gamma(n+\frac{y}{x})}{xn!\Gamma(j+1+\frac{y}{x})}\mathcal{H}_j\right] - \frac{yK!\Gamma(n+\frac{y}{x})}{xn!\Gamma(K+1+\frac{y}{x})}\right\}, \tag{4.13}$$

for all $n = K-1, \ldots, 1$. Finally, we obtain

$$P_K^- = \frac{1}{1+\sum_{n=1}^{K}\mathcal{H}_n}.$$

Example 4.2 *(Deterministic distribution) Assume the inter-arrival time is deterministic with mean $1/\lambda = d$,*

$$a(t) = \delta(t-d) = \{ \begin{array}{ll} \infty, & t = d, \\ 0, & t \neq d. \end{array}$$

We can determine

$$P_{m,n} = \begin{cases} \binom{m+1}{n} e^{-\frac{xn}{y}} (1-e^{-\frac{x}{y}})^{m+1-n}, & 0 \leq n \leq m < K, \\ \binom{K}{n} e^{-\frac{xn}{y}} (1-e^{-\frac{x}{y}})^{K-n}, & 0 \leq n \leq m = K, \\ e^{-\frac{xn}{y}}, & 1 \leq m+1 = n \leq K, \\ 0, & 1 \leq m+1 < n \leq K. \end{cases} \tag{4.14}$$

From the above $P_{m,n}$'s, we can construct a K by K matrix as (4.39), where we can easily observe that

$$P^-_{K-1} = (e^{\frac{Kx}{y}} - 1) P^-_K. \tag{4.15}$$

From iteratively backward computation, we can find

$$P^-_n = \mathcal{H}_{n+1} P^-_K, \ \forall n = K-1, K-2, \ldots, 1, 0,$$

where

$$\mathcal{H}_K = e^{\frac{Kx}{y}} - 1 \tag{4.16}$$

and

$$\begin{aligned} \mathcal{H}_n \;\; = \;\; & e^{\frac{nx}{y}} \left\{ \mathcal{H}_{n+1} - \left[\sum_{j=n+1}^{K} \binom{j}{n} e^{-\frac{nx}{y}} (1-e^{-\frac{x}{y}})^{j-n} \mathcal{H}_j \right] \right. \\ & \left. - \binom{K}{n} e^{-\frac{nx}{y}} (1-e^{-\frac{x}{y}})^{K-n} \right\}, \end{aligned} \tag{4.17}$$

for all $n = K-1, \ldots, 1$. Finally, we obtain

$$P^-_K = \frac{1}{1 + \sum_{n=1}^{K} \mathcal{H}_n}.$$

Example 4.3 *(Erlang-r distribution) Assume the inter-arrival time is Erlang-r, for* *positive integers $r \geq 2$, with mean $1/\lambda$. Then*

$$P_{m,n} = \begin{cases} \binom{m+1}{n} (\frac{ry}{x})^r \sum_{l=0}^{m+1-n} \binom{m+1-n}{l} \frac{(-1)^{m+2-n}}{(m+\frac{ry}{x})^r}, & 0 \leq n \leq m < K, \\ \binom{K}{n} (\frac{ry}{x})^r \sum_{l=0}^{K-n} \binom{K-n}{l} \frac{(-1)^{K+1-n}}{(K-1+\frac{ry}{x})^r}, & 0 \leq n \leq m = K, \\ (\frac{ry}{x})^r \frac{(-1)}{(n-1+\frac{ry}{x})^r}, & 1 \leq m+1 = n \leq K, \\ 0, & 1 \leq m+1 < n \leq K. \end{cases} \tag{4.18}$$

From the above $P_{m,n}$'s, we can construct a K by K matrix as (4.39), where we can easily observe that

$$P^-_{K-1} = (-1)\left[\left(1+\frac{Kx}{y}\right)^r - 1\right] P^-_K. \tag{4.19}$$

From iteratively backward computation, we can find

$$P^-_n = \mathcal{H}_{n+1} P^-_K, \ \forall n = K-1, K-2, \ldots, 1, 0,$$

where

$$\mathcal{H}_K = \left(1+\frac{Kx}{y}\right)^r - 1 \tag{4.20}$$

and

$$\begin{aligned} &\mathcal{H}_n \\ = \ &(-1)\left[\frac{(n-1)x}{ry}+1\right]^r \left\{ \mathcal{H}_{n+1} - \left[\sum_{j=n+1}^{K} \binom{j}{n} \left(\frac{ry}{x}\right)^r \sum_{l=0}^{j-1} \binom{j-1}{l} \frac{(-1)^j}{(j-1+\frac{ry}{x})^r} \mathcal{H}_j\right] \right. \\ &\left. - \binom{K}{n} \left(\frac{ry}{x}\right)^r \sum_{l=0}^{K-n} \binom{K-n}{l} \frac{(-1)^{K+1-n}}{(K-1+\frac{ry}{x})^r} \right\}, \end{aligned} \tag{4.21}$$

for all $n = K-1, \ldots, 1$. Finally, we obtain

$$P^-_K = \frac{1}{1+\sum_{n=1}^K \mathcal{H}_n}.$$

Remark 4.1 *In real-world communication networks, it becomes difficult or even impossible to numerically compute the blocking probability $\mathcal{P}(x, K, y)$ for huge number of virtual paths, K. As mentioned in [11], the main drawback with exact methods of analyzing the $GI/G/K/K$ queue is the often-excessive computation times required.*

Indeed, many problems become intractable with small to medium-sized values of K. In our case, the difficulty comes from numerically solving the system of equations in (4.2) with complex elements (4.3) as the number of virtual paths is large.

4.2 Approximation of the Blocking Probability with Heavy-Traffic Limits

The aim of this section is to provide an approximation for the blocking probability as the number of virtual paths is huge. We revise the approach that was adopted in Kim and Choi [63] to derive an asymptotic analysis of blocking probability as K tends to infinity. In our case, both the roots of the characteristic equation and the stationary probabilities depend on the available number of virtual paths (servers), K.

4.2.1 Heavy-Traffic Limits

For real telecommunication systems, the case of a heavy load parameter is the most interesting in practice [13, 54]. Because multi-server systems often have a very large number of servers, it is natural to look for insight into system performance by considering asymptotic behaviors as the number of servers is allowed to increase [57, 70, 126]. Network managers may be interested in the behavior of the blocking probability in heavy loaded systems [10, 57, 96]. Here, we study the blocking probability under $GI/M/K/K$ in which the number of available virtual paths, K, is a very large value.

Choi et al. [34] and Kim and Choi [63] obtained some results related to the $GI/M/K/n$ and $GI^X/M/K/n$ queues with batch size X, where K is the fixed number of servers and n is the variable number of waiting places. As the waiting places n increases to infinity, Choi et al. [34] obtained the estimation for the convergence rate of the stationary $GI/M/K/n$ queue-length distribution to the stationary queue-length distribution of the $GI/M/K$ queueing system. In [63], Kim and Choi provided an analysis of the blocking probability in the $GI^X/M/K/n$ queueing systems. Recently, Abramov [1] provides an asymptotic analysis of the blocking probability of the $GI/M/K/n$ queue as the waiting place n approaches to infinity. However, in those literatures, the number of servers K is fixed and hence the traffic intensity is also fixed.

The most commonly used limit theorem for large-scale queueing systems is that of Halfin and Whitt [48], who considered the $GI/M/K$ queue as $K \to \infty$ and $\rho_K \to 1$ such that

$$(1 - \rho_K)\sqrt{K} \to \gamma \tag{4.22}$$

with $-\infty < \gamma < \infty$. For the $M/M/K$ queue with $\rho_K < 1$, they showed that the steady-state probability that a customer must wait in the queue approaches to a limit κ with $0 < \kappa < 1$ as $K \to \infty$ if and only if $0 < \gamma < \infty$. For the $GI/M/K$ queue, they showed that a properly centered and normalized version of the queue length process converges to a one-dimensional diffusion.

In those queueing models, diffusion approximations for stochastic processes are quite common, e.g., [30, 50, 70, 124]. Such limits were established for the $GI/M/K/\infty$ queueing model (with renewal arrival process, exponential service times, K servers, and unlimited waiting room) by Halfin and Whitt [126]. They considered a sequence of queues indexed by the number of servers, K, and let $K \to \infty$ with the traffic intensities ρ_K converging to one, the critical value for stability [48]. Note that the parameter ρ_K is assumed to depend on K. The diffusion approximations can be obtained by heuristic methods and limit theorems involving a sequence of queueing systems under heavy traffic conditions [47, 126, 127].

Here, we consider a sequence of queueing models indexed by the number of virtual paths, K, and let both $K \to \infty$ and $\lambda_K \to \infty$ such that the traffic intensity $\rho_K \to 1$ from below. From a user's perspective, the higher the allocated bandwidth x, the higher the user's utility. So we assume that the higher the bandwidth allocation x arises, the higher the arrival rate λ_K becomes. That is, network managers allocate more bandwidth will increase the users' arrival rate [125]. In practice, network managers may handle congestion control by reducing the bandwidth allocated to each connection for regulating the arrival rate λ_K during peak periods [124]. The following definitions are given and will be used throughout the whole context of this thesis.

Remark 4.2 *According to Definition 2.2, we refine the traffic intensity of queueing system indexed by K virtual paths. Namely, the traffic intensity of system is*

$$\rho_K \triangleq \frac{\lambda_K}{K\mu} = \frac{\lambda_K \sigma}{Kx}, \tag{4.23}$$

where $\sigma > 0$ is the average connection volume and $x > 0$ is the allocated bandwidth.

Assumption 4.1 *As the number of virtual paths, K, increases to infinity, we assume the traffic intensities ρ_K approaches to 1 from below, i.e.,*

$$\lim_{K\to\infty} \rho_K = 1. \tag{4.24}$$

Remark 4.3 *Assume that $\lambda_K = K\mu - \gamma\mu\sqrt{K}$, with $0 < \gamma < \sqrt{K}$. Then the traffic intensity is formulated as follows*

$$\rho_K = 1 - \frac{\gamma}{\sqrt{K}}. \tag{4.25}$$

In such a case, there exists an interesting nondegenerate limit in Halfin-Whitt heavy traffic regimes, namely, $\rho_K \to 1$ and $(1-\rho_K)\sqrt{K} \to \gamma$ as $K \to \infty$.

Our objective is to estimate the blocking probability that all K virtual paths are occupied. In real-world communication networks, it becomes difficult to compute numerically the blocking probability $\mathcal{P}(\rho_K, K)$ for large K even though by computers. As mentioned in [11], the main drawback with exact methods of analyzing the $GI/G/K/K$ queue is the often-excessive computation times required. Indeed, many problems become intractable with small to medium-sized values of K.

In the current and following sections, the $GI/M/K/K$ queueing systems is considered as the number of virtual paths K increases to infinity, where the traffic intensity ρ_K depends on K. We provide an approximation for the blocking probability by solving a root of a characteristic equation with the stationary probability of $GI/M/\infty$ queues in the next section. If the traffic intensity $\rho_K < 1$, then the root $0 < \Phi_K < 1$ exists and hence blocking probabilities can be estimated. Once we obtain the root Φ_K and its limit (to be discussed later), the blocking probability $\mathcal{P}(\rho_K, K)$ can be estimated immediately for any value of K. Computational effort for this approximation is much less than the one for determining the exact value of the blocking probability $\mathcal{P}(\rho_K, K)$ in $GI/M/K/K$ queueing systems as K is large when solving the matrix computation is considered.

4.2.2 Roots of the Characteristic Equations

Here, we assume that the traffic intensity $\rho_K < 1$ for all K. In real-word communication networks, it means that the allocated links are sufficiently large to meet the increasing traffic demand y. In such a case, the maximum throughput Θ exceeds the traffic demand y. It implies that the allocated bandwidth x is sufficiently large, i.e.,

$$x > \frac{y}{K},$$

by definition of the maximum throughput $\Theta = Kx$ given in (2.4). Note that the condition $x > y/K$ implies $\rho_K < 1$.

First, we consider the scenario that there are infinite virtual paths in system, i.e., $GI/M/\infty$ queue [102]. Let $P_n^{(\infty)}$, $n = 0, 1, 2, \ldots$, be the stationary probability that the number of connections occupying those links is n just before arrivals. Let $\beta_n(K)$, $n = 0, 1, 2, \ldots$, be the probability that number of connections which have been served during an inter-arrival time is n when all K virtual paths are busy. Then $\beta_n(K)$ is given by

$$\beta_n(K) = \int_0^\infty e^{-K\mu t} \frac{(K\mu t)^n}{n!} dA(t), \; n = 0, 1, 2, \ldots, \tag{4.26}$$

where $A(t)$ is the c.d.f. of inter-arrival times. From [102], we have the following lemma.

Lemma 4.1 *If the c.d.f. $A(t)$ is non-lattice, then the integral*

$$\beta_n(K) = \int_0^\infty e^{-K\mu t} \frac{(K\mu t)^n}{n!} dA(t), \tag{4.27}$$

exists for each $n = 0, 1, 2, \ldots$.

The generating function of $\beta_i(K)$, $i = 0, 1, 2, \ldots$, is given by

$$\sum_{n=0}^{\infty} \beta_n(K) z^n = A^*(K\mu - K\mu z), \ |z| \leq 1. \tag{4.28}$$

Next, we consider the characteristic equation

$$z - A^*(K\mu - K\mu z) = 0, \ |z| \leq 1, \tag{4.29}$$

that is,

$$z = \int_0^\infty e^{-(K\mu - K\mu z)t} dA(t). \tag{4.30}$$

Let Φ_K denote the least in absolute value root of the characteristic equation (4.30). It is well-known that the root Φ_K belongs to the open interval $(0, 1)$ if the traffic intensity $\rho_K < 1$, and it is equal to 1 otherwise [1, 105, 102]. We note here that the roots of the characteristic equation and the stationary probabilities $\beta_n(K)$ depend on the allocated bandwidth x.

Lemma 4.2 *Let $A^*(z)$ be the Laplace-Stiltjes' transform of the c.d.f. $A(t)$. If $A(t)$ is non-lattice, then there exists a positive real number $0 < \Phi_K < 1$ such that*

$$\Phi_K - A^*(K\mu - K\mu\Phi_K) = 0, \tag{4.31}$$

equivalently, which may be written as

$$\Phi_K - \sum_{n=0}^{\infty} \beta_n(K) \Phi_K^n = 0. \tag{4.32}$$

Proof. Let $Y(z) = A^*(K\mu - K\mu z)$. Since $Y(0) = \beta_0 > 0$, $Y(1) = 1$, and $Y'(1) = 1/\rho_K > 1$, there exists at least one real root between 0 and 1 for the characteristic equation (4.29). For real z, $0 < z \leq 1$, by changing variables via $z = e^s$, $-\infty < s \leq 0$, (4.29) becomes

$$e^s = Y(e^s), \ -\infty < s \leq 0, \tag{4.33}$$

and so

$$s = \ln Y(e^s), \ -\infty < s \leq 0. \tag{4.34}$$

Note that, Hölder's inequality gives

$$Y(e^{(ps_1+(1-p)s_2)}) \leq (Y(e^{s_1}))^p \, (Y(e^{s_2}))^{1-p} \tag{4.35}$$

for all $s_1, \ s_2 \leq 0$ and $0 \leq p \leq 1$. Thus

$$\ln Y(e^{ps_1+(1-p)s_2}) \leq p \ln Y(e^{s_1}) + (1-p) \ln Y(e^{s_2}) \tag{4.36}$$

for all $s_1, \ s_2 \leq 0$ and $0 \leq p \leq 1$. Therefore, the right hand side of (4.34) is convex. Then (4.34) has exactly one negative root and the root is simple. Thus, (4.29) has exactly one real root between 0 and 1, and the root is also simple. Denote by Φ_K the real root of (4.29) between 0 and 1. By Rouché's theorem, the number of zeros of $z - A^*(K\mu - K\mu z)$ and z, counted by multiplicities, on $\{z \in \mathbb{C} : |z| < \eta\}$ are the same for any η, $\Phi_K < \eta < 1$. It is easy to see that $z - A^*(K\mu - K\mu z)$ has no other zeros on $\{z \in \mathbb{C} : |z| = \Phi_K\}$ except the simple zero Φ_K. Thus, we conclude that the characteristic equation (4.29) has exactly one root Φ_K on $\{z \in \mathbb{C} : |z| < 1\}$. □

Remark 4.4 *For each $\varepsilon > 0$ and $\rho_K < 1$, there exists a positive integer N_{ε,ρ_K} such that $\beta_n(K) < \varepsilon$ for all $n > N_{\varepsilon,\rho_K}$. Hence, the root $0 < \Phi_K < 1$ may be solved approximately from the following characteristic equation:*

$$\Phi_K - \sum_{n=0}^{N_{\varepsilon,\rho_K}} \beta_n(K) \Phi_K^n = 0 \tag{4.37}$$

because $\beta_n(K) \approx 0$ for all $n > N_{\varepsilon,\rho_K}$. Moreover, we find that (i.) if the inter-arrival time follows Erlang distribution, it gives $\Phi_K < \rho_K < 1$; (ii.) if the inter-arrival time follows Exponential distribution, it gives $\Phi_K = \rho_K < 1$; (iii.) if the inter-arrival time follows Hyperexponential distribution, it gives $\Phi_K > \rho_K$.

Proposition 4.1 *If the root $0 < \Phi_K < 1$ exists for each positive integer K, then the limit of the roots $(\Phi_K)^K$ exists as $K \gg 1$, and*

$$\lim_{K\to\infty} (\Phi_K)^K = \begin{cases} 0, & \textit{if } \limsup_{K\to\infty} \Phi_K < 1, \\ 1, & \textit{if } \limsup_{K\to\infty} \Phi_K = 1. \end{cases} \tag{4.38}$$

Proof. By the root test in real analysis, if we have $\limsup_{K\to\infty} \sqrt[K]{(\Phi_K)^K} = \limsup_{K\to\infty} \Phi_K < 1$, it gives that the series $\sum_K (\Phi_K)^K$ converges, and hence

$$\lim_{K\to\infty} (\Phi_K)^K = 0.$$

Otherwise, it gives $\lim_{K\to\infty}(\Phi_K)^K = 1$ if the condition $\limsup_{K\to\infty} \Phi_K = 1$ holds. □

4.2.3 Estimating the Blocking Probability

After finding the root $0 < \Phi_K < 1$ of the characteristic equation (4.32), the blocking probability can be estimated by investigating the asymptotic behavior of the stationary probabilities of K connections in service, as $K \to \infty$.

Let $\mathbf{P}^{(K)} = (P_0^{(K)}, \ldots, P_K^{(K)})$ be the stationary probability vector of connections in service. It is easy to see that $\mathcal{P}(\rho_K, K) = P_K^{(K)}$. Let $\mathbf{T}^{(K)} = [P_{m,n}]_{(K+1)\times(K+1)}$ be the one-step transition probability matrix of the embedded Markov chain. The one-step transition probability matrix $\mathbf{T}^{(K)}$ can be represented as follows

$$\begin{bmatrix} P_{0,0}^{(\infty)} & P_{0,1}^{(\infty)} & 0 & \cdots & 0 \\ P_{1,0}^{(\infty)} & P_{1,1}^{(\infty)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ P_{K-2,0}^{(\infty)} & P_{K-2,1}^{(\infty)} & \cdots & P_{K-2,K-1}^{(\infty)} & 0 \\ P_{K-1,0}^{(\infty)} & P_{K-1,1}^{(\infty)} & \cdots & P_{K-1,K-1}^{(\infty)} & \beta_0(K-1) \\ P_{K,0} & P_{K,1} & \cdots & P_{K,K-1} & \beta_0(K)+\beta_1(K) \end{bmatrix}, \tag{4.39}$$

where those elements $P_{m,n}^{(\infty)}$, $0 \le m \le K-1$, $0 \le n \le K-1$, and $P_{K,n}$, $0 \le n \le K-1$, can be explicitly determined in terms of model parameters. It can be found that the stationary probability vector $\mathbf{P}^{(K)} = (P_0^{(K)}, \ldots, P_K^{(K)})$ is the unique solution of $\mathbf{P}^{(K)}\mathbf{T}^{(K)} = \mathbf{P}^{(K)}$ and $\sum_{n=0}^{K} P_n^{(K)} = 1$. By applying the coupling method in [102] on the limiting distribution of $\mathbf{P}^{(K)}$ as $K \to \infty$, it gives the following lemma, which will be borrowed for further development in this thesis.

Lemma 4.3 *The stationary distribution of* $\mathbf{P}^{(K)} = (P_0^{(K)}, \ldots, P_K^{(K)})$ *goes weakly to that of the* $GI/M/\infty$ *queue as* $K \to \infty$*, i.e.,*

$$\lim_{K\to\infty} P_n^{(K)} = P_n^{(\infty)},\ n = 0, 1, \ldots, K. \tag{4.40}$$

Now, to investigate the asymptotic behavior of the probability $P_K^{(K)}$ as $K \to \infty$, we introduce a transformation

$$\tilde{\pi}(P_K^{(K)}) = (\tilde{\pi}_0(P_K^{(K)}), \tilde{\pi}_1(P_K^{(K)}), \ldots, \tilde{\pi}_K(P_K^{(K)}))$$

as follows:

$$\tilde{\pi}_0(P_K^{(K)}) = P_K^{(K)} \tag{4.41}$$

and

$$\tilde{\pi}_n(P_K^{(K)}) = \frac{1}{\beta_0(K)}\left[\tilde{\pi}_{n-1}(P_K^{(K)}) - P_K^{(K)} b_{n-1}(1) - \sum_{j=1}^{n-1} \tilde{\pi}_j(P_K^{(K)})\beta_{n-i}\right],\ n \geq 1, \tag{4.42}$$

where $b_0(1) = \beta_0(K) + \beta_1(K)$ and $b_n(1) = 0$, $n \geq 1$. Similar to the derivation in [63], it gives the following lemma.

Lemma 4.4 *Let* $\tilde{\pi}(\cdot) : [0,1]^{K+1} \to [0,1]^{K+1}$ *be a transformation defined as (4.41) and (4.42). Then*

$$P_{K-n}^{(K)} = \tilde{\pi}_n(P_K^{(K)}),\ n = 0, 1, \ldots, K. \tag{4.43}$$

By the definition of $\tilde{\pi}_n(P_K)$, it is immediate to derive that

$$\tilde{\pi}_n(P_K^{(K)}) = P_K^{(K)} T_n,\ n \geq 0, \tag{4.44}$$

where

$$T_0 = 1 \tag{4.45}$$

and

$$T_n = \frac{(1 - \beta_0(K) - \beta_1(K))(1 - \beta_1(K))^{n-1}}{\beta_0(K)^n},\ n = 1, \ldots, K. \tag{4.46}$$

From Lemma 4.4, when $n = K$, we have

$$P_0^{(K)} = \tilde{\pi}_K(P_K^{(K)}) = P_K^{(K)} T_K. \tag{4.47}$$

Hence, if T_K is invertible, we have

$$\mathcal{P}(\rho_K, K) = P_K^{(K)} = P_0^{(K)} \frac{\beta_0(K)^K}{(1-\beta_0(K)-\beta_1(K))(1-\beta_1(K))^{K-1}}. \tag{4.48}$$

Proposition 4.2 *If $A(t)$ follows one of the exponential, deterministic, or Erlang-r distributions, then the limit of*

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)}$$

exists as $K \to \infty$. Namely,

$$\lim_{K\to\infty} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} = C_1, \tag{4.49}$$

where C_1 is a constant number.

Proof. From (4.26), we can determine $\beta_0(K)$ and $\beta_1(K)$ for different examples of inter-arrival time distributions. By Corollary 4.1, Corollary 4.3, and Corollary 4.5, there exist those limit $\lim_{K\to\infty} \beta_n(K)$ for $n = 0, 1$. Because the limits of $\beta_0(K)$ and $\beta_1(K)$ exist as $K \to \infty$, and $1 - \beta_0(K) - \beta_1(K) > 0$ for all K, the limit of

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)}$$

exists as $K \to \infty$ and equals to a constant number. □

Proposition 4.3 *If $A(t)$ is non-lattice and $(1-\rho_K)\sqrt{K} \to \gamma$ as $K \to \infty$, then the limit of the sequence*

$$\left(\frac{\beta_0(K)}{1-\beta_1(K)}\right)^K$$

exists as $K \gg 1$ and equals to a constant number C_2. Namely, there exists a positive number $\varepsilon > 0$ such that

$$\left(\frac{\beta_0(K)}{1-\beta_1(K)}\right)^K = O(\rho_K^p), \ \ K \gg 1, \tag{4.50}$$

for all $0 < p < \varepsilon$.

Proof. It is obvious that the following inequalities hold

$$0 \leq \frac{\beta_0(K)}{1-\beta_1(K)} = \frac{\beta_0(K)}{\beta_0(K)+\sum_{n=2}^{\infty}\beta_n(K)} \leq 1.$$

Then, we have

$$0 \leq (\frac{\beta_0(K)}{1-\beta_1(K)})^K < 1.$$

Hence, the sequence

$$(\frac{\beta_0(K)}{1-\beta_1(K)})^K$$

is bounded. Moreover, it can be derived that

$$1 \geq (\frac{\beta_0(2)}{1-\beta_1(2)})^1 \geq (\frac{\beta_0(K)}{1-\beta_1(K)})^K$$

for all $K \gg 1$, and we have

$$(\frac{\beta_0(K)}{1-\beta_1(K)})^K$$

decreases when increasing $K \gg 1$. Therefore, it have been shown that $[\beta_0(K)/(1-\beta_1(K))]^K$ is bounded and monotone as $K \gg 1$. So, the limit of

$$(\frac{\beta_0(K)}{1-\beta_1(K)})^K$$

exists as $K \gg 1$ and equals to a constant number. □

Theorem 4.1 *Consider $GI/M/K/K$ queueing systems with non-lattice c.d.f $A(t)$ of inter-arrival times. Assume that $(1-\rho_K)\sqrt{K} \to \gamma$ as $K \to \infty$. Then, as $K \gg 1$, we have the approximation of the blocking probability*

$$\mathcal{P}(\rho_K, K) \approx P_0^{\infty} C_1 C_2. \tag{4.51}$$

Proof. From (4.48), it can be derived that

$$\mathcal{P}(\rho_K, K) = P_0^{(K)} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} (\frac{\beta_0(K)}{1-\beta_1(K)})^K. \tag{4.52}$$

First, by Lemma 4.3, it gives that $P_0^K \to P_0^{\infty}$ as $K \to \infty$. Next, by Proposition 4.2, there exists a constant number C_1 such that

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} \to C_1$$

as $K \to \infty$. In addition, by Proposition 4.3, we have

$$(\frac{\beta_0(K)}{1-\beta_1(K)})^K \to C_2$$

as $K \gg 1$. Hence, the approximation for the blocking probability can be determined as $K \gg 1$. □

Remark 4.5 *Consider the scenario that there are infinite available virtual paths in the network, and both inter-arrival times and sojourn times of connections in that traffic class follow exponential distributions, that is, $M/M/\infty$ queueing systems. It gives*

$$P_0^{(\infty)} \to e^{-\frac{\lambda_K}{\mu}},$$

where

$$P_n^{(\infty)} \to \frac{(\lambda_K/\mu)^n e^{-\lambda_K/\mu}}{n!}$$

as $K \to \infty$.

As the number of virtual paths, K, is large, we provide an approximation for the blocking probability $\mathcal{P}(\rho_K, K)$. Through Lemma 4.2, we can find exactly one real root of (4.29) between 0 and 1, which is denoted by Φ_K. Then, by Proposition 4.2 and Proposition 4.3, we can compute the limit C_1 and C_2 with probabilities $\beta_0(K)$ and $\beta_1(K)$. Finally, by Theorem 4.1, we can estimate efficiently the blocking probability of the $GI/M/K/K$ queueing system for large K with limit C_1, limit C_2, and the stationary probability $P_0^{(\infty)}$.

4.3 Three Illustrative Examples

In this section, we determine the probability $\beta_n(K)$, limit constants C_1 and C_2, and the blocking probability $\mathcal{P}(\rho_K, K)$ under three common distributions for inter-arrival time: exponential, deterministic and Erlang-r distributions.

Example 4.4 *(Exponential distribution) Suppose that the inter-arrival time is exponentially distributed with parameter λ. Then*

$$\begin{aligned}
\beta_n(K) &= \int_0^\infty e^{-K\mu t}\frac{(K\mu t)^n}{n!}\lambda e^{-\lambda t}dt \\
&= \frac{\lambda(K\mu)^n}{n!}\int_0^\infty e^{-(\lambda+K\mu)t}t^n dt \\
&= \frac{\lambda}{n!(1+\rho_K)^n}\int_0^\infty e^{-\lambda(1+\frac{1}{\rho_K})t}(\lambda(1+\frac{1}{\rho_K})t)^n dt \\
&= \frac{\rho_K}{n!(1+\rho_K)^{n+1}}\int_0^\infty e^{-s}s^n ds \quad (\textit{let } s=\lambda t(1+\rho_K)) \\
&= \frac{\rho_K}{n!(1+\rho_K)^{n+1}}\Gamma(n+1) \\
&= \frac{\rho_K}{(1+\rho_K)^{n+1}}, \qquad (4.53)
\end{aligned}$$

for $n = 0, 1, 2, \ldots$, where $\Gamma(n+1) = n!$ is the Gamma function. From (4.53), we derive that

$$\beta_0(K) = \frac{\rho_K}{1+\rho_K}, \qquad (4.54)$$

and

$$\beta_n(K) = \frac{1}{1+\rho_K}\beta_{n-1}(K) = \frac{\rho_K}{(1+\rho_K)^{n+1}} \qquad (4.55)$$

for $n = 1, 2, \ldots$.

From Lemma 4.2, the characteristic equation (4.32) becomes

$$\Phi_K - \sum_{n=0}^\infty \frac{\rho_K}{(1+\rho_K)^{n+1}}\Phi_K^n = 0$$

which implies

$$\Phi_K - \frac{\rho_K}{1+\rho_K-\Phi_K} = 0$$

because we have $\Phi_K/(1+\rho_K) < 1$ and

$$\sum_{n=0}^\infty (\frac{\Phi_K}{1+\rho_K})^n = \frac{1+\rho_K}{1+\rho_K-\Phi_K}.$$

Then we obtain

$$\Phi_K^2 - (1+\rho_K)\Phi_K + \rho_K = 0,$$

which implies that the unique root between 0 and 1 is

$$\Phi_K = \rho_K. \qquad (4.56)$$

Therefore, if $\rho_K < 1$, we determine the root $\Phi_K = \rho_K$ in the case that the inter-arrival time follows the exponential distribution.

Corollary 4.1 *If $A(t)$ is the c.d.f. of exponential inter-arrival times and $\rho_K < 1$ for all K, then the limit of the probability $\beta_n(K)$ exists as $K \to \infty$, and*

$$\lim_{K\to\infty} \beta_n(K) = \frac{1}{2^{n+1}}, \tag{4.57}$$

for $n = 0, 1, 2, \ldots$.

Proof. From (4.55), we derive

$$\lim_{K\to\infty} \beta_n(K) = \lim_{K\to\infty} \frac{\rho_K}{(1+\rho_K)^{n+1}} = \frac{\lim_{K\to\infty} \rho_K}{\lim_{K\to\infty}(1+\rho_K)^{n+1}} = \frac{1}{2^{n+1}} \tag{4.58}$$

with the help of assumption that traffic intensities ρ_K approaches to 1 from below as $K \to \infty$. □

Corollary 4.2 *If $A(t)$ is the c.d.f. of exponential inter-arrival times and $\rho_K < 1$ for all K, then the limit of the roots Φ_K exists as $K \to \infty$, and*

$$\lim_{K\to\infty} \Phi_K = 1. \tag{4.59}$$

Proof. From (4.56), we determine the root $\Phi_K = \rho_K$ if the inter-arrival time follows the exponential distribution. By the assumption that $\rho_K \to 1$ as $K \to \infty$, we have $\Phi_K = \rho_K \to 1$. □

Example 4.5 *(Deterministic case) Assume the inter-arrival time is deterministic with constant $1/\lambda = d$,*

$$a(t) = \delta(t-d) = \{ \begin{array}{ll} \infty, & t = d, \\ 0, & t \neq d. \end{array}$$

In this case, the number of served connections during an inter-arrival time d follows a Poisson distribution with mean $1/\rho_K = K\mu d$. So, the probability $\beta_n(K)$ is determined as the distribution with parameter $K\mu d$, i.e.,

$$\beta_n(K) \;=\; \frac{(K\mu d)^n e^{-K\mu d}}{n!} \tag{4.60}$$

for $n = 0, 1, 2, \ldots$. *From (4.60), we derive that*

$$\beta_0(K) = e^{-1/\rho_K}, \tag{4.61}$$

and

$$\beta_n(K) = \frac{1}{n\rho_K}\beta_{n-1}(K) = \frac{e^{-1/\rho_K}}{n!\rho_K^n} \tag{4.62}$$

for $n = 1, 2, \ldots$.

From Lemma 4.2, the characteristic equation (4.32) becomes

$$\Phi_K - \sum_{n=0}^{\infty} \frac{(1/\rho_K)^n e^{-1/\rho_K}}{n!}\Phi_K^n = 0$$

which implies

$$\Phi_K - e^{-1/\rho_K} e^{\Phi_K/\rho_K} = 0$$

because

$$\sum_{n=0}^{\infty} \frac{(\Phi_K/\rho_K)^n}{n!} = e^{\Phi_K/\rho_K}.$$

By solving the equation

$$\Phi_K = e^{(\Phi_K - 1)/\rho_K}$$

it implies the root

$$\Phi_K = -\rho_K ProductLog[-\frac{e^{-1/\rho_K}}{\rho_K}], \tag{4.63}$$

where $ProductLog[z]$ *is the Lambert's W function (also called the Omega function or product logarithm) which gives the branches of the inverse relation of the function* $z = W(z)e^{W(z)}$. *The value of (4.63) can be determined numerically by the commercial software Mathematica [128].*

Corollary 4.3 *If the inter-arrival time is constant and* $\rho_K < 1$ *for all* K, *then the limit of the probability* $\beta_n(K)$ *exists as* $K \to \infty$, *and*

$$\lim_{K\to\infty} \beta_n(K) = \frac{e^{-1}}{n!}, \tag{4.64}$$

for $n = 0, 1, 2, \ldots$.

Proof. From (4.62), we derive

$$\lim_{K\to\infty}\beta_n(K) = \lim_{K\to\infty}\frac{e^{-1/\rho_K}}{n!\rho^n(K)} = \frac{\lim_{K\to\infty} e^{-1/\rho_K}}{n!\lim_{K\to\infty}\rho^n(K)} = \frac{e^{-1}}{n!} \tag{4.65}$$

with the help of assumption that traffic intensities ρ_K approaches to 1 from below as $K \to \infty$. □

Corollary 4.4 *If the inter-arrival time is constant and $\rho_K < 1$ for all K, then the limit of the roots Φ_K exists as $K \to \infty$, and*

$$\lim_{K\to\infty}\Phi_K = 1. \tag{4.66}$$

Proof. From the properties of the Lambert's W function $ProductLog[z]$ in (4.63), we find that

$$\begin{aligned}\lim_{K\to\infty} ProductLog[-\frac{e^{-1/\rho_K}}{\rho_K}] &= ProductLog[\lim_{K\to\infty}-\frac{e^{-1/\rho_K}}{\rho_K}] \\ &= ProductLog[-e^{-1}] \\ &= -1. \end{aligned}\tag{4.67}$$

Then, by the assumption that $\rho_K \to 1$ as $K \to \infty$, we can determine the limit of the root

$$\lim_{K\to\infty}\Phi_K = \lim_{K\to\infty} -\rho_K \lim_{K\to\infty} ProductLog[-\frac{e^{-1/\rho_K}}{\rho_K}] = 1.$$

□

Example 4.6 *(Erlang-r distribution) Assume the inter-arrival time is Erlang-r with mean $1/\lambda$. Then*

$$\begin{aligned}\beta_n(K) &= \int_0^\infty e^{-K\mu t}\frac{(K\mu t)^n}{n!}\frac{\lambda^r t^{r-1}e^{-\lambda t}}{(r-1)!}dt \\ &= \frac{\lambda^{n+r}}{n!(r-1)!\rho_K^n}\int_0^\infty e^{-\lambda(1+1/\rho_K)t}t^{n+r-1}dt \\ &= \frac{\lambda\rho_K^{r-1}}{n!(r-1)!(1+\rho_K)^{n+r-1}}\int_0^\infty e^{-\lambda(1+1/\rho_K)t}(\lambda(1+\frac{1}{\rho_K})t)^{n+r-1}dt \\ &= \frac{\rho_K^r}{n!(r-1)!(1+\rho_K)^{n+r}}\int_0^\infty e^{-s}s^{n+r-1}ds \quad (let\ s = \lambda t(1+\tfrac{1}{\rho_K})) \\ &= \frac{\rho_K^r}{n!(r-1)!(1+\rho_K)^{n+r}}\Gamma(n+r) \\ &= \frac{\rho_K^r}{nB(n,r)(1+\rho_K)^{n+r}}, \end{aligned}\tag{4.68}$$

for $n = 0, 1, 2, \ldots,$ *where* $B(n, r) = \Gamma(n)\Gamma(r)/\Gamma(n + r)$ *is the Beta function. From (4.68), it is derived that*

$$\beta_0(K) = (\frac{\rho_K}{1 + \rho_K})^r \tag{4.69}$$

and

$$\beta_n(K) = \frac{n + r - 1}{n(1 + \rho_K)}\beta_{n-1}(K) = \frac{\rho_K^r}{n!(r-1)!(1 + \rho_K)^{n+r}}\Gamma(n + r) \tag{4.70}$$

for $n = 1, 2, \ldots.$

From Lemma 4.2 and $\Phi_K/(1 + \rho_K) < 1$*, the characteristic equation (4.32) becomes*

$$\Phi_K - \sum_{n=0}^{\infty} \frac{\rho_K^r}{n!(r-1)!(1 + \rho_K)^{n+r}}\Gamma(n + r)\Phi_K^n = 0$$

which implies

$$\Phi_K - \sum_{n=0}^{\infty} \frac{\rho_K^r \Phi_K^n}{nB(n, r)(1 + \rho_K)^{n+r}} = 0. \tag{4.71}$$

Recall Newton's binomial series from Mathematical Analysis [98], if $z < 1$*, it gives*

$$\sum_{n=0}^{\infty} C_n^{n+r-1} z^n = \frac{1}{(1 - z)^r}, \tag{4.72}$$

where $C_n^{n+r-1} = (n + r - 1)!/(n!(r - 1)!)$ *is the binomial coefficient. By the above Newton's binomial series, if we reconsider (4.69), (4.71) can be rewritten as*

$$\Phi_K - \beta_0(K) \sum_{n=0}^{\infty} C_n^{n+r-1} (\frac{\Phi_K}{1 + \rho_K})^n = 0,$$

which implies

$$\Phi_K(1 - \frac{\Phi_K}{1 + \rho_K})^r = \beta_0(K) \tag{4.73}$$

because $\Phi_K/(1 + \rho_K) < 1$*. Note that we shall determine the root* Φ_K *in the open interval* $(0, 1)$*, and it is trivial that* $\Phi_K/(1 + \rho_K) \neq 1$ *since* $\Phi_K < 1$*.*

In the case of $r = 1$*, (4.73) becomes*

$$\Phi_K^2 - (1 + \rho_K)\Phi_K + \rho_K = 0,$$

and we have the root

$$\Phi_K = \rho_K,$$

which is the same result given in the example of the Exponential inter-arrival time distribution.

In the case of $r = 2$, *(4.73) becomes*

$$\Phi_K^3 - 2(1+\rho_K)\Phi_K^2 + (1+\rho_K)^2\Phi_K - \rho_K^2 = 0,$$

and we have the root

$$\Phi_K = \frac{1 + 2\rho_K - \sqrt{1+4\rho_K}}{2},$$

which lies in the interval $(0, 1)$.

For any given $r \geq 3$ *and* $0 < \rho_K < 1$, *it is impossible to find the exact value of root* Φ_K *in (4.73). Hence, in the following, we determine the upper and lower bounds for the root* Φ_K. *Consider a real-value function* G *of* s *as follows:*

$$G(s) = s(1 - \frac{s}{1+\rho_K})^r - (\frac{\rho_K}{1+\rho_K})^r,$$

where $0 \leq s \leq 1$, *and the first derivative of* G *with respect to* s *is*

$$G'(s) = (1 - \frac{s}{1+\rho_K})^r - \frac{rs}{1+\rho_K}(1 - \frac{s}{1+\rho_K})^{r-1}.$$

Then, we have

$$G(0) = -(\frac{\rho_K}{1+\rho_K})^r < 0,$$

$$G(1) = (1 - \frac{1}{1+\rho_K})^r - (\frac{\rho_K}{1+\rho_K})^r = 0,$$

$$G'(0) = 1 > 0,$$

and

$$G'(1) = (\frac{\rho_K}{1+\rho_K})^{r-1}\frac{\rho_K - r}{1+\rho_K} < 0$$

given that $\rho_K < 1 \leq r$. *If* $\hat{s}$ *is a zero of* $G'(\hat{s}) = 0$ *and* $\hat{s}/(1+\rho_K) \neq 1$, *then it implies that*

$$(1 - \frac{\hat{s}}{1+\rho_K})^r = \frac{r\hat{s}}{1+\rho_K}(1 - \frac{\hat{s}}{1+\rho_K})^{r-1}$$

becomes

$$1 - \frac{\hat{s}}{1+\rho_K} = \frac{r\hat{s}}{1+\rho_K}$$

by the condition $\hat{s} < 1 < 1 + \rho_K$*, and we obtain*

$$\hat{s} = \frac{1+\rho_K}{1+r} < 1.$$

Therefore, we find that

$$\Phi_K < \frac{1+\rho_K}{1+r}.$$

Moreover, it shows

$$G(\rho_K) = \rho_K(\frac{1}{1+\rho_K})^r - (\frac{\rho_K}{1+\rho_K})^r = \frac{\rho_K - \rho_K^r}{(1+\rho_K)^r} > 0,$$

Hence, the range of the root Φ_K *is given as follows:*

$$0 < \Phi_K < \min\{\frac{1+\rho_K}{1+r}, \rho_K\}. \tag{4.74}$$

Corollary 4.5 *If* $A(t)$ *is the c.d.f. of Erlang-r inter-arrival times for* $r < \infty$ *and* $\rho_K < 1$ *for all* K*, then the limit of the probability* $\beta_n(K)$ *exists as* $K \to \infty$*, and*

$$\lim_{K\to\infty} \beta_n(K) = \frac{\Gamma(n+r)}{n!(r-1)!2^{n+r}}, \tag{4.75}$$

for $n = 0, 1, 2, \ldots$.

Proof. From (4.70), we have

$$\begin{aligned}
& \lim_{K\to\infty} \beta_n(K) \\
= & \lim_{K\to\infty} \frac{\rho_K^r}{n!(r-1)!(1+\rho_K)^{n+r}}\Gamma(n+r) \\
= & \frac{\Gamma(n+r)}{n!(r-1)!} \frac{\lim_{K\to\infty}\rho^r(K)}{\lim_{K\to\infty}(1+\rho_K)^{n+r}} \\
= & \frac{\Gamma(n+r)}{n!(r-1)!2^{n+r}}
\end{aligned} \tag{4.76}$$

with the help of assumption that traffic intensities ρ_K approaches to 1 from below as $K \to \infty$. □

Corollary 4.6 *If* $A(t)$ *is the c.d.f. of Erlang-r inter-arrival times,* $r < \infty$*, and* $\rho_K < 1$ *for all* K*, then the limit of the roots* Φ_K *exists as* $K \to \infty$*, and*

$$0 < \lim_{K\to\infty} \Phi_K < \frac{2}{1+r} \tag{4.77}$$

for $r = 1, 2, \ldots$.

Proof. By the assumption that $\rho_K \to 1$ as $K \to \infty$, it gives

$$\lim_{K\to\infty} \frac{1+\rho_K}{1+r} = \frac{2}{1+r}$$

for $r = 1, 2, \ldots$. Moreover, for all $r = 1, 2, \ldots$, it holds that

$$\min\{\frac{2}{1+r}, 1\} = \frac{2}{1+r}.$$

Hence, from (4.74), we find the range of the limit of the root in the case of the Erlang-r distributions. □

For each non-lattice $A(t)$ and fixed positive integer K, there exists a sufficiently large integer N_K such that $\beta_n(K)$ is non-increasing for all $n \geq N_K$.

Proposition 4.4 *Under the condition that the c.d.f. of inter-arrival times, $A(t)$, is non-lattice, we find that (i.) if $A(t)$ is the c.d.f. of exponential inter-arrival times, the probability $\beta_n(K)$ is non-increasing for all $n \geq 0$; (ii.) if the inter-arrival time is constant, there exists an integer $N_d = \lceil 1/\rho_K \rceil$ such that the probability $\beta_n(K)$ is non-increasing for all $n \geq N_d$, where the ceiling function $\lceil x \rceil$ outputs the smallest integer greater than or equal to x; (iii.) if $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, there exists an integer*

$$N_r = \lceil \frac{r-1}{\rho_K} \rceil$$

such that the probability $\beta_n(K)$ is non-increasing for all $n \geq N_r$.

Proof. (Case i.) In the case of exponential inter-arrival times, we have

$$\beta_n(K) = \int_0^\infty e^{-K\mu t} \frac{(K\mu t)^n}{n!} \lambda e^{-\lambda t} dt = \frac{\rho_K}{(1+\rho_K)^{n+1}}.$$

Because $\rho_K > 0$ for each positive integer K, it is clear that

$$\beta_n(K) = \frac{\rho_K}{(1+\rho_K)^{n+1}} \geq \frac{\rho_K}{(1+\rho_K)^{n+2}} = \beta_{n+1}(K)$$

for all $n = 0, 1, 2, \ldots$. So, $\beta_n(K)$ is non-increasing for all $n \geq 0$.

(Case ii.) If the inter-arrival time is constant, we have

$$\beta_n(K) = \frac{e^{-1/\rho_K}}{n!\rho_K^n}$$

for $n = 0, 1, 2, \ldots$. If $n \geq N_d = \lceil 1/\rho_K \rceil$, it implies $n\rho_K \geq 1$ and it can be derived that

$$\beta_{n-1}(K) = \frac{e^{-1/\rho_K}}{(n-1)!\rho_K^{n-1}} \geq \frac{e^{-1/\rho_K}}{n!\rho_K^n} = \beta_n(K).$$

Hence, $\beta_n(K)$ is non-increasing for all $n \geq N_d = \lceil 1/\rho_K \rceil$.

(Case iii.) In the case of Erlang-r inter-arrival times for a fixed positive integer r, we have

$$\beta_n(K) = \int_0^\infty e^{-K\mu t}\frac{(K\mu t)^n}{n!}\frac{\lambda^r t^{r-1}e^{-\lambda t}}{(r-1)!}dt = \frac{\rho_K^r}{nB(n,r)(1+\rho_K)^{n+r}},$$

where $B(n, r) = \Gamma(n)\Gamma(r)/\Gamma(n + r)$ is the Beta function and $\Gamma(n) = (n-1)!$ is the Gamma function [98]. It can be derived that

$$\beta_n(K) = \frac{n+r-1}{n(1+\rho_K)}\beta_{n-1}(K)$$

for $n = 1, 2, \ldots$. If $n \geq N_r = \lceil \frac{r-1}{\rho_K} \rceil$, it implies $\beta_n(K) \leq \beta_{n-1}(K)$. Therefore, $\beta_n(K)$ is non-increasing for all $n \geq N_r = \lceil \frac{r-1}{\rho_K} \rceil$. □

Proposition 4.5 *Under assumptions of Proposition 4.2, we have the limit in (4.49) as follows. (i.) If $A(t)$ is the c.d.f. of exponential inter-arrival times, it gives the limit*

$$C_1^{Exp} = 3. \tag{4.78}$$

(ii.) If the inter-arrival time is constant, it gives the limit

$$C_1^{Det} = \frac{e-1}{e-2}. \tag{4.79}$$

(iii.) If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, it gives the limit

$$C_1^{Erlang-r} = \frac{2^{r+1}-r}{2^{r+1}-2-r}, \tag{4.80}$$

for positive integer $r \geq 2$.

Proof. (Case i.) If $A(t)$ is the c.d.f. of exponential inter-arrival times, we have $\beta_0(K) = \rho_K/(1 + \rho_K)$ and $\beta_1(K) = \rho_K/(1 + \rho_K)^2$. Then we derive that

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} = \frac{1-\frac{\rho_K}{(1+\rho_K)^2}}{1-\frac{\rho_K}{1+\rho_K}-\frac{\rho_K}{(1+\rho_K)^2}} = \rho_K^2 + \rho_K + 1. \tag{4.81}$$

Because $\rho_K \to 1$ as $K \to \infty$, it gives the following limit

$$C_1^{Exp} = \lim_{K\to\infty} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} = \lim_{K\to\infty}(\rho_K^2+\rho_K+1) = 3.$$

(Case ii.) If the inter-arrival time is constant, we have $\beta_0(K) = e^{-1/\rho_K}$ and $\beta_1(K) = e^{-1/\rho_K}/\rho_K$. Then we derive that

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} = \frac{1-\frac{e^{-1/\rho_K}}{\rho_K}}{1-e^{-1/\rho_K}-\frac{e^{-1/\rho_K}}{\rho_K}} = \frac{\rho_K e^{1/\rho_K}-1}{\rho_K e^{1/\rho_K}-\rho_K-1}. \tag{4.82}$$

Hence, as $K \to \infty$, it gives the following limit

$$C_1^{Det} = \lim_{K\to\infty} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} = \lim_{K\to\infty} \frac{\rho_K e^{1/\rho_K}-1}{\rho_K e^{1/\rho_K}-\rho_K-1} = \frac{e-1}{e-2}.$$

(Case iii.) If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, for positive integer $r \geq 2$, it gives

$$\beta_0(K) = \left(\frac{\rho_K}{1+\rho_K}\right)^r$$

and

$$\beta_1(K) = \frac{\rho_K^r\Gamma(r)}{(r-1)!(1+\rho_K)^{r+1}} = \frac{r\rho_K^r}{(1+\rho_K)^{r+1}}.$$

Then we have

$$\begin{aligned} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} &= \frac{1-\frac{r\rho_K^r}{(1+\rho_K)^{r+1}}}{1-(\frac{\rho_K}{1+\rho_K})^r-\frac{r\rho_K^r}{(1+\rho_K)^{r+1}}} \\ &= \frac{(1+\rho_K)^{r+1}-r\rho_K^r}{(1+\rho_K)^{r+1}-\rho_K^r(1+\rho_K)-r\rho_K^r}. \end{aligned} \tag{4.83}$$

Hence, it gives the following limit

$$\begin{aligned} C_1^{Erlang-r} &= \lim_{K\to\infty} \frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)} \\ &= \frac{(1+\rho_K)^{r+1}-r\rho_K^r}{(1+\rho_K)^{r+1}-\rho_K^r(1+\rho_K)-r\rho_K^r} \\ &= \frac{2^{r+1}-r}{2^{r+1}-2-r}. \end{aligned} \tag{4.84}$$

□

Proposition 4.6 *Under assumptions of Proposition 4.3, we have the constant number in (4.50) as follows. (i.) If $A(t)$ is the c.d.f. of exponential inter-arrival times, it gives*

$$C_2^{Exp} \approx \left(\frac{\rho_K + \rho_K^2}{1 + \rho_K + \rho_K^2}\right)^K, \ K \gg 1. \tag{4.85}$$

(ii.) If the inter-arrival time is constant, it gives

$$C_2^{Det} \approx \left(\frac{\rho_K}{\rho_K e^{1/\rho_K} - 1}\right)^K, \ K \gg 1. \tag{4.86}$$

(iii.) If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, it gives

$$C_2^{Erlang-r} \approx \left(\frac{\rho_K^r(1 + \rho_K)}{(1 + \rho_K)^{r+1} - r\rho_K^r}\right)^K, \ K \gg 1. \tag{4.87}$$

Proof. (Case i.) If $A(t)$ is the c.d.f. of exponential inter-arrival times, we have $\beta_0(K) = \rho_K/(1 + \rho_K)$ and $\beta_1(K) = \rho_K/(1 + \rho_K)^2$. Then it can be derived that

$$(\frac{\beta_0(K)}{1 - \beta_1(K)})^K = \left(\frac{\rho_K + \rho_K^2}{1 + \rho_K + \rho_K^2}\right)^K. \tag{4.88}$$

Because $\rho_K \to 1$ as $K \gg 1$, it gives the following limit

$$C_2^{Exp} \approx (\frac{\beta_0(K)}{1 - \beta_1(K)})^K = \left(\frac{\rho_K + \rho_K^2}{1 + \rho_K + \rho_K^2}\right)^K.$$

(Case ii.) If the inter-arrival time is constant, it gives $\beta_0(K) = e^{-1/\rho_K}$ and $\beta_1(K) = e^{-1/\rho_K}/\rho_K$. Then we have

$$(\frac{\beta_0(K)}{1 - \beta_1(K)})^K = \left(\frac{\rho_K}{\rho_K e^{1/\rho_K} - 1}\right)^K. \tag{4.89}$$

It gives the limit

$$C_2^{Det} \approx (\frac{\beta_0(K)}{1 - \beta_1(K)})^K = \left(\frac{\rho_K}{\rho_K e^{1/\rho_K} - 1}\right)^K$$

as $K \gg 1$.

(Case iii.) If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, for positive integer $r > 1$, we have $\beta_0(K) = [\rho_K/(1+\rho_K)]^r$ and $\beta_1(K) = r\rho_K^r/(1+\rho_K)^{r+1}$. Then it gives

$$(\frac{\beta_0(K)}{1 - \beta_1(K)})^K = \left(\frac{\rho_K^r(1 + \rho_K)}{(1 + \rho_K)^{r+1} - r\rho_K^r}\right)^K. \tag{4.90}$$

As $K \gg 1$, it gives the following limit

$$C_2^{Erlang-r} \approx (\frac{\beta_0(K)}{1-\beta_1(K)})^K = \left(\frac{\rho_K^r(1+\rho_K)}{(1+\rho_K)^{r+1} - r\rho_K^r}\right)^K .$$

□

Proposition 4.7 *Consider three queueing systems with inter-arrival times of exponential, deterministic and Erlang-r, $r \geq 2$, distributions individually. Given the traffic intensity $\rho_K = 1 - \frac{\gamma}{\sqrt{K}} \to 1$ from below, for $0 < \gamma < \sqrt{K}$, as the number of virtual paths $K \gg 1$, we have*

$$\mathcal{P}^{Exp}(\rho_K, K) \geq \mathcal{P}^{Det}(\rho_K, K) \geq \mathcal{P}^{Erlang}(\rho_K, K). \tag{4.91}$$

Proof. By Theorem 4.1, the blocking probability can be determined as $\mathcal{P}(\rho_K, K) \approx P_0^\infty C_1 C_2$, where the probability P_0^∞ is constant given fixed K and ρ_K in $GI/M/\infty$ queues. From Proposition 4.5, we have shown that the sequence of

$$\frac{1-\beta_1(K)}{1-\beta_0(K)-\beta_1(K)}$$

converges to $C_1^{Exp} = 3$ as $K \gg 1$ for the exponential inter-arrival times. In addition, as $K \gg 1$, the sequence converges to $C_1^{Det} = (e-1)/(e-2)$ for the deterministic inter-arrival times, and it converges to $C_1^{Erlang-r} = (2^{r+1}-r)/(2^{r+1}-2-r)$ for the Erlang-r inter-arrival times. In can be easily checked that

$$3 \geq \frac{e-1}{e-2} \geq \frac{2^{r+1}-r}{2^{r+1}-2-r} \geq \frac{2^{r+2}-(r+1)}{2^{r+2}-2-(r+1)} \tag{4.92}$$

for all positive integers $r \geq 2$. Hence, we have the inequality $C_1^{Exp} \geq C_1^{Det} \geq C_1^{Erlang-r}$. Next, for comparison of limit C_2^{Exp}, C_2^{Det} and $C_2^{Erlang-r}$, we consider the term

$$\frac{\beta_0(K)}{1-\beta_1(K)}$$

in (4.50) for exponential, deterministic and Erlang-r, inter-arrival times individually. It can be derived that the term $\beta_0(K)/(1-\beta_1(K))$ for exponential distributions is the largest, and that is the least for Erlang-r distributions. Then, we have

$C_2^{Exp} \geq C_2^{Det} \geq C_2^{Erlang-r}$. Therefore, the inequality (4.91) holds by applying the approximation (4.51) in Theorem 4.1. $\square$

In this chapter, we have introduced an analytic approach for determining the blocking probability of connections on communication networks. As the number of virtual paths is large, the approximations is provided. We derive an asymptotic analysis of the blocking probability in terms of the roots of the characteristic equation and the stationary probabilities of the corresponding $GI/M/\infty$ queue. Blocking probabilities are estimated under assumptions of exponential, deterministic and Erlang-r distributions for the inter-arrival times, individually. For the class of problems studied, it is concluded that the approximation is adequate for practical purposes.

Chapter 5

Bandwidth Allocation of Two Revenue Management Schemes

In this chapter, we provide the solution analysis of those two revenue management schemes introduced in Chapter 2. The revenue functions (2.10) and (2.12) are studied through those monotone and convex properties of the blocking probability and expected path occupancy.

5.1 Monotone and Convex Properties of Two Revenue Functions

First, we study the relation between revenue function (2.10) and allocated bandwidth. As mentioned in [130], there exists no closed-form algebraic expression of the optimal solution in Revenue Management Scheme I. Yacoubi et al. [130] plotted the revenue function (2.10) only and solved it numerically. To investigate the objective function (2.10), we derive and prove the monotonicity of the revenue function $F_i(x_i, K_i, y_i)$ in (2.10) with respect to model parameters x_i and K_i, individually, in the following results.

Theorem 5.1 *Let $\sigma_i > 0$ and $b_i^{\min} \geq 0$ be the mean connection volume and the minimum bandwidth requirement of class i, respectively. Given costs $c_i^t > 0$ and $c_i^b > 0$ in (2.10), if the allocated bandwidth*

$$x_i \geq \max\{\sqrt{\frac{c_i^t \sigma_i}{c_i^b}},\ b_i^{\min}\} \tag{5.1}$$

holds for class $i \in \mathbb{M}$, the long-run average revenue $F_i(x_i, K_i, y_i)$ is increasing in bandwidth x_i, given $K_i \geq 1$ and $y_i > 0$.

Proof. From the minimum bandwidth requirement (2.1), it holds that feasible solution $x_i \geq b_i^{\min}$ for $i \in \mathbb{M}$. From the condition $x_i \geq \sqrt{c_i^t \sigma_i / c_i^b}$, it implies that

$$c_i^b \lambda_i - c_i^t \frac{y_i}{x_i^2} \geq 0,$$

where the traffic demand $y_i = \lambda_i \sigma_i$. From (3.5), we obtain that

$$\frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial x_i} = -\frac{y_i}{x_i^2}(1 - \mathcal{P}(x_i, K_i, y_i)) + \frac{y_i}{x_i}(-\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i}). \tag{5.2}$$

By Proposition 3.1, we know that $\partial \mathcal{P}(x_i, K_i, y_i)/\partial x_i < 0$. Hence, if $c_i^b \lambda_i \geq c_i^t y_i / x_i^2$, the first derivative of (2.10) with respect to x_i is

$$\begin{aligned} & \frac{\partial F_i(x_i, K_i, y_i)}{\partial x_i} \\ = \ & c_i^t \frac{\partial \mathcal{L}(x_i, K_i, y_i)}{\partial x_i} + c_i^b \lambda_i \left[(1 - \mathcal{P}(x_i, K_i, y_i)) - x_i \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i}\right] \\ = \ & (1 - \mathcal{P}(x_i, K_i, y_i)) \left[c_i^b \lambda_i - c_i^t \frac{y_i}{x_i^2}\right] - \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} \left[c_i^b \lambda_i x_i + c_i^t \frac{y_i}{x_i}\right] \\ \geq \ & 0, \end{aligned} \tag{5.3}$$

for all $K_i \geq 1$ and $y_i > 0$. So, the long-run average revenue $F_i(x_i, K_i, y_i)$ is an increasing function of bandwidth x_i given that the inequality (5.1) holds. □

Theorem 5.2 *The long-run average revenue $F_i(x_i, K_i, y_i)$ is an increasing function of the number of virtual paths K_i for each class $i \in \mathbb{M}$, given $x_i > b_i^{\min}$ and $y_i > 0$.*

Proof. From Proposition 3.4, we know that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is a decreasing function of the number of virtual paths K_i. Moreover, the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is an increasing function of K_i. Given $x_i > b_i^{\min}$ and $y_i > 0$, the increment of long-run average revenue (2.10) with respect to K_i is

$$\begin{aligned} & F_i(x_i, K_i+1, y_i) - F_i(x_i, K_i, y_i) \\ = \;& c_i^t \left[\mathcal{L}(x_i, K_i+1, y_i) - \mathcal{L}(x_i, K_i, y_i)\right] + c_i^b \lambda_i x_i \left[\mathcal{P}(x_i, K_i, y_i) - \mathcal{P}(x_i, K_i+1, y_i)\right] \\ \geq \;& 0, \end{aligned} \tag{5.4}$$

for all $K_i \geq 1$. So, the long-run average revenue $F_i(x_i, K_i, y_i)$ is increasing in K_i for class $i \in \mathbb{M}$. □

Theorem 5.1 implies that the objective function of Revenue Management Scheme I is increasing in bandwidth x_i. Theorem 5.2 shows that the long-run average revenue $F_i(x_i, K_i, y_i)$ in Revenue Management Scheme I is also an increasing function of the number of virtual paths K_i. Those structural results on the long-run average revenue function can be helpful in the problem of maximizing the long-run average reward in communication networks with dynamic pricing [59, 2, 133].

Next, in the following results, we prove the monotonicity and convexity of the profit function $G_i(x_i, K_i, y_i)$.

Theorem 5.3 *The profit function $G_i(x_i, K_i, y_i)$ is increasing in bandwidth x_i, given $K_i \geq 1$ and $y_i > 0$.*

Proof. By Proposition 3.1, we have proved that $\partial \mathcal{P}(x_i, K_i, y_i)/\partial x_i < 0$. It can be derived that the first partial derivative of (2.12) with respect to x_i is

$$\frac{\partial G_i(x_i, K_i, y_i)}{\partial x_i} = \frac{p_i K_i}{x_i \log(a_i/r_i)} - q_i \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} \geq 0, \tag{5.5}$$

for all $K_i \geq 1$ and $y_i > 0$. So, the profit function $G_i(x_i, K_i, y_i)$ is an increasing function of bandwidth x_i. □

Theorem 5.4 *The profit function $G_i(x_i, K_i, y_i)$ is an decreasing function of the traffic demand y_i for each class $i \in \mathbb{M}$, given $x_i > b_i^{\min}$ and $K_i \geq 1$.*

Proof. By Proposition 3.3, we know that $\partial\mathcal{P}(x_i, K_i, y_i)/\partial y_i > 0$. Then the first partial derivative of (2.12) with respect to y_i is

$$\frac{\partial G_i(x_i, K_i, y_i)}{\partial y_i} = -q_i \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial y_i} \leq 0 \tag{5.6}$$

for all $K_i \geq 1$ and $x_i > b_i^{\min}$. So, $G_i(x_i, K_i, y_i)$ is an decreasing function of traffic demand y_i. □

Theorem 5.5 *The profit function $G_i(x_i, K_i, y_i)$ is an increasing function of the number of virtual paths K_i for each class $i \in \mathbb{M}$, given $x_i > b_i^{\min}$ and $y_i > 0$.*

Proof. From Proposition 3.4, we know that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is a decreasing function of the number of virtual paths K_i. Given $x_i > b_i^{\min}$ and $y_i > 0$, the increment of the profit function (2.12) with respect to K_i is

$$\begin{aligned} & G_i(x_i, K_i + 1, y_i) - G_i(x_i, K_i, y_i) \\ = \ & p_i f_i(x_i) + q_i \left[\mathcal{P}(x_i, K_i, y_i) - \mathcal{P}(x_i, K_i + 1, y_i)\right] \geq 0 \end{aligned} \tag{5.7}$$

for all $K_i \geq 1$. So, the profit function $G_i(x_i, K_i, y_i)$ is increasing in K_i for class $i \in \mathbb{M}$. □

Theorem 5.6 *For each $K_i \geq 1$ and $y_i > 0$, there exists a region $\mathbb{S}_i$ of positive real numbers such that the profit function $G_i(x_i, K_i, y_i)$ is concave in bandwidth x_i for all $x_i \in \mathbb{S}_i$.*

Proof. From Proposition 3.2, there exists a region $\mathbb{S}_i$ of positive real numbers such that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex in bandwidth x_i for all $x_i \in \mathbb{S}_i$. That is, for all $x_i \in \mathbb{S}_i$, we have $\partial^2\mathcal{P}(x_i, K_i, y_i)/\partial x^2 \geq 0$. Then the second derivative of $G_i(x_i, K_i, y_i)$ with respect to x_i is

$$\frac{\partial^2 G_i(x_i, K_i, y_i)}{\partial x_i^2} = -\frac{p_i K_i}{x_i^2 \log(a_i/r_i)} - q_i \frac{\partial^2 \mathcal{P}(x_i, K_i, y_i)}{\partial x_i^2} \leq 0 \tag{5.8}$$

for all $K_i \geq 1$ and $y_i > 0$. So, the profit function $G_i(x_i, K_i, y_i)$ is concave in bandwidth x_i for all $x_i \in \mathbb{S}_i$. □

Remark 5.1 *If network managers allocate bandwidth in a specific region $\mathbb{S}_i$ such that $\partial^2\mathcal{P}(x_i, K_i, y_i)/\partial x_i^2 \geq 0$, the profit function $G_i(x_i, K_i, y_i)$ is concave in bandwidth x_i, which implies that the marginal profit is decreasing in such region $\mathbb{S}_i$. Otherwise, network managers could gain insight into the convexity of $G_i(x_i, K_i, y_i)$ by determining the opportunity cost q_i if bandwidth is allocated such that $\partial^2\mathcal{P}(x_i, K_i, y_i)/\partial x_i^2 < 0$. That is, in the case of $\partial^2\mathcal{P}(x_i, K_i, y_i)/\partial x_i^2 < 0$, $G_i(x_i, K_i, y_i)$ becomes convex in bandwidth x_i if the opportunity cost q_i is sufficiently large, i.e.,*

$$q_i > \frac{p_i K_i}{x_i^2 \log(a_i/r_i)(-\partial^2\mathcal{P}(x_i, K_i, y_i)/\partial x_i^2)};$$

otherwise, $G_i(x_i, K_i, y_i)$ is still concave in bandwidth x_i.

5.2 Solution Analysis and a Solution Algorithm

In this section, we present the optimality conditions for Revenue Management Scheme I and Revenue Management Scheme II, respectively. By introducing the Lagrangian multiplier and Lagrangian function [85, 94], the optimality conditions for these two revenue management schemes can be derived. First, for Revenue Management Scheme I, we denote the Lagrangian multiplier by v_1, and let $\psi_1(\mathbf{x}, v_1)$ be the Lagrangian function, where $\mathbf{x} = (x_1, \ldots, x_m)$. The Lagrangian function $\psi_1(\mathbf{x}, v_1)$ is formulated as

$$\psi_1(\mathbf{x}, v_1) = \sum_{i\in\mathbb{M}} w_i F_i(x_i, K_i, y_i) + v_1(\sum_{i\in\mathbb{M}} K_i c_i x_i - B). \tag{5.9}$$

Then, we can obtain

$$\frac{\partial\psi_1(\mathbf{x}, v_1)}{\partial x_i} = w_i \frac{\partial F_i(x_i, K_i, y_i)}{\partial x_i} + v_1 K_i c_i \tag{5.10}$$

for all $i \in \mathbb{M}$, where $\partial F_i(x_i, K_i, y_i)/\partial x_i$ is determined in (5.3), and

$$\frac{\partial\psi_1(\mathbf{x}, v_1)}{\partial v_1} = \sum_{i\in\mathbb{M}} K_i c_i x_i - B. \tag{5.11}$$

From $\partial\psi_1(\mathbf{x}, v_1)/\partial x_i = 0$ and $\partial\psi_1(\mathbf{x}, v_1)/\partial v_1 = 0$, we obtain the following Proposition 5.1. We can determine the optimal solution of Revenue Management Scheme I by solving the system of equations (5.12) and (5.13) in Proposition 5.1.

Proposition 5.1 *In solving the optimal solutions for Revenue Management Scheme I, the optimal bandwidth allocation* $\boldsymbol{x} = (x_1, \ldots, x_m)$ *and Lagrangian multiplier* v_1 *satisfy*

$$v_1 = -\frac{w_i}{K_i c_i} \frac{\partial F_i(x_i, K_i, y_i)}{\partial x_i}, \ \forall i \in \mathbb{M}, \tag{5.12}$$

and

$$x_i = \frac{B - \sum_{j \neq i \in \mathbb{M}} K_j c_j x_j}{K_i c_i}, \ \forall i \in \mathbb{M}, \tag{5.13}$$

where $\partial F_i(x_i, K_i, y_i)/\partial x_i$ *is determined in (5.3).*

Proof. The Lagrangian relaxation formulation for Revenue Management Scheme I is $\psi_1(\mathbf{x}, v_1)$ given in (5.9). From $\partial \psi_1(\mathbf{x}, v_1)/\partial x_i = 0$, we have

$$w_i \frac{\partial F_i(x_i, K_i, y_i)}{\partial x_i} + v_1 K_i c_i = 0,$$

and it implies (5.12). From $\partial \psi_1(\mathbf{x}, v_1)/\partial v_1 = 0$, we obtain $\sum_{i \in \mathbb{M}} K_i c_i x_i = B$, which implies (5.13). □

Next, we introduce the Lagrangian multiplier v_2 and Lagrangian function $\psi_2(\mathbf{x}, v_2)$ for Revenue Management Scheme II. The Lagrangian function $\psi_2(\mathbf{x}, v_2)$ is defined as

$$\begin{aligned} & \psi_2(\mathbf{x}, v_2) \\ = & \sum_{i \in \mathbb{M}} w_i G_i(x_i, K_i, y_i) + v_2(\sum_{i \in \mathbb{M}} K_i c_i x_i - B) \\ = & \sum_{i \in \mathbb{M}} w_i p_i K_i f_i(x_i) - \sum_{i \in \mathbb{M}} w_i q_i \mathcal{P}_i(x_i, K_i, y_i) + v_2(\sum_{i \in \mathbb{M}} K_i c_i x_i - B). \end{aligned} \tag{5.14}$$

Then, we can obtain

$$\frac{\partial \psi_2(\mathbf{x}, v_2)}{\partial x_i} = \frac{w_i p_i K_i}{x_i \log(a_i/r_i)} - w_i q_i \frac{\partial \mathcal{P}_i(x_i, K_i, y_i)}{\partial x_i} + v_2 K_i c_i \tag{5.15}$$

for all $i \in \mathbb{M}$, and

$$\frac{\partial \psi_2(\mathbf{x}, v_2)}{\partial v_2} = \sum_{i \in \mathbb{M}} K_i c_i x_i - B. \tag{5.16}$$

From $\partial \psi_2(\mathbf{x}, v_2)/\partial x_i = 0$ and $\partial \psi_2(\mathbf{x}, v_2)/\partial v_2 = 0$, we can determine the optimal solution of Revenue Management Scheme II in the following Proposition 5.2.

Proposition 5.2 *In solving the optimal solutions for Revenue Management Scheme II, the optimal bandwidth allocation $\boldsymbol{x} = (x_1, \ldots, x_m)$ and Lagrangian multiplier v_2 satisfy*

$$v_2 = -\frac{w_i}{K_i c_i}\left[\frac{p_i K_i}{x_i \log(a_i/r_i)} - q_i(\frac{y_i}{x_i} - K_i)\frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i+1}}\right], \ \forall i \in \mathbb{M}, \tag{5.17}$$

and

$$x_i = \frac{B - \sum_{j \neq i \in \mathbb{M}} K_j c_j x_j}{K_i c_i}, \ \forall i \in \mathbb{M}. \tag{5.18}$$

Proof. The Lagrangian relaxation formulation for Revenue Management Scheme II is $\psi_2(\mathbf{x}, v_2)$ given in (5.14). From $\partial \psi_2(\mathbf{x}, v_2)/\partial x_i = 0$ in (5.15), we have

$$\frac{w_i p_i K_i}{x_i \log(a_i/r_i)} - w_i q_i \frac{\partial \mathcal{P}_i(x_i, K_i, y_i)}{\partial x_i} + v_2 K_i c_i = 0,$$

and it implies (5.17). From $\partial \psi_2(\mathbf{x}, v_2)/\partial v_2 = 0$ in (5.16), we obtain (5.18). □

In practice, it is complicated to solve the system of equations (5.17) and (5.18) because the blocking probability $\mathcal{P}_i(x_i, K_i, y_i)$ in (3.4) is a nonlinear function of bandwidth x_i. If we omit temporarily the consideration of blocking probability $\mathcal{P}_i(x_i, K_i, y_i)$ in the profit function $G_i(x_i, K_i, y_i)$ defined in (2.12), the objective function of Revenue Management Scheme II will be reduced to the weighted sum of utility functions only. Then, Revenue Management Scheme II can be reduced to a utility maximization problem as studied in [44]. It implies that the objective function of Revenue Management Scheme II can be rewritten according to [44] as follows:

$$\sum_{i \in \mathbb{M}} w_i p_i K_i f_i(x_i) = \sum_{i \in \mathbb{M}} K_i \left(w_i p_i \log_{\frac{a_i}{r_i}} \frac{x_i}{r_i} \right) = \sum_{i \in \mathbb{M}} K_i \left(\mathcal{D}_i \log x_i - \mathcal{C}_i \right),$$

where $\mathcal{D}_i = w_i p_i / \log(a_i/r_i)$ and $\mathcal{C}_i = w_i p_i \log r_i / \log(a_i/r_i)$ are constant for all $i \in \mathbb{M}$. Hence, the utility maximization problem under budget constraint can be formulated as follows:

$$\begin{aligned} \max \quad & \sum_{i \in \mathbb{M}} K_i \left(\mathcal{D}_i \log x_i - \mathcal{C}_i \right) \\ \text{s.t.} \quad & \text{budget constraint (2.2).} \end{aligned} \tag{5.19}$$

Then we may relax the result in Proposition 5.2 to the following Corollary 5.1, which is an optimal solution to the utility maximization problem (5.19).

Corollary 5.1 *In the case of omitting the blocking probability from Revenue Management Scheme II, the optimal bandwidth allocation* $(\hat{x}_1, \ldots, \hat{x}_m)$ *and Lagrangian multiplier* $\hat{v}_2$ *for the utility maximization problem (5.19) satisfy*

$$\hat{v}_2 = -\frac{\sum_{i \in \mathbb{M}} K_i \mathcal{D}_i}{B}, \tag{5.20}$$

and

$$\hat{x}_i = \frac{B\mathcal{D}_i}{c_i \sum_{i \in \mathbb{M}} K_i \mathcal{D}_i}, \ \forall i \in \mathbb{M}, \tag{5.21}$$

where $\mathcal{D}_i = w_i p_i / \log(a_i/r_i)$ *is constant.*

Proof. In the case of omitting the blocking probability from Revenue Management Scheme II, we let the term $q_i \mathcal{P}_i(x_i, K_i, y_i)$ equal zero in (5.14). From $\partial \psi_2(\mathbf{x}, v_2)/\partial x_i = 0$, we can determine

$$\hat{x}_i = -\frac{\mathcal{D}_i}{\hat{v}_2 c_i}, \ \forall i \in \mathbb{M}, \tag{5.22}$$

where $\mathcal{D}_i = w_i p_i / \log(a_i/r_i)$. Then we take $\hat{x}_i$ to substitute for x_i in (5.18). Solving the equation (5.18), we can obtain the Lagrangian multiplier $\hat{v}_2$ in (5.20). Thus, by combining equations (5.20) and (5.22), we have the optimal bandwidth $\hat{x}_i$ in (5.21). □

Note that Corollary 5.1 is the same as the result of Proposition 3 studied in Guan et al. [44]. The feasible set of the utility maximization problem in Guan et al. [44] is the same as that of Revenue Management Scheme II. Observe that the utility function $f_i(x_i)$ is increasing in bandwidth x_i. Moreover, in Proposition 3.1, we have shown that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is a decreasing function of bandwidth x_i. It implies that the objective function of Revenue Management Scheme II is increasing in bandwidth x_i. Hence, we can determine the optimal solution of Revenue Management Scheme II efficiently from the optimal solution of the utility maximization problem (5.19).

According to Corollary 5.1, we present an optimal solution $\hat{x}_i$ to the utility maximization problem (5.19) given parameters K_i, c_i and p_i corresponding to each class i, $\forall i \in \mathbb{M}$. Let $\hat{\mathbf{x}}$ and $\mathbf{x}$ denote vectors with their component $\hat{x}_i$ and x_i respectively.

For the Lagrangian function $\psi_2(\mathbf{x}, v_2)$ defined in (5.14), we denote its gradient vector by $\nabla\psi_2(\mathbf{x}, v_2)$ and Hessian matrix by $\nabla^2\psi_2(\mathbf{x}, v_2)$ at a point $(\mathbf{x}, v_2) \in \mathbb{R}^{m+1}$. By using (5.15) and (5.16), we can determine

$$\frac{\partial^2\psi_2(\mathbf{x}, v_2)}{\partial x_i^2} = \frac{-K_i\mathcal{D}_i}{x_i^2} - w_i q_i \frac{\partial^2\mathcal{P}_i(x_i, K_i, y_i)}{\partial x_i^2}, \tag{5.23}$$

$$\frac{\partial^2\psi_2(\mathbf{x}, v_2)}{\partial x_i \partial v_2} = \frac{\partial^2\psi_2(\mathbf{x}, v_2)}{\partial v_2 \partial x_i} = K_i c_i, \tag{5.24}$$

and

$$\frac{\partial^2\psi_2(\mathbf{x}, v_2)}{\partial v_2^2} = 0. \tag{5.25}$$

From the above equations, we can compute the gradient vector $\nabla\psi_2(\mathbf{x}, v_2)$ and Hessian matrix $\nabla^2\psi_2(\mathbf{x}, v_2)$. Therefore, for a given $\hat{\mathbf{x}}$ obtained from Corollary 5.1, we can determine an optimal objective value and the corresponding optimal bandwidth $\mathbf{x} = (x_1, \ldots, x_m)$ as shown in (5.17) and (5.18) by using the following algorithm.

A solution algorithm: *(Newton's method for optimal bandwidths of Management Scheme II.)*

Initialization Set iteration $k = 0$. Apply Corollary 5.1, we set the initial solution $x_i^0 = B\mathcal{D}_i/(c_i \sum_{i\in\mathbb{M}} K_i\mathcal{D}_i)$, $\forall i \in \mathbb{M}$ and $v_2^0 = -\sum_{i\in\mathbb{M}} K_i\mathcal{D}_i/B$. Set $\mathbf{x}^0 = (x_1^0, \ldots, x_m^0)$, and choose a sufficiently small number $\xi > 0$.

Step 1 Compute the gradient vector $\nabla\psi_2(\mathbf{x}^k, v_2^k)$ at point $(\mathbf{x}^k, v_2^k)$.

Step 2 If $\nabla\psi_2(\mathbf{x}^k, v_2^k) = 0$ or

$$\|\psi_2(\mathbf{x}^k, v_2^k) - \psi_2(\mathbf{x}^{k-1}, v_2^{k-1})\| < \xi,$$

then stop with an optimal (or near optimal) solution $\mathbf{x}^k$ and v_2^k. Otherwise, go to the next step.

Step 3 Compute the Hessian matrix $\nabla^2\psi_2(\mathbf{x}^k, v_2^k)$ and its inverse

$$\mathbf{H}^k = (\nabla^2\psi_2(\mathbf{x}^k, v_2^k))^{-1}.$$

Step 4 Perform the translation

$$(\mathbf{x}^{k+1}, v_2^{k+1}) \leftarrow (\mathbf{x}^k, v_2^k) - \mathbf{H}^k \nabla\psi_2(\mathbf{x}^k, v_2^k).$$

Set $k \leftarrow k+1$ and go to Step 1.

This algorithm defines a series of $(\mathbf{x}^k, v_2^k)$'s, starting from an initial solution $(\hat{\mathbf{x}}, \hat{v}_2)$, such that the sequence converges toward an optimal solution which satisfies $\nabla\psi_2 = 0$. The initial solution $(\hat{\mathbf{x}}, \hat{v}_2)$ obtained from Corollary 5.1 is a feasible solution to Revenue Management Scheme II because it satisfies the budget constraint (2.2). In Step 1, the gradient vector $\nabla\psi_2(\mathbf{x}^k, v_2^k)$ can be easily calculated through equations (5.15) and (5.16) given a point $(\mathbf{x}^k, v_2^k)$. From equations (5.23)-(5.25), we can construct the symmetric Hessian matrix $\nabla^2\psi_2(\mathbf{x}^k, v_2^k)$ and determine its inverse $\mathbf{H}^k$ in Step 3, where $\nabla^2\psi_2(\mathbf{x}^k, v_2^k)$ and $\mathbf{H}^k$ are $(m+1)$-by-$(m+1)$ matrices. Note that the computational time needed to find $\nabla^2\psi_2(\mathbf{x}^k, v_2^k)$ and $\mathbf{H}^k$ is short because the total number of traffic classes, m, is small in the core network. The descent direction is given by $-\mathbf{H}^k\nabla\psi_2(\mathbf{x}^k, v_2^k)$ at the point $(\mathbf{x}^k, v_2^k)$. With the good initial point $(\hat{\mathbf{x}}, \hat{v}_2)$, the sequence $(\mathbf{x}^k, v_2^k)$ defined in Step 4 converges to the root of $\nabla\psi_2(\mathbf{x}, v_2) = 0$. The stopping criterion in Step 2 implies that we terminate the algorithm with an optimal solution which satisfies the optimal condition in Proposition 5.2. Or, we terminate the algorithm early when the improvement of $\psi_2(\mathbf{x}^k, v_2^k)$ in (5.14) is less than a sufficiently small number ξ. The convergence of the algorithm is quadratic, and it converges quickly to the (near) optimal solution because the initial point $(\hat{\mathbf{x}}, \hat{v}_2)$ is easily directed to the root of $\nabla\psi_2 = 0$ since the monotone property is applied to the case.

5.3 Three Elasticities of Blocking

In this section, we present connection's blocking elasticity to bandwidth for each traffic class to illustrate how the expressions of $\mathcal{P}(x_i, K_i, y_i)$ can be used with some applications. The term elasticity was introduced to the networking research community by Shenker [100]. Based on the investigation of elasticity, one can develop distributed pricing algorithm that takes user's elasticity into consideration [79, 99, 6, 132]. By using the concept of elasticity, we can define the elasticity of blocking probability with respect to bandwidth as follows.

Definition 5.1 *The* ***bandwidth elasticity of blocking*** *is defined as*

$$\mathcal{E}_i^b \triangleq \frac{\triangle \mathcal{P}(x_i, K_i, y_i)/\mathcal{P}(x_i, K_i, y_i)}{\triangle x_i / x_i}, \tag{5.26}$$

where x_i is the allocated bandwidth, $\triangle x_i$ is the change of allocated bandwidth, $\mathcal{P}(x_i, K_i, y_i)$ is the blocking probability given predetermined number of virtual paths K_i, and $\triangle \mathcal{P}(x_i, K_i, y_i)$ is the change of the blocking probability.

The bandwidth elasticity of blocking (5.26) can be rewritten as

$$\mathcal{E}_i^b = \frac{x_i}{\mathcal{P}(x_i, K_i, y_i)} \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i}. \tag{5.27}$$

The elasticity $\mathcal{E}_i^b$ represents the percentage change of the blocking probability in response to a percent change of bandwidth. Proposition 5.3 shows the phenomenon that the blocking probability will decrease if the allocated bandwidth increases.

Proposition 5.3 *The bandwidth elasticity of blocking $\mathcal{E}_i^b$ is negative for all $x_i > 0$.*

Proof. By Proposition 3.1, we obtain

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} < 0.$$

Hence, the result can be directly derived through (5.27). □

As $K_i \gg 1$, Proposition 5.4 shows that the bandwidth elasticity of blocking $\mathcal{E}_i^b$ will decrease when the allocated bandwidth x_i increases.

Proposition 5.4 *The bandwidth elasticity of blocking $\mathcal{E}_i^b$ is decreasing in bandwidth x_i for all $x_i > 0$ as $K_i \gg 1$.*

Proof. We use (3.20) as the blocking probability $\mathcal{P}(x_i, K_i, y_i)$, and its first derivative with respect to x_i is

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial x_i} = (\frac{y_i}{x_i} - K_i)\frac{y_i^{K_i} e^{-y_i/x_i}}{K_i! x_i^{K_i+1}}$$

as the number of end-ro-end paths $K_i \gg 0$. Thus, the bandwidth elasticity of blocking can be approximated as

$$\mathcal{E}_i^b = \frac{y_i}{x_i} - K_i \tag{5.28}$$

with the help of (5.27) and (3.20). From equation (5.28), the first derivative of $\mathcal{E}_i^b$ with respect to x_i is

$$\frac{\partial \mathcal{E}_i^b}{\partial x_i} = -\frac{y_i}{x_i^2} < 0$$

for all $x_i > 0$ and $y_i > 0$. Hence, the bandwidth elasticity of blocking $\mathcal{E}_i^b$ is decreasing when increasing bandwidth x_i. □

As applications of elasticity in economics, we present the demand elasticity of blocking $\mathcal{E}_i^d$ and the capacity elasticity of blocking $\mathcal{E}_i^c$ for each traffic class $i \in M$ in the following. The **demand elasticity of blocking** is formulated as

$$\mathcal{E}_i^d \triangleq \frac{\triangle \mathcal{P}(x_i, K_i, y_i)/\mathcal{P}(x_i, K_i, y_i)}{\triangle y_i / y_i} = \frac{y_i}{\mathcal{P}(x_i, K_i, y_i)} \frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial y_i}, \tag{5.29}$$

where $\triangle y_i$ is the change in the traffic demand. In addition, the **capacity elasticity of blocking** is written as

$$\mathcal{E}_i^c \triangleq \frac{\triangle \mathcal{P}(x_i, K_i, y_i)/\mathcal{P}(x_i, K_i, y_i)}{\triangle K_i / K_i} = \frac{K_i}{\mathcal{P}(x_i, K_i, y_i)} \frac{\triangle \mathcal{P}(x_i, K_i, y_i)}{\triangle K_i}, \tag{5.30}$$

where $\triangle K_i$ is the change in the number of virtual paths. Similarly, the properties of the demand elasticity of blocking $\mathcal{E}_i^d$ and the capacity elasticity of blocking $\mathcal{E}_i^c$ are derived as follows.

Proposition 5.5 *The demand elasticity of blocking $\mathcal{E}_i^d$ is non-negative for all traffic demand $y_i \geq 0$.*

Proof. In Proposition 3.3, we have proved that

$$\frac{\partial \mathcal{P}(x_i, K_i, y_i)}{\partial y_i} > 0.$$

By the definition of the demand elasticity of blocking $\mathcal{E}_i^d$, it is obvious that $\mathcal{E}_i^d$ is non-negative as the traffic demand $y_i \geq 0$. $\square$

Proposition 5.6 *The capacity elasticity of blocking $\mathcal{E}_i^c$ is non-positive for all $K_i \geq 1$ and decreasing as the number of virtual paths K_i increases.*

Proof. In Proposition 3.4, we have showed that

$$\frac{\triangle \mathcal{P}(x_i, K_i, y_i)}{\triangle K_i} < 0.$$

Moreover, the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex in K_i for fixed $x_i > 0$ and $y_i > 0$. By the definition of the capacity elasticity of blocking $\mathcal{E}_i^c$, it implies that $\mathcal{E}_i^c$ is non-positive and decreasing as the number of virtual paths K_i increases. $\square$

Proposition 5.5 infers that the blocking probability will increase as the traffic demand increases. Proposition 5.6 concludes that the blocking probability is decreasing as increasing the number of virtual paths. Those properties can be easily derived from the results of Proposition 3.3 and Proposition 3.4.

Chapter 6

Numerical Results

6.1 Sensitivity Analysis of Two Revenue Management Schemes under Budget Control

In this section, we present numerical results to show the optimal bandwidth allocation of Revenue Management Schemes I and II for different traffic classes under the budget constraint. Here, we select four traffic classes as test examples from statistical data monitored at the Cooperative Association for Internet Data Analysis (CAIDA) [106]. There are four different traffic classes according to their QoS requirements for different applications, as proposed by UMTS [108]. Connections of class 4 have the highest priority, and traffic class 1 is given the lowest priority. The number of connections in the high-priority traffic class is often less than that in the low-priority class, but the traffic demand and bandwidth requirement of high-priority traffic class are always larger than those of low-priority traffic class. Those parameters are summarized in Table 6.1, including class weight w_i, minimum bandwidth requirement $b_i^{\min}$ (Mbps), aspiration level a_i (Mbps), reservation level r_i (Mbps), the average cost c_i (cents) of one unit bandwidth through class i's virtual paths, number of virtual paths K_i, mean occurrence rate λ_i, the connection volume σ_i (Mb), cost charged for using per unit of bandwidth c_i^b (cents), cost per unit of sojourn time c_i^t (cents), payoff p_i (cents) and opportunity cost q_i (cents).

Table 6.1: Characteristics of four traffic classes.

i	w_i	$b_i^{\min}$	a_i	r_i	y_i	c_i	K_i	λ_i	σ_i	c_i^t	c_i^b	p_i	q_i
1	0.1	0.8	1.32	0.84	59.1	1.0	96	96	0.62	4	130	130	4
2	0.2	1.0	1.85	1.18	96.5	1.3	75	75	1.29	6	180	180	6
3	0.3	2.0	2.79	1.73	185.4	1.8	42	42	4.41	12	320	320	12
4	0.4	2.5	3.58	2.28	324.3	2.5	27	27	12.01	20	500	500	20

The allocated bandwidth is determined by solving Revenue Management Scheme I and Revenue Management Scheme II, respectively. Table 6.2 shows those optimal bandwidth allocation and optimal values of Management Scheme I and Management II, where the total budget B varies from \$500 to \$1,000. Numerical results are obtained by programming the solution algorithm with Matlab [107]. For each available budget B, according to (5.21), the initial solutions $x_i^0 = B\mathcal{D}_i/(c_i \sum_{i=1}^{4} K_i \mathcal{D}_i)$ for all $i = 1, \ldots, 4$, are given in the optimization process. It can be observed from Table 6.2 that those bandwidth allocations and optimal values are increasing when enlarging the available budget. From Fig. 6.1 to Fig. 6.3, we graphically illustrate the effect of changing budget B on optimal bandwidth allocation and optimal values of those two schemes. In Fig. 6.1, it shows that almost all the available resources are allocated in the direction of class 3 in Management Scheme I, and the others are allocated to satisfy the minimum bandwidth requirements only. However, in Management Scheme II, it can be seen in Fig. 6.2 that the available resource is allocated to all classes proportionally. Both optimal values of Management Scheme I and Management Scheme II are increasing in the total budget B. In Fig. 6.3, we find that there exists an inflection point such that the optimal revenues of Management Scheme I are concave up when the budget B is smaller than the inflection point, and the optimal revenue is concave down if the budget exceeds the inflection point. However, the optimal profit of Management Scheme II is expressed in logarithmic form. The marginal (optimal) profit obtained by solving Management Scheme II is decreasing with respect to the available budget B.

Table 6.2: Budget versus optimal bandwidth allocation x_i.

	Scheme I		Scheme II	
Budget B	Optimal bandwidth allocation (x_1, x_2, x_3, x_4)	Revenue F	Optimal bandwidth allocation (x_1, x_2, x_3, x_4)	Profit G
500	(0.80, 1.00, 2.08, 2.50)	10,141	(0.80, 1.00, 2.00, 2.59)	1,412
550	(0.80, 1.00, 2.74, 2.50)	12,898	(0.80, 1.08, 2.08, 3.11)	4,480
600	(0.80, 1.00, 3.40, 2.50)	16,254	(0.80, 1.20, 2.30, 3.44)	7,187
650	(0.80, 1.00, 4.06, 2.50)	19,997	(0.80, 1.31, 2.52, 3.77)	9,646
700	(0.80, 1.00, 4.72, 2.50)	23,787	(0.80, 1.42, 2.74, 4.10)	11,900
750	(0.80, 1.00, 5.38, 2.50)	27,284	(0.80, 1.54, 2.97, 4.43)	13,980
800	(0.80, 1.00, 6.04, 2.50)	30,379	(0.80, 1.65, 3.19, 4.76)	15,911
850	(0.80, 1.00, 6.71, 2.50)	33,203	(0.82, 1.76, 3.40, 5.07)	17,714
900	(0.80, 1.00, 7.37, 2.50)	35,911	(0.87, 1.87, 3.60, 5.37)	19,412
950	(0.80, 1.00, 8.03, 2.50)	38,584	(0.92, 1.97, 3.80, 5.67)	21,018
1,000	(0.80, 1.00, 8.69, 2.50)	41,247	(0.97, 2.07, 4.00, 5.97)	22,542

Next, we compare bandwidth sharing policies between Scheme I and Scheme II by showing the blocking probability and budget ratio for four traffic classes. Table 6.3 presents those blocking probabilities determined by those optimal bandwidth allocation of two management schemes. For Management Scheme I, it can be observed from Fig. 6.4 that the blocking probability of class 3 is decreasing when the budget B increases from \$500 to \$1,000 while other classes' blocking probabilities remain unimproved. However, in Management Scheme II, Fig. 6.5 shows that those blocking probabilities of four traffic classes are decreasing proportionally when increasing the total budget. The phenomena of changes on system utilizations can be observed from Fig. 6.6 to Fig. 6.11 for two revenue management schemes under budget control. It shows those utilization levels are decreasing when increasing the total budget. System utilizations are decreasing proportionally in Management Scheme II

Given a budget B, the budget ratio for each traffic class is determined from the

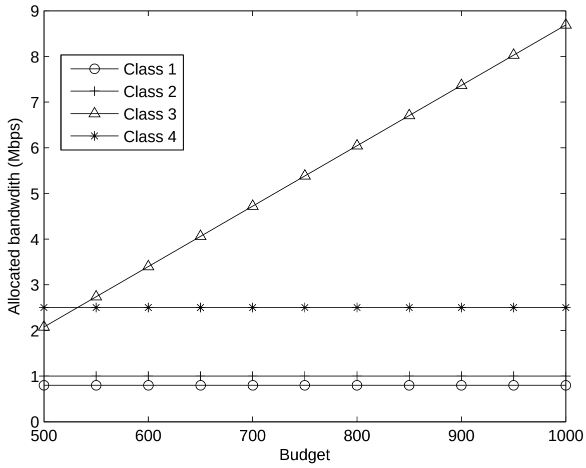

Figure 6.1: Budget versus optimal bandwidth allocation of Revenue Management Scheme I.

definition in (2.14) as follows: $B_i = K_i c_i x_i / \sum_{i=1}^{4} K_i c_i x_i$, for $i = 1, \ldots, 4$, where bandwidth x_i is determined by solving Management Scheme I and Management Scheme II, respectively. Table 6.4 summarizes those numerical results of the budget ratio for four traffic classes when increasing the available budget B from \$500 to \$1,000. It can be seen from Fig. 6.12 that most of the available budget is allocated to class 3, and the other classes only get the minimum bandwidth to satisfy the feasibility. This is because the marginal improvement of the objective function in Management Scheme I is the largest in the direction of class 3, i.e., $\partial w_3 F_3(x_3, K_3, y_3)/\partial x_3 \geq \partial w_i F_i(x_i, K_i, y_i)/\partial x_i$ for $i = 1, 2, 4$. Fig. 6.13 shows that all the budget is allocated to four traffic classes proportionally. The budget ratios for four traffic classes are almost invariable when varying budget B from \$500 to \$1,000.

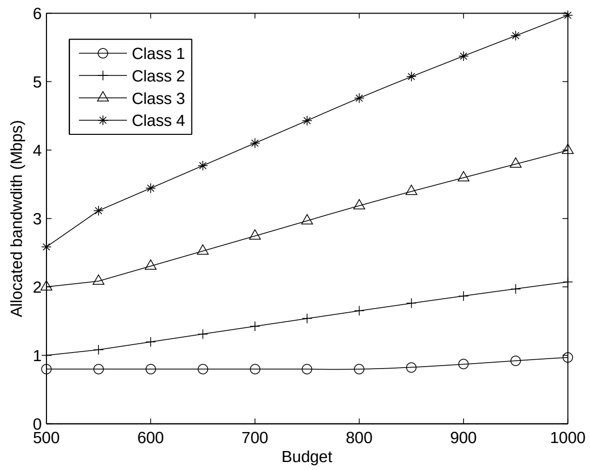

Figure 6.2: Budget versus optimal bandwidth allocation of Revenue Management Scheme II.

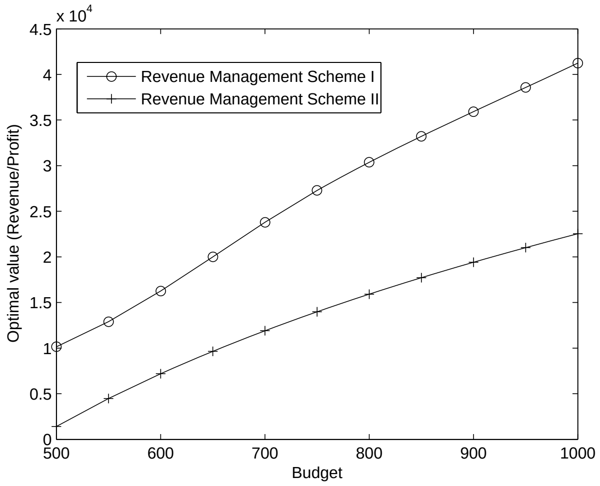

Figure 6.3: Budget versus optimal values of two revenue management schemes.

Table 6.3: Budget versus blocking probability of allocated bandwidth.

Budget B	Blocking probability in Scheme I $(\mathcal{P}(x_1, K_1, y_1), \mathcal{P}(x_2, K_2, y_2), \mathcal{P}(x_3, K_3, y_3), \mathcal{P}(x_4, K_4, y_4))$	Blocking probability in Scheme II $(\mathcal{P}(x_1, K_1, y_1), \mathcal{P}(x_2, K_2, y_2), \mathcal{P}(x_3, K_3, y_3), \mathcal{P}(x_4, K_4, y_4))$
500	(0.0020, 0.2504, 0.5389, 0.7938)	(0.0020, 0.2504, 0.5553, 0.7869)
550	(0.0020, 0.2504, 0.4004, 0.7938)	(0.0020, 0.1977, 0.5372, 0.7440)
600	(0.0020, 0.2504, 0.2720, 0.7938)	(0.0020, 0.1311, 0.4903, 0.7174)
650	(0.0020, 0.2504, 0.1616, 0.7938)	(0.0020, 0.0768, 0.4442, 0.6910)
700	(0.0020, 0.2504, 0.0795, 0.7938)	(0.0020, 0.0383, 0.3989, 0.6648)
750	(0.0020, 0.2504, 0.0311, 0.7938)	(0.0020, 0.0158, 0.3547, 0.6386)
800	(0.0020, 0.2504, 0.0096, 0.7938)	(0.0020, 0.0054, 0.3119, 0.6127)
850	(0.0020, 0.2504, 0.0025, 0.7938)	(0.0010, 0.0016, 0.2725, 0.5881)
900	(0.0020, 0.2504, 0.0006, 0.7938)	(0.0002, 0.0005, 0.2368, 0.5649)
950	(0.0020, 0.2504, 0.0001, 0.7938)	(0.0000, 0.0001, 0.2030, 0.5419)
1,000	(0.0020, 0.2504, 0.0000, 0.7938)	(0.0000, 0.0000, 0.1714, 0.5191)

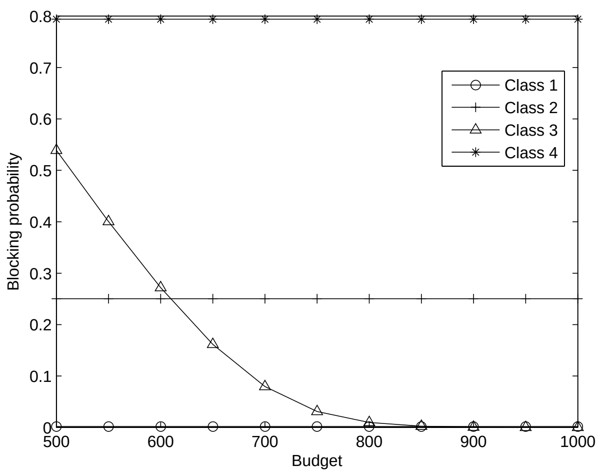

Figure 6.4: Budget versus blocking probability given by optimal solutions of Revenue Management Scheme I.

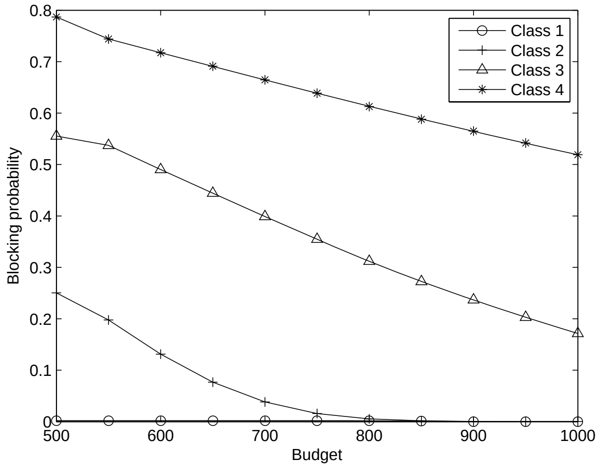

Figure 6.5: Budget versus blocking probability given by optimal solutions of Revenue Management Scheme II.

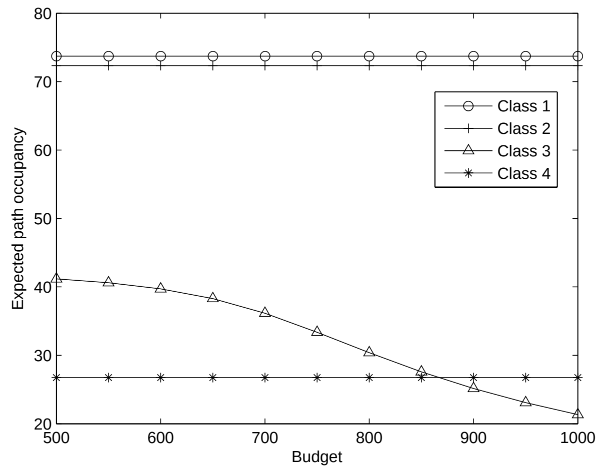

Figure 6.6: Budget versus expected path occupancy given by optimal solutions of Revenue Management Scheme I.

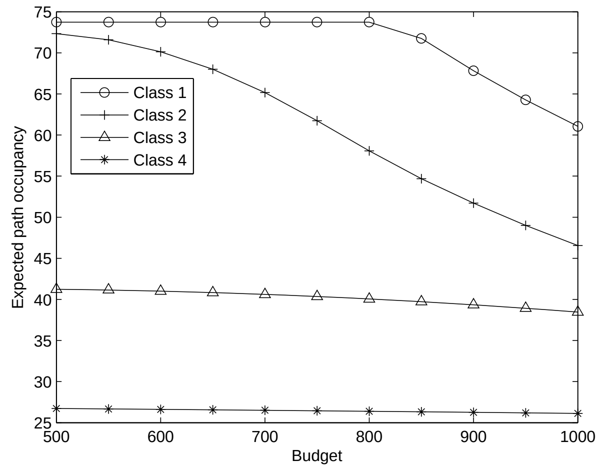

Figure 6.7: Budget versus expected path occupancy given by optimal solutions of Revenue Management Scheme II.

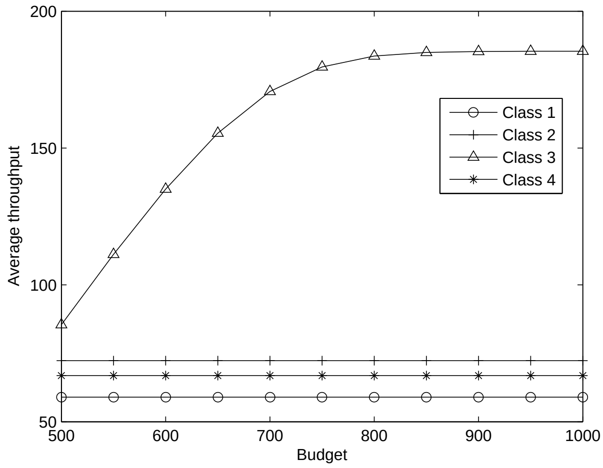

Figure 6.8: Budget versus average throughput given by optimal solutions of Revenue Management Scheme I.

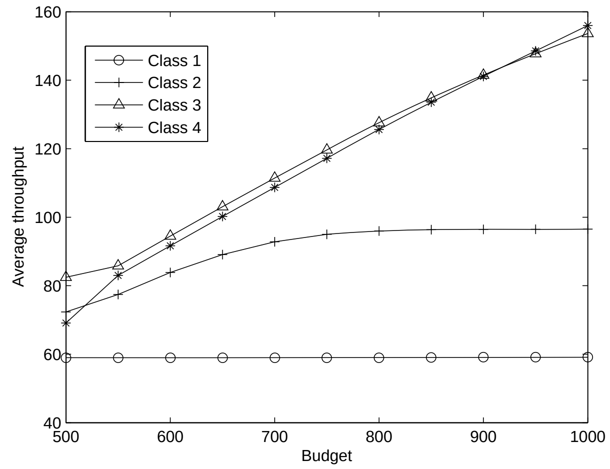

Figure 6.9: Budget versus average throughput given by optimal solutions of Revenue Management Scheme II.

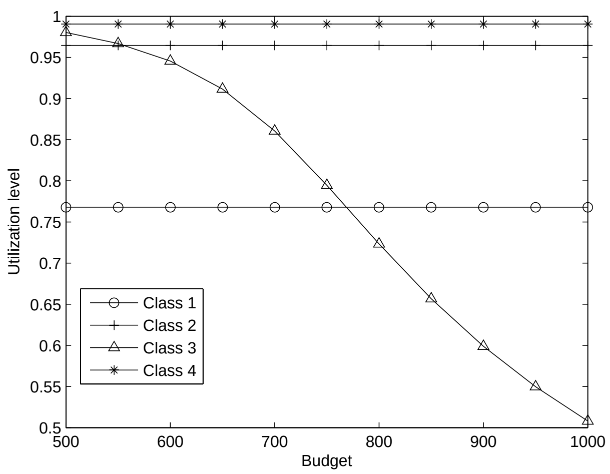

Figure 6.10: Budget versus utilization level given by optimal solutions of Revenue Management Scheme I.

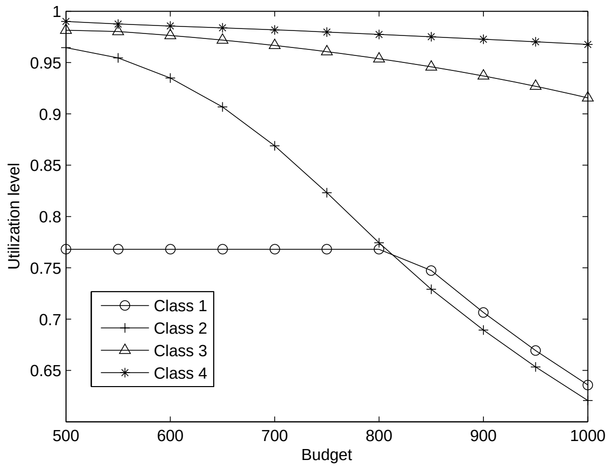

Figure 6.11: Budget versus utilization level given by optimal solutions of Revenue Management Scheme II.

Table 6.4: Budget versus budget ratio for four traffic classes.

Budget B	Budget ratio in Scheme I (B_1, B_2, B_3, B_4)	Budget ratio in Scheme II (B_1, B_2, B_3, B_4)
500	(15.36%, 19.50%, 31.39%, 33.75%)	(15.36%, 19.50%, 30.24%, 34.90%)
550	(13.96%, 17.73%, 37.63%, 30.68%)	(13.96%, 19.17%, 28.65%, 38.22%)
600	(12.80%, 16.25%, 42.82%, 28.13%)	(12.80%, 19.43%, 29.03%, 38.73%)
650	(11.82%, 15.00%, 47.22%, 25.96%)	(11.82%, 19.65%, 29.36%, 39.17%)
700	(10.97%, 13.93%, 50.99%, 24.11%)	(10.97%, 19.84%, 29.64%, 39.55%)
750	(10.24%, 13.00%, 54.26%, 22.50%)	(10.24%, 20.00%, 29.89%, 39.87%)
800	(9.60%, 12.19%, 57.12%, 21.09%)	(9.60%, 20.14%, 30.10%, 40.16%)
850	(9.04%, 11.47%, 59.64%, 19.85%)	(9.29%, 20.21%, 30.20%, 40.29%)
900	(8.53%, 10.83%, 61.88%, 18.75%)	(9.29%, 20.21%, 30.20%, 40.29%)
950	(8.08%, 10.26%, 63.89%, 17.76%)	(9.29%, 20.21%, 30.20%, 40.29%)
1,000	(7.68%, 9.75%, 65.69%, 16.88%)	(9.29%, 20.21%, 30.20%, 40.29%)

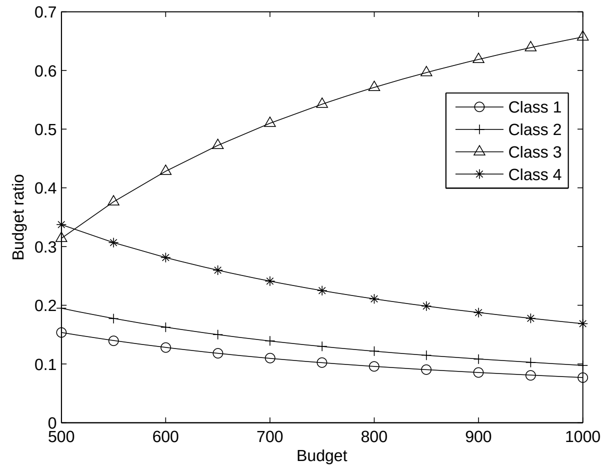

Figure 6.12: Budget versus budget ratio given by optimal solutions of Revenue Management Scheme I.

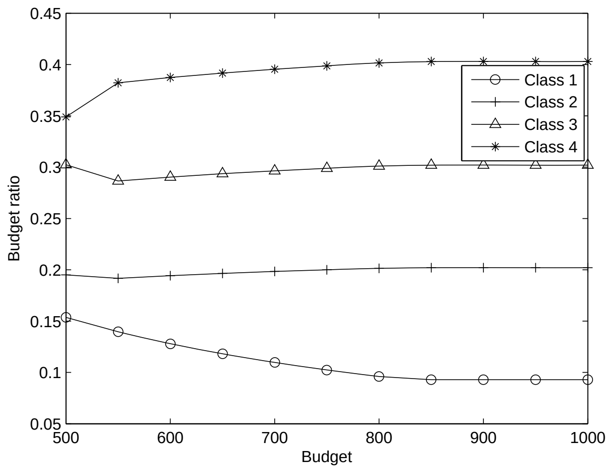

Figure 6.13: Budget versus budget ratio given by optimal solutions of Revenue Management Scheme II.

6.2 Monotonicity and Convexity of Blocking Probability and System Utilization

In this section, we present sensitivity analysis of the blocking probability and system utilizations to illustrate numerically those monotone and convex properties.

6.2.1 Effect of Increasing Bandwidth

First, we observe the effect of changing bandwidth x_i on the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ defined in (3.4). To conduct the sensitivity analysis, we check the bandwidth x_i from 0.1 Mbps to 8 Mbps, and other parameters remain fixed as listed in Table 6.1. It can be seen from Fig. 6.14 that the blocking probability is decreasing and convex when increasing bandwidth, which are consistent with those theoretical results given in Proposition 3.1 and Proposition 3.2.

Next, we show the effect of changing bandwidth x_i on the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$, average throughput $\overline{\Theta}_i$ and utilization level U_i, respectively. These three performance measures have been represented as functions of the blocking probability according to (3.5), (2.8) and (2.9). From Fig. 6.15, we find that the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ in (3.5) is a decreasing function of bandwidth x_i. In addition, it can be seen from class 1 or class 2 in Fig. 6.15 that there exists an inflection point $\widetilde{x}_i$ such that for all $x_i \leq (\geq)\widetilde{x}_i$, the expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ is concave (convex) in bandwidth x_i. Those monotone and convex properties have been summarized in Proposition 3.5 and Remark 3.2. Furthermore, it can be observed from Fig. 6.16 that average throughput $\overline{\Theta}_i$ defined in (2.8) is increasing in bandwidth x_i for four traffic classes. In Fig. 6.17, it shows the effect of changing bandwidth on the utilization levels of preset virtual paths for four traffic classes. Proposition 3.5 infers that those utilization levels U_i defined in (2.9) are decreasing when enlarging bandwidth x_i, which can be seen numerically in Fig. 6.17. Moreover, it can be observed clearly from class 1 or class 2 that there exists an inflection point

$\widetilde{x}_i$ such that U_i is concave (convex) in bandwidth x_i for all $x_i < (>) \widetilde{x}_i$.

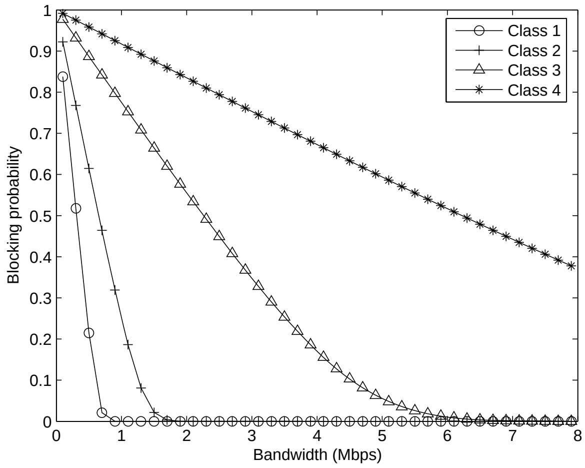

Figure 6.14: Blocking probability $\mathcal{P}(x_i, K_i, y_i)$ versus bandwidth x_i for four traffic classes.

In the following, we present numerical analysis of revenue function (2.10) in Management Scheme I and profit function (2.12) in Management Scheme II, individually. As increasing bandwidth, those numerical results are shown in Fig. 6.18 and Fig. 6.19. The results can graphically illustrate monotone and convex relationships which have been proven in Theorem 5.1-Theorem 5.6.

It has been proven in Theorem 5.1 that the objective function $F_i(x_i, K_i, y_i)$ of Revenue Management Scheme I is increasing in bandwidth x_i if it satisfies the inequality (5.1). It can be seen from Fig. 6.18 that $F_1(x_1, K_1, y_1)$ is increasing in bandwidth x_1 for all bandwidth $x_1 \geq \max\{\sqrt{c_1^t \sigma_1 / c_1^b}, b_1^{\min}\} = 0.8$ Mbps. Similarly, $F_2(x_2, K_2, y_2)$ is increasing in bandwidth x_2 for all bandwidth $x_2 \geq 1$ Mbps, and so on. We find that the convexity of average revenue (2.10) fluctuates when bandwidth x_i is small corresponding to other system parameters. Proposition 3.2 infers that, for each traffic class i, there exists a region $\mathbb{S}_i$ of bandwidth such that the blocking probability $\mathcal{P}(x_i, K_i, y_i)$ is convex (concave) for all $x_i \in (\notin)\ \mathbb{S}_i$, where the region $\mathbb{S}_i$ can be constructed from the proof of Proposition 3.2. From numerical experiments, we find that if the budget B or bandwidth x_i is large enough, those revenue function

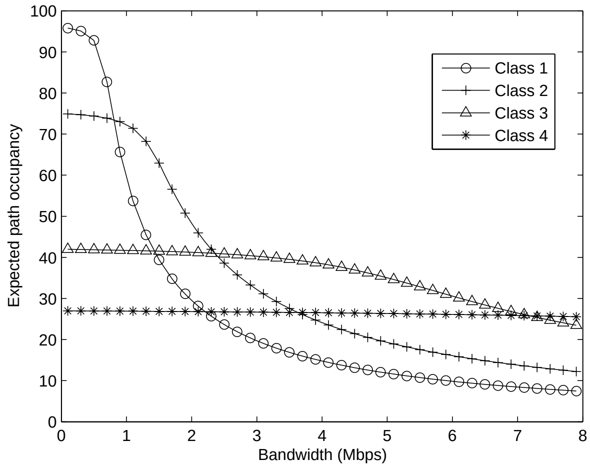

Figure 6.15: Expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ versus bandwidth x_i for four traffic classes.

$F_i(x_i, K_i, y_i)$ will become increasing and concave.

Finally, we illustrate the effect on the profit $G_i(x_i, K_i, y_i)$ in Revenue Management Scheme II when increasing bandwidth x_i. It can be observed from Fig. 6.19 that the economic profit $G_i(x_i, K_i, y_i)$ defined in (2.12) increases for all bandwidth x_i, which has already been proved in Theorem 5.3. Theorem 5.6 infers that the profit $G_i(x_i, K_i, y_i)$ is concave for all $x_i \leq 8$ Mbps, which can be seen obviously in Fig. 6.19.

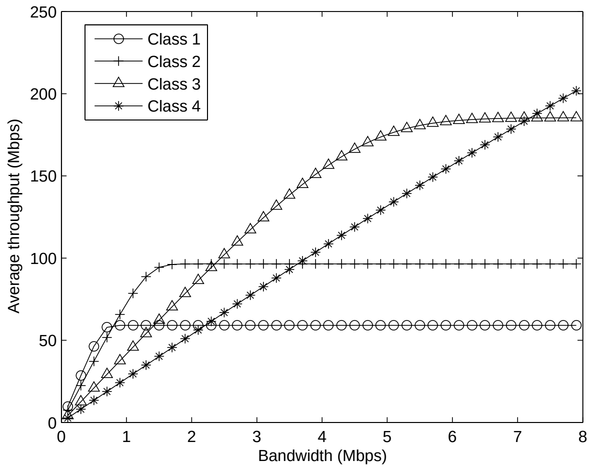

Figure 6.16: Average throughput $\overline{\Theta}_i$ versus bandwidth x_i for four traffic classes.

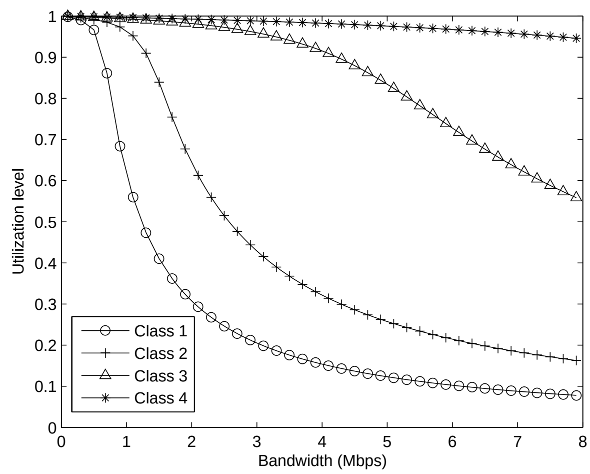

Figure 6.17: Utilization level U_i versus bandwidth x_i for four traffic classes.

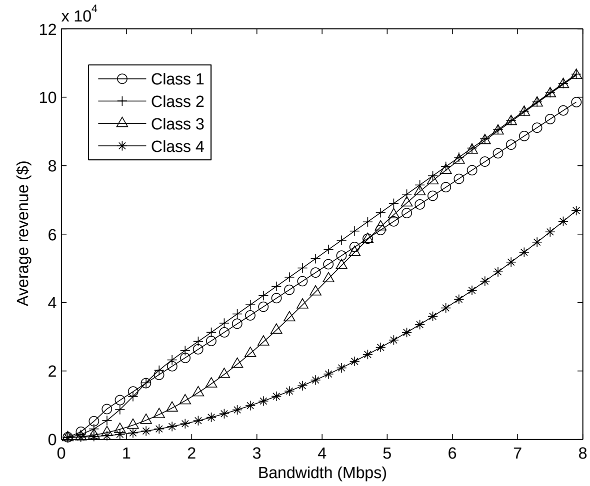

Figure 6.18: Average revenue $F_i(x_i, K_i, y_i)$ versus bandwidth x_i for four traffic classes.

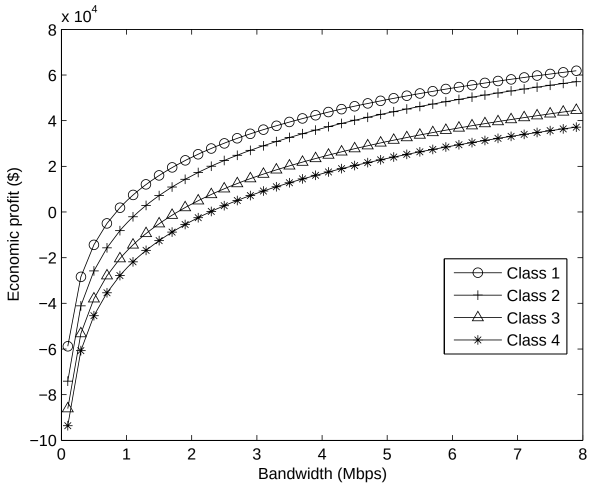

Figure 6.19: Economic profit $G_i(x_i, K_i, y_i)$ versus bandwidth x_i for four traffic classes.

6.2.2 Effect of Increasing Traffic Demand

Next, we focus on the effect of varying y_i on $\mathcal{P}(x_i, K_i, y_i)$ and system utilizations for four traffic classes. In the following, we check the traffic demand y_i from 50 Mbps to 500 Mbps to show the behavior of blocking probability $\mathcal{P}(x_i, K_i, y_i)$ and system utilizations from light traffic to heavy traffic. Here, we take the allocated bandwidth listed in Table 6.2 for four traffic classes as follows: $x_1 = 0.82$ Mbps, $x_2 = 1.76$ Mbps, $x_3 = 3.40$ Mbps, $x_4 = 5.07$ Mbps. And the available number of virtual paths K_i for four traffic classes are $K_1 = 96$, $K_2 = 75$, $K_3 = 42$ and $K_4 = 27$, which are given in Table 6.1.

Those numerical results are given in Figs. 6.20–6.23. Fig. 6.20 shows that the blocking probability is increasing when the traffic demand y_i increases. Meanwhile, we observe that those figures of blocking probabilities are concave down when the traffic demand y_i is greater than the maximal throughput $K_i x_i$. Otherwise, those figures are concave up. Those phenomena are summarized in Proposition 3.3 and Remark 3.4. Figs. 6.21–6.23 show the effect of varying traffic demand y_i on the utilization levels for four traffic classes. Fig. 6.23 shows the utilization levels U_i are increasing and concave in y_i when y_i increases for four traffic classes.

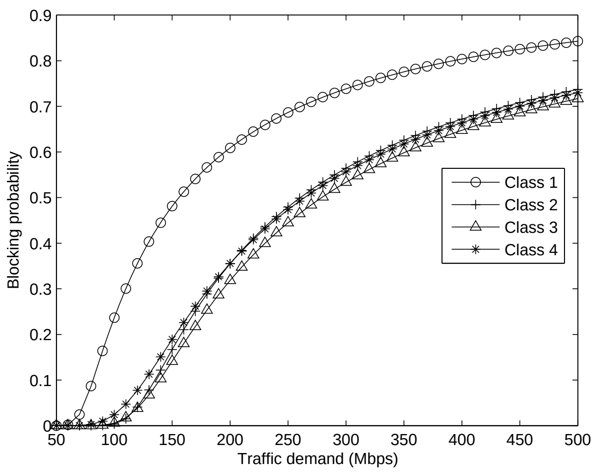

Figure 6.20: Blocking probability $P_{\text{block}}(x_i, K_i, y_i)$ versus traffic demand y_i for four traffic classes.

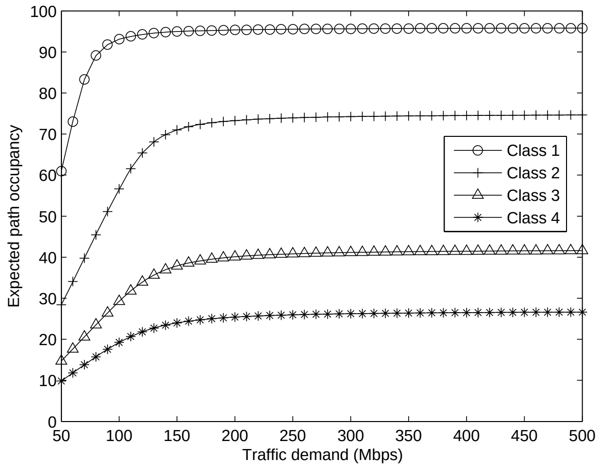

Figure 6.21: Expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ versus traffic demand y_i for four traffic classes.

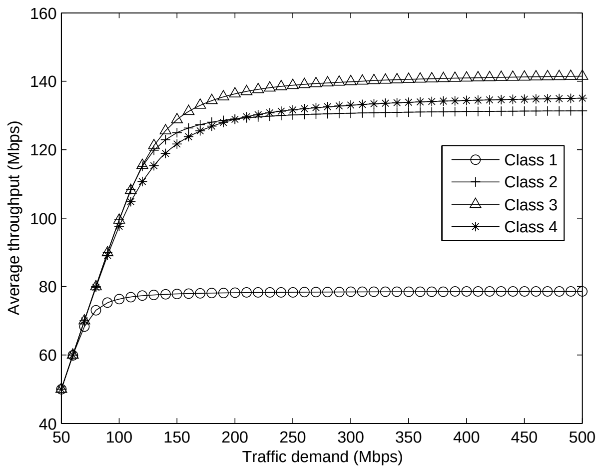

Figure 6.22: Average throughput $\overline{\Theta}_i$ versus traffic demand y_i for four traffic classes.

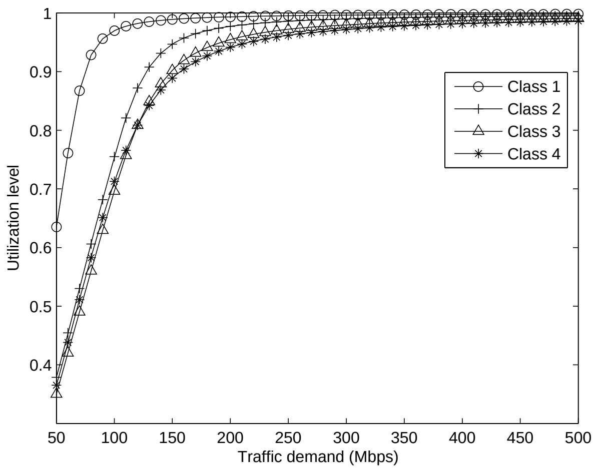

Figure 6.23: Utilization level U_i versus traffic demand y_i for four traffic classes.

6.2.3 Effect of Increasing Number of Virtual Paths

Finally, we observe the effect of increasing the available number of virtual paths K_i on $P_{\text{block}}(x_i, K_i, y_i)$ and individually. Here, the number of virtual paths K_i varies from 20 to 120, for all $i = 1, \ldots, 4$. We take the allocated bandwidth listed in Table 6.2 for four traffic classes as follows: $x_1 = 0.82$ Mbps, $x_2 = 1.76$ Mbps, $x_3 = 3.40$ Mbps, $x_4 = 5.07$ Mbps. In addition, the traffic demand y_i for four traffic classes are $y_1 = 59.1$ Mbps, $y_2 = 96.5$ Mbps, $y_3 = 185.4$ Mbps and $y_4 = 324.3$ Mbps, which are given in Table 6.1.

Numerical results of increasing the number of virtual paths K_i are depicted from Fig. 6.24 to Fig. 6.27. Fig. 6.24 shows that the blocking probability is decreasing and convex when the number of paths K_i increases. Figs. 6.25–6.27 show the effect of varying K_i on the utilization levels for four traffic classes. The utilization level of each class is increasing and concave when K_i increases.

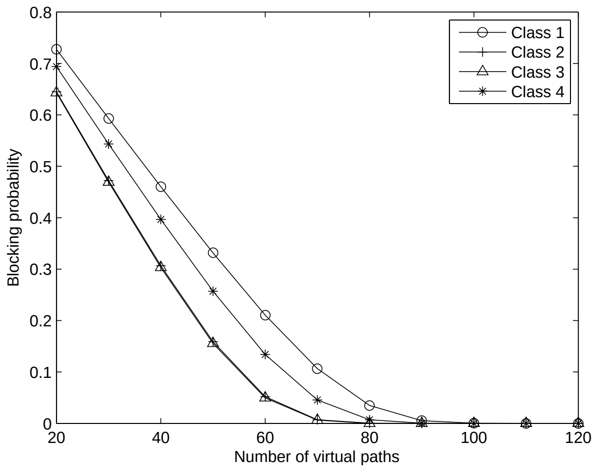

Figure 6.24: Blocking probability $P_{\text{block}}(x_i, K_i, y_i)$ versus a number of virtual paths K_i for four traffic classes.

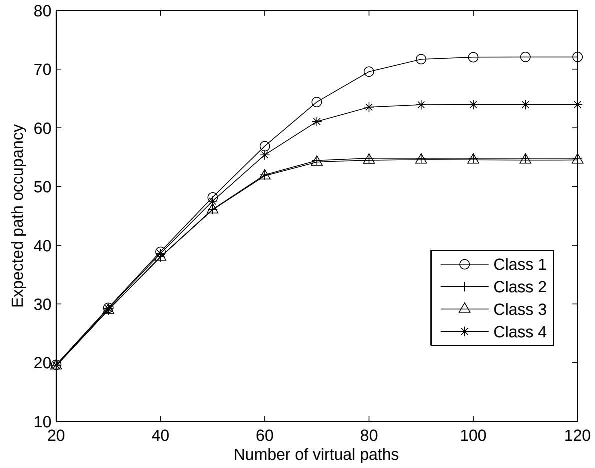

Figure 6.25: Expected path occupancy $\mathcal{L}(x_i, K_i, y_i)$ versus a number of virtual paths K_i for four traffic classes.

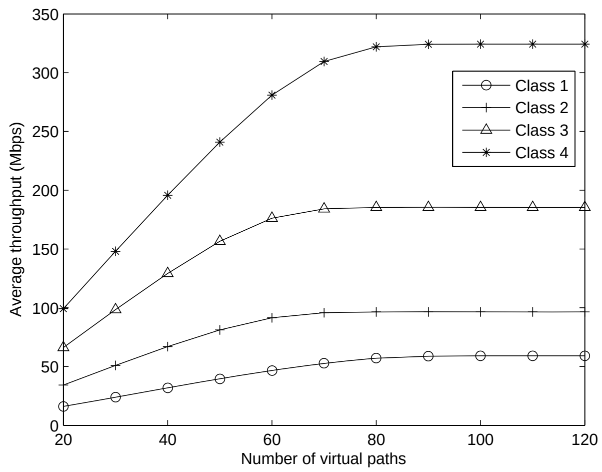

Figure 6.26: Average throughput $\overline{\Theta}_i$ versus a number of virtual paths K_i for four traffic classes.

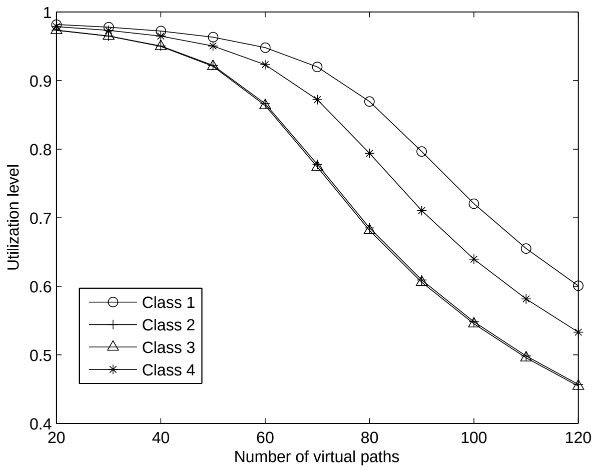

Figure 6.27: Utilization level U_i versus a number of virtual paths K_i for four traffic classes.

6.3 Numerical Illustrations of Heavy-Traffic Limits

We compare numerical results of probability $\beta_n(K)$ and blocking probability $\mathcal{P}(\rho_K, K)$ in cases of Exponential, Deterministic, Erlang-2 and Erlang-3 inter-arrival times. Here, the traffic intensity $\rho_K = 1 - \frac{\gamma}{\sqrt{K}}$ is applied from (4.22), where $\gamma = 0.35$ is given. The number of virtual paths, K, varies from 50 to 1,500.

In Fig. 6.28, it can be observed that the probability $\beta_0(K)$ of Exponential inter-arrival times is always the largest among four examples, and $\beta_0(K)$ of Erlang-3 inter-arrival times is the least. Moreover, the probability $\beta_0(K)$ is increasing as the number of servers K increases.

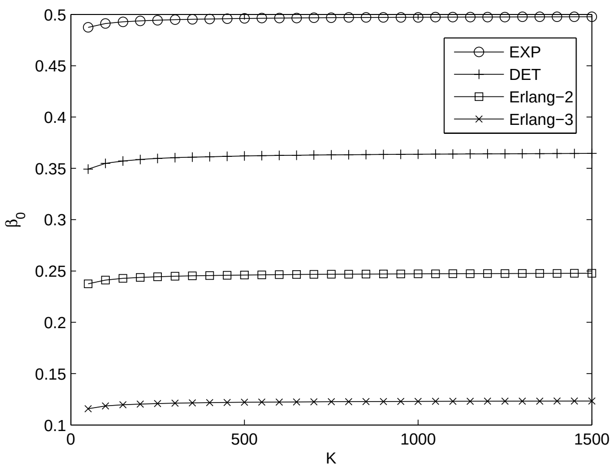

Figure 6.28: Probability $\beta_0(K)$ versus a number of virtual paths K.

Regarding the probability $\beta_1(K)$, we find that the Deterministic case is the largest and the Erlang-3 case is the least, which is shown in Fig. 6.29. It is also observed that the probability $\beta_1(K)$ is increasing as the number of servers K increases.

From Fig. 6.30, it can be seen that the limit C_1 is increasing as the number of servers K becomes larger. We find that the limit C_1 determined from Exponential

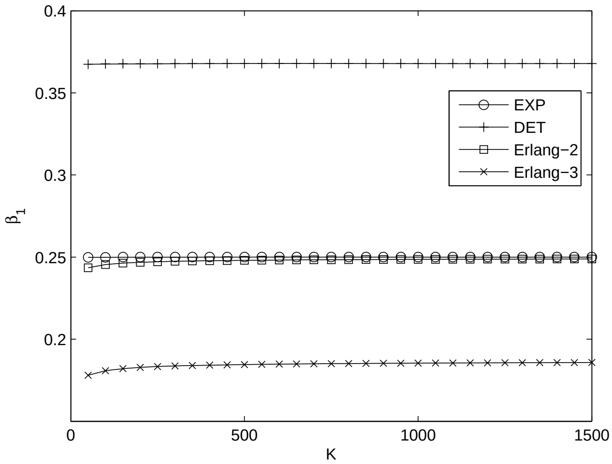

Figure 6.29: Probability $\beta_1(K)$ versus a number of virtual paths K.

inter-arrival times is always larger than that of Deterministic case, and the limit C_1 of Erlang-3 case is the least.

It can be seen in Fig. 6.31 that the number C_2 is very small for all four examples. Numerical results are observed with the logarithm of C_2, i.e., $log(C_2)$. It implies that the blocking probability becomes very small, which is shown in Fig. 6.32. Hence, the blocking probability is also depicted with the logarithm $log(\mathcal{P}(\rho_K, K))$. We find that C_2 determined from Exponential inter-arrival times is always larger than others, and C_2 of Erlang-3 case is the least. Similarly, the blocking probability under Exponential inter-arrival time distribution is the largest, and the blocking probability determined from Erlang-3 inter-arrival times is the least. The order of those blocking probabilities has been shown theoretically in Proposition 4.7.

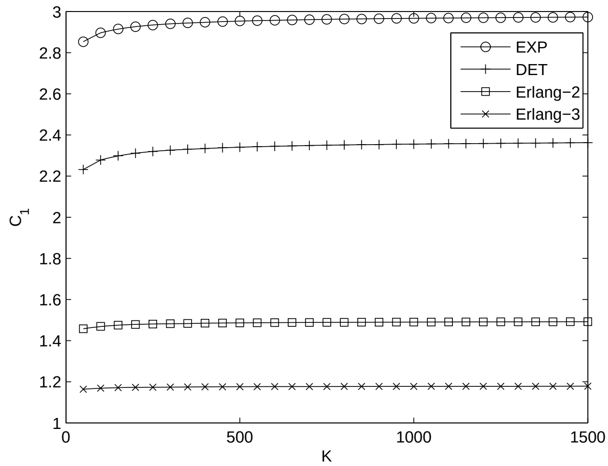

Figure 6.30: Limit C_1 versus a number of virtual paths K.

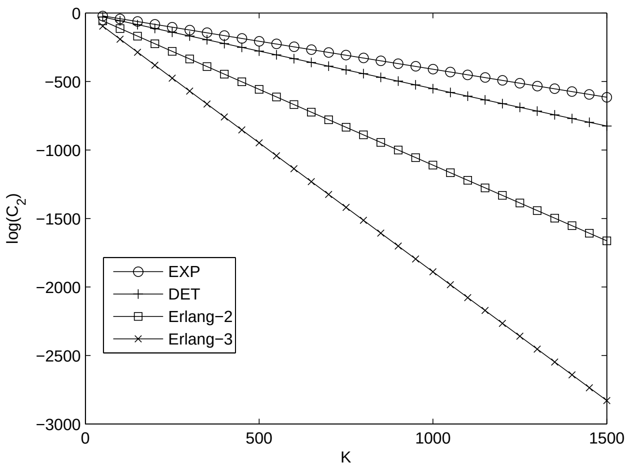

Figure 6.31: The logarithm of C_2 versus a number of virtual paths K.

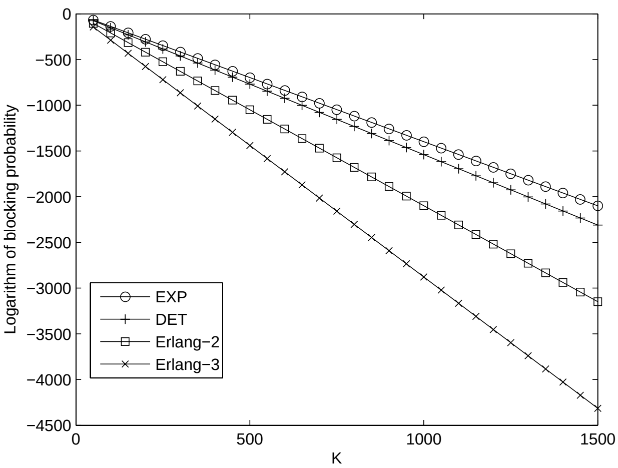

Figure 6.32: The logarithm of blocking probability $log(\mathcal{P}(\rho_K, K))$ versus a number of virtual paths K.

6.4 Sensitivity Analysis of Elasticities

In this section, we investigate the effect on three elasticities when increasing bandwidth allocation x_i, the traffic demand y_i and the number of virtual paths K_i for six traffic classes. The bandwidth allocation x_i varies from 0.1 Mbps to 30 Mbps, K_i varies from 1 to 50, and y_i varies from 0 Mbps to 2500 Mbps to show those phenomenons from light traffic to heavy traffic. Here, six traffic classes are selected as test examples from CAIDA [106], where $y_1 = 22.44$ megabit per second (Mbps) for traffic class 1, $y_2 = 76.08$ Mbps for traffic class 2, $y_3 = 145.71$ Mbps for traffic class 3, $y_4 = 347.41$ Mbps for traffic class 4, $y_5 = 476.93$ Mbps for traffic class 5, and $y_6 = 1.09$ Gigabit per second (Gbps) for traffic class 6. The bandwidth for traffic class 1 (E1 lines) is $x_1 = 2.048$ Mbps, $x_2 = 6.17$ Mbps for traffic class 2 (T2 lines), $x_3 = 11.21$ Mbps for traffic class 3, $x_4 = 21.45$ Mbps for traffic class 4, $x_5 = 22.08$ Mbps for traffic class 5, and $x_6 = 27.90$ Mbps for traffic class 6. Meanwhile, the numbers of virtual paths for those six traffic classes are $K_1 = 7$, $K_2 = 10$, $K_3 = 13$, $K_4 = 18$, $K_5 = 27$ and $K_6 = 50$. we use $c_1^t = 25$ cents per second, $c_1^b = 5$ cents per second, $c_2^t = 26$ cents per second, $c_2^b = 6$ cents per second, $c_3^t = 30$ cents per second, $c_3^b = 7$ cents per second, $c_4^t = 35$ cents per second, $c_4^b = 10$ cents per second, $c_5^t = 38$ cents per second, $c_5^b = 14$ cents per second, $c_6^t = 40$ cents per second, $c_6^b = 17$ cents per second.

Those numerical results are shown from Fig. 6.33 to Fig. 6.35. The bandwidth elasticity of blocking and the capacity elasticity of blocking are negative and decreasing for those six classes, which are shown in Fig. 6.33 and Fig. 6.35. Fig. 6.34 shows the change of demand elasticity of blocking as the traffic demand increases. The demand elasticity of blocking is positive and concave, but it becomes convex when the traffic demand is larger than certain inflection point.

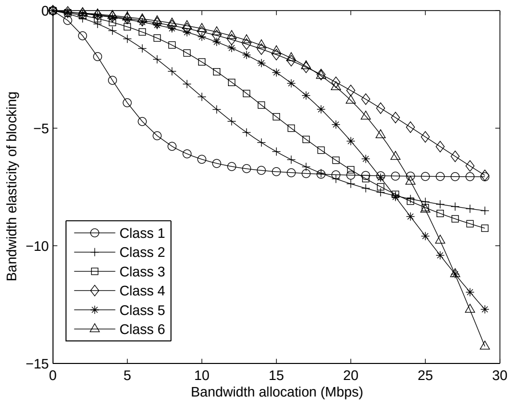

Figure 6.33: Bandwidth elasticity of blocking versus bandwidth for six classes.

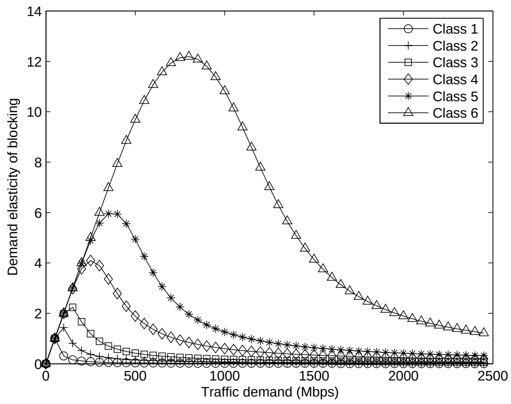

Figure 6.34: Demand elasticity of blocking versus traffic demand for six classes.

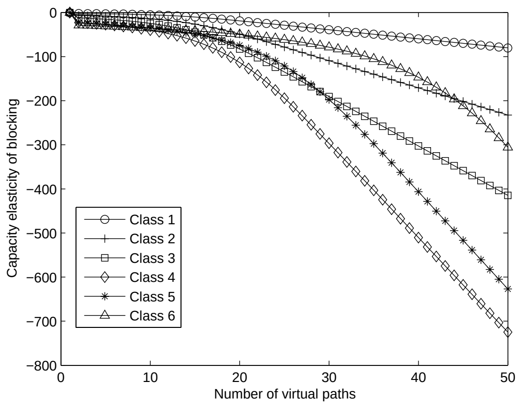

Figure 6.35: Capacity elasticity of blocking versus the number of virtual paths for six classes.

6.5 Discussions

Two revenue management schemes have been investigated theoretically and numerically to determine the amount of bandwidth required by a connection for each traffic class. Given network users' willingness-to-pay and other system parameters, our aim is to determine the bandwidth allocation that maximizes the average revenue/profit for the ISP under the budget constraint.

In Management Scheme I, almost all the resources are allocated to only one class whose marginal revenue is the largest, and the remainder are allocated to other classes to meet their feasibility only. That is, most of the available budget is allocated to certain traffic class i with the largest marginal improvement $\partial w_i F_i(x_i, K_i, y_i)/\partial x_i$ in Management Scheme I. Network managers may apply Management Scheme I to allocate limited resources among competing classes in order to maximize the weighted sum of average revenue. On the other hand, by solving Management Scheme II, all resources are allocated proportionally to four traffic classes. With the help of the utility function in (2.11), we can achieve the proportional fairness by allocating bandwidth through Management Scheme II.

To investigate these two bandwidth allocation policies, monotone and convex properties of the revenue/profit function as well as the blocking probability have been proven theoretically in previous chapters and illustrated numerically in this chapter. Those phenomena in numerical experiments are consistent with theoretical results. In practice, those results may help network managers to determine their optimal/acceptable bandwidth allocation according to one of those two management schemes [134].

Chapter 7

Conclusions and Future Works

7.1 Conclusions

In this thesis, we consider the revenue management problems on communication networks with multi-class traffic under the budget constraint. Two revenue management schemes have been investigated through the monotone and convex properties of the blocking probability and expected path occupancy of connections. The blocking probability is a measure of risk that the network service providers can not provide a 100% guaranteed availability for the stochastic traffic demand. We analyze the sensitivity of the blocking probability with model parameters, where the parameters change one-at-a-time. Under general assumptions, we have proved that the blocking probability is directionally (i) decreasing in bandwidth, (ii) convex in bandwidth for specific regions, (iii) increasing in traffic demand, and (iv) decreasing in the number of virtual paths. We also demonstrate the monotone and convex relations among those model parameters and expected path occupancy. Furthermore, we prove that for a fixed number of virtual paths, the blocking probability is increasing and convex in traffic intensity for specific regions.

The optimality conditions are derived to obtain an optimal bandwidth allocation for two revenue management schemes. A solution algorithm is also developed to allocate limited budget among competing traffic classes. We have conducted the sensitivity analysis of the average revenue function and the economic profit function for a given traffic class by changing bandwidth allocation, traffic demand and the available number of virtual paths respectively. Those results have also been verified with numerical examples interpreting the blocking probability, utilization level, average revenue, etc. The relationship between blocking probability and bandwidth allocation can help network managers to design network pricing mechanisms for sharing bandwidth in terms of blocking/congestion costs.

As the number of virtual paths is huge, approximation for the blocking probability of connections is also provided. We derive an asymptotic analysis of the blocking probability in terms of the roots of the characteristic equation and the stationary probabilities of the corresponding $GI/M/\infty$ queue. In conclusive results, blocking probabilities are estimated analytically with heavy traffic under assumptions of exponential, deterministic and Erlang-r inter-arrival time distributions, individually. Numerical experiments verify the approximation which is extremely simple yet fairly good in its performance. For the class of problems studied with different parameters, it is concluded that the approximation is adequate for practical purposes.

7.2 Future Works and Potential Applications

The contribution of this thesis is the analysis of monotone and convex relations among model parameters and performance measures of interest. We derive the relationship between the blocking probability and allocated bandwidth under the budget constraint. The blocking probability of connections for each traffic class is formulated as a function of allocated bandwidth, traffic demand and the available number of virtual paths. Monotone and convex properties of the blocking probability are shown in both theoretical construction and numerical examples. The results of this work can help in the operational processes involved in the efficient set-up and usage of a core network under the budget constraint, e.g., network design and provisioning purposes.

One application of the relationship between blocking probability and bandwidth allocation may be referred to as designing network pricing mechanisms for sharing bandwidth in terms of blocking/congestion costs, whose examples were given by Yacoubi et al. [130] and Anderson et al. [6], etc. The closed-form expression of the blocking probability in terms of bandwidth can also be used to investigate the optimal buffer size in capacitated communication systems so that the blocking probability is kept below a specific threshold [103]. Another application of this work is used to consider the admission control in networks under different bandwidth sharing policies including throughput maximization, max-min fairness, proportional fairness and balanced fairness, etc. Interested readers may refer to Bonald et al. [22], Nilsson and Pióro [88], Jordan [61], etc.

In addition, we present three elasticities to investigate the effect of changing model parameters on the average revenue in analysis of economic models. The sensitivity results derived here could be used to guide development of congestion-based pricing of network resources, and to adjust bandwidth in the optimal proportion in response to changes in desired levels of blocking probability. Future work will be conducted in the direction of further investigation for the network revenue management schemes. Additional works would have to be done in the future to make such an approach practical, e.g., design of reservation protocols, scheduling policies, measurement algorithms, and feedback algorithms to guarantee convergence.

References

[1] Abramov, V.M., Asymptotic analysis of loss probabilities in $GI/M/m/n$ queueing systems as n increases to infinity, *Quality Technology and Quantitative Management*, Vol. 4, pp. 379–393, 2007.

[2] Aktaran-Kalayci, T., Ayhan, H., Sensitivity of optimal prices to system parameters in a steady-state service facility, *European Journal of Operational Research*, Vol. 193, pp. 120–128, 2009.

[3] Alkahtani, A.M.S., Woodward, M.E., Al-Begain, K., Prioritised best effort routing with four quality of service metrics applying the concept of the analytic hierarchy process, *Computers and Operations Research*, Vol. 33, pp. 559–580, 2006.

[4] Al-Manthari, B., Ali, N.A. Nasser, N., Hassanein, H., Dynamic multiple-frame bandwidth provisioning with fairness and revenue considerations for broadband wireless access systems, *Performance Evaluation*, 2011, doi:10.1016/j.peva.2010.11.001

[5] Amiri, A., Barkhi, R., The combinatorial bandwidth packing problem, *European Journal of Operational Research*, Vol. 208, No. 1, pp. 37–45, 2011.

[6] Anderson, E., Kelly, F.P., Steinberg, R., A contract and balancing mechanism for sharing capacity in a communication network, *Management Science*, Vol. 52, No. 1, pp. 39–53, 2006.

[7] Andrade, R., Lisser, A., Maculan, N., and Plateau, G., B&B frameworks for the capacity expansion of high speed telecommunication networks under uncertainty, *Annals of Operations Research*, Vol. 140, pp. 49–65, 2005.

[8] Antunes, N., Fricker, C., Robert, P., Tibi, D., Analysis of loss networks with routing, *Annals of Applied Probability*, Vol. 16, No. 4, pp. 2007–2026, 2006.

[9] Apostolopoulos, G., Tripathi, S.K., On the effectiveness of path pre-computation in reducing the processing cost of on-demand QoS path computation, *Proc. 3rd IEEE Symposium Computers and Communications*, pp. 42–46, 1998.

[10] Atar, R., A diffusion model of scheduling control in queueing systems with many servers, *Annals of Applied Probability*, Vol. 15, pp. 820–852, 2005.

[11] Atkinson, J.B., Two new heuristics for the $GI/G/n/0$ queueing loss system with examples based on the two-phase Coxian distribution, *Journal of the Operational Research Society*, Vol. 60, No. 6, pp. 818–830, 2009.

[12] Atov, I., Tran, H.T., Harris, R.J., OPQR-G: Algorithm for efficient QoS partition and routing in multiservice IP Networks, *Computer Communications*, Vol. 28, pp. 1987–1996, 2005.

[13] Ayesta, U., Mandjes, M., Bandwidth-sharing networks under a diffusion scaling, *Annals of Operations Research*, Vol. 170, pp. 41–58, 2009.

[14] Bai, Y., Ito, M.R., Class-based packet scheduling to improve QoS for IP video, *Telecommunication Systems*, Vol. 29, No. 1, pp. 47–60, 2005.

[15] Bakry, S.H., A new method for computing Erlang-B formula, *Computers and Mathematics with Applications*, Vol. 19, No. 2, pp. 73–74, 1990.

[16] Berger, A.W., Kogan, Y., Dimensioning bandwidth for elastic traffic in high-speed data networks, *IEEE/ACM Transactions on Networking*, Vol. 8, No. 5, pp. 643–654, 2000.

[17] Bertsekas, D.P., Gallager, R., *Data Networks*, 2nd ed., Prentice Hall, New Jersey, 1992.

[18] Bertsekas, D.P., *Network Optimization*, Athena Scientific, 1998.

[19] Bogomolnaia, A., Holzman, R., Moulin, H., Sharing the cost of a capacity network, *Mathematics of Operations Research*, Vol. 35, No. 1, pp. 173–192, 2010.

[20] Bonald, T., Proutière, A., Insensitive bandwidth sharing in data networks, *Queueing Systems*, Vol. 44, No. 1, pp. 69–100, 2003.

[21] Bonald., T., Proutière, A., On performance bounds for balanced fairness, *Performance Evaluation*, Vol. 55, pp. 25–50, 2004.

[22] Bonald, T., Massoulié, L., Proutière, A., Virtamo, J., A queueing analysis of max-min fairness, proportional fairness and balanced fairness, *Queueing Systems*, Vol. 53, pp. 65–84, 2006.

[23] Bonnans, J.F., Haddou, M., Asymptotic analysis of congested communication networks, *Mathematics of Operations Research*, Vol. 25, No. 3, pp. 409–426, 2000.

[24] Borovkov, A.A., *Stochastic Processes in Queueing Theory*, New York: Springer-Verlag, 1976.

[25] Boutaba, R., Szeto, W., Iraqi, Y., DORA: Efficient routing for MPLS traffic engineering, *Journal of Network and Systems Management*, Vol. 10, No. 3, pp. 309–325, 2002.

[26] Bruni, C., Priscoli, F.D., Koch, G., Marchetti, I., Resource management in network dynamics: An optimal approach to the admission control problem, *Computers and Mathematics with Applications*, Vol. 59, pp. 305–318, 2010.

[27] Bruno, R., Conti, M., Pinizzotto, A., Routing Internet traffic in heterogeneous mesh networks: Analysis and algorithms, *Performance Evaluation*, 2011, doi:10.1016/j.peva.2011.01.006

[28] Cancela, H., Rodríguez-Bocca, P., Tuffin, B., End-to-end availability-dependent pricing of network services, *Annals of Operations Research*, Vol. 157, pp. 61–71, 2008.

[29] Catanzaro, D., Gourdin, E., Labbé, M., Özsoy, F.A., A branch-and-cut algorithm for the partitioning-hub location-routing problem, *Computers and Operations Research*, Vol. 38, No. 2, pp. 539–549, 2011.

[30] Chen, H., Zhang, H., Diffusion approximations for some multiclass queueing networks with FIFO service disciplines, *Mathematics of Operations Research*, Vol. 25, No. 4, pp. 679–707, 2000.

[31] Cheng, Y.H., Luh, H., Wang, C.H., Modeling on weighted utilizations of network dimensioning problems, *International Journal of Operations Research*, Vol. 7, No. 1, pp. 41–52, 2010.

[32] Chiu, D.M., Tam, A.S.W., Fairness of traffic controls for inelastic flows in the Internet, *Computer Networks*, Vol. 51, pp. 2938–2957, 2007.

[33] Cho, H., Girard, A., Rosenberg, C., On the advantages of optimal end-to-end QoS budget partitioning, *Telecommunication Systems*, Vol. 34, pp. 91–106, 2007.

[34] Choi, B.D., Kim, B., Kim, J., Wee, I.S., Exact convergence rate for the distributions of $GI/M/c/K$ queue as K tends to infinity, *Queueing Systems*,Vol. 44, pp. 125–136, 2003.

[35] Choi, D.W., Kim, N.K., Chae, K.C., A two-moment approximation for the $GI/G/c$ queue with finite capacity, *Informs Journal on Computing*, Vol. 17, No. 1, pp. 75–81, 2005.

[36] Correa, J.R., Schulz, A.S., Stier-Moses, N.E., Fast, fair, and efficient flows in networks, *Operations Research*, Vol. 55, No. 2, pp. 215–225, 2007.

[37] Dutta, M.Kr., Chaubey, V.K., Performance analysis of all-optical WDM network with wavelength converter using Erlang C traffic model, *Communications in Computer and Information Science*, Vol. 70, pp. 238–244, 2010.

[38] Erera, A.L., Daganzo, C.F., Lovell, D.J., The access-control problem on capacitated FIFO networks with unique O-D paths is hard, *Operations Research*, Vol. 50, No. 4, pp. 736–743, 2002.

[39] Esteves, J.S., Craveirinha, J., Cardoso, D., Computing Erlang-B function derivatives in the number of servers, *Communications in Statistics – Stochastic Models*, Vol. 11, No. 2, pp. 311–331, 1995.

[40] Faragó, A., Efficient blocking probability computation of complex traffic flows for network dimensioning, *Computers and Operations Research*, Vol. 35, No. 12, pp. 3834–3847, 2008.

[41] Fashandi, S., Gharan, S.O., Khandani, A.K., Path diversity over packet switched networks: Performance analysis and rate allocation, *IEEE/ACM Transactions on Networking*, Vol. 18, No. 5, pp. 1373–1386, 2010.

[42] Gerstel, O., Sasaki, G., A general framework for service availability for bandwidth-efficient connection-oriented networks, *IEEE/ACM Transactions on Networking*, Vol. 18, No. 3, pp. 985–995, 2010.

[43] Gozdecki, J., Jajszczyk, A., Stankiewicz, R., Quality of service terminology in IP networks, *IEEE Communications Magazine*, Vol. 41, No. 3, pp. 153–159, 2003.

[44] Guan, Y., Yang, W., Owen, H., Blough, D.M., A pricing approach for bandwidth allocation in differentiated service networks, *Computers and Operations Research*, Vol. 35, pp. 3769–3786, 2008.

[45] Guérin, R.A., Orda, A., QoS routing in networks with inaccurate information: Theory and algorithms, *IEEE/ACM Transactions on Networking*, Vol. 7, No. 3, pp. 350–364, 1999.

[46] Güven, T., La, R.J., Shayman, M.A., Bhattacharjee, B., A unified framework for multipath routing for unicast and multicast traffic, *IEEE/ACM Transactions on Networking*, Vol. 16, No. 5, pp. 1038–1051, 2008.

[47] Halachmi, B., Franta, W.R., A diffusion approximation to multi-server queue, *Management Science*, Vol. 24, No. 5, pp. 522–529, 1978.

[48] Halfin, S., Whitt, W., Heavy-traffic limits for queues with many exponential servers, *Operations Research*, Vol. 29, No. 3, pp. 567–588, 1981.

[49] Harel, A., Convexity properties of the Erlang loss formula, *Operations Research*, Vol. 38, No. 3, pp. 499–505, 1990.

[50] Harrison, J.M., Zeevi, A., Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime, *Operations Research*, Vol. 52, No. 2, pp. 243–257, 2004.

[51] Helber, S., Schimmelpfeng, K., Stolletz, R., Lagershausen, S., Using linear programming to analyze and optimize stochastic flow lines, *Annals of Operations Research*, Vol. 182, No. 1, pp. 193–211, 2011.

[52] Hernández-Orallo, E., Vila-Carbo, J., Efficient QoS routing for differentiated services EF flows, *Proc. 10th IEEE Symposium Computers and Communications*, pp. 91–96, 2005.

[53] van Hoesel, S., Optimization in telecommunication networks, *Statistica Neerlandica*, Vol. 59, No. 2, pp. 180–205, 2005.

[54] Huang, Q., Ko, K.T., Iversen, V.B., A new convolution algorithm for loss probability analysis in multiservice networks, *Performance Evaluation*, Vol. 68, No. 2, pp. 76–87, 2011.

[55] Izady, N., Worthington, D., Approximate analysis of non-stationary loss queues and networks of loss queues with general service time distributions, *European Journal of Operational Research*, 2011, doi:10.1016/j.ejor.2011.03.029

[56] Jagers, A.A., Van Doorn, E.A., On the continued Erlang loss function, *Operations Research Letters*, Vol. 5, No. 1, pp. 43–46, 1986.

[57] Jarvis, J.P., Approximating the equilibrium behavior of multi-server loss systems, *Management Science*, Vol. 31, No. 2, pp. 235–239, 1985.

[58] Jeong, S.B., Kim, S., Lee, H., Data traffic scheduling algorithm for multiuser OFDM system with adaptive modulation considering fairness among users, *Computers and Operations Research*, Vol. 32, No. 7, pp. 1723–1737, 2005.

[59] Jin, N., Jordan, S., The effect of bandwidth and buffer pricing on resource allocation and QoS, *Computer Networks*, Vol. 46, pp. 53–71, 2004.

[60] Johari, R., Tan, D.K.H., End-to-end congestion control for the Internet: Delays and stability, *IEEE/ACM Transactions on Networking*, Vol. 9, No. 6, pp. 818–832, 2001.

[61] Jordan, S., A recursive algorithm for bandwidth partitioning, *IEEE Transactions on Communications*, Vol. 58, No. 4, pp. 1026–1030, 2010.

[62] Kim, H.S., Shroff, N.B., Loss probability calculations and asymptotic analysis for finite buffer multiplexers, *IEEE/ACM Transactions on Networking*, Vol. 9, No. 6, pp. 755–768, 2001.

[63] Kim, B., Choi, B.D., Asymptotic analysis and simple approximation of the loss probability of the $GI^X/M/c/K$ queue, *Performance Evaluation*, Vol. 54, pp. 331–356, 2003.

[64] Kimura, T., A consistent diffusion approximation for finite-capacity multiserver queues, *Mathematical and Computer Modelling*, Vol. 38, pp. 1313–1324, 2003.

[65] Kelly, F.P., Charging and rate control for elastic traffic, *European Transactions on Telecommunications*, Vol. 8, pp. 33–37, 1997.

[66] Kelly, F.P., Maulloo, A.K., Tan, D.K.H., Rate control for communication networks: Shadow prices, proportional fairness and stability, *Journal of the Operational Research Society*, Vol. 49, pp. 237–252, 1998.

[67] Kelly, F.P., Mathematical modelling of the Internet, *Mathematics Unlimited - 2001 and Beyond*, Springer-Verlag, Berlin, pp. 685-702, 2001.

[68] Kelly, F.P., Fairness and stability of end-to-end congestion control, *European Journal of Control*, Vol. 9, pp. 159–176, 2003.

[69] Kelly, F.P., Williams, R.J., Fluid model for a network operating under a fair bandwidth-sharing policy, *Annals of Applied Probability*, Vol. 14, No. 3, pp. 1055–1083, 2004.

[70] Kimura, T., A consistent diffusion approximation for finite-capacity multiserver queues, *Mathematical and Computer Modelling*, Vol. 38, pp. 1313–1324, 2003.

[71] Krile, S., Perakovic, D., Load control for overloaded MPLS/DiffServ networks during SLA negotiation, *International Journal of Communications, Network and System Sciences*, Vol. 2, No. 5, pp. 422–432, 2009.

[72] Kumar, N., Saraph, G., End-to-end QoS in interdomain routing, *Proc. International Conference on Networking and Services*, pp. 82-82, 2006.

[73] Lien, Y.N., Jang, H.C., Tsai, T.C., Luh, H., BBQ: A QoS management infrastructure for All-IP networks, *Communications of Institute of Information and Computing Machinery: Mobile Communications and Wireless Networks*, Vol. 7, No. 1, pp. 89–115, 2004.

[74] Lin, Y.K., Yeh, C.T., Using minimal cuts to optimize network reliability for a stochastic computer network subject to assignment budget, *Computers and Operations Research*, Vol. 38, No. 8, pp. 1175–1187, 2011.

[75] Lee, C.Y., Cho, H.K., Discrete bandwidth allocation considering fairness and transmission load in multicast networks, *Computers and Operations Research*, Vol. 34, No. 3, pp. 884–899, 2007.

[76] Low, S.H., Lapsley, D.E., Optimization flow control–Part I: Basic algorithm and convergence, *IEEE/ACM Transactions on Networking*, Vol. 7, No. 6, pp. 861–874, 1999.

[77] Low, S.H., A duality model of TCP and queue management algorithms, *IEEE/ACM Transactions on Networking*, Vol. 11, No. 4, pp. 525–536, 2003.

[78] Luh, H., Wang, C.H., Proportional bandwidth allocation for unicasting in All-IP networks, *Proc. 2nd Sino-International Symposium on Probability, Statistics, and Quantitative Management*, pp. 111–130, 2005.

[79] Maglaras, C., Zeevi, A., Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations, *Management Science*, Vol. 49, No. 8, pp. 1018–1038, 2003.

[80] Maglaras, C., Zeevi, A., Pricing and design of differentiated services: Approximate analysis and structural insights, *Operations Research*, Vol. 53, No. 2, pp. 242–262, 2005.

[81] Massoulié, L., Roberts, J., Bandwidth sharing and admission control for elastic traffic, *Telecommunication Systems*, Vol. 15, pp. 185–201, 2000.

[82] Massoulié, L., Roberts, J., Bandwidth sharing: objectives and algorithms, *IEEE/ACM Transactions on Networking*, Vol. 10, No. 3, pp. 320–328, 2002.

[83] Messerli, E., Proof of a convexity property of the Erlang B formula, *Bell System Technical Journal*, Vol. 51, pp. 951–953, 1972.

[84] Mingozzi, E., Stea, G., Callejo-Rodríguez, M.A., Enríquez-Gabeiras, J., García-de-Blas, G., Ramón-Salquero, F.J., Burakowski, W., Beben, A., Sliwinski, J., Tarasiuk, H., Dugeon, O., Diaz, M., Baresse, L., Monteiro, E., EuQoS: End-to-end quality of service over heterogeneous networks, *Computer Communications*, Vol. 32, pp. 1355–1370, 2009.

[85] Minoux, M., *Mathematical Programming: Theory and Algorithms*, Wiley, Chichester, 1986.

[86] Mo, J., Walrand, J., Fair end-to-end window-based congestion control, *IEEE/ACM Transactions on Networking*, Vol. 8, No. 5, pp. 556–567, 2000.

[87] Nain, P., Qualitative properties of the Erlang blocking model with heterogeneous user requirements, *Queueing Systems*, Vol. 6, No. 2, pp. 189–206, 1990.

[88] Nilsson, P., Pióro, M., Solving dimensioning tasks for proportionally fair networks carrying elastic traffic, *Performance Evaluation*, Vol. 49, pp. 371–386, 2002.

[89] Ogryczak, W., Śliwiński, T., Wierzbicki, A., Fair resource allocation schemes and network dimensioning problems, *Journal of Telecommunications and Information Technology*, Vol. 3, pp. 34–42, 2003.

[90] Orda, A., Routing with end-to-end QoS guarantees in broadband networks, *IEEE/ACM Transactions on Networking*, Vol. 7, No. 3, pp. 365–374, 1999.

[91] Orda, A., Sprintson, A., Precomputation schemes for QoS routing, *IEEE/ACM Transactions on Networking*, Vol. 11, No. 4, pp. 578–591, 2003.

[92] Paschalidis, I.Ch., Liu, Y., Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects, *IEEE/ACM Transactions on Networking*, Vol. 10, No. 3, pp. 425–438, 2002.

[93] Pióro, M., Malicskó, G., Fodor, G., Optimal link capacity dimensioning in proportionally fair networks, *Lecture Notes in Computer Science*, Vol. 2345, pp.277–288, 2002.

[94] Rardin, R.L., *Optimization in Operations Research*, Prentice-Hall, 1998.

[95] Roberts, J.W., A survey on statistical bandwidth sharing, *Computer Networks*, Vol. 45, pp. 319–332, 2004.

[96] Rodrigo, M.V., Latouche, G. Remiche, M.A., Blocking probability computation in reversible Markovian bufferless multi-server systems, *Performance Evaluation*, Vol. 67, No. 3, pp. 121–140, 2010.

[97] Ross, S.M., *Stochastic Processes*, New York: Wiley, 1983.

[98] Rudin, W., *Principles of Mathematical Analysis*, McGraw-Hill Science/Engineering/Math, 3rd edition, 1976.

[99] Shakkottai, S., Srikant, R., Economics of network pricing with multiple ISPs, *IEEE/ACM Transactions on Networking*, Vol. 14, No. 6, pp. 1233–1245, 2006.

[100] Shenker, S., Fundamental design issues for the future Internet, *IEEE Journal of Selected Areas in Communications*, Vol. 13, pp. 1176–1188, 1995.

[101] Shetty, N., Schwartz, G., Walrand, J., Internet QoS and regulations, *IEEE/ACM Transactions on Networking*, Vol. 18, No. 6, pp. 1725–1737, 2010.

[102] Simonot, F., A comparison of the stationary distributions of $GI/M/c/n$ and $GI/M/c$, *Journal of Applied Probability*, Vol. 35, No. 2, pp. 510–515, 1998.

[103] Smith, J.M., $M/G/c/K$ blocking probability models and system performance, *Performance Evaluation*, Vol. 52, pp. 237–267, 2003.

[104] Stockman, A.C., *Introduction to Economics*, 2nd ed., Dryden Press, 1999.

[105] Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, New York/London (1962)

[106] The Cooperative Association for Internet Data Analysis (CAIDA). [Online]. Available: http://www.caida.org/home

[107] The MathWorks Company, *MATLAB The Language of Technical Computing: Using MALTAB*, Version 6, 2002.

[108] The UMTS Forun: Enabling UMTS/Third Generation Services and Applications. UMTS Forun Report, Vol. 11, 2000.

[109] Thomas, P., Teneketzis, D., Mackie-Mason, J.K., A market-based approach to optimal resource allocation in integrated-services connection-oriented networks, *Operations Research*, Vol. 50, No. 4, pp. 603–616, 2002.

[110] Vijaya Laxmi, P., Gupta, U.C., Analysis of finite-buffer multi-server queues with group arrivals: $GI^X/M/c/N$, *Queueing Systems*, Vol. 36, pp. 125–140, 2000.

[111] Wang, C.H., Luh, H., A precomputation-based scheme for QoS routing and fair bandwidth allocation, *Lecture Notes in Computer Science*, Vol. 4297, pp. 595–606, 2006.

[112] Wang, C.H., Luh, H., Network dimensioning problems of applying achievement functions. In: Zhang, X.S., Liu, D.G., Wu, L.Y. (Eds.), *Lecture Notes in Operations Research–Operations Research and Its Applications*, Vol. 6, pp. 35–59, World Publishing Corporation, 2006.

[113] Wang, C.H., Yue, W., Luh, H., Performance evaluation of predetermined bandwidth allocation for heterogeneous networks, *IEICE Technical Report*, Vol. 107, No. 6, pp. 37–42, 2007.

[114] Wang, C.H., Luh, H., Two-phase modeling of QoS routing in communication networks, *Proc. 16th International Conference on Computer Communications and Networks (ICCCN'07)*, pp. 1210–1216, 2007.

[115] Wang, C.H., Luh, H., Bandwidth allocation and QoS routing for heterogeneous networks, *Applied Mathematical Sciences*, Vol. 1, No. 43, pp. 2139–2151, 2007.

[116] Wang, C.H., Luh, H., A fair QoS scheme for bandwidth allocation by precomputation-based approach, *International Journal of Information and Management Sciences*, Vol. 19, No. 3, pp. 391–412, 2008.

[117] Wang, C.H., Luh, H., A utility maximization scheme for end-to-end QoS routing in communication networks with heterogeneous service classes, *Proc. IEEE International Conference on Service Operations and Logistics, and Informatics (IEEE/SOLI'2008)*, pp. 488–493, 2008.

[118] Wang, C.H., Luh, H., Mathematical models of QoS management for communication networks. In: Álvarez, M.P. (Eds.), *Leading-Edge Applied Mathematical Modeling Research*, Chapter 8, pp. 295–317, Nova Science Publishers Inc., 2008.

[119] Wang, C.H., Luh, H., Blocking probabilities of multiple classes in IP networks with QoS routing. In: Yue, W., Takahashi, Y., Takagi, H. (Eds.), *Advances in Queueing Theory and Network Applications*, Chapter 16, pp. 281–290, Springer, 2009.

[120] Wang, C.H., Luh, H., A fast branch and bound based algorithm for bandwidth allocation and QoS routing on class-based IP networks, *Proc. 1st World Congress on Global Optimization in Engineering and Science (WCGO2009)*, Hunan, China, June 1–5, 2009. Full paper is available in CD-ROM and ISTP database.

[121] Wang, C.H., Luh, H., Measuring average revenue with bandwidth and traffic demand on communication networks, *Proc. International Conference on Computational Intelligence and Software Engineering (CiSE 2009)*, Wuhan, China, December 11–13, 2009. Full paper is available in CD-ROM, IEEE Xplore and ISTP databases.

[122] Wang, C.H., Luh, H., Blocking probability on end-to-end QoS networks under budget control, working paper, submitted to *Performance Evaluation*, 2010.

[123] Wang, C.H., Luh, H., Analysis of bandwidth allocation on end-to-end QoS networks under budget control, *Computers and Mathematics with Applications*, 2011, doi:10.1016/j.camwa.2011.05.024

[124] Whitt, W., Heavy-traffic approximations for service systems with blocking, *AT&T Bell Laboratories Technical Journal*, Vol. 63, No. 5, pp. 689–708, 1984.

[125] Whitt, W., How multiserver queues scale with growing congestion-dependent demand, *Operations Research*, Vol. 51, No. 4, pp. 531–542, 2003.

[126] Whitt, W., A diffusion approximation for the $G/GI/n/m$ queue, *Operations Research*, Vol. 52, No. 6, pp. 922–941, 2004.

[127] Whitt, W., Heavy-traffic limits for the $G/H_2^*/n/m$ queue, *Mathematics of Operations Research*, Vol. 30, No. 1, pp. 1–27, 2005.

[128] Wolfram, S., *The Mathematica Book*, Wolfram Media, Inc., 5th edition, 2004.

[129] Xiao, X., Ni, L.M., Internet QoS: A big picture, *IEEE Network*, Vol. 13, No. 2, pp. 8–18, 1999.

[130] Yacoubi, M., Emelianenko, M., Gautam, N., Pricing in next generation networks: A queuing model to guarantee QoS, *Performance Evaluation*, Vol. 52, pp. 59–84, 2003.

[131] Ye, H.Q., Qu, J., Stability of data networks: Stationary and bursty models. *Operations Research*, Vol. 53, No. 1, pp. 107–125, 2005.

[132] Yuksel, M., Kalyanaraman, S., Elasticity considerations for optimal pricing of networks, *Proc. IEEE Symposium on Computer Communications*, Vol. I, pp. 163–168, 2003.

[133] Zachariadis, G., Barria, J.A., Dynamic pricing and resource allocation using revenue management for multiservice networks, *IEEE/ACM Transactions on Networking*, Vol. 5, No. 4, pp. 215–226, 2008.

[134] Zukerman, M., Mammadov, M., Tan, L., Ouveysi, I., Andrew, L.L.H., To be fair or efficient or a bit of both, *Computers and Operations Research*, Vol. 35, No. 12, pp. 3787–3806, 2008.

Appendix. Table of Notations

Notations:

B	The limited budget
$\mathbb{M}$	An index set consisting of m traffic classes
x_i	The bandwidth allocated to each connection of traffic class $i \in \mathbb{M}$
K_i	The available number of virtual paths of traffic class $i \in \mathbb{M}$
$\mathbf{x}$	The vector $(x_1, \ldots, x_m)$
$\vec{K}$	The vector $(K_1, \ldots, K_m)$
$b_i^{\min}$	The minimum bandwidth requirement of connections in traffic class $i \in \mathbb{M}$
c_i	The average cost of one unit bandwidth on virtual paths for traffic class $i \in \mathbb{M}$
$p_{i,j}$	A virtual (end-to-end) path for connection j of traffic class $i \in \mathbb{M}$
λ_i	The mean arrival rate of connections of traffic class $i \in \mathbb{M}$
σ_i	The average connection volume of connections in traffic class $i \in \mathbb{M}$
y_i	The traffic demand of connections in traffic class $i \in \mathbb{M}$
$\frac{1}{\mu_i}$	The mean sojourn time of connections on virtual paths for traffic class $i \in \mathbb{M}$
ρ_i	The traffic intensity of virtual paths for traffic class $i \in \mathbb{M}$
Θ_i	The maximum throughput of traffic class $i \in \mathbb{M}$
$\overline{\Theta}_i$	The average throughput of traffic class $i \in \mathbb{M}$
B_i	The budget ratio of class i
U_i	The utilization level of virtual paths for traffic class $i \in \mathbb{M}$
$\Omega(\vec{K}, B)$	The total available bandwidth purchased with limited budget B for preset numbers of virtual paths $\vec{K}$ on the core network
$f_i(x_i)$	The utility function of traffic class $i \in \mathbb{M}$
$F_i(x_i, K_i, y_i)$	The average revenue function of traffic class $i \in \mathbb{M}$
$G_i(x_i, K_i, y_i)$	The economic profit function of traffic class $i \in \mathbb{M}$
P_n	The steady-state occupancy probability of n connections at arbitrary epoch
P_n^-	The steady-state occupancy probability of n connections at pre-arrival epoch
$P_{m,n}$	The one-step transition probability from state m to state n
$\mathcal{P}(x_i, K_i, y_i)$	The blocking probability of incoming connections for traffic class $i \in \mathbb{M}$
$\mathcal{L}(x_i, K_i, y_i)$	The expected path occupancy in the steady state for traffic class $i \in \mathbb{M}$

Notations:

$\mathcal{E}_i^b$	The bandwidth elasticity of blocking for traffic class $i \in \mathbb{M}$
$\mathcal{E}_i^d$	The demand elasticity of blocking for traffic class $i \in \mathbb{M}$
$\mathcal{E}_i^c$	The capacity elasticity of blocking for traffic class $i \in \mathbb{M}$
ρ_K	The traffic intensity of queueing system indexed by K virtual paths
λ_K	The mean arrival rate of queueing system indexed by K virtual paths
Φ_K	The root of the characteristic equation indexed by K virtual paths
$\beta_n(K)$	The probability that number of connections served during an inter-arrival time is n when all K virtual paths are busy
$\mathcal{P}(\rho_K, K)$	The blocking probability of queueing system indexed by K virtual paths