JOINT SERVICE AND PRICE COMPETITIONS FACING NAIVE CUSTOMERS

LI LI

M.Phil

The Hong Kong Polytechnic University

2010

The Hong Kong Polytechnic University

The Department of Logistics and Maritime Studies

JOINT SERVICE AND PRICE COMPETITIONS FACING NAIVE CUSTOMERS

LI Li

A thesis submitted in partial fulfillment of the requirements

for the Degree of Master of Philosophy

August 2009

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Li LI (Name of student)

ABSTRACT

In this paper, we consider a system consisting of two service providers each with its queue. Customers are unaware of the service rates and are pragmatic in service selections. They each choose a queue to enter based on prices and actual queue lengths upon arrival and can in real time change queues before entering service. Under such customer behavior assumptions, we first characterize the steady state distributions for the queue lengths, for given service rates and prices at the two service providers, and then investigate a game in which the two service providers competitively select service rates and prices. The results underlie our exploration of the interplay between the two competition modes. We also compare system performance with those in existing literature that model customer behaviors in a different way than that in this paper, and find that the service providers tend to select lower service rates but earn higher profits when the customers are unaware of service rates than when they are aware of such information; but the uninformed customers are expected to spend more time waiting in line. Customers' state-dependent service selection upon arrival and jockeying between the queues aggravate service providers' capacity under investment and further lengthen customers' duration of stay.

PUBLICATIONS

1. Price and Service Rate Competition via Dynamic Demand Allocation. INFORMS MSOM Conference 2009, MIT, USA.
2. Price and Service Rate Competition via Dynamic Demand Allocation. 2009 CORS/INFORMS International Meeting, Toronto, Canada.
3. Joint Service and Price Competitions Facing Pragmatic Customers. Working Paper.

ACKNOWLEDGEMENTS

Thank you for reading the acknowledgment. It is my great pleasure to thank all those who have supported and contributed so much to the completion of this thesis.

First, I would like to take this opportunity to express my most sincere gratitude to my chief supervisor, Dr. Li Jiang, and my co-supervisor, Prof. Liming Liu, for their patient guidance, constant support and encouragement throughout my study and research.

I am very grateful to The Hong Kong Polytechnic University for the financial support and a conductive learning environment. I specially thank all of the academic staff and administrative staff in the Department of Logistics and Maritime Studies.

I would like to thank my friends Lixian Fan, Jingbo Yin, Jing Xu, Jie Min, and a group of research students in the Faculty of Business. Their brilliant ideas and suggestions are helpful to my research.

Finally, I want to express my deepest appreciation to my family. Without their support and understanding, I would not have finished my research work.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODCTION

The service sector encompasses a vast spectrum of business activities including trade, hotels and restaurants, communication, corporate services and more. It is a major driver of economic growth in today's world. Its proportions in global GDP and labor market have grown from below 30% in 1950's to nearly 70% (> 90% in some countries) in early 21st century. The continuously growing customer demands and expanding service networks expose service providers to fierce competitions. Service speed and price, in their different formats, have long been the two main instruments for service providers to compete for market shares. It is imperative for managers to understand the interplay between the two modes of competitions in service and price to align operations and marketing initiatives. In academic arena, existing literature on competitive service systems has mainly focused on service speed, whereas the price is not treated as a decision variable. Quoted or expected waiting time is often a primary measure whereby arriving customers select servers according to predetermined rules. Customers modeled in past papers have the privilege to learn the information of the servers' capacity investments and their prevailing service rates; and are sophisticated to derive the expected waiting time or lead time, when not committed by the service providers, to make server selection accordingly. Moreover, since server selection is based on expected

waiting times, customers stay with the queue they pick upon arrival until service completion, no matter how the actual services take place.

In real life, however, customers may not occupy the information stand to be aware of the service rates at the servers, and they are *naive* or *non-strategic* spot-utility maximizers to choose servers based on available information alone. This paper contributes to the literature on service operation in a competitive system by modeling the behaviors of such naive customers and evaluating the impacts on the servers' service capacities and price decisions as well as the experience of the customers. We build our exploration in a service system that consists of two servers each with a separate queue, and the customers choose servers based on queue lengths and service charges. The customers will not statistically infer the servers' service rates by use of average waiting time and queue length, and choose servers accordingly. One customer selects a server and enters its queue upon arrival. And when waiting for service, he can change queues in real time before entering into service provided that he feels such a move is beneficial.

For given service rates and prices at the two servers, we apply difference equations to capture the state transitions and derive the steady-state distributions for the queue lengths in closed forms. Then, we explore the competitions between servers in both service and price in a two-stage game setting. Service capacity takes time to build and its investment decision is usually made early. So, we let the two servers simultaneously build service capacities in the first stage. Their service rates, once set up, are known to each other. It is not a very unrealistic assumption. Since the two service providers work in the same industry and serve in the same market, they can in one way or another learn of the equipment and labor investments by each other to infer service

capacities. Informed of the service rates at each other, the two servers simultaneously determine prices in the second stage to affect the market demand.

We provide a complete characterization of the equilibrium for the two-stage game that demands us to tackle a non-trivial system with both continuous and discrete elements. We show that neither server chooses to be competitive in price unless its service is slow, and the faster server always overprices its opponent. Neither server holds absolute competitiveness in both price and service. In a symmetric system where the servers incur identical investment costs, they will forfeit price as a competitive instrument and each charge the maximum allowable price. Numerical results shed more insights on the servers' behaviors in an asymmetric setting where price competition plays a more influential role.

We have also compared our finding to those obtained in selected existing literature, and observe that the servers invest less in capacities and earn higher profits when customers are not informed of the servers' service rates and naive to pick servers based on queue lengths than when they are informed and choose servers based on expected waiting times. Customer sophistication has its root in the availability of the service rate information. Unaware of such information, the naive customers spend more time waiting in line on expectation than their sophisticated counterparts. Moreover, the state-independent customer allocations in which customers make real-time queue change improves the capacity utilization of the servers, inducing them to further lower service rates but reap in higher profits.

The remainder of this paper is organized as follows. We review literature in section 2. In section 3, we introduce the model, discuss customers' dynamic choice

process, and derive closed-form solutions of the steady-state distribution for queue lengths. We analyze price competition for given service rates in section 4; and explore service competition to reveal the complete equilibrium outcomes in section 5. In section 6, we do a performance comparison of the model in this paper to those in selected existing literature under different assumptions of information availability and customer behaviors. Finally, we conclude the paper in section 7.

CHAPTER 2

LITERATURE REVIEW

2.1 Dynamic Customer Allocation

In our work, *naïve or pragmatic* customers are allocated in a dynamic way, which means arriving customer is allocated to servers depending on the current work load of the servers. Our research is related to the stream of work on dynamic customer allocation in multi-server queueing system. Naor (1969) explores the behavior of strategic customers who aim to minimize their cost in a queueing system. He studies a single queue model where a customer can decide whether or not to join a queue. He finds that self serving behavior on part of the customer can lead to over congestion. Kotiah and Slater (1973) consider two general queueing settings. One is that two servers have their own queues and the other is that two servers share a single queue. They compare the steady-state performances and show numerically that the two-server-single-queue model has shorter mean queue length and lead time. Rubinovitch (1985) studies a special system of two heterogeneous servers that share a single queue, the M/M2 queue setting. The customer at the front of waiting line is randomly assigned to an idle server with equal probability. And he characterizes the system performances of this model. Singh (1970) compares the performances of M/M/2 queueing systems consisting with

homogenous and heterogeneous servers. Customer balking is allowed in his model. Konheim et al. (1981) analyze a different queueing setting in which a server splits its service capacity between two queues. If one queue is empty, full service capacity is granted to the other queue, while if neither queue is empty, the server divides its capacity to each queue equally. Flatto and Hahn (1984,1985) explore the two-server queueing system where customers are allocated to each server according to a specific generated probability which is related to the queue lengths of servers.

Shortest queue policy is a widely applied customer allocation policy in existing queueing literatures. It means that the arriving customer joins in the queue with shortest queue length. Haight (1958) applies differential-difference equations to study the steady state performance of a two-queue system with observable queue lengths, where customers join the shortest queue upon their arrival. Grassman (1980) carries out a numerical study on system performance under the shortest queue policy with limited state space. Knessl et al. (1986) investigate the steady-state distribution of the numbers of customers in a two-queue system under the shortest queue policy and characterize some properties of the solutions. Nakamura (1989) extends Knessl's work to a threshold-type scheduling, in which arriving customers are sent to the buffer of the faster server as far as the difference of the two queues does not exceed the threshold value. Zhao and Grassman (1990) analyze the case when jockeying is permitted under the shortest queue policy. Besides the shorted queue policy, Houck (1987) studies the shortest delay policy, in which the arriving customer sent to the queue with shortest expected lead time. He conducts numerical study on the system performances. Hassin and Haviv (2003) conduct a comprehensive survey of customer allocation policies,

customer behaviors and servers operations in queueing systems. Readers can refer to their book for an excellent overview.

2.1 Competitive Queueing System

There are several papers that investigate strategic servers in competitive queueing systems. Kalai et al. (1992) analyze a system consisting of two servers where customers join a single queue by the first-come-first-served (FIFS) rule. Each customer is immediately allocated to an idle server or is placed in a queue waiting for the first available server. If both servers are idle upon his arrival, customer will choose each server with equal probability. To maximize the expected profit, each server decides service rate with a cost increasing in capacity and exogenously given service fees. Kalai et al. (1992) discuss the equilibrium service rates under this queueing model. Their equilibrium analyzing methodology, however, is not correct. In the same multiple-server and single-queue setting, Li (1992) investigates the role of inventory in delivery time-based competition. He shows that the competition can breed a demand of produce-to-stock, and that delivery-time competition increases the customer's welfare while decreasing the producer's welfare. Christ and Avi-Itzhak (2002) extend the model of Kalai et al. (1992) to allow customer balking and they characterize the equilibrium service rates for this game. Bell and Stidham (1983) examine the queueing system where customers are allocated to minimize their expected lead time. They compare the results under individual and social optimization criteria and find that individuals fail to consider the externalities on the others, such as the inconvenience and overcongestion of the

faster server. Lee and Cohen (1985) study the competitive allocation policy of customers to servers by agents who wish to minimize customers' expected waiting time. Gilbert and Weng (1998) develop a service rate competition model where each server maintains a single queue. They expect both servers to have the same expected system time and provide a state-independent allocation policy for arriving customers based on that assumption. Cachon and Zhang (2007) show how different customer allocations, which include state-dependent and state-independent policies, can induce competition among servers to achieve different system performances. They discuss how customers obtain shortest lead time through choosing allocation policies. In all these studies, however, customers pay a fixed fee to each server, and allocation policies are not influenced by price.

Levhari and Luski (1978) model an M/M/2 service system in which customers choose the server on the basis of price and expected waiting time. They analyze the price competition while the capacity levels are fixed. Chen and Wan (2003) release the requirements of this model to allow firms providing different values of service and having nonidentical unit costs of waiting. Davison (1988) explores the competition among multiple servers which have fixed service rate and decide their prices. But in his model, customers have imperfect information. Li and Lee (1994) present a model of market competition in which customers make choices with the expected utilities of each server with exogenous decided service rates. Firms compete on prices and Customer utility is a function of price, quality and delivery speed. However, in all of these papers, the service rates are exogenous. Hence, the firms compete only on price. In our model,

we analyze both price and service rate decisions, which is a significant distinction between the existing papers.

There are also several papers investigate the decisions of price and capacity for a monopolistic single-server queueing system to seek the maximum system value: Dewan and Mendelson (1990), Mendelson and Whang (1990), Stidham (1992), Stidham and Rump (1998). Taking into account both customers' delay cost and servers' capacity cost, they study the internal pricing and capacity selection to seek maximum system value. So and Song (1998) consider the same problem while maximizing a firm's profit other than system value.

As for the multiple-server price and capacity competition papers, De Vany and Savings (1983) firstly address a richer type of competition in which service providers compete with price and service rate. In their model, price and service rate decisions are made at the same time. Customers choose server by the full price which consists of price and expected waiting time. Deneckere and Peck (1995) consider the similar model in which a large number of firms choose prices and capacities simultaneously, and customers select firms based on expected utility. However, they assume that customer can only access one firm, and do not allow interfirm price dispersion which means the firms in their model are homogenous. Ha et al. (2003) studies two suppliers in a supply chain serving on buyer, in which pricing and delivery frequency decisions are made in a three-stage competitive game. They assume deterministic demand and state-independent rule for customers to select service. Reitman (1991) examines competitive capacity and pricing decisions when customers make choices on price and delay time. He gives numerical results for the capacity and the price in the equilibrium. Lederer and Li (1997)

include scheduling as a strategic variable together with price and production rate for different types of customers with heterogeneous delay costs. So (2000) analyzes the situation when firms use prices and delivery time guarantees to compete for market share. In their model, demand rate functions are specified as a special type of logit models. Cachon and Harker (2002) develop a model where customers are price- and time-sensitive, and service providers face economy of scale. They also investigate the impact of outsourcing on competition. Armony and Haviv (2003) study a duopoly competition model in which customers make choice by full price, which includes the service fee plus expected waiting costs. Allon and Federgruen (2007) consider different types of competition which depend on industry dynamics through which the firms make strategic choices of service charge and capacity by various sequences. However, the competition in those papers is exogenous, while in our model it is determined by the customers via a dynamic allocation policy inside the service system.

CHAPTER 3

THE MODLE

We consider a two-server system in which each server maintains a separate queue. Customers arrive to the system according to a Poisson process at rate λ. The service time at server i is exponentially distributed with rate μ_i, for $i = 1,2$. We let $\rho \equiv \frac{\lambda}{\mu_1 + \mu_2}$ be the system load factor, and $A \equiv \lambda - \mu_1 - \mu_2$ the difference between arrival rate and aggregate service rate. It is easy to verify $A > 0$ iff $\rho > 1$. To attain its service rate μ_i, server i incurs capacity cost $C(\mu_i)$, non-negative, increasing, and convex. Service cost is normalized to zero. Server i charges p_i to each customer who uses its service. We will interchangeably use price and service charge in this paper.

To describe system dynamics, we let $n_i(t)$, for $i = 1,2$ and $t \geq 0$, be the number of customers in the queue for server i at time t; and $n_i = n_i(\infty)$ be its stationary counterpart when the system has been in operation for a long time. We are interested in the steady state (n_1, n_2).

Customers are *spot* utility maximizers. We let the utility, U_i, that a customer obtains by choosing server i, with service charge p_i and queue length, n_i, take the form bellow:

$$U_i = -\beta_1 n_i - \beta_2 p_i, \text{ for } i = 1,2,^{[1]} \tag{3.1}$$

where, β_1 and β_2 are, respectively, the marginal disutilities for queue length and service charge. We can normalize the utility function, as given in (1), to

$$U_i = -n_i - \beta p_i, \text{ for } i = 1,2, \tag{3.2}$$

and simply call β the marginal disutility.

For given prices by the two servers, we define $B \equiv [\beta(p_1 - p_2)]$, where $[x]$ is the largest integer no greater than x; and call it *markup*. Markup takes integer values only. As we will show later, the effect of price competition on customers' distribution between the two queues is solely captured by B. Server 1 is price competitive if $B < 0$, whereas server 2 is price competitive if $B > 0$. The two servers do not engage in price competition if $B = 0$. When one server is price competitive, we say this server occupies a *strong* price position and its opponent a *weak* price position. The value of markup determines the price positions of the two servers relative to each other.

Informed of service charges and queue lengths for both servers, an arriving customer selects a queue to join by the following rule: He will join

i) queue i when $U_i > U_j$, for $i, j = 1,2$ and $i \neq j$;

ii) the queue for the server that charges lower price when $U_1 = U_2$ but $p_1 \neq p_2$;

iii) the queue for either server with equal probability if $U_1 = U_2$ and $p_1 = p_2$.

[1] The utility function defined in (3.1), while it assumes negative values captures the key features for customers' server selection. We can use utility functions like U=exp(-$\beta_1 n$-$\beta_2 p$) to derive the same results.

We let $\overline{P}$ be the maximum service charge that customers can accept. If one server charges a price higher than $\overline{P}$, then no customers will choose its service. Therefore, $\overline{P}$ can be thought of as the value of service to the customers.

When a customer is waiting in one queue before he enters into service, we allow him to switch to the other queue *dynamically*, but require that he gain strictly higher utility from the move. A customer switching from one queue to the other, regardless of his position in the original queue, has to move to the end of the destination queue. Observe that whenever opportunity arises for existing customers to switch queues, it happens after a server has just completed one service, and it must be the customer at the end of one queue who makes the first move. We assume that the customers do not incur cost to switch from one queue to the other. Incorporating such switching cost into our model is analytically not difficult and does not alter the qualitative nature of the results, but makes the expressions be more tedious.

3.1 Steady-State Distribution

Given the service rates and charges at the two servers, we derive the steady-state distributions by analyzing state transitions, taking into consideration the customers' strategic behaviors in selecting queues upon arrival and changing queues while waiting.

When $\rho \geq 1$, or $\mu_1 + \mu_2 \leq \lambda$, the aggregate service capacity is insufficient to serve the customers in entirety. As a result, the queue length of each server will go to infinity in the long run, no matter how much the servers charge for their services, and each server will be completely occupied.

To derive the steady-state distribution, we focus on $\mu_1 + \mu_2 > \lambda$, or $\rho < 1$, to avoid customer jammed. We denote the steady-state probability for state (n_1, n_2), where n_i is the number of customers in queue i, as $\pi(n_1, n_2)$; and the marginal probability of n customers in queue i, as $\pi_i(n)$, for $i = 1,2$, i.e. $\pi_1(n) = \sum_{n_2=0}^{\infty} \pi(n, n_2)$ and $\pi_2(n) = \sum_{n_1=0}^{\infty} \pi(n_1, n)$. In the following, we consider three scenarios with respect to the relative magnitudes of p_1 and p_2 to derive the steady-state distribution and, particularly, the probability that each server is idle, which constitutes the building block for equilibrium analysis.

3.1.1. Servers Charge Identical Prices: $p_1 = p_2 = p$

Since the two servers charge the same prices, the queue length is the only factor influencing customers' server selection. In this case, the difference in the lengths of the queues for the two servers should not be larger than 1. To see this, note that if n customers are in queue 1 and n+2 customers are in queue 2, the last customer in queue 2 can gain higher utility, from $U_2 = -n - 2 - \beta p$ to $U_1 = -n - 1 - \beta p$, by moving to the end of queue 1. On the other hand, a customer other than the last one in queue 2 is unable to obtain higher utility by changing queues. Thus, the system state will change from $(n, n+2)$ to $(n+1, n+1)$ to reach a temporary stability. As a consequence, the system state must take either one of three forms: (n, n), $(n, n+1)$, and $(n+1, n)$ for $n \geq 0$. The difference equations can be expressed as follows:

$$\pi(0,0)\lambda = \pi(1,0)\mu_1 + \pi(0,1)\mu_2 \tag{3.3}$$

$$\pi(1,0)(\lambda+\mu_1) = \pi(1,1)\mu_2 + \pi(0,0)\frac{\lambda}{2} \tag{3.4}$$

$$\pi(0,1)(\lambda+\mu_2) = \pi(1,1)\mu_1 + \pi(0,0)\frac{\lambda}{2} \tag{3.5}$$

when $n \geq 1$,

$$\pi(n,n)(\lambda+\mu_1+\mu_2) = [\pi(n,n+1)+\pi(n+1,n)](\mu_1+\mu_2) + \pi(n,n-1)\lambda + \pi(n-1,n)\lambda \tag{3.6}$$

$$\pi(n+1,n)(\lambda+\mu_1+\mu_2) = \pi(n+1,n+1)\mu_2 + \pi(n,n)\frac{\lambda}{2} \tag{3.7}$$

$$\pi(n,n+1)(\lambda+\mu_1+\mu_2) = \pi(n+1,n+1)\mu_1 + \pi(n,n)\frac{\lambda}{2} \tag{3.8}$$

$$\sum_{n_1}^{\infty}\sum_{n_2}^{\infty}\pi(n_1,n_2) = 1 \tag{3.9}$$

Solving equations (3.3) to (3.8), we can derive that:

$$\pi(1,0) = \pi(0,0)\frac{\lambda}{2\mu_1} \tag{3.10}$$

$$\pi(0,1) = \pi(0,0)\frac{\lambda}{2\mu_2} \tag{3.11}$$

$$\pi(1,1) = \pi(0,0)\frac{\lambda^2}{2\mu_1\mu_2} \tag{3.12}$$

For $n \geq 1$,

$$\pi(n+1,n+1) = \pi(n,n)\rho^2 \tag{3.13}$$

$$\pi(n+1,n) = \pi(n,n-1)\rho^2 \tag{3.14}$$

$$\pi(n,n+1) = \pi(n-1,n)\rho^2 \tag{3.15}$$

Substituting equations (3.10) to (3.15) into (3.9), we can derive that:

$$\pi(0,0) = \frac{2\mu_1\mu_2(\mu_1+\mu_2-\lambda)}{\mu_1^2\lambda+\mu_2^2\lambda+2\mu_1^2\mu_2+2\mu_2^2\mu_1} \tag{3.16}$$

$$\pi(1,0) = \frac{\lambda\mu_2(\mu_1+\mu_2-\lambda)}{\mu_1^2\lambda+\mu_2^2\lambda+2\mu_1^2\mu_2+2\mu_2^2\mu_1} \tag{3.17}$$

$$\pi(0,1) = \frac{\lambda\mu_1(\mu_1+\mu_2-\lambda)}{\mu_1^2\lambda+\mu_2^2\lambda+2\mu_1^2\mu_2+2\mu_2^2\mu_1} \tag{3.18}$$

The probability that server i is idle is:

$$\pi_i(0) = \frac{\mu_i(\mu_1+\mu_2-\lambda)(\lambda+2\mu_j)}{\mu_1^2\lambda+\mu_2^2\lambda+2\mu_1^2\mu_2+2\mu_2^2\mu_1}, \quad \text{for } i,j=1,2 \text{ and } i \neq j. \tag{3.19}$$

3.1.2. Server 1 Charges Higher Prices: $p_1 > p_2$

Now that service 1 charges a higher price than server 2, it is at a disadvantageous position to attract customers, since a customer will choose server 1 only when the queue length at server 2 is sufficiently long, i.e., $n_2 > n_1 + \beta(p_1 - p_2)$. When there is no customer in server 1, the queue length for server 2 can be as long as $B+1$ before a customer may consider server 1. On the other hand, when there are customers in both queues, by the same logic as that for the case when $p_1 = p_2$, we can argue that while server 2 has a longer queue than server 1, the two queue lengths will not differ by more than $B+1$ in the steady state. The difference equations are as follows:

$$\pi(0,n_2)=\pi(0,0)(\frac{\lambda}{\mu_2})^{n_2} \quad (0\le n_2\le B) \tag{3.20}$$

$$\pi(0,B)(\lambda+\mu_2)=\pi(0,B-1)\lambda+\pi(0,B+1)\mu_2 \tag{3.21}$$

$$\pi(0,B+1)(\lambda+\mu_2)=\pi(0,B)\lambda+\pi(1,B+1)(\mu_1+\mu_2) \tag{3.22}$$

For $n_1\ge 1$,

$$\pi(n_1,n_1+B)(\lambda+\mu_2+\mu_1)=\pi(n_1-1,n_1+B)\lambda+\pi(n_1,n_1+B+1)(\mu_1+\mu_2) \tag{3.23}$$

$$\pi(n_1,n_1+B)=\pi(n_1-1,n_1+B)\rho \tag{3.24}$$

$$\pi(n_1,n_1+B+1)(\lambda+\mu_2+\mu_1)=\pi(n_1+1,n_1+B+1)(\mu_1+\mu_2)+\pi(n_1,n_1+B)\lambda \tag{3.25}$$

$$\pi(n_1,n_1+B+1)=\pi(n_1-1,n_1+B)\rho^2 \tag{3.26}$$

$$\pi(n_1+1,n_1+B+1)=\pi(n_1,n_1+B)\rho^2 \tag{3.27}$$

$$\sum_{n_1}^{\infty}\sum_{n_2}^{\infty}\pi(n_1,n_2)=1 \tag{3.28}$$

Solving the equations (3.20) to (3.28), we can derive that:

$$\pi(0,0)=\frac{\mu_2^{B+1}(\lambda-\mu_1-\mu_2)(\mu_2-\lambda)}{\mu_1\lambda^{B+2}+\mu_2^{B+2}(\lambda-\mu_1-\mu_2)} \tag{3.29}$$

The probability that a server is idle is:

$$\pi_1(0)=\frac{(\lambda-\mu_1-\mu_2)(\mu_2^{B+2}-\lambda^{B+2})}{\mu_1\lambda^{B+2}+\mu_2^{B+2}(\lambda-\mu_1-\mu_2)} \tag{3.30}$$

$$\pi_2(0)=\frac{\mu_2^{B+1}(\lambda-\mu_1-\mu_2)(\mu_2-\lambda)}{\mu_1\lambda^{B+2}+\mu_2^{B+2}(\lambda-\mu_1-\mu_2)} \tag{3.31}$$

3.1.3. Server 1 Charges Lower Prices: $p_1 < p_2$

Server 1 has price advantage over server 2 when it charges a lower price. Arriving or existing customers will select server 2 only when there are too many customers already waiting for server 1, i.e., $n_1 > n_2 + \beta(p_2 - p_1)$. Similar to the case when $p_1 > p_2$, we can apply the similar logic and analyzing method to list the difference equations:

$$\pi(n_1,0) = \pi(0,0)(\frac{\lambda}{\mu_1})^{n_1} \quad (0 \le n_1 \le -B) \tag{3.32}$$

$$\pi(-B,0)(\lambda+\mu_1) = \pi(-B-1,0)\lambda + \pi(-B+1,0)\mu_1 \tag{3.33}$$

$$\pi(-B+1,0)(\lambda+\mu_1) = \pi(-B,0)\lambda + \pi(-B+1,1)(\mu_1+\mu_2) \tag{3.34}$$

For $n_2 \ge 1$,

$$\pi(n_2-B,n_2)(\lambda+\mu_1+\mu_2) = \pi(n_2-B,n_2-1)\lambda + \pi(n_2-B+1,n_2)(\mu_1+\mu_2) \tag{3.35}$$

$$\pi(n_2-B,n_2) = \pi(n_2-B,n_2-1)\rho \tag{3.36}$$

$$\pi(n_2-B+1,n_2)(\lambda+\mu_1+\mu_2) = \pi(n_2-B+1,n_2+1)(\mu_1+\mu_2) + \pi(n_2-B,n_2)\lambda \tag{3.37}$$

$$\pi(n_2-B+1,n_2) = \pi(n_2-B,n_2-1)\rho^2 \tag{3.38}$$

$$\pi(n_2-B+1,n_2+1) = \pi(n_2-B,n_2)\rho^2 \tag{3.39}$$

$$\sum_{n_1}^{\infty}\sum_{n_2}^{\infty}\pi(n_1,n_2) = 1 \tag{3.40}$$

Solving the equations (3.32) to (3.40), we can derive that:

$$\pi(0,0) = \frac{\mu_1^{-B+1}(\lambda-\mu_1-\mu_2)(\mu_1-\lambda)}{\mu_2\lambda^{-B+2} + \mu_1^{-B+2}(\lambda-\mu_1-\mu_2)} \tag{3.41}$$

The probability that a server is idle is:

$$\pi_1(0) = \frac{\mu_1^{-B+1}(\lambda-\mu_1-\mu_2)(\mu_1-\lambda)}{\mu_2\lambda^{-B+2} + \mu_1^{-B+2}(\lambda-\mu_1-\mu_2)} \tag{3.42}$$

$$\pi_2(0) = \frac{(\lambda - \mu_1 - \mu_2)(\mu_1^{-B+2} - \lambda^{-B+2})}{\mu_2 \lambda^{-B+2} + \mu_1^{-B+2}(\lambda - \mu_1 - \mu_2)} \tag{3.43}$$

With these steady-state distributions for the number of customers in the two queues, for given service rates and charges, we are ready to investigate the competition between the servers in price and service.

CHAPTER 4

PRICE COMPETITION GAME

Each server maximizes its own profit by investing in a service rate and charging a price. We formulate a two-stage game. In the first stage, the two servers simultaneously invest in their capacities to attain service rates. In the second stage, knowing the service rates at each other, they charge prices simultaneously. The customers are unaware of the exact service rates, but are informed of service charges and can observe queue lengths. We apply backward induction to first analyze price competition game for given service rates; and then explore service competition to obtain the complete equilibrium scenario.

4.1 Profit Functions

For given service rates at the two servers, we now explore the value of price competition to the serves. The price at one server, say i, determines the unit revenue for each service it provides and in the meantime competes against the service charge by server $j \neq i$. Higher service charge increases revenue from each completed service, but causes the server to lose customers to its opponent.

If $\mu_1+\mu_2 \le \lambda$, or $\rho \ge 1$, the aggregate service capacity is insufficient to handle all the customers so that both queues will be fully occupied no matter how the customers choose and switch between queues. The expected profit of server i is:

$$\Pi_i = p_i \cdot \mu_i - C(\mu_i), \text{ for } i = 1,2. \tag{4.1}$$

Where $C(\mu_i)$ is the capacity cost, and $C(0)=0$, $C'(\cdot)>0$, and $C''(\cdot)>0$ are assumed. Service charges do not exert substantial effects on the distribution of customers between the queues. To maximize its own profit, each server will charge the highest price that customers can accept. i.e., $\overline{P}$.

We next consider the situation when $\rho<1$, for which the steady-state queue-length distributions were derived in section 3.1. The expected number of customers that server i, for $i=1,2$, serves is $[1-\pi_i(0)]\mu_i$; and we can write its expected profit as

$$\Pi_i(p_i \mid p_j) = p_i \cdot [1-\pi_i(0)] \cdot \mu_i - C(\mu_i) = p_i \cdot f_i(B) \cdot \mu_i - C(\mu_i), \tag{4.2}$$

where, $f_i(B) \equiv 1-\pi_i(0)$ is the fraction of time that sever i is busy, and we call it the *occupancy rate*. Larger occupancy rate implies larger market share. By the results in Chapter 3, we can express $f_i(B)$ as:

$$f_1(B)=\begin{cases} \dfrac{\lambda^{B+2}(\lambda-\mu_2)}{\mu_1\lambda^{B+2}+\mu_2^{B+2}A}, & B>0 \\ \dfrac{\lambda(\mu_2^2+\mu_1\mu_2+\lambda\mu_1)}{\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2}, & B=0 \\ \dfrac{\mu_2\lambda^{-B+2}+\lambda\mu_1^{-B+1}A}{\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A}, & B<0 \end{cases}; \text{ and } f_2(B)=\begin{cases} \dfrac{\mu_1\lambda^{B+2}+\lambda\mu_2^{B+1}A}{\mu_1\lambda^{B+2}+\mu_2^{B+2}A}, & B>0 \\ \dfrac{\lambda(\mu_1^2+\mu_1\mu_2+\lambda\mu_2)}{\lambda\mu_1^2+\lambda\mu_2^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2}, & B=0 \\ \dfrac{\lambda^{-B+2}(\lambda-\mu_1)}{\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A}, & B<0 \end{cases}.$$

For given service rates at the two servers, $f_i(B)$ is solely a function of markup, $B = [\beta(p_1 - p_2)]$, which we know determines the price positions of the two servers relative to each other. Given $p_2 \geq 0$, server 1 can choose any price in $\left(p_2 + \frac{B-1}{\beta}, p_2 + \frac{B}{\beta}\right]$ for $B \in I$ to attain the same occupancy rate. If it targets a markup of B with respect to p_2, then it will charge $p_1 = p_2 + \frac{B}{\beta}$ to reap in the highest unit revenue. Server 1's problem of finding the best price is then equivalent to that of finding the best markup. Server 2's price decision for given server 1's price is similar. Note, however, to target a markup of $B \in I$, server 2 *marks down* its price relative to that at server 1 and charges $p_2 = p_1 - \frac{B}{\beta}$, for given p_1. Lemma 4.1 shows the sensitivities of occupancy rates with respect to markup and service rates.

Lemma 4.1: *Given the service rates at the two servers, μ_1 and μ_2, and customer arrival rate λ, with $\mu_1 + \mu_2 < \lambda$,*

1. *$f_1(B)$ decreases in B, and $f_2(B)$ increases in B;*
2. *$f_i(B)$ decreases in μ_1 and μ_2, for $i = 1,2$.*

Part 1) of Lemma 4.1 quantifies the intuition that, as the markup increases, the server that charges a higher price is less frequently patronized by customers. Part 2) shows that, for given markup, the occupancy rates of both servers decrease as either server raises its service speed. A higher service rate at one server exerts a direct impact on reducing the amount of time that this server spends serving customers. Meanwhile, faster service offers this server a competitive edge to attract customers that may

otherwise have chosen the other server, thus downsizing the pool of customers that go to its opponent and exerting an indirect impact on lowering the opponent's occupancy rate.

To further investigate price competition and, particularly, the effects of markup on the profits of the two servers, we define, for $i = 1,2$, and $B \in I$,

$$H_i(B) \equiv \frac{B-1}{\beta} \cdot \varepsilon_i^f (B-1), \tag{4.3}$$

where, $\varepsilon_i^f (B-1) = \dfrac{Bf_i(B) - (B-1)f_i(B-1)}{(B-1)f_i(B-1)} \cdot \dfrac{f_i(B-1)}{\left| f_i(B-1) - f_i(B) \right|}$ is the elasticity of the part of server i's expected revenue that is attributable to markup, i.e., the indicator for price competition, with respect to its occupancy rate, evaluated at $B-1$. That is, $\varepsilon_i^f (B-1)$ is the percentage change in $(B-1)f_i(B-1)$ relative to the percentage change in $f_i(B-1)$. We can think of $H_i(B)$ as the marginal change in server i's revenue due to a marginal change in its occupancy rate that is associated with an increase in markup from $B-1$ to B.

Consider server 1 first. For given p_2, as markup rises from $B-1$ to B, its profit change can be expressed by:

$$\Delta_1 = \left(p_2 + \frac{B}{\beta} \right) f_1(B)\mu_1 - \left(p_2 + \frac{B-1}{\beta} \right) f_1(B-1)\mu_1 = \mu_1 \left[f_1(B-1) - f_1(B) \right] \left(H_1(B) - p_2 \right),$$

where, by (4.3) and Lemma 4.1, $H_1(B) \equiv \dfrac{1}{\beta} \cdot \dfrac{Bf_1(B) - (B-1)f_1(B-1)}{f_1(B-1) - f_1(B)}$.

As server 1 raises markup from $B-1$ to B, its occupancy rate will drop and customers will turn to server 2. Server 1 incurs a loss of p_2 for marginal reduction in occupancy rate that is however accompanied by a marginal change of $H_1(B)$ in revenue.

If $H_1(B) < 0$, a higher markup lowers server 1's revenue. $H_1(B) - p_2 < 0$, and Δ_1 is negative so that server 1 will not raise markup to B from $B-1$, for any price at server 2. If $H_1(B) > 0$, as markup increases, server 1 earns higher revenue that may be able to make up its loss due to reduced occupancy rate and bring higher profit. For any $B \in I$, $H_1(B)$ sets a threshold for p_2 such that server 1 earns higher profit by raising markup from $B-1$ to B when $p_2 < H_1(B)$.

Similarly, for any given p_1, as markup rises from $B-1$ to B, server 2's price is lowered from $p_1 - \frac{B-1}{\beta}$ to $p_1 - \frac{B}{\beta}$, and the change in its profit is:

$$\Delta_2 = \left(p_1 - \frac{B}{\beta}\right) f_2(B)\mu_2 - \left(p_1 - \frac{B-1}{\beta}\right) f_2(B-1)\mu_2 = \mu_2 \left[f_2(B) - f_2(B-1)\right] \cdot (p_1 - H_2(B)),$$

where, by (4.3) and Lemma 4.1, $H_2(B) = \frac{1}{\beta} \cdot \frac{B f_2(B) - (B-1) f_2(B-1)}{f_2(B) - f_2(B-1)}$. Server 2 weighs the gain due to higher occupancy rate against the change in revenue as markup incrementally changes. If $H_2(B) < 0$, then $\Delta_2 > 0$ and it is beneficial for server 2 to raise markup from $B-1$ to B, for any price by server 1. If $H_2(B) > 0$, it is possible that server 1 reaps in lower profit by lowering price. $H_2(B)$ sets the threshold for p_1 such that server 2 earns higher profit by raising markup from $B-1$ to B whenever $p_1 > H_2(B)$.

In Lemma 4.2, we characterize the properties of $H_i(B)$.

Lemma 4.2: *Let $H_i(B)$ be as defined in (4.3). Then*

1. $H_1(B)$ decreases in B when $\mu_2 > \lambda$; decreases in B for $B \le 0$ but increases in

B for $B > 0$ when $\mu_2 < \lambda$.

2. *$H_2(B)$ increases in B when $\mu_1 > \lambda$; decreases in B for $B \le 0$ but increases in B for $B > 0$ when $\mu_1 < \lambda$.*

Lemma 4.2 shows that the sensitivities of $H_i(B)$ with respect to B are influenced by the opponents' service rates and reflects the interplay between the two modes of competition in price and service. The service rates at the servers build up the platform upon which they engage in price competition. Suppose server 2 has enough capacity to serve all customers, its speedy service makes server 1 receive relatively limited contribution from its own service capacity on its market share and resort to price as a vital weapon. Raising markup weakens server 1's price position; although it may still result in higher revenue, it is a less and less attractive strategic choice.

On the other hand, if server 2 is unable to serve the entire market on its own, server 1's service capacity can earn it decent market share when it charges the same price as server 2. Its price position affects the value of higher markups. When its price position is strong ($B < 0$), raising markup will deprive it of this advantage and the marginal contribution to revenue by higher markup will reduce. When server 1 is in a weak price position ($B > 0$) (and relies mainly on service capacity to earn market share), raising markups, while further weakening its price position, can be a lucrative option due to the higher profit margin thus generated, especially when server 2's price is so low that the marginal loss due to lowered market share is weak. By Lemma 4.2 and the fact that $H_i(0) > 0$, Corollary 1 is straightforward.

Corollary 4.1: *Let $H_i(B)$ be as defined in (4.3). Then*

1. *When* $\mu_2 > \lambda$, *there exists* $B_1 > 0$ *such that* $H_1(B) > 0$ *for* $B < B_1$ *and* $H_1(B) \le 0$ *otherwise; when* $\mu_2 < \lambda$, $H_1(B) > 0$.

2. *When* $\mu_1 > \lambda$, *there exists* $B_2 < 0$ *such that* $H_2(B) > 0$ *for* $B > B_2$ *and* $H_2(B) \le 0$ *otherwise; when* $\mu_1 < \lambda$, $H_2(B) > 0$.

Hence, server 1 has the potential to earn higher profit by imposing higher markup and weakening its price competitiveness, unless server 2 has high service capacity ($\mu_2 > \lambda$) and is highly price competitive ($B \ge B_1 > 0$) as well, i.e., server 2 is competitive in both price and service. Similarly, server 2 has the potential to earn higher profit by imposing higher markup to strengthen price competitiveness, unless server 1 is very fast ($\mu_1 > \lambda$) and is highly price competitive ($B < B_2 < 0$) as well, i.e., server 1 is competitive in both price and service.

We now examine server 1's profit function Π_1 for given p_2. Analysis of server 2's profit function for given server 1's price is in a similar vein. Π_1 assumes different functional forms when $p_1 \in [p_2, \overline{P}]$ and $p_1 \in [0, p_2]$, since server 1's price positions relative to server 2 differ in the two cases.

Case 1: $p_1 \in [p_2, \overline{P}]$

By (3.31) and (4.2), we can write the expected profit of server 1 as

$$\Pi_1(p_1 \mid p_2) = p_1 \cdot \left[\frac{(\lambda - \mu_2)\lambda^{B+2}}{\mu_1 \lambda^{B+2} + \mu_2^{B+2}(\lambda - \mu_1 - \mu_2)} \right] \cdot \mu_1 - c_1 \cdot \mu_1^2, \text{ for } p_1 \in [p_2, \overline{P}]. \qquad (4.4)$$

Lemma 4.3 shows the properties of $\Pi_1(p_1 \mid p_2)$ in this case.

Lemma 4.3: *Suppose* $p_1 \in [p_2, \overline{P}]$. *For given* $p_2 \ge 0$*:*

1. *When* $\mu_2 > \lambda$, $\Pi_1(p_1 \mid p_2)$ *increases in* p_1 *for* $p_2 \le p_1 \le p_2 + B_U$; *but decreases in* p_1 *for* $p_2 + B_U < p_1 \le \overline{P}$, *where* B_U *satisfies* $H_1(B_U + 1) < p_2 \le H_1(B_U)$.
2. *When* $\mu_2 < \lambda$, $\Pi_1(p_1 \mid p_2)$ *decreases in* p_1 *for* $p_2 \le p_1 \le p_2 + B_D$; *but increases in* p_1 *for* $p_2 + B_D < p_1 \le \overline{P}$, *where* B_D *satisfies* $H_1(B_D) < p_2 \le H_1(B_D + 1)$.

Now that server 1 is in weak price position, i.e., $B > 0$, the service capacity at server 2 influences how its profit change as it further weakens price competitiveness. When server 2 has sufficiently high service capacity ($\mu_2 > \lambda$), server 1 mainly competes in price with server 2. The price charged by server 2 determines the marginal loss in profit due to smaller market share associated with an increase in markup. When server 1's markup is low, raising it may bring about higher revenue to offset the loss due to lower occupancy rate and earns server 1 higher profit. When server 1's markup is high, the loss in market share will eat into the gain in revenue and lower its profit. B_U defined in part 1) of Lemma 4.3, for given p_2, establishes a threshold markup above which server 1's profit ceases to increase. On the other hand, when server 2 has insufficient service capacity, i.e., $\mu_2 < \lambda$, server 1's service capacity can win it decent share of market demand when it charges the same price as server 2. Raising price will cost it market share and may hurt its profit. The drop in profit will continue until its price is high enough to make the negative marginal loss in market share insubstantial; B_D in part 2) of Lemma 4.3 defines threshold markup in this case.

Case 2: $p_1 \in [0, p_2]$

By (3.42) and (4.2), when server 1 charges a lower price than server 2, its expected profit can be written as:

$$\Pi_1(p_1 \mid p_2) = p_1 \cdot \left[\frac{\mu_2 \lambda^{-B+2} + \lambda \mu_1^{-B+1}(\lambda - \mu_1 - \mu_2)}{\mu_2 \lambda^{-B+2} + \mu_1^{-B+2}(\lambda - \mu_1 - \mu_2)} \right] \cdot \mu_1 - c_1 \cdot \mu_1^2, \text{ for } p_1 \in [0, p_2]. \quad (4.5)$$

Lemma 4.4 shows the functional shape of $\Pi_1(p_1 \mid p_2)$ in this case.

Lemma 4.4: *For given* p_2, $\Pi_1(p_1 \mid p_2)$ *is increasing in* p_1 *for* $0 \le p_1 \le p_2 + B_L$*; but is decreasing in* p_1 *for* $p_2 + B_L < p_1 \le p_2$*, where* B_L *satisfies* $H_1(B_L + 1) < p_2 \le H_1(B_L)$.

Server 1 is price competitive by charging a lower price than server 2. In this situation, by Lemma 2, as server 1 raises price with associated increase in markup, a weakened price position stills bring about higher revenue but marginal contribution decreases. Provided that server 2's price, p_2, is not too high, the marginal increase in revenue can offset the marginal loss due to reduced market share to bring higher profit to server 1; otherwise server 1 suffers profit loss by raising markup. Lemma 4 identifies a threshold B_L for given p_2, such that server 1 earns higher profit as it raises markup till B_L and loses profit afterwards. The profit function then displays a quasi-concave pattern.

4.2 Best-Response Price

We derive the equilibrium prices by the two servers by use of the best-response price curve. The properties for the profits of the two servers help us derive the best-response price for one server, given the price charged by the other server. To ease the expressions, we introduce four thresholds values for markup B.

$$\overline{B}_0 \equiv \left\{ B \in I^+ : \Pi_1\left(\frac{B}{\beta} \mid 0\right) \geq \Pi_1\left(\frac{B+1}{\beta} \mid 0\right) \ \ \& \ \ \Pi_1\left(\frac{B}{\beta} \mid 0\right) \geq \Pi_1\left(\frac{B-1}{\beta} \mid 0\right) \right\} \tag{4.6}$$

$$\underline{B}_0 \equiv \left\{ B \in I^- : \Pi_2\left(-\frac{B}{\beta} \mid 0\right) \geq \Pi_2\left(-\frac{B-1}{\beta} \mid 0\right) \ \ \& \ \ \Pi_2\left(-\frac{B}{\beta} \mid 0\right) \geq \Pi_2\left(-\frac{B+1}{\beta} \mid 0\right) \right\} \tag{4.7}$$

$$\overline{B}_P \equiv \left\{ B \in I^+ : \Pi_2\left(\frac{\overline{P}-B}{\beta} \mid \overline{P}\right) \geq \Pi_2\left(\frac{\overline{P}-B+1}{\beta} \mid \overline{P}\right) \ \ \& \ \ \Pi_2\left(\frac{\overline{P}-B}{\beta} \mid \overline{P}\right) \geq \Pi_2\left(\frac{\overline{P}-B-1}{\beta} \mid \overline{P}\right) \right\} \tag{4.8}$$

$$\underline{B}_P \equiv \left\{ B \in I^- : \Pi_1\left(\frac{\overline{P}+B}{\beta} \mid \overline{P}\right) \geq \Pi_1\left(\frac{\overline{P}+B-1}{\beta} \mid \overline{P}\right) \ \ \& \ \ \Pi_1\left(\frac{\overline{P}+B}{\beta} \mid \overline{P}\right) \geq \Pi_1\left(\frac{\overline{P}+B+1}{\beta} \mid \overline{P}\right) \right\} \tag{4.9}$$

In words, $\overline{B}_0$ and $\underline{B}_P$ are the best markups for server 1 when sever 2 charges zero price and $\overline{P}$, respectively. As we will show later, they set the upper and lower bounds on the best-response markup by server 1. Similarly, $\underline{B}_0$ and $\overline{B}_P$ are the best markups for server 2 when server 1 charges zero price and $\overline{P}$, respectively; and set the upper and lower bounds on the best-response markup by server 2.

4.2.1 Best-Response Price of Server 1

The best-response price by server 1 for given p_2 is implied in Proposition 4.1 for different service speeds at server 2. We use the convention: $x \wedge y \equiv \min\{x, y\}$, $x \vee y \equiv \max\{x, y\}$.

Proposition 4.1: *Suppose that the two servers select service rates μ_1 and μ_2, for given price p_2 by server 2, the best-response price of server 1 is:*

1) When $\mu_2 > \lambda$,

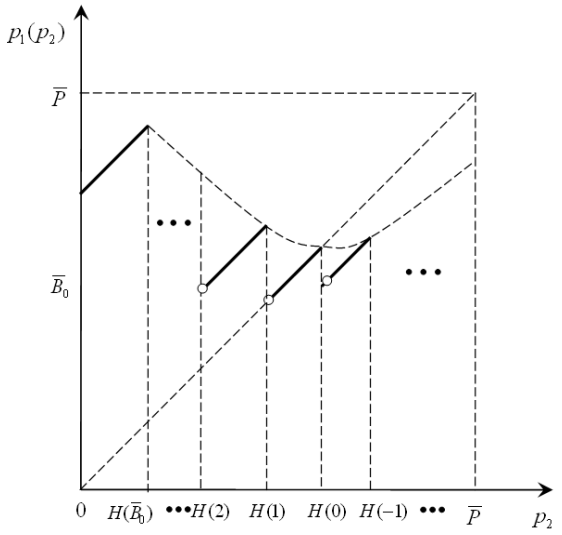
$$p_1(p_2)=\left(\left(p_2+\frac{B}{\beta}\right)\wedge\overline{P}\right)\vee 0,\quad if\ \ H_1(B+1)<p_2\le H_1(B),\ for\ B\in[\underline{B}_P,\overline{B}_0]$$

2) *When* $\mu_2<\lambda$,

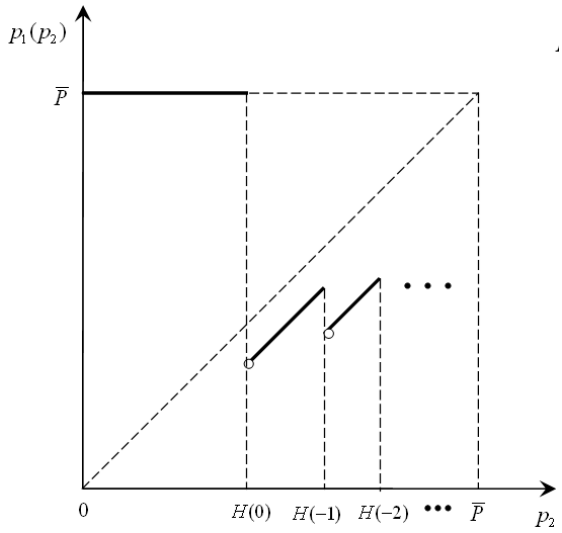
$$p_1(p_2)=\begin{cases}\overline{P} & 0<p_2\le H_1(0)\\ \vdots & \\ \left(p_2+\frac{B}{\beta}\right)\vee 0 & H_1(B+1)<p_2\le H_1(B)\qquad \underline{B}_P\le B<0\end{cases}$$

3) $p_1(p_2)-p_2$ *decreases in* p_2.

Figure 4.1 plots server 1's best-response curves shown in Proposition 1. Server 1's best-response price is piece-wise linear with p_2, but not continuous with the breakpoints at $H_1(B)$ for $B\in I$ due to discrete nature of markup. For server 1, server 2's price determines the marginal loss in profit due to reduced occupancy rate. Its best-response markup is the largest B at which marginal increase in revenue, $H_1(B)$, is larger than p_2. The maximum allowable price and zero enforce upper and lower bounds respectively on the value of its price. Suppose that server 2 is fast enough to serve all that customers. When p_2 is low, i.e., $p_2<H_1(0)$, the profit loss due to lowered occupancy rate associated with a weaker price position is low to server 1, who is safe to charge a higher price to enjoy higher unit revenue. As p_2 increases, the contribution to profit from market share is larger and the room for server 1 to raise price is smaller so that it will enforce lower markup. When p_2 is high, i.e., $p_2>H_1(0)$, server 1 tends to be price competitive by charging lower price than server 2.



a. $\mu_2 > \lambda$



b. $\mu_2 < \lambda$

Figure 4.1. Best Response Curve of Server 1

If server 2 does not have enough service capacity, server 1 has more flexibility to strike balance between the two competition instruments in service rate and price. When p_2 is low, i.e., $0 < p_2 \le H_1(0)$, the loss in profit due to lowered occupancy rate with higher markup is weak, server 1 tends to enforce the highest possible markup and relies more on service to compete in market. When server 2 charges a high price, i.e.,

$p_2 > H_1(0)$, the contribution to profit by occupancy rate is high, which induces server 1 to seek strong price position. As shown in Figure 4.1, server 1 will charge a lower price than server 2, with specific markup value determined by $H_1(B)$.

An important property of the best response price curve is that server 1's markup with respect to p_2, $p_1(p_2) - p_2$, decreases with p_2. It reaches the maximum value $\overline{B}_0$ when $p_2 = 0$, and the minimum value $\underline{B}_P$ when $p_2 = \overline{P}$. Server 1 tends to charge high if server 2's price is low, when the high profit margin exerts a stronger effect on its profit even at the cost of smaller market share. As server 2 bids higher, reducing markup can induce more customers to visit server 1, but its absolute service charge does not necessarily drop, and its profit will eventually increase. Therefore, as its opponent raises price, further competing on price is a less attractive decision by one server.

4.2.2 Best-Response Price of Server 2

The best-response price of server 2, for given price by server 1, can be derived in the similar vein as that for server 1. For completeness, we show the results in Proposition 4.2.

Proposition 4.2: *Suppose that the two servers select service rates μ_1 and μ_2, for given p_1 by server 1, the best response of server 2 is*

1) When $\mu_1 > \lambda$,

$$p_2(p_1) = \left(\left(p_1 - \frac{B}{\beta} \right) \wedge \overline{P} \right) \vee 0, \quad if \quad H_2(B) < p_1 \le H_2(B+1), \quad B \in [\underline{B}_0, \overline{B}_P].$$

2) *When* $\mu_1 < \lambda$,

$$p_2(p_1) = \begin{cases} \overline{P} & 0 < p_1 \le H_2(1) \\ \vdots & \\ (p_1 - \dfrac{B}{\beta}) \vee 0 & H_2(B) < p_1 \le H_2(B+1) \qquad 0 < B \le \overline{B}_P \end{cases}.$$

4.3 Price Equilibrium

By the best-response price functions by the two servers, we can establish the existence and the specific forms for the equilibrium prices, which are characterized in Theorem 4.1.

Theorem 4.1: *Suppose that the two servers set service rates at* μ_1 *and* μ_2 *respectively. The equilibrium of the price competition game,* (p_1^*, p_2^*), *is:*

1) *When* $\mu_1 > \lambda$ *and* $\mu_2 > \lambda$,

$$p_1^* = 0 \vee \left(\left(H_1(B) + \frac{B}{\beta} \right) \wedge H_2(B+1) \wedge \overline{P} \right) \quad \text{and} \quad p_2^* = p_1^* - \frac{B}{\beta} \quad , \quad \text{if} \quad Y(B) > 0 \quad \text{and}$$

$Y(B+1) < 0$ *for* $\underline{B} \le B \le \overline{B}$, *where* $Y(B) = H_1(B) + \dfrac{B}{\beta} - H_2(B)$, $\underline{B} \equiv (\underline{B}_0) \vee (\underline{B}_P)$, *and*

$\overline{B} \equiv \overline{B}_0 \wedge \overline{B}_P$.

2) *When* $\mu_1 > \lambda$ *and* $\mu_2 < \lambda$,

$p_1^* = \overline{P}$ *and* $p_2^* = \left(\overline{P} - \dfrac{\overline{B}_P}{\beta} \right) \vee 0$, *if* $\overline{P} - \overline{B}_P \le H_1(0)$ *and* $H_2(\overline{B}_P) < \overline{P} \le H_2(\overline{B}_P + 1)$.

3) *When* $\mu_1 < \lambda$ *and* $\mu_2 > \lambda$,

$p_1^* = \left(\overline{P} + \frac{\underline{B}_P}{\beta} \right) \vee 0$ *and* $p_2^* = \overline{P}$, *if* $\overline{P} + \frac{\underline{B}_P}{\beta} \le H_2(1)$ *and* $H_1(\underline{B}_P + 1) \le \overline{P} \le H_1(\underline{B}_P)$.

4) *When* $\mu_1 < \lambda$ *and* $\mu_2 < \lambda$,

a. $p_1^* = \overline{P}$ *and* $p_2^* = \left(\overline{P} - \frac{\overline{B}_P}{\beta} \right) \vee 0$, *if* $\overline{P} - \overline{B}_P \le H_1(0)$ *and* $H_2(\overline{B}_P) < \overline{P} \le H_2(\overline{B}_P + 1)$.

b. $p_1^* = p_2^* = \overline{P}$, *if* $\overline{P} \le H_2(1)$ *and* $\overline{P} \le H_1(0)$.

c. $p_1^* = \left(\overline{P} + \frac{\underline{B}_P}{\beta} \right) \vee 0$ *and* $p_2^* = \overline{P}$, *if* $\overline{P} + \frac{\underline{B}_P}{\beta} \le H_2(1)$ *and* $H_1(\underline{B}_P + 1) < \overline{P} \le H_1(\underline{B}_P)$.

Theorem 4.1 has embedded in it a relationship between price and service competitions, which we specify in Proposition 4.3.

Proposition 4.3: *In the equilibrium of price competition,* $p_i^* > p_j^*$ *only if* $\mu_i > \mu_j$, *for* $i, j = 1,2$ *and* $i \neq j$.

Proposition 3 establishes that it is not to the best interest of a server to outperform its opponent in both speed and price. A server, if it chooses to overcharge its opponent and lose price competitiveness, must be faster in service. On the other hand, if the two servers set the same service rates, then price competition does not bring additional value to them and they will each charge the maximum allowable price. Hence, price and service competitiveness do not co-exist, but instead are substitutable to each other.

As shown in Theorem 4.1, the specific price equilibrium is critically influenced by the service rates. To explore further on the effects of service rates, we consider two situations, one in which each server has enough capacity to serve the entire market, and the other in which at least one server does not have enough capacity.

4.3.1 Both Servers Have Enough Capacity $\mu_1 > \lambda,\ \mu_2 > \lambda$

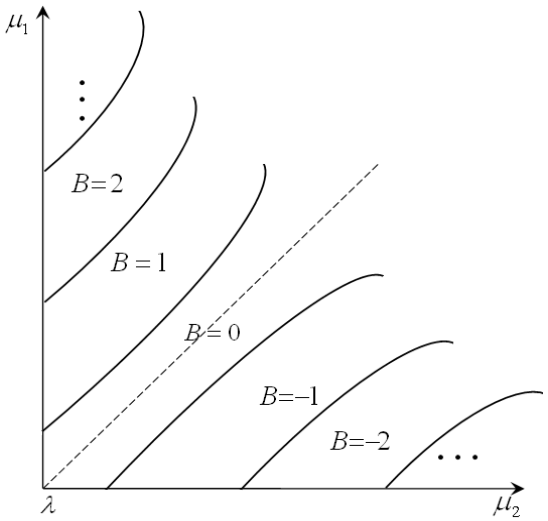
When both servers are fast in serving the customers, prices are vital for them to compete in the market. The equilibrium condition given in part i) of Theorem 1 involves a new function, $Y(B) = H_1(B) + \frac{B}{\beta} - H_2(B)$, to delimit regions for specific outcomes. We consider B as a continuous variable for the moment. For given markup B, we know that $H_1(B)$ is *the* price for p_2 to which server 1's best price is $p_1 = H_1(B) + \frac{B}{\beta}$; whereas $H_2(B)$ is *the* price for p_1 to which server 2 earns the best profit by accepting that markup. So, for a specific B to arise in equilibrium, both servers have to agree on the relative price positions it implies, and the two prices of $H_1(B) + \frac{B}{\beta}$ and $H_2(B)$ must be equal. Now that B is discrete, $Y(B)$, as the difference between the two prices, delimits the equilibrium partition. Lemma 4.5 shows some properties of $Y(B)$.

Lemma 4.5: *Let $Y(B)$ be defined in Theorem 1, then:*

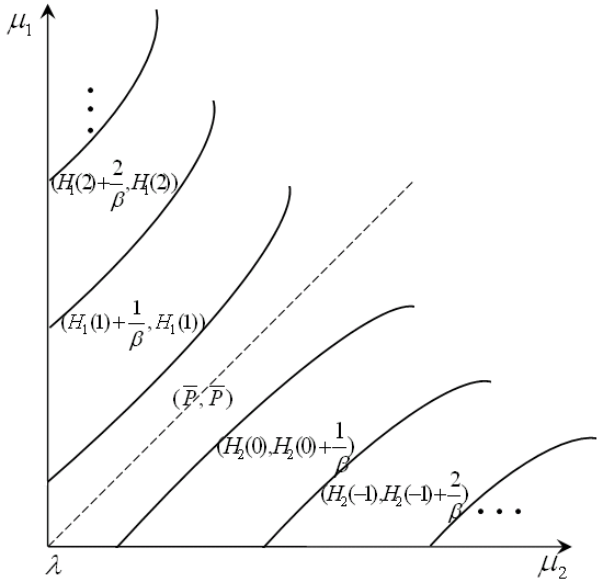
1. *It decreases in B.*
2. *When $\mu_1 > \lambda$ and $\mu_2 > \lambda$, it increases in μ_1, and decreases in μ_2.*

3. When $\mu_1 > \mu_2 > \lambda$, Given μ_2 and B, there exists $\bar{\mu}_1(\mu_2, B)$ at which $Y(B) = 0$; and $\bar{\mu}_1(\mu_2, B)$ increases in both μ_2 and B.

By parts 1) and 2) of Lemma 4.5 and the equilibrium characterization in Theorem 4.1, the service rates at the two servers influence the equilibrium markup. Server 2 tends to be more price competitive, i.e., the value of B increases, as the service at server 1 is faster, i.e., μ_1 increases, or its own service is slower, i.e., μ_2 decreases. On the other hand, server 1 will be more price competitive, i.e., the value of B decreases, as the service at server 2 becomes faster, or its own service slower. This echoes our earlier finding that one server is price competitive only when it is not service competitive.



a. Markups



b. Prices

Figure 4.2. Price Equilibrium, when $\mu_1 > \lambda$ and $\mu_2 > \lambda$

Consider the specific situation when server 1 is faster than server 2, i.e., $\mu_1 > \mu_2 > \lambda$. The analysis for the case $\mu_2 > \mu_1 > \lambda$ follows by symmetry. By part 3) of Lemma 4.5, we can rewrite the existence condition in part i) of Theorem 4.1 as $\bar{\mu}_1(\mu_2, B+1) > \mu_1 > \bar{\mu}_1(\mu_2, B)$. Figure 4.2 demonstrates price equilibrium. The discrete markup causes the equilibria to reside in strips, while the equilibrium pattern is symmetric around equal-service-rate line. For given service rate at server 2, μ_2, the equilibrium markup is $B \in I$ when the service rate at server 1 is in $(\bar{\mu}_1(\mu_2, B+1), \bar{\mu}_1(\mu_2, B))$. When the two servers have comparable service rates, they will each charge the maximum allowable price, i..e, further price competition is of no value to the servers. Given the service rate at server 2, as server 1 raises service rate above μ_2, server 2 charges lower price and the markup non-decreases, relegating server 1 to a weaker price position. Note that server 2 always charges threshold price $H_1(B)$ to

induce server 1 to accept markup B and charge $H_1(B)+\frac{B}{\beta}$. Similar observations apply when server 2 has higher service rate than server 1.

4.3.2 At Least One Server Lacks Enough Capacity

There are three possible cases: 1) $\mu_1 > \lambda$ and $\mu_2 < \lambda$; 2) $\mu_1 < \lambda$ and $\mu_2 > \lambda$; 3) $\mu_1 < \lambda$ and $\mu_2 < \lambda$. We first consider the case where server 1 has enough capacity, whereas server 2 does not. Server 2 will rely on lowering price to compete for market share, whereas server 1, thanks to its advantageous position in service speed, will weigh the values of engaging in one more layer of price competition. Part ii) of Theorem 1 shows, if equilibrium exists, server 1, by charging the highest possible price to enjoy the best unit revenue, will not turn to price competition and server 2 makes best-response price decision, with specific value dependent on the prevailing service rates, or the extent of service competition.

For the sake of partitioning equilibrium space, we define

$$Z(\mu_1,\mu_2,\overline{B}_P) = (H_1(0)+\overline{B}_P/\beta) \wedge H_2(\overline{B}_P+1). \tag{4.10}$$

The existence condition, as given in part (ii) of Theorem 1, can be rewritten as

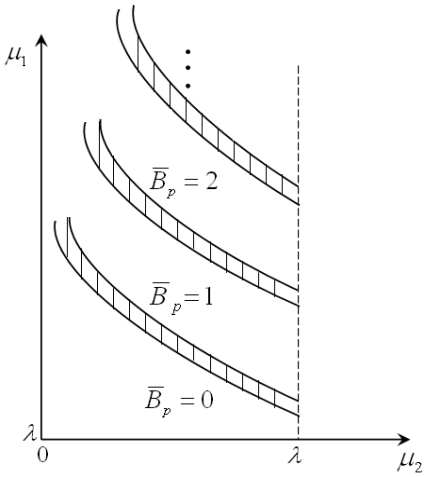
$$H_2(\overline{B}_P) < \overline{P} < Z(\mu_1,\mu_2,\overline{B}_P).$$

Recall that, by (4.8), $\overline{B}_P$ is server 2's optimal markup when server 1 charges $\overline{P}$. $Z(\mu_1,\mu_2,\overline{B}_P)$, as defined in (4.10), locates the specific value for $\overline{B}_P$ and identify existence of pure-strategy equilibrium.

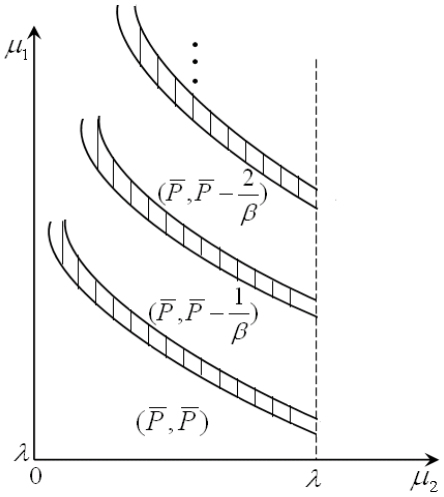
Lemma 4.6: *Let* $Z(\mu_1, \mu_2, \overline{B}_P)$ *be defined in (4.10).*

1. *It is decreasing in* μ_1 *and* μ_2.
2. *For given* $\overline{P}$ *and* μ_2, *there exists* $\mu_z(\overline{B}_P)$ *such that* $Z(\mu_1, \mu_2, \overline{B}_P) \geq \overline{P}$ *for* $\mu_1 \leq \mu_z(\overline{B}_P)$ *and* $Z(\mu_1, \mu_2, \overline{B}_P) < \overline{P}$ *otherwise; and* $\mu_G(\overline{B}_P)$ *such that* $H_2(\overline{B}_P) \geq \overline{P}$ *for* $\mu_1 \leq \mu_G(\overline{B}_P)$ *and* $H_2(\overline{B}_P) < \overline{P}$. $\mu_G(\overline{B}_P)$ *and* $\mu_z(\overline{B}_P)$ *decrease in* μ_2 *and increasing in* $\overline{B}_P$.

By Lemma 4.6, for given $\overline{P}$ and μ_2, price equilibrium in pure strategy exists when $\mu_G(\overline{B}_P) < \mu_1 \leq \mu_z(\overline{B}_P)$. Figure 4.3 demonstrates this condition as well as the equilibrium in this case. While equilibrium markup, if exists, appears in strips, it is likely that pure-strategy price equilibrium do not exist, which happens when the service rates fall in the shaded areas. Mixed-strategy price equilibria exist in these situations, but it is difficult to derive. So we choose not to pursue in that direction. Excluding the non-existence areas, the general trend is that, as server 1 speeds up, server 2 will choose to be more price competitive.



a. Markups



b. Prices

Figure 4.3. Price Equilibrium, when $\mu_1 > \lambda$ and $\mu_2 < \lambda$

The price equilibrium for case 2) is in a similar vein. In case 3) where neither server has enough capacity, both servers will pick the maximum service charge in any equilibrium. So they purely rely on service competition to share the market by charging the highest price to reap in the best unit revenue. Combining the results for all the scenarios, we can partition the whole space for the service rates by $\overline{\mu}_1(B, \mu_2)$, $\mu_z(B)$ and $\mu_G(B)$, with specific equilibrium for each area in Figure 4.4 and Figure 4.5.

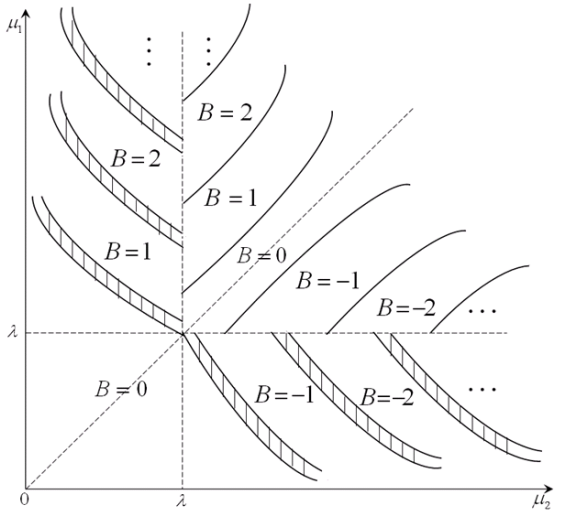


Figure 4.4. Markup, Given Service Rate

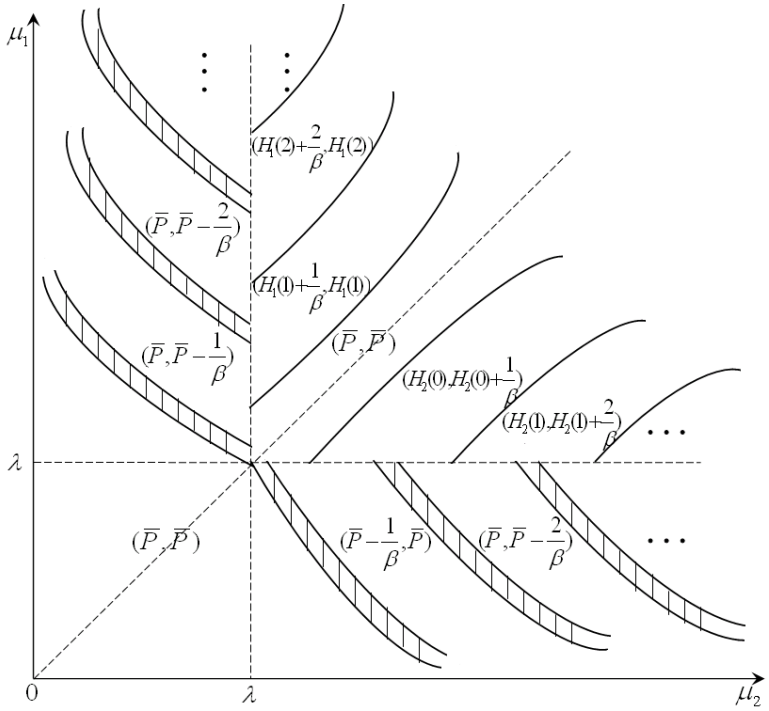


Figure 4.5. Equilibrium Price, Given Service Rate

CHAPTER 5

SERVICE RATE COMPETITION GAME

By the equilibrium for the price competition, derived in Chapter 4, the two servers engage in a game to invest in their respective service capacities. In this stage, servers choose service rates μ_1 and μ_2 simultaneously to maximize their profits. With the results of equilibrium prices and the equilibrium existence conditions in Chapter 4, we can solve the service rate competition game.

5.1 Symmetric System

To facilitate the equilibrium expressions, we consider a specific functional form for the capacity cost, $C_i(\mu_i) = c_i\mu_i^2$, and $c_i > 0$ for $i = 1, 2$. For a symmetric system, where the two servers incur the same marginal capacity costs, that is $c_1 = c_2 = c$. By substituting the equilibrium price derived in Chapter 4 into profit functions, we can completely characterize equilibrium service rates to obtain the system behavior, as shown in Proposition 5.1.

Proposition 5.1: *In a symmetric system where the two servers have the same investment costs, i.e.,* $c_1 = c_2 = c$, *in the equilibrium, if* $c < \frac{2\overline{P}}{\lambda}$, *then*

1. *The service rate at server* i, μ_i^*, *for* $i=1,2$, *is:*

$$\mu_1^* = \mu_2^* = \frac{\sqrt[3]{4K}}{12c} + \frac{\sqrt[3]{16}\lambda^2 c}{12\sqrt[3]{K}} - \frac{1}{6}\lambda, \tag{5.1}$$

where $a \equiv \sqrt{\overline{P}(27\overline{P} - 4c\lambda)}$ *and* $K \equiv \lambda^2 c^2[27\overline{P} - 2c\lambda + 3\sqrt{3}a]$. μ_i^* *decreases with capacity cost* c, *increases with the maximum price* $\overline{P}$ *and arrival rate* λ.

2. *The price charged by server* i, p_i^*, *for* $i=1,2$, *is* $p_1^* = p_2^* = \overline{P}$.

3. *The profit of server* i, Π_i^*, *for* $i=1,2$, *is:*

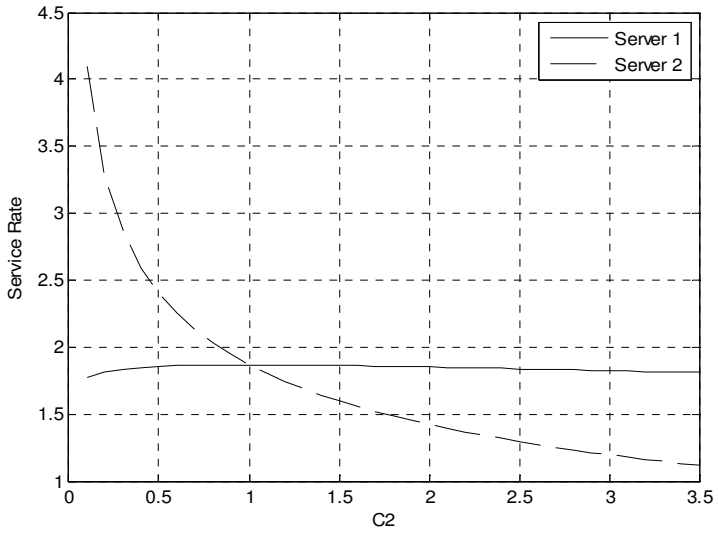
$$\Pi_1^* = \Pi_2^* = \frac{\overline{P}\lambda}{2} - c\left[\frac{\sqrt[3]{4K}}{12c} + \frac{\sqrt[3]{16}\lambda^2 c}{12\sqrt[3]{K}} - \frac{1}{6}\lambda\right]^2. \tag{5.2}$$

The equilibrium is symmetric across servers in symmetric systems, and the resultant aggregate service rate is higher than arrival rate. Note that, by charging maximum allowable prices, the two servers do not engage in price competition, forfeiting price as a weapon to compete for customers, but rely on service speed to share the market. Lower capacity cost allows servers to build up higher service rates. Now that each server charges $\overline{P}$, as its value increases, the unit revenue from a completed service is higher, which makes the servers able to afford the increase in capacity cost to build higher service capacities. Further, as more customers arrive, the servers speed up their services.

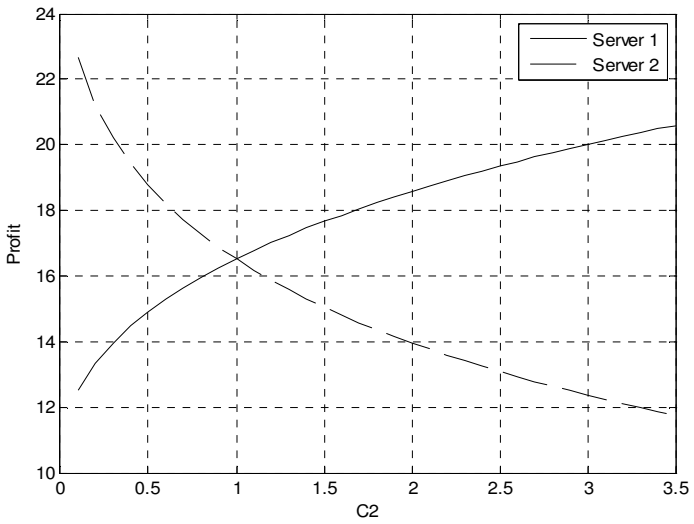
5.2 Asymmetric System

For the asymmetric system, two servers' margin costs are not identical, that is $c_1 \neq c_2$. It is hard to derive closed-form analytical solutions of the equilibrium service rates. However, we can always conduct comprehensive set of numerical studies to explore further insights.

We first investigate the effects of capacity costs on the prices, service rates, and profits, by fixing server 1's capacity cost and varying that at server 2. Note that, opposite to what we observed in symmetric system, the two servers may engage in price competition to complement service competition in the asymmetric system, depending on how price influences customers' utility, captured by marginal disutility β.



a) μ_1^* **vs** c_2



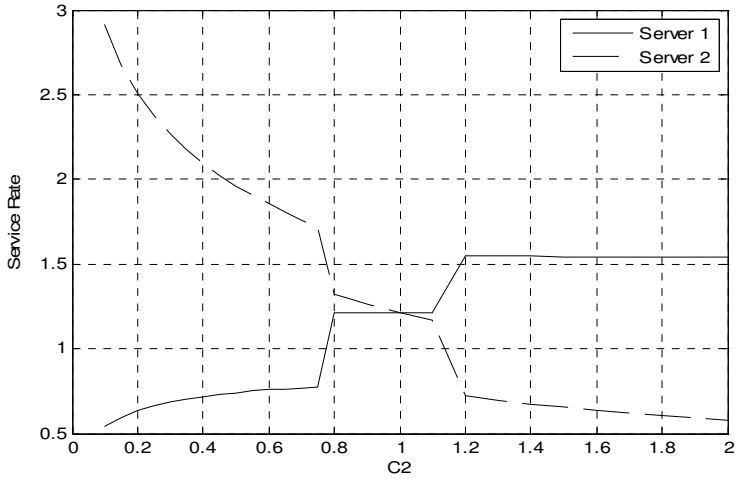
b) Π_i^* **vs** c_2

Figure 5.1. Effects of Capacity Costs, when Customers are Price Insensitive

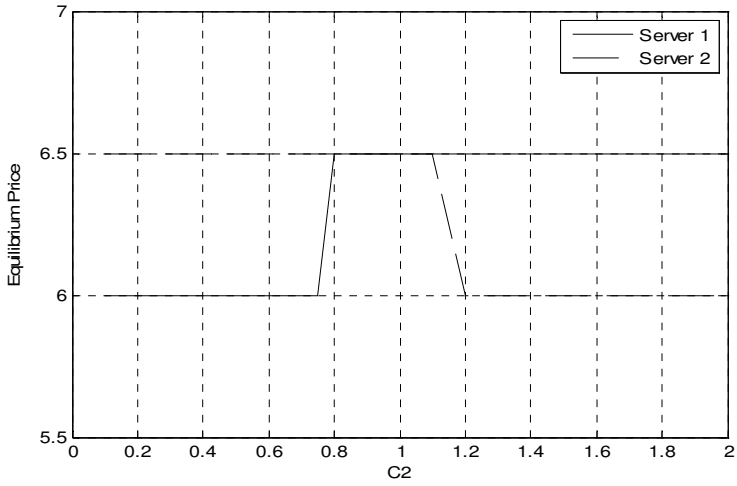
Figure 5.1 reveals typical outcomes when marginal disutility is low, $\beta = 0.2$. The other parameters for the data in this figure are $\lambda = 2.0$, $c_1 = 1.0$, and $\overline{P} = 20$. When customer is insensitive to price, the servers choose not to compete in price, but resort to the highest allowable price $p_1^* = p_2^* = \overline{P}$. It is intuitive that the server with higher capacity cost invests in lower service rate, i.e., $\mu_i^* \geq \mu_j^*$ iff $c_j \geq c_i$, for $i, j = 1,2$ and $i \neq j$. As shown in Figure 5.1, as server 2' capacity cost, c_2, increases, it lowers service rate. The effect of c_2 on server 1's service is, however, not monotone. As c_2 increases in $[0, c_1)$ when server 1 incurs higher capacity cost, server 1 has tendencies to set higher service rate as its cost disadvantage shrinks, until its capacity cost is equal to that at the opponent. As c_2 increases in $[c_1, +\infty)$, server 1 has lower capacity cost and can outperform its opponent in service speed. However, the marginal contribution to revenue by further increasing service rate may not be high enough to offset the increase in

capacity cost. Figure 5.1.a shows that server 1 tends to lower service rate, provided it maintains faster service speed. Moreover, as shown in Figure 5.1.b, server 2's capacity cost increases, server 1's profit strictly increases while server 2's profit decreases; the two servers earn the same profit when their capacity costs are identical.

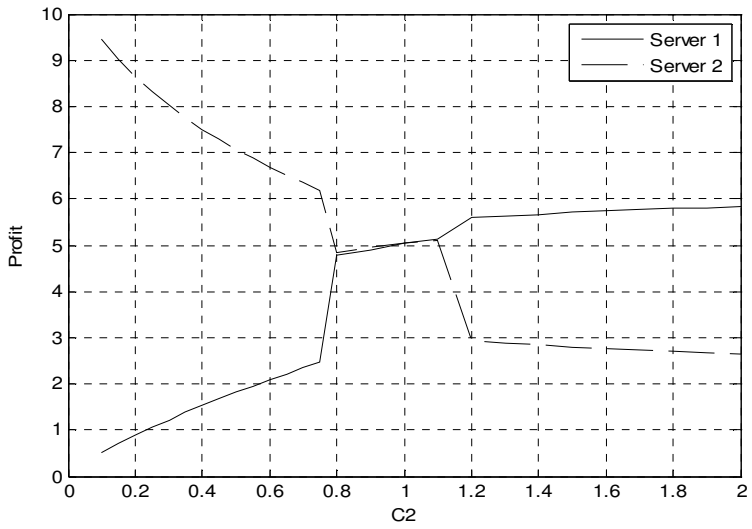
We next examine the situation when customers are sensitive to prices. Figure 5.2 shows the typical outcome when the marginal disutility is large. The parameters to draw this figure are $\lambda = 2.0$, $c_1 = 1.0$, $\beta = 2$, and $\overline{P} = 6.5$. Part a. and b. show the effects of server 2's capacity cost on service rates and prices at the servers. The two servers make similar service capacity decisions and do not quite engage in price competition when their capacity costs are comparable to each other. When c_2 is low enough, server 2 has the cost advantage to build higher service capacity and overprice server 1, which poses itself at a weak price position. When c_2 is high enough, however, server 2 builds lower service rate and underprices server 1. Hence, when their capacity costs differ substantially from each other, the servers rely on substitutable competitive weapons: the one with cost advantage relies on fast service and the other on low price. This echoes our finding that the competitiveness in price and service do not co-exist in the service system.



a) μ_i^* **vs** c_2



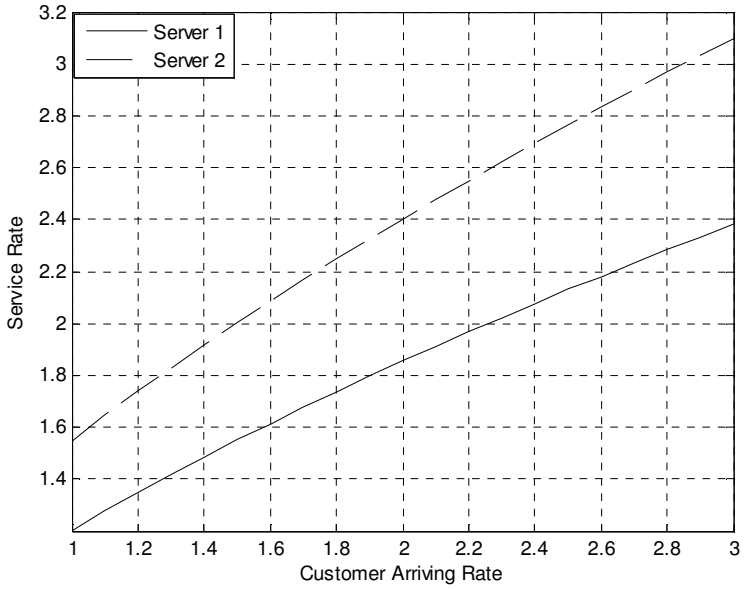
b) p_i^* **vs** c_2



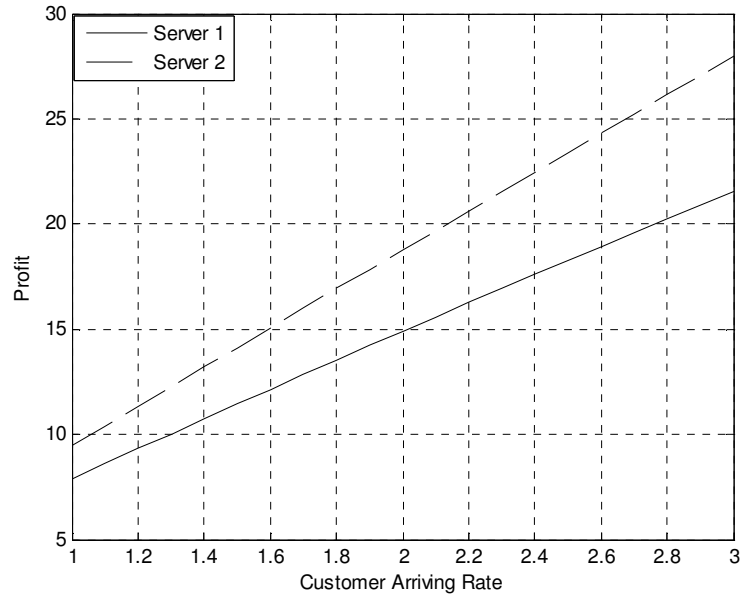
c) Π_i^* **vs** c_2

Figure 5.2. Effects of Capacity Costs, when Customers are Price Sensitive

The population size of customers also affects the servers' competition outcomes. To explore its effects in the asymmetric service system, we vary customer arrival rate to examine the effects on servers' behaviors. Intuitively, the servers invest more in service capacities as more customers arrive and higher customer arrival rate brings about higher profits to them, which are demonstrated in Figure 5.3, where the parameters to generate the data are $\overline{P} = 2.0$, $c_1 = 1.0$, $c_2 = 0.5$ and $\beta = 0.2$.



a) μ_1^* **vs** c_2

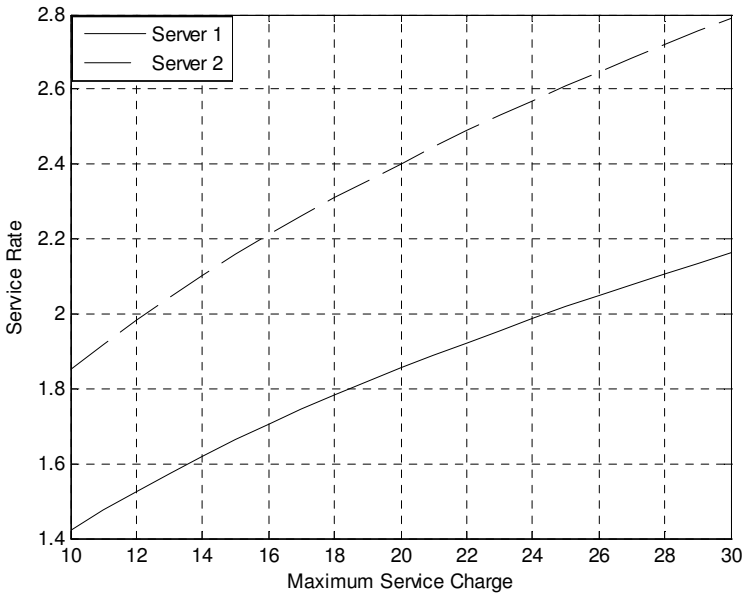


b) Π_i^* **vs** c_2

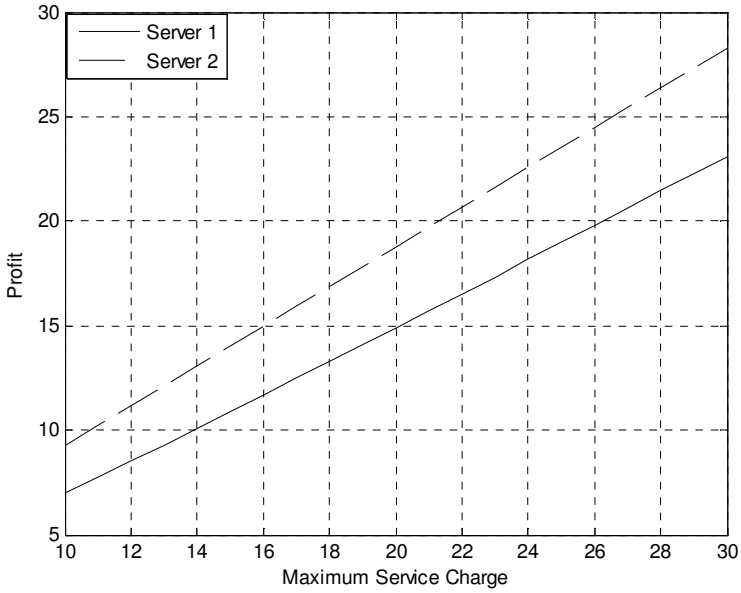
Figure 5.3. The Effect of Customer Arrival Rate

Figure 5.4 illustrates how service rates changes with the maximum service charge, $\overline{P}$, based on the problem instance: $\lambda = 2.0$, $c_1 = 1.0$, $c_2 = 0.5$ and $\beta = 0.2$. As

we discussed before, the maximum charge $\overline{P}$ reflects the customers' service valuation. While server 2 invests more in service capacity and earns higher profit than server 1 due to its lower capacity cost, both servers make more capacity investments and earn higher profits when their services are more valuable to customers. A plausible explanation is that: as the customers value their services more, the servers have stronger incentives to raise service charges. In the competitive setting, however, a higher service charge may cost a server its market share, so servers each make more investment in capacity to provide faster service with the mounted investment cost paid off by the higher service charge.



a) μ_1^* **vs** c_2



b) Π_i^* **vs** c_2

Figure 5.4. The Effects of Maximum Service Charged

CHAPTER 6

PERFORMANCE COMPARISIONS

Our approach to model customer behaviors in the service system is different from those in the existing literature. Customers in our model are uninformed of the service providers' service rates and are "pragmatic" decision makers to select service provider based on observed queue lengths; while customers in the past papers are informational rich and sophisticated enough to derive expected waiting times, whereby they can choose service providers more intelligently. Different levels of information held by customers with their corresponding service selection behaviors critically influence the service providers' service investments and price decisions. Moreover, we assumed that customers select between the two service providers in real time, while several models in the existing literature have customers allocated to service providers by a system manager under certain policy not influenced by the actual system states.

In this section, to examine the effects of customer behaviors on system performance, we compare our findings to those in Li and Lee (1994), hereinafter referred to as LL (1994), So (2000), Bell and Stidham (1983), hereinafter referred to as BS (1983), and Gilbert and Weng (1998), hereinafter referred to as GW (1998). The customer selection criterion in So (2000), BS (1983), and GW (1998) are captured by allocation formulas, while those in LL (1994) and this paper by utility functions.

We first provide the relevant results before making comparisons. LL (1994) assumes customers select service providers based on price and expected waiting time that depends on queue length and service rate; and the utility that one customer obtains by selecting service provider i takes the form:

$$U_i = \begin{cases} -r(n_i+1)/\mu_i - \beta p_i & if\ p_i \le \overline{P} \\ -\infty & Otherwise \end{cases} \text{ for } i = 1,2, \qquad (6.1)$$

where p_i is the price charged by server i, $\overline{P}$ the customers' reservation price (similar to the maximum allowable price in our model), β and r are respectively the marginal disutilities of price and expected waiting time. The resulting utility function is similar in nature to the one we use, as given in (3.1). However, while LL (1994) also makes use of the observable queue lengths, it differs fundamentally from our model by assuming customers are aware of the service rates and incorporating that information into the utility function.

With service rate information available to all the customers, instead of actual system state such as queue length, some earlier models apply state-independent allocation policies to distribute customers between the service providers to achieve target long-run performances. So (2000) assumes that customers select service providers by expected lead-time and price, and presents the demand for each service provider by a multiplicative competitive interaction (MCI) model:

$$\lambda_i = \lambda\left(\frac{p_i^{-\beta} t_i^{-1}}{p_i^{-\beta} t_i^{-1} + p_j^{-\beta} t_j^{-1}}\right) \text{ for } i = 1,2, \qquad (6.2)$$

where p_i is server i's price, t_i is the expected lead time and β is price elasticity.

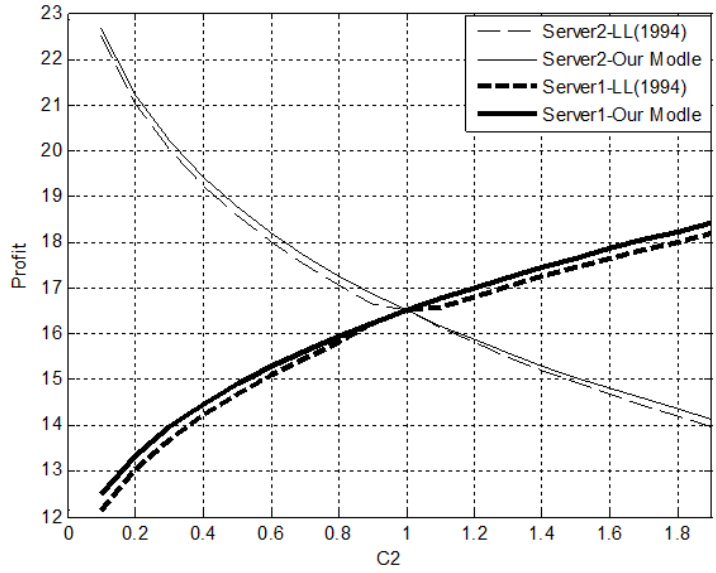
In other papers, customers do not select service providers; system managers allocate them according to certain allocation rule that depends on service rates, while price is usually not treated as a decision factor. BS (1983) designs a rule to minimize the customer's expected lead time under which the demand for each service provider, λ_i, takes the form:

$$\lambda_i = \mu_i - (\frac{\sqrt{\mu_i}}{\sqrt{\mu_i} + \sqrt{\mu_j}})(\mu_i + \mu_j - \lambda) \text{ for } i = 1, 2 . \quad (6.3)$$

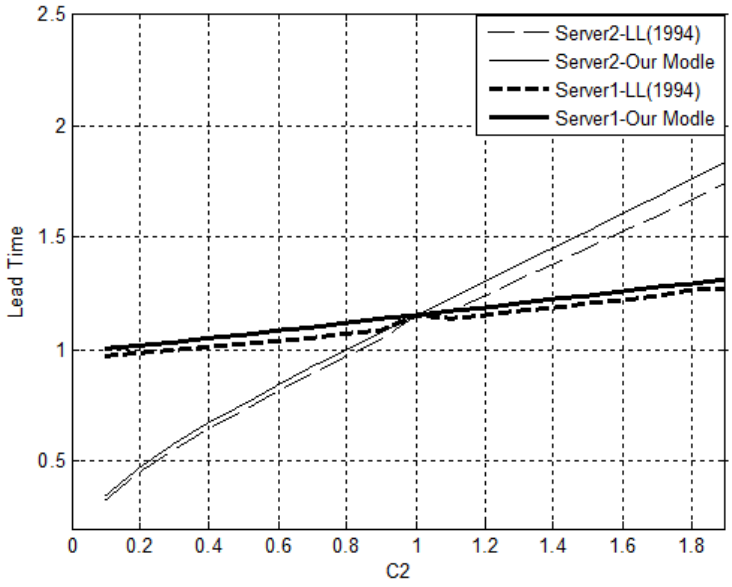
GW (1998) assumes that a system manager aims to equalize the customers' expected waiting times at the two service providers, and the demand to each of them is

$$\lambda_i = \begin{cases} \lambda & \lambda + \mu_j \le \mu_i \\ (\mu_i - \frac{1}{2}(\mu_i + \mu_j - \lambda))^+ & otherwise \end{cases} \text{ for } i, j = 1, 2 \text{ and } i \neq j \quad (6.4)$$

To screen out the effects of the availability of service rate information to customers on system performance, we compare our model with LL (1994), with real-time service selection assumed in both models. Figure 6.1 displays the typical outcomes where the service providers incur different capacity costs. The base parameter values for Figure 6.1 are $c_1 = 1.0$, $\lambda = 2.0$, $\overline{P} = 20$, $r = 1$, and $\beta = 1.0$.



a) Π_i^* **vs** c_2



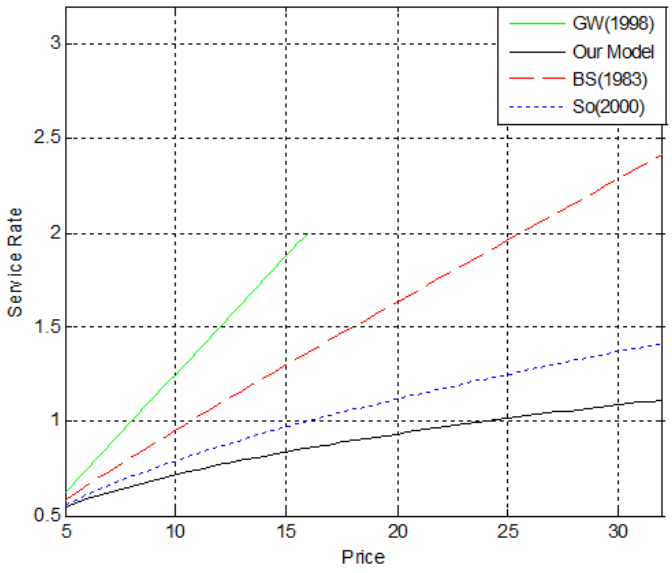
b) t_i^* **vs** c_2

Figure 6.1. Effects of Availability of Service Rate Information

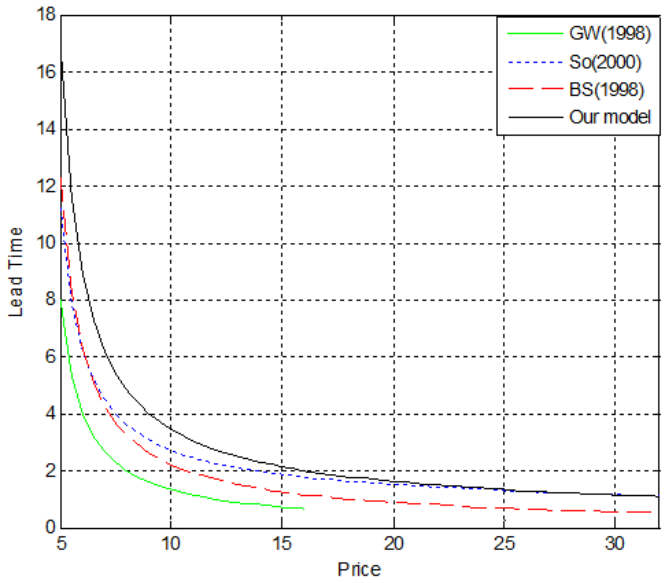
Observe that the service providers invest in lower service rates but earn higher profits, while customers spend longer time waiting in line in our model than in LL (1994). It is mainly attributed to the different information stands held by customers in

the two models. The customers in our model are unaware of the service rates and their pragmatic service selections rely on actual queue lengths and prices. Knowing this, the service providers do not have strong incentives to invest in service and would let the random nature do its work. On the contrary, the customers, as modeled in LL (1994), are aware of the service rates and use them directly in service selections. As a result, each service provider is prompted to raise investment; but the higher capacity cost is not paid off by revenue, and its profit suffers.

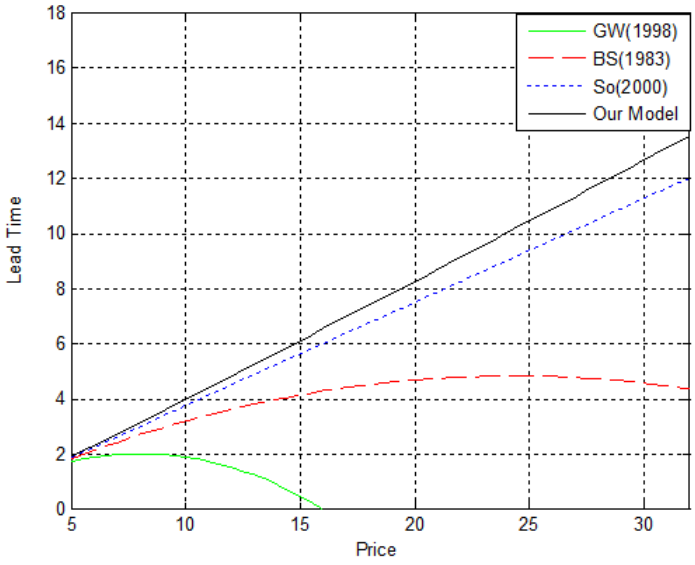
To study the effects of customers' real-time service selection on system performance, we compare our model to So (2000), GW (1998), and BS (1983). Figure 6.2 displays typical comparison outcomes for the symmetric settings with base parameters $c = 2.0$, $\lambda = 1.0$, and $b = 1.0$.



a) μ^* **vs** $\overline{P}$



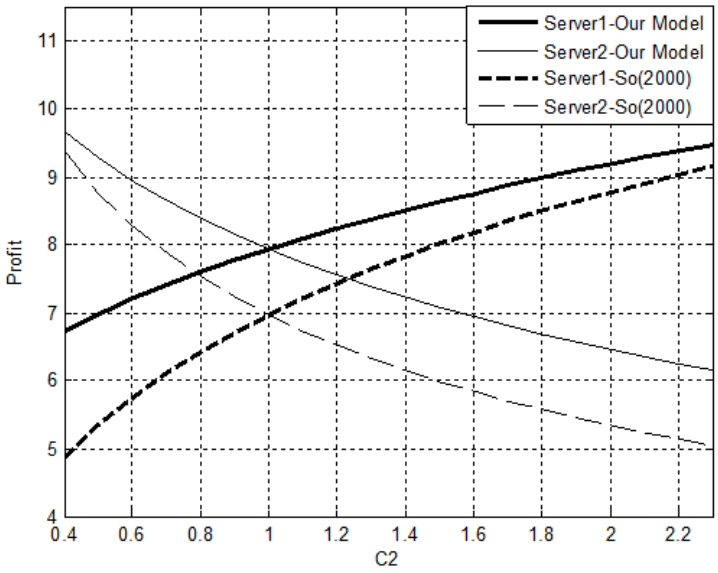
b) t^* **vs** $\overline{P}$



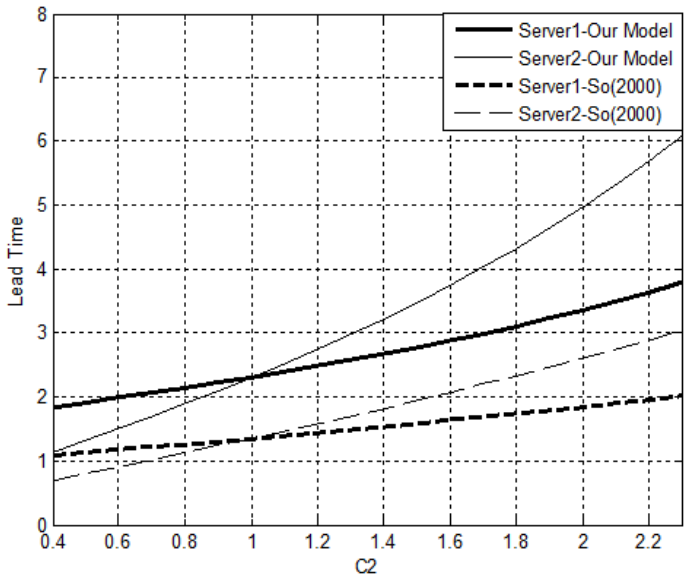
c) Π^* **vs** $\overline{P}$

Figure 6.2. Effects of Customers' Service Selection Behavior in Symmetric Systems

Figure 6.3 displays the comparison results of asymmetric systems with base parameter values $c_1 = 1.0$, $\lambda = 1.0$, $\overline{P} = 20$, and $\beta = 1.0$. Since BS (1983) and GW (1998) focus on symmetric systems, only So (2000) is examined here.



a) Π_i^* **vs** c_2



b) t_i^* **vs** c_2

Figure 6.3. Effects of Customers' Service Selection Behavior in Asymmetric Systems

In both symmetric and asymmetric systems, service providers invest less in service rates, earn higher profits, but customers spend more time waiting in line in our model than in either of the selected models. Notice that the differences, as revealed

between Figures 8 and 10, are attributed to both the availability of service rate information to customers and their service selection behaviors. The performance gaps in Figure 6.3 are notably wider than those in Figure 6.1, which implies that, besides availability of service rate information, customers' real-time service selection exerts an additional effect to lower the service provider's capacity investments and lengthen the customers' stay in system. On top of the differences originating from information availability, customers' state-dependent service selection upon arrival and jockeying between queues, as assumed in our model, lowers the chance for waiting customers and idle service providers to co-exist, particularly when service providers charge comparable prices. As a consequence, service providers that face customers making real-time service selections expect to achieve a higher capacity utilization than those whose customers commit to one service provider in advance. So they further lower capacity investments in service and detain customers longer, but reap in higher profits.

CHAPTER 7

CONCLUDING REMARKS

In this work, we have analyzed a service system consisting of two servers each with its own queue. Customers' utility from one server is influenced by its service charge and the number of customers waiting in its queue. An arrival customer chooses to join the queue for a server that gives higher utility. A customer waiting for service in one queue can change to the other queue as long as such a move can bring strictly higher utility to him. We call these naïve or pragmatic behaviors by customers who are unaware of service rate information. To our best knowledge, this is the first paper to model pragmatic consumer behaviors in a competitive service system.

For given service rates and charges at the two servers, we derive the steady-state distributions for the queue lengths in closed forms, taking into consideration customers' information stand and server-selection behaviors. Then we explore the servers' joint competition in service and price, and study the interplay between these two modes of competition. We have conducted a complete analysis for a symmetric system, and turn to a set of numerical studies to gain more understandings for an asymmetric system. Among the interesting results, we find that neither server will try to hold competitive advantages in both price and service. In a symmetric system, the two servers will forfeit price as a competitive instrument and rely solely on service speeds to compete in the market. We have also compared the results in our model to those in existing literature, which assumes that the customers are aware of the servers' service rates to make

sophisticated server selections; and found that the service providers invest less in service rates and earn higher profits but the customers spend longer time in line when the customers lack server information and make pragmatic decisions as modeled in our work.

Several directions seem promising for future research. One extension is to study different kind of capacity cost functions. In this paper, we do not consider the operational cost. In reality, however, operation cost exists when servers provide service and usually it is not linear with demand. Therefore, the operational cost and scale of economy can be considered to explore further insights. We only study the duopoly setting so a second extension would be considering the generalization of multiple servers' competition (more than 2 servers).

APPENDIX: Mathematical Proofs

Proof of Lemma 4.1:

Part 1)

We only prove the monotonicity of $f_1(B)$. The proof of $f_2(B)$ is similar and we omit the proof process here.

Let $A=\lambda-\mu_1-\mu_2<0$, $m_1=\frac{\lambda}{\mu_1}$, $m_2=\frac{\lambda}{\mu_2}$, $D=(m_1m_2-m_1-m_2)<0$.

For $B>0$, $f_1(B)=\frac{\lambda^{B+2}(\lambda-\mu_2)}{\mu_1\lambda^{B+2}+\mu_2^{B+2}A}>0$, and $f_1(B-1)=\frac{\lambda^{B+1}(\lambda-\mu_2)}{\mu_1\lambda^{B+1}+\mu_2^{B+1}A}>0$.

So we have $\frac{\lambda-\mu_2}{\mu_1\lambda^{B+2}+\mu_2^{B+2}A}>0$, and $\frac{\lambda-\mu_2}{\mu_1\lambda^{B+1}+\mu_2^{B+1}A}>0$.

$$f_1(B)-f_1(B-1)=\frac{\lambda^{B+1}\mu_2^{B+1}(\lambda-\mu_2)^2A}{(\mu_1\lambda^{B+2}+\mu_2^{B+2}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}<0.$$

For $B<0$, $f_1(B)=\frac{\mu_2\lambda^{-B+2}+\lambda\mu_1^{-B+1}A}{\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A}>0$, and $f_1(B-1)=\frac{\mu_2\lambda^{-B+3}+\lambda\mu_1^{-B+3}A}{\mu_2\lambda^{-B+3}+\mu_1^{-B+3}A}>0$.

$$f_1(B)-f_1(B-1)=\frac{A\mu_2(\lambda-\mu_1)^2\mu_1^{-B+1}\lambda^{-B+2}}{(\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A)(\mu_2\lambda^{-B+3}+\mu_1^{-B+3}A)}<0.$$

$$f_1(0)-f_1(1)=\frac{(-A)\mu_2^2\lambda(\mu_1\mu_2+2\lambda\mu_1+\lambda\mu_2+\mu_2^2+\lambda^2)}{(\mu_1\lambda^2+\lambda\mu_1\mu_2+\mu_2^3+\mu_1\mu_2^2)(\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2)}>0.$$

$$f_1(-1)-f_1(0)=\frac{(-A)\mu_1\mu_2\lambda(\mu_1\mu_2+2\lambda\mu_2+\lambda\mu_1+\mu_1^2+\lambda^2)}{(\mu_2\lambda^2+\lambda\mu_1\mu_2+\mu_1^3+\mu_2\mu_1^2)(\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2)}>0.$$

So $f_1(B)$ decreases with B.

Part 2)

When $B>0$,

$$f_1(B)=\frac{\lambda^{B+2}(\lambda-\mu_2)}{\mu_1\lambda^{B+2}+\mu_2^{B+2}A}=\frac{\lambda^{B+2}(\lambda-\mu_2)}{\mu_1(\lambda^{B+2}-\mu_2^{B+2})+\mu_2^{B+2}(\lambda-\mu_2)}=\frac{\lambda^{B+2}}{\mu_1(\lambda^{B+1}+\cdots+\mu_2^{B+1})+\mu_2^{B+2}}$$

It is easy to see that $f_1(B)$ decreases with μ_1.

Also we can rewrite that $f_1(B)=\frac{m_1m_2^{B+2}(m_2-1)}{(m_2^{B+3}+D)}>0$, so $\frac{(m_2-1)}{(m_2^n+D)}>0$.

Therefore, $\frac{\partial f_1(B)}{\partial m_2}=\frac{m_1(m_2-1)m_2^{B+1}[(m_2^{B+2}+D)+(m_2^{B+1}+D)\cdots(m_2+D)]}{(m_2^{B+3}+D)^2}>0$. And

we can know that $f_1(B)$ decreases with μ_2.

With the same logic and method, we can prove that when $B<0$, $f_1(B)$ also

decreases with μ_1 and μ_2.

When *B=0*, $f_1(0)=\frac{\lambda(\mu_2^2+\mu_1\mu_2+\lambda\mu_1)}{\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2}$.

$\frac{\partial f_1(0)}{\partial\mu_1}=\frac{-\lambda(\lambda+2\mu_2)(2\mu_2^2\mu_1+\mu_1^2\mu_2+\mu_2^3+\mu_1^2\lambda-\mu_2^2\lambda)}{(\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2)^2}$. Since $\lambda<\mu_1+\mu_2$, we can

know that $\mu_2^2\lambda<\mu_2^3+\mu_1\mu_2^2$. So $2\mu_2^2\mu_1+\mu_1^2\mu_2+\mu_2^3+\mu_1^2\lambda-\mu_2^2\lambda>0$ and $\frac{\partial f_1(0)}{\partial\mu_1}<0$.

Also, $\frac{\partial f_1(0)}{\partial\mu_2}=\frac{-\lambda(\lambda+2\mu_1)(2\mu_1^2\mu_2+\mu_2^2\mu_1+\mu_1^3+\mu_2^2\lambda-\mu_1^2\lambda)}{(\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2)^2}<0$. Thus, we can

conclude that $f_1(B)$ decreases with μ_1 and μ_2. The proof for $f_2(B)$ is similar so

we omit the process here.

Proof of Lemma 4.2

Part 1)

With $A=\lambda-\mu_1-\mu_2<0$, define $k(0)=\frac{\lambda^2(\lambda-\mu_2)}{\mu_1\lambda^2+\mu_2^2A}$, and $l(0)=\frac{\lambda^2\mu_2+\lambda\mu_1A}{\mu_1\lambda^2+\mu_2^2A}$.

$$H_1(B+1)-H_1(B)=\frac{1}{\beta}\left\{\frac{f_1(B)}{f_1(B)-f_1(B+1)}-B-1-[\frac{f_1(B-1)}{f_1(B-1)-f_1(B)}-B]\right\}$$

$$=\frac{1}{\beta}\left\{\frac{2f_1(B+1)f_1(B-1)-f_1(B+1)f_1(B)-f_1(B-1)f_1(B)}{[f_1(B)-f_1(B+1)][f_1(B-1)-f_1(B)]}\right\}$$

$$\equiv\frac{1}{\beta}\frac{N}{D},$$

where $N=2f_1(B+1)f_1(B-1)-f_1(B+1)f_1(B)-f_1(B-1)f_1(B)$, and

$D=[f_1(B)-f_1(B+1)][f_1(B-1)-f_1(B)]$.

In the following, we consider two cases.

Case 1) $B\geq 0$

$$N=\frac{2\lambda^{2B+4}(\lambda-\mu_2)^2}{(\mu_1\lambda^{B+3}+\mu_2^{B+3}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}-\frac{\lambda^{B+2}(\lambda-\mu_2)^2}{(\mu_1\lambda^{B+2}+\mu_2^{B+2}A)}(\frac{\lambda^{B+3}}{\mu_1\lambda^{B+3}+\mu_2^{B+3}A}+\frac{\lambda^{B+1}}{\mu_1\lambda^{B+1}+\mu_2^{B+1}A})$$

$$=\frac{2\lambda^{2B+4}(\lambda-\mu_2)^2}{(\mu_1\lambda^{B+3}+\mu_2^{B+3}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}-\frac{\lambda^{2B+3}(\lambda-\mu_2)^2}{(\mu_1\lambda^{B+2}+\mu_2^{B+2}A)}[\frac{2\mu_1\lambda^{B+3}+\mu_2^{B+1}(\mu_2^2+\lambda^2)A}{(\mu_1\lambda^{B+3}+\mu_2^{B+3}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}]$$

$$=\frac{2(-A)\lambda^{2B+3}\mu_2^{B+1}(\lambda-\mu_2)^4}{(\mu_1\lambda^{B+3}+\mu_2^{B+3}A)(\mu_1\lambda^{B+2}+\mu_2^{B+2}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}$$

By the proofs for Lemma 1, $\frac{(\lambda-\mu_2)^3}{(\mu_1\lambda^{B+3}+\mu_2^{B+3}A)(\mu_1\lambda^{B+2}+\mu_2^{B+2}A)(\mu_1\lambda^{B+1}+\mu_2^{B+1}A)}>0$.

It is easy to verify that $N>0$ if $\mu_2<\lambda$; but $N\leq 0$ otherwise.

Since $D=[f_1(B)-f_1(B+1)][f_1(B-1)-f_1(B)]>0$, we can then conclude that, for $B>0$, $H_1(B+1)<H_1(B)$ if $\mu_2>\lambda$; and $H_1(B+1)\geq H_1(B)$, otherwise.

Moreover, $H_1(1)-H_1(0)=\frac{1}{\beta}\left\{\frac{2f_1(1)f_1(-1)-f_1(0)(f_1(1)+f_1(-1))}{[f_1(0)-f_1(1)][f_1(-1)-f_1(0)]}\right\}$.

We can show that:

$$f_1(0)-k(0)=\frac{-\lambda\mu_2^2A^2}{(\mu_2^2+\mu_1\mu_2+\mu_1\lambda)(\lambda\mu_2^2+\lambda\mu_1^2+2\mu_2^2\mu_1+2\mu_1^2\mu_2)}<0.$$

So $2f_1(1)f_1(-1)-f_1(0)[f_1(1)+f_1(-1)]>2f_1(1)f_1(-1)-k(0)[f_1(1)+f_1(-1)]$.

$$2f_1(1)f_1(-1)-k(0)[f_1(1)+f_1(-1)]$$

$$=\frac{A\mu_2\lambda^3(\mu_2^2\mu_1+\mu_2^2\lambda-\mu_2\lambda^2-\mu_1\mu_2\lambda+\mu_1^2\mu_2-\mu_1\lambda^2-\mu_1^2\lambda-\lambda^3)}{(\mu_2^2+\mu_1\mu_2+\mu_1\lambda)(\mu_2^3+\mu_2^2\mu_1+\mu_1\mu_2\lambda+\mu_1\lambda^2)(\mu_1^3+\mu_1^2\mu_2+\mu_1\mu_2\lambda+\mu_2\lambda^2)}.$$

If $\mu_2<\lambda$, $\mu_2^2\mu_1+\mu_2^2\lambda-\mu_2\lambda^2-\mu_1\mu_2\lambda+\mu_1^2\mu_2-\mu_1\lambda^2-\mu_1^2\lambda-\lambda^3<-\mu_2\lambda^2-\mu_1\mu_2\lambda<0$,

so that $2f_1(1)f_1(-1)-k(0)[f_1(1)+f_1(-1)]>0$,

and $2f_1(1)f_1(-1)-f_1(0)[f_1(1)+f_1(-1)]>0$, with $H_1(1)>H_1(0)$. Therefore, if $\mu_2<\lambda$, then $H_1(B)$ is increasing in B when $B\geq 0$.

Also, $f_1(0)-l(0)=\frac{\mu_1^2\mu_2^2A^2}{(\mu_1^2+\mu_1\mu_2+\mu_2\lambda)(\lambda\mu_1^2+\lambda\mu_2^2+2\mu_1^2\mu_2+2\mu_2^2\mu_1)}>0$.

So $2f_1(1)f_1(-1)-f_1(0)[f_1(1)+f_1(-1)]<2f_1(1)f_1(-1)-l(0)[f_1(1)+f_1(-1)]$.

$$2f_1(1)f_1(-1)-l(0)[f_1(1)+f_1(-1)]$$
$$=\frac{A\lambda^2\mu_2(\lambda\mu_2^3+\mu_1\mu_2^3+2\mu_1^2\mu_2^2+2\mu_1\mu_2^2\lambda+\mu_1^2\lambda^2+2\mu_1\mu_2\lambda^2+2\mu_2\mu_1^2\lambda+\mu_2\mu_1^3+\mu_2\lambda^3+\mu_1^2\lambda^2-\mu_1\lambda^3+\lambda\mu_1^3)}{(\mu_1^2+\mu_1\mu_2+\mu_2\lambda)(\mu_2^3+\mu_2^2\mu_1+\mu_1\mu_2\lambda+\mu_1\lambda^2)(\mu_1^3+\mu_1^2\mu_2+\mu_1\mu_2\lambda+\mu_2\lambda^2)}$$

If $\mu_2>\lambda$, then $\mu_1\mu_2^3-\mu_1\lambda^3>0$, and $2f_1(1)f_1(-1)-l(0)[f_1(1)+f_1(-1)]<0$. So,

$2f_1(1)f_1(-1)-f_1(0)[f_1(1)+f_1(-1)]<0$ and hence $H_1(1)<H_1(0)$.

Therefore, if $\mu_2>\lambda$, $H_1(B)$ is decreasing in B when $B\geq 0$.

Case 2) $B<0$

$$f_1(B)=\frac{\mu_2\lambda^{-B+2}+\lambda\mu_1^{-B+1}A}{\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A}.$$

$$N=\frac{2(\mu_2\lambda^{-B+3}+\lambda\mu_1^{-B+2}A)(\mu_2\lambda^{-B+1}+\lambda\mu_1^{-B}A)}{(\mu_2\lambda^{-B+3}+\mu_1^{-B+3}A)(\mu_2\lambda^{-B+1}+\mu_1^{-B+1}A)}-(\frac{\mu_2\lambda^{-B+2}+\lambda\mu_1^{-B+1}A}{\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A})(\frac{\mu_2\lambda^{-B+3}+\lambda\mu_1^{-B+2}A}{\mu_2\lambda^{-B+3}+\mu_1^{-B+3}A}+\frac{\mu_2\lambda^{-B+1}+\lambda\mu_1^{-B}A}{\mu_2\lambda^{-B+1}+\mu_1^{-B+1}A})$$

$$=\frac{A\lambda^{-B+2}\mu_1^{-B+3}\mu_2(\lambda-\mu_1)\left[\mu_2\lambda^{-B+1}-\mu_1^{-B+1}(\frac{\lambda^2}{\mu_1^2}-\frac{\lambda}{\mu_1}+1)A\right]}{\left[\mu_2(\frac{\lambda}{\mu_1})^{-B+3}+A\right]\left[\mu_2(\frac{\lambda}{\mu_1})^{-B+2}+A\right][\mu_2(\frac{\lambda}{\mu_1})^{-B+1}+A]}$$

Since $(\lambda-\mu_1)$ and $(\mu_2\lambda^{-B+2}+\mu_1^{-B+2}A)$ have the same signs, we have

$$\frac{(\lambda-\mu_1)}{\left[\mu_2(\frac{\lambda}{\mu_1})^{-B+3}+A\right]\left[\mu_2(\frac{\lambda}{\mu_1})^{-B+2}+A\right][\mu_2(\frac{\lambda}{\mu_1})^{-B+1}+A]}>0.$$

$(\frac{\lambda}{\mu_1})^2-\frac{\lambda}{\mu_1}+1>0$ and $A<0$, so that $\mu_2\lambda^{-B+1}-\mu_1^{-B+1}\left[\left(\frac{\lambda}{\mu_1}\right)^2-\frac{\lambda}{\mu_1}+1\right]A>0$.

Thus, $N<0$. $D=[f_1(B-1)-f_1(B)][f_1(B)-f_1(B+1)]>0$, so $H_1(B+1)<H_1(B)$.

Therefore, $H_1(B)$ is decreasing in B when $B<0$.

Combine the above two cases, we conclude that when $\mu_2>\lambda$, $H_1(B)$ is decreasing in B; when $\mu_2<\lambda$, $H_1(B)$ is increasing in B for $B\geq 0$, but is decreasing in B for $B<0$.

Part 2) The proof of $H_2(B)$'s property is similar to that of $H_1(B)$, we omit the details.

Proof of Lemma 4.3

When $p_1\in[p_2,\overline{P}]$, for given μ_1,μ_2 and p_2, if $B/\beta\leq p_1-p_2<(B+1)/\beta$, it is easy to see that Π_1 is linear and increasing in p_1. So $p_1=p_2+B/\beta$ for $B\in I^+$.

$$\Pi_1(p_2+B/\beta\,|\,p_2)-\Pi_1(p_2+(B+1)/\beta\,|\,p_2)=\mu_1[f_1(B)-f_1(B+1)][p_2-H_1(B+1)]$$

$$\Pi_1(p_2+B/\beta\,|\,p_2)-\Pi_1(p_2+(B-1)/\beta\,|\,p_2)=\mu_1[f_1(B-1)-f_1(B)][H_1(B)-p_2]$$

By Lemma 4.2, $H_1(B)$ decreases in B when $\mu_2>\lambda$. If $H_1(B+1)<p_2\leq H_1(B)$, then $\Pi_1(p_2+B/\beta\,|\,p_2)>\Pi_1(p_2+(B+1)/\beta\,|\,p_2)$

and $\Pi_1(p_2 + B/\beta \mid p_2) \geq \Pi_1(p_2 + (B-1)/\beta \mid p_2)$.

Let $B_U = \{B \in I^+ : H_1(B+1) < p_2 \leq H_1(B)\}$.

$\Pi_1(p_1)$ reaches the maximum when $p_1 = p_2 + B_U/\beta$. Thus, $\Pi_1(p_1)$ increases in p_1 if $p_2 \leq p_1 \leq p_2 + B_U/\beta$; and decreases in p_1, if $p_2 + B_U/\beta < p_1 \leq \overline{P}$.

By Lemma 2, $H_1(B)$ increases in B when $\mu_2 < \lambda$. If $H_1(B) < p_2 \leq H_1(B+1)$, then $\Pi_1(p_2 + B/\beta \mid p_2) \leq \Pi_1(p_2 + (B+1)/\beta \mid p_2)$

and $\Pi_1(p_2 + B/\beta \mid p_2) < \Pi_1(p_2 + (B-1)/\beta \mid p_2)$.

Let $B_D = \{B \in I^+ : H_1(B) < p_2 \leq H_1(B+1)\}$.

$\Pi_1(p_1)$ reaches the minimum when $p_1 = p_2 + B_D/\beta$. Thus, $\Pi_1(p_1)$ decreases in p_1 for $p_2 \leq p_1 \leq p_2 + B_D/\beta$; and increases in p_1 for $p_2 + B_D/\beta < p_1 \leq \overline{P}$.

Proof of Lemma 4.4

When $p_1 \in [0, p_2]$, for given μ_1, μ_2 and p_2, if $B/\beta \leq p_1 - p_2 < (B+1)/\beta$, it is easy to observe that Π_1 is linear and increasing in p_1. So $p_1 = p_2 + B/\beta$ for $B \in I^-$.

$\Pi_1(p_2 + B/\beta \mid p_2) - \Pi_1(p_2 + (B+1)/\beta \mid p_2) = \mu_1[f_1(B) - f_1(B+1)][p_2 - H_1(B+1)]$

$\Pi_1(p_2 + B/\beta \mid p_2) - \Pi_1(p_2 + (B-1)/\beta \mid p_2) = \mu_1[f_1(B-1) - f_1(B)][H_1(B) - p_2]$

By Lemma 4.2, $H_1(B)$ decreases in B. Thus, if $H_1(B+1) < p_2 \leq H_1(B)$,

$\Pi_1(p_2 + B/\beta \mid p_2) > \Pi_1(p_2 + (B+1)/\beta \mid p_2)$

and $\Pi_1(p_2 + B/\beta \mid p_2) \geq \Pi_1(p_2 + (B-1)/\beta \mid p_2)$.

Let $B_L = \{B \in I^- : H_1(B+1) < p_2 \leq H_1(B)\}$.

$\Pi_1(p_1)$ reaches the maximum when $p_1 = p_2 + B_L/\beta$. Thus, $\Pi_1(p_1)$ increases in

p_1 if $0 < p_1 \le p_2 + B_L/\beta$; and decreases in p_1, if $p_2 + B_L/\beta < p_1 \le p_2$.

Proof of Proposition 4.1

1) $\mu_2 > \lambda$

For given p_2, μ_1 and μ_2, let the best response of server 1 be $p_1(p_2) = p_2 + B(p_2)/\beta$.

By Lemma 4.3, if $H_1(B+1) < p_2 \le H_1(B)$, server 1's best price is $p_1 = p_2 + B/\beta$.

So $B(p_2) = \{B \in I : H_1(B+1) < p_2 \le H_1(B)\}$. As $H_1(B)$ decreases in B, $B(p_2)$ decreases in p_2. $B(p_2)$ reaches it maximum at $p_2 = 0$ with value $\overline{B}_0$, and reaches its minimum at $p_2 = \overline{P}$ with value $\underline{B}_P$.

2) $\mu_2 < \lambda$

We consider the cases when $p_1 \ge p_2$ and $p_1 < p_2$ separately, and combine them to get the best response of server 1.

Case 1) $p_1 \ge p_2$, $B \equiv [\beta(p_1 - p_2)] \ge 0$.

$\Pi_1(p_2 + B/\beta \mid p_2) - \Pi_1(p_2 + (B-1)/\beta \mid p_2) = \mu_1[f_1(B-1) - f_1(B)][H_1(B) - p_2]$.

If $0 < p_2 \le H_1(0)$, $\Pi_1(p_2 + B/\beta \mid p_2) > \Pi_1(p_2 + (B-1)/\beta \mid p_2) \cdots > \Pi_1(p_2 \mid p_2)$. So there is only one local maximum for $\Pi_1(p_1 \mid p_2)$ at $\overline{P}$.

If $p_2 > H_1(0)$, there might exist two local maxima for $\Pi_1(p_1 \mid p_2)$, one at p_2 and the other at $\overline{P}$.

But $\Pi_1(p_2 \mid p_2) - \Pi_1(p_2 - 1/\beta \mid p_2) = \mu_1[f_1(-1) - f_1(0)][H_1(0) - p_2] < 0$.

So at this time, the optimal price is $p_1(p_2) = p_2 - 1/\beta$.

Hence, $p_1(p_2) = \overline{P}$ when $0 < p_2 \le H_1(0)$; and $p_1(p_2) = p_2 + B(p_2)/\beta$ for

$B(p_2) < 0$, when $p_2 > H_1(0)$.

Case 2) $p_1 < p_2$

The proof is similar to the case of $\mu_2 > \lambda$. So we omit the proof process here.

Combining the two cases, we have the best response function of server 1.

3) $p_1(p_2) - p_2 = B(p_2)/\beta$. $B(p_2)/\beta$ is decreasing in p_2 is obvious by its forms.

Proof of Proposition 4.2

The proof for Proposition 2 is similar to that for Proposition 1, we omit the proof process and details.

Proof of Theorem 4.1:

Here we only prove the first and second cases. The proofs for the remaining two cases are similar and hence omitted.

1) $\mu_1 > \lambda$ *and* $\mu_2 > \lambda$

By Proposition 1, if $p_1 \geq p_2$, we can get the best response function of two servers:

$$p_1(p_2) = ((p_2 + B/\beta) \wedge \overline{P}) \vee 0, \quad if \quad H_1(B+1) < p_2 \leq H_1(B),\ B \in [\underline{B}_P, \overline{B}_0]$$

$$p_2(p_1) = ((p_1 - B/\beta) \wedge \overline{P}) \vee 0, \quad if \quad H_2(B) < p_1 \leq H_2(B+1),\ B \in [\underline{B}_0, \overline{B}_p]$$

Only if $B_1 = B_2 = B$ and the best response curves of two servers have intersections, the price equilibrium might exist. At this time, the best response functions are:

$p_1 = p_2 + B/\beta$ for $H_1(B+1) < p_2 \leq H_1(B)$;

$p_2 = p_1 - B/\beta$ for $H_2(B) < p_1 \leq H_2(B+1)$.

The necessary and sufficient condition of the price equilibrium to exist is that the two lines must have intersections or share a common part. That is:

$H_1(B)+B/\beta-H_2(B)>0$ and $H_1(B+1)+(B+1)/\beta-H_2(B+1)<0$;

or $Y(B)>0$ and $Y(B+1)<0$.

Under such conditions, the equilibrium prices are: $p_1^* = p_2^* + \frac{B}{\beta}$ and

$p_1^* = 0 \vee ((H_1(B)+B/\beta) \wedge H_2(B+1) \wedge \overline{P})$.

2) $\mu_1 > \lambda$ *and* $\mu_2 < \lambda$

In this case, $p_1^* \geq p_2^*$. The best response functions of two servers are:

$p_1(p_2) = \overline{P}$, *if* $0 < p_2 \leq H_1(0)$, and

$$p_2(p_1) = \begin{cases} (p_1 - B/\beta) \vee 0, \text{ if } H_2(B) < p_1 \leq H_2(B+1) & (0 \leq B < \overline{B}_P) \\ \vdots \\ (p_1 - \overline{B}_P/\beta) \vee 0, \text{ if } H_2(\overline{B}_P) < p_1 \leq \overline{P} \leq H_2(\overline{B}_P+1) \end{cases}$$

Only when $B = \overline{B}_P$ do the best response curves of the two servers have intersections, and price equilibrium exist. For this case, the best response functions are:

$p_1 = \overline{P}$, for $0 < p_2 \leq H_1(0)$; and

$p_2 = p_1 - \overline{B}_P/\beta$, for $H_2(\overline{B}_P) < p_1 \leq \overline{P} \leq H_2(\overline{B}_P+1)$.

The two curves must have intersections or share a common part. That is:

$\overline{P} - \overline{B}_P \leq H_1(0)$ and $H_2(\overline{B}_P) < \overline{P} \leq H_2(\overline{B}_P+1)$

Under such conditions, $p_1^* = \overline{P}$ and $p_2^* = (\overline{P} - \overline{B}_P/\beta) \vee 0$.

By the same logic and analyzing methods, we can find the equilibrium prices together with their existence conditions for the other cases.

Proof of Proposition 4.3:

Let $m_1 = \frac{\lambda}{\mu_1}$, $m_2 = \frac{\lambda}{\mu_2}$. Suppose that $p_1^* > p_2^*$, that is $p_1^* = p_2^* + B/\beta$, for $B \geq 1$.

When $\mu_2 > \lambda$ and $\mu_1 > \lambda$, the condition for the equilibrium to exist is: $Y(B) > 0$ and $Y(B+1) < 0$. Assume at equilibrium, $\mu_1 \leq \mu_2$. We will show later that $Y(B)$ decreases in B and increases in μ_1. When $B = 1$, $\mu_1 = \mu_2$, $Y(B)$ reaches its maximum value.

$Y(1) = \frac{1}{\beta}\left[\frac{2f_1(0) - m_1}{f_1(0) - f_1(1)} - 1\right] = -\frac{1}{\beta} < 0$, which does not satisfy existence condition $Y(1) > 0$, and contradicts the assumption that it is the equilibrium. Thus, only if $\mu_1 > \mu_2$, $p_1^* > p_2^*$.

When $\mu_2 > \lambda$ and $\mu_1 < \lambda$, by the analysis above, we can show that at equilibrium, p_1^* can not be larger than p_2^*. And, similarly for the case when $\mu_2 < \lambda$ and $\mu_1 > \lambda$.

When $\mu_2 < \lambda$ and $\mu_1 < \lambda$, if $p_1^* > p_2^*$, by the best response functions:

$0 < p_2 \leq H_1(0)$ and $H_2(\overline{B}_P) < p_1 \leq H_2(\overline{B}_P + 1)$, for $\overline{B}_P > 0$.

As shown later, $H_2(\overline{B}_P)$ increases in $\overline{B}_P$, its lowest value is $H_2(1) < p_1$. $H_2(1)$ is then lowest point on server 1's best response curve. If equilibrium exists, the two best response curves must have intersections. That is, $0 < H_1(0) + 1/\beta - H_2(1)$.

When $\mu_1 \leq \mu_2 < \lambda$, $H_1(0) < H_1(1)$, and by the above proof, $H_1(0) + 1/\beta - H_2(1) < H_1(1) + 1/\beta - H_2(1) < 0$. It means the equilibrium can not exist in this case, which contradicts the assumption.

Thus, $p_1^* > p_2^*$ only if $\mu_1 > \mu_2$.

Proof of Lemma 4.5:

Part 1)

When $B > 0$, for given μ_1, μ_2, $Y(B) = H_1(B) - \frac{1}{\beta}\frac{f_2(B-1)}{f_2(B) - f_2(B-1)}$.

Let $k(B) = \frac{f_2(B-1)}{f_2(B) - f_2(B-1)}$.

$$k(B+1) - k(B) = \frac{f_2(B)}{f_2(B+1) - f_2(B)} - \frac{f_2(B-1)}{f_2(B) - f_2(B-1)} = \frac{f_2^2(B) - f_2(B-1)f_2(B+1)}{[f_2(B+1) - f_2(B)][f_2(B) - f_2(B-1)]}$$

$$f_2^2(B) - f_2(B-1)f_2(B+1) = \frac{\lambda^2(\mu_1\lambda^{B+1} + \mu_2^B A)^2}{(\mu_1\lambda^{B+2} + \mu_2^{B+2}A)^2} - \frac{\lambda^2(\mu_1\lambda^{B+2} + \mu_2^{B+1}A)(\mu_1\lambda^B + \mu_2^{B-1}A)}{(\mu_1\lambda^{B+3} + \mu_2^{B+3}A)(\mu_1\lambda^{B+1} + \mu_2^{B+1}A)}$$

$$= \frac{(-A)\lambda^{B+2}\mu_1\mu_2^B(\mu_1\lambda^{B+3/2} + A\mu_2^{B+3/2})(\mu_1\lambda^{B+3/2} - A\mu_2^{B+3/2})(\lambda - \mu_2)^3}{(\mu_1\lambda^{B+3} + \mu_2^{B+3}A)(\mu_1\lambda^{B+1} + \mu_2^{B+1}A)(\mu_1\lambda^{B+2} + \mu_2^{B+2}A)^2}$$

Since $A < 0$, $\mu_1\lambda^{B+3/2} - A\mu_2^{B+3/2} > 0$.

By our earlier proofs, $\lambda - \mu_1$ and $\mu_2\lambda^x + \mu_1^x A$ have the same signs for $x > 0$.

Thus, the value of $\mu_1\lambda^{B+3/2} + A\mu_2^{B+3/2}$ is between $\mu_1\lambda^{B+1} + A\mu_2^{B+1}$ and $\mu_1\lambda^{B+2} + A\mu_2^{B+2}$. Then, $\frac{(\mu_1\lambda^{B+3/2} + A\mu_2^{B+3/2})(\lambda - \mu_2)^3}{(\mu_1\lambda^{B+3} + \mu_2^{B+3}A)(\mu_1\lambda^{B+1} + \mu_2^{B+1}A)(\mu_1\lambda^{B+2} + \mu_2^{B+2}A)^2} > 0$.

$f_2^2(B) - f_2(B-1)f_2(B+1) > 0$ and $k(B+1) > k(B)$. $k(B)$ decreases in B. By Lemma 4.2, $H_1(B)$ decreases in B. So $Y(B) = H_1(B) - k(B)$ decreases in B for $B \geq 0$. The proof to show $Y(B)$ decreases in B when $B < 0$ is similar, and we omit the details.

Part 2)

Let $m_1 = \frac{\lambda}{\mu_1} < 1$, $m_2 = \frac{\lambda}{\mu_2} < 1$, $D = (m_1 m_2 - m_1 - m_2) < 0$.

We first prove for the case when $B > 0$.

The system handles all customers, so that $f_1(B)\mu_1 + f_2(B)\mu_2 = \lambda$, or $f_2(B) = \frac{\lambda - \mu_1 f_1(B)}{\mu_2}$.

By substitution , we have $Y(B) = \frac{1}{\beta}\left[\frac{2f_1(B-1) - m_1}{f_1(B-1) - f_1(B)} - B\right]$.

We define $J(\mu_1) = \frac{2f_1(B-1) - m_1}{f_1(B-1) - f_1(B)}$.

$$J(\mu_1) = -\left[\frac{\mu_2^{B+2} - \lambda^{B+2}}{\mu_2^{B+2}} + \frac{(\mu_2 - \lambda)\lambda^{B+2}}{(\mu_1 + \mu_2 - \lambda)\mu_2^{B+2}}\right] \cdot \left\{\lambda[(\frac{\lambda}{\mu_2})^{B+2} + 2(\frac{\lambda}{\mu_2})^B + 1)] + \frac{\lambda(\mu_2 - \lambda)}{\mu_1}\right\}$$

$$\frac{\partial J(\mu_1)}{\partial \mu_1} = \frac{(\mu_2 - \lambda)\lambda^{B+2}}{(\mu_1 + \mu_2 - \lambda)^2 \mu_2^{B+2}} \cdot \left\{\lambda[(\frac{\lambda}{\mu_2})^{B+2} + 2(\frac{\lambda}{\mu_2})^B + 1)] + \frac{\lambda(\mu_2 - \lambda)}{\mu_1}\right\}$$
$$+ \left[\frac{\mu_2^{B+2} - \lambda^{B+2}}{\mu_2^{B+2}} + \frac{(\mu_2 - \lambda)\lambda^{B+2}}{(\mu_1 + \mu_2 - \lambda)\mu_2^{B+2}}\right] \cdot \frac{\lambda(\mu_2 - \lambda)}{\mu_1^2}$$

Since $\mu_2 > \lambda$, $\frac{(\mu_2 - \lambda)\lambda^{B+2}}{(\mu_1 + \mu_2 - \lambda)^2 \mu_2^{B+2}} \left\{\lambda[(\frac{\lambda}{\mu_2})^{B+2} + 2(\frac{\lambda}{\mu_2})^B + 1)] + \frac{\lambda(\mu_2 - \lambda)}{\mu_1}\right\} > 0$ and $\left[\frac{\mu_2^{B+2} - \lambda^{B+2}}{\mu_2^{B+2}} + \frac{(\mu_2 - \lambda)\lambda^{B+2}}{(\mu_1 + \mu_2 - \lambda)\mu_2^{B+2}}\right] \frac{\lambda(\mu_2 - \lambda)}{\mu_1^2} > 0$. So $\frac{\partial J(\mu_1)}{\partial \mu_1} > 0$, and $Y(B)$ increases in μ_1.

Let $k(\mu_2) = \frac{2f_1(B-1)}{f_1(B-1) - f_1(B)}$, and $l(\mu_2) = \frac{m_1}{f_1(B-1) - f_1(B)}$. Then $Y(B) = k(\mu_2) - l(\mu_2)$.

Since $f_1(B) = \frac{m_1 m_2^{B+2}(m_2 - 1)}{(m_2^{B+3} + D)} > 0$, $k(\mu_2) = \frac{2f_1(B-1)}{f_1(B-1) - f_1(B)} = \frac{2(m_2^{B+3} + D)}{(1 - m_2)D}$.

$$\frac{\partial k}{\partial m_2} = \frac{2[(B+3)m_2^{B+2} + m_1 - 1](1 - m_2)D + (2D - 1)(m_2^{B+3} + D)]}{(1 - m_2)^2 D^2}$$

The numerator is $2D[(B+1)m_2^{B+2}(1-m_2)+m_2^{B+2}-1+m_2^{B+2}+D]-2(m_2^{B+3}+D)$.

Since $m_2<1$, $m_2^{B+2}<m_2^{B+1}<\cdots<m_2$, and $m_2^{B+3}+D<m_2^{B+2}+D<\cdots<0$, we have

$(B+1)m_2^{B+2}(1-m_2)+m_2^{B+2}-1=(1-m_2)[(B+1)m_2^{B+2}-m_2^{B+1}-m_2^{B}\cdots-1]<0$.

So the numerator is positive. As the denominator is always positive, $\frac{\partial k}{\partial m_2}>0$.

As $m_2=\frac{\lambda}{\mu_2}<1$, $\frac{\partial k}{\partial \mu_2}<0$, i.e., $k(\mu_2)$ decreases with μ_2.

Next we consider $l(\mu_2)$ and examine its denominator only since its numerator in insensitive to μ_2.

$$\frac{\partial f_1(B)}{\partial m_2}=\frac{m_1(m_2-1)m_2^{B+1}[(m_2^{B+2}+D)+(m_2^{B+1}+D)\cdots(m_2+D)]}{(m_2^{B+3}+D)^2}.$$

Let $X(B+2)=(m_2^{B+2}+D)+(m_2^{B+1}+D)\cdots(m_2+D)$, and it is easy to show $X(B+2)<0$.

$$\frac{\partial f_1(B)}{\partial m_2}-\frac{\partial f_1(B-1)}{\partial m_2}=\frac{m_1(m_2-1)m_2^{B}\left[(m_2-1)X(B+1)(D^2-m_2^{2B+5})+(m_2^{B+2}+D)^3\right]}{\left(m_2^{B+3}+D\right)^2\left(m_2^{B+2}+D\right)^2}$$

$$(m_2-1)X(B+1)(D^2-m_2^{2B+5})+(m_2^{B+2}+D)^3>(m_2-1)X(B+1)(D^2-m_2^{2B+4})+(m_2^{B+2}+D)^3.$$

$$(m_2-1)X(B+1)(D^2-m_2^{2B+4})+(m_2^{B+2}+D)^3$$

$$=(m_2^{B+2}+D)\left[(m_2-1)X(B+1)D+X^2(B+2)-X(B+1)X(B+3)\right]$$

$$X^2(B+2)-X(B+1)X(B+3)=m_2^{B+3}-D^2+[2-(B+1)(m_2-1)]m_2D$$

$m_2^{B+3}-D^2=(m_2^{(B+3)/2}-D)(m_2^{(B+3)/2}+D)<0$, and $[2-(B+1)(m_2-1)]m_2D<0$.

We can conclude that $X^2(B+2)-X(B+1)X(B+3)<0$.

$m_2^{B+2}+D<0$, and $(m_2-1)X(B+1)D<0$.

So $(m_2^{B+2}+D)[(m_2-1)X(B+1)D+X(B+2)-X(B+1)X(B+3)>0$.

Since $(m_2-1)X(B+1)(D^2-m_2^{2B+5})+(m_2^{B+2}+D)^3>0$, $\frac{\partial f_1(B)}{\partial m_2}-\frac{\partial f_1(B-1)}{\partial m_2}<0$ for $m_2<1$.

Therefore, $\frac{\partial f_1(B)}{\partial \mu_2}-\frac{\partial f_1(B-1)}{\partial \mu_2}>0$, and hence $\frac{\partial l(\mu_2)}{\partial \mu_2}>0$, i.e., $l(\mu_2)$ increases in μ_2. As a consequence, $Y(B)=k(\mu_2)-l(\mu_2)$ is a decreasing function of μ_2.

The proofs to show that $Y(B)$ increases with μ_1 and μ_2 when $B>0$. As for the case when $B<0$, the proofs are similar and we omit the details.

Part 3)

By Theorem 1, when $\mu_1>\lambda$ and $\mu_2>\lambda$, the existence condition for the price equilibrium is $Y(B)>0$ and $Y(B+1)<0$ ($B\in I$). $Y(B)$ increases with μ_1. Then, for given μ_2 and B, there is a unique solution of μ_1 to $Y(B)=0$. Let it be $\bar{\mu}_1(\mu_2,B)$, so that $Y(B)>0$ when $\mu_1>\bar{\mu}_1(\mu_2,B)$. For given B, applying the Implicit Function Theorem to $Y(B,\bar{\mu}_1,\mu_2)=0$, we have $\frac{\partial \bar{\mu}_1(\mu_2,B)}{\partial \mu_2}=\frac{\frac{\partial Y(B)}{\partial \mu_2}}{\frac{\partial Y(B)}{\partial \mu_1}}$.

Since $\frac{\partial Y(B)}{\partial \mu_2}<0$ and $\frac{\partial Y(B)}{\partial \mu_1}>0$, $\frac{\partial \bar{\mu}_1(\mu_2,B)}{\partial \mu_2}>0$, i.e., $\bar{\mu}_1(\mu_2,B)$ increases in μ_2. The proof to show that $\bar{\mu}_1(\mu_2,B)$ increases in B is similar.

Proof of Lemma 4.6

Part 1)

Let $m_1 = \frac{\lambda}{\mu_1} < 1$, $m_2 = \frac{\lambda}{\mu_2} > 1$, $D = (m_1 m_2 - m_1 - m_2) < 0$.

Define $l(B) = \frac{f_2(B-1)}{f_2(B)} = \frac{(m_2^{B+3} + D)(m_2^{B+1} + D)}{(m_2^{B+2} + D)^2}$, for any $B \in I^+$.

$$\frac{\partial l(B)}{\partial m_1} = \frac{\frac{\partial f_2(B-1)}{\partial m_1} f_2(B) - \frac{\partial f_2(B)}{\partial m_1} f_2(B-1)}{f_2^2(B)} = \frac{m_2^{B+1}(m_2-1)^3(m_2^{B+3} - D)}{(m_2^{B+3} + D)^3}.$$

By our earlier proof, $(m_2 - 1)(m_2^{B+2} + D) > 0$, so that $\frac{\partial l(B)}{\partial m_1} > 0$.

$$H_2(B) = \frac{f_2(B-1)}{f_2(B) - f_2(B-1)} + B,$$

and $$\frac{\partial H_2(B)}{\partial m_1} = \frac{\frac{\partial f_2(B-1)}{\partial m_1} f_2(B) - \frac{\partial f_2(B)}{\partial m_1} f_2(B-1)}{[f_2(B) - f_2(B-1)]^2} > 0.$$

So $H_2(B)$ increases with m_1, and hence decreases with μ_1.

With the same logic as that in the proof of Lemma 5, we can show $\frac{\partial f_2(B-1)}{\partial m_2} > \frac{\partial f_2(B)}{\partial m_2} > 0$ when $\mu_2 < \lambda$. Since $f_2(B) > f_2(B-1)$, we can obtain that:

$$\frac{\partial f_2(B-1)}{\partial m_1} f_2(B) - \frac{\partial f_2(B)}{\partial m_1} f_2(B-1) > \left[\frac{\partial f_2(B-1)}{\partial m_1} - \frac{\partial f_2(B)}{\partial m_1}\right] f_2(B-1) > 0.$$

That is $\frac{\partial H_2(B)}{\partial m_2} > 0$. $H_2(B)$ increases with m_2, and hence decreases with μ_2.

Thus, $H_2(\overline{B}_P + 1)$ decreases with μ_1 and μ_2.

Similarly, we can prove that $H_1(0)$ decreases with both μ_1 and μ_2.

Therefore, $Z(\mu_1, \mu_2, \overline{B}_P)$ is decreasing in μ_1 and μ_2.

Part 2)

We have proved that $Z(\mu_1,\mu_2,\overline{B}_P)$ is decreasing in μ_1. For given $\overline{P}$ and μ_2, there exists a unique solution of μ_1 which satisfies $Z(\mu_1,\mu_2,\overline{B}_P)-\overline{P}=0$. Let this solution be $\mu_z(\overline{B}_P)$. Then, $Z(\mu_1,\mu_2,\overline{B}_P)\geq\overline{P}$ if $\mu_1\leq\mu_z(\overline{B}_P)$, and $Z(\mu_1,\mu_2,\overline{B}_P)<\overline{P}$ if $\mu_1>\mu_z(\overline{B}_P)$. Since $H_2(\overline{B}_P)$ decreases with μ_1, there is a unique solution to $\mu_1=\mu_G(\overline{B}_P)$ s.t. $H_2(\overline{B}_P)\geq\overline{P}$ for $\mu_1\leq\mu_G(\overline{B}_P)$ and $H_2(\overline{B}_P)<\overline{P}$. By the Implicit Function Theorem, both $\mu_G(\overline{B}_P)$ and $\mu_z(\overline{B}_P)$ decrease in μ_2 and increase in $\overline{B}_P$.

Proof of Proposition 5.1

In a symmetric system, the two servers have the same capacity investment costs, i.e., $c_1=c_2=c$. We substitute the equilibrium prices into the profit functions, and find that the equilibrium service rates exist when $p_1^*=p_2^*=\overline{P}$ and the corresponding service rates are:

$$\mu_1^*=\mu_2^*=\frac{\sqrt[3]{4K}}{12c}+\frac{\sqrt[3]{16}\lambda^2c}{12\sqrt[3]{K}}-\frac{1}{6}\lambda,$$

where $a\equiv\sqrt{\overline{P}(27\overline{P}-4c\lambda)}$ and $K\equiv\lambda^2c^2[27\overline{P}-2c\lambda+3\sqrt{3}a]$.

Substituting the equilibrium prices and service rates in the profit function, we have

$$\Pi_1^*=\Pi_2^*=\frac{\overline{P}\lambda}{2}-c\left[\frac{\sqrt[3]{4K}}{12c}+\frac{\sqrt[3]{16}\lambda^2c}{12\sqrt[3]{K}}-\frac{1}{6}\lambda\right]^2.$$

In the symmetric system, we know that $\mu_1^*=\mu_2^*$, and $\mu_1^*+\mu_2^*>\lambda$, so $\mu_1^*>\frac{\lambda}{2}$.

Since the profit will not be negative, that is $\frac{\overline{P}\lambda}{2}-c(\mu_i^*)^2>0$. Therefore, we can obtain that $c<\frac{2\overline{P}}{\lambda}$.

REFERENCES

1. Allon, G., and A. Federgruen. 2007. "Competition in Service Industries". *Operations Research*, 55(1), 37-55.

2. Armony, M. and M. Haviv. 2003. "Price and delay competition between two service providers". *European Economic Review*, 147(1), 32-50.

3. Bell, C. and S. Stidham. 1983. "Individual versus social optimization in the allocation of customers to alternative servers". *Management Science*, 29(7), 821-839.

4. Cachon, G. P. and P. T. Harker. 2002. "Competition and outsourcing with scale economies". *Management Science*, 48(10), 1314-1333.

5. Cachon, G. P. and F. Zhang. 2007. "Obtaining fast service in a queueing system via performance-based allocation of demand". *Management Science*, 53(3) 408-420.

6. Chen, H., and Y. W. Wan. 2003. "Price competition of make-to order firms". *IIE Transaction*, 35(9), 817-832.

7. Christ, D. and B. Avi-Itzhak. 2002. "Strategic equilibrium for a pair of competing servers with convex cost and balking". *Management Science*, 48(6), 813-820.

8. Davidson, C. 1988. "Equilibrium in servicing industries: an economic application of queueing theory". *Journal of Business*, 61(3), 347-367.

9. De Vany, A. and T. Savings. 1983. "The economics of quality". *Journal of Political*

Economy, 91(6). 979-1000.

10. Deneckere, R. and J. Peck. 1995. "Competition over price and service rate when demand is stochastic: a strategy analysis". *RAND Journal of Economics*, 26(1), 148-162.
11. Dewan, S and H. Mendelson. 1990. "User delay costs and internal pricing for a service facility". *Management Science*, 36(12), 1502-1517.
12. Flatto, L. and H. McKean. 1984. "Two parallel queues created by arrivals with two demand-I". *SIAM Journal on Applied Mathematics*, 44(5), 1041-1053.
13. Flatto, L. and H. McKean. 1984. "Two parallel queues created by arrivals with two demand-II". *SIAM Journal on Applied Mathematics*, 45(5), 861-878.
14. Gilbert, S. M. and Z. K. Weng. 1998. "Incentive effects favor nonconsolidating queues in a service system: the principle-agent perspective". *Management Science*, 44(12) 1662-1669.
15. Grassmann, W. K. 1980. "Transient and steady state results for two parallel queues". *The International Journal of Management Science*, 8(1), 105-112.
16. Ha, A. Y., Li, L. and S. M. Ng. 2003. "Price and delivery logistics competition in a supply chain". *Management Science*, 49(9), 1139-1153.
17. Haight, F. A. 1958. "Two queues in parallel". *Biometrica*, 45(34) 401-410.
18. Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.
19. Houck, D. J. 1987. "Comparison of policies for routing customers to parallel queueing systems". *Operations Research*, 35(2), 306-310.

20. Kalai, E., M. I. Kamien, and M. Rubinovitch. 1992. “Optimal service speeds in a competitive environment”. *Management Science*. 38(8) 1154-1163.

21. Knessl, C., J. M. Bernard, Z. Svhuss, and C. Tier. 1986. “Two parallel queues with dynamic routing”. *IEEE Transactions on Communications*, 34(12), 1170-1175.

22. Konheim, A. G., I. Meilijson, and A. Melkman. 1981. “Processor-sharing of two parallel lines”. *Journal of Applied Probability*, 18, 952-956.

23. Kotiah, T. C. T. and N. B. Slater. 1973. “On two-server queues with two types of customers”. *Operations Research*, 21(2), 597-603.

24. Lederer, P. J. and L. Li. 1997. “Pricing, production, scheduling, and delivery time competition”. *Operations Research*, 45(3), 407-420.

25. Lee, H. and M. Cohen. 1985. “Multi-agent customer allocation in a stochastic service system”. *Management Science*, 31(6), 752-763.

26. Levhari, D., and I. Luski. 1978. “Duopoly pricing and waiting lines”. *European Economic Review*, 11(1), 17-35.

27. Li, L. 1992. “The role of inventory in delivery-time competition”. *Management Science*, 38(2), 182-197.

28. Li, L. and Y. S. Lee. 1994. “Pricing and delivery-time performance in a competitive environment”. *Management Science*, 40(5) 633-646.

29. Mendelson, H. and S. Whang. 1990. “Optimal incentive-compatible priority pricing for the M/M/1 queue”. *Operations Research*. 38(5), 870-883.

30. Nakamura, M., I. Sasase, and S. Mori. 1989. “Two parallel queues with dynamic routing under a threshold-type scheduling”. *IEEE GLOBECON*, 1445-1449.

31. Naor, P. 1969. "On the regulation of queue size by levying tolls". *Economitrica*, 37, 15-24.

32. Owen, G. 1982. *Game Theory*, (Second Ed.), Academic Press, New York.

33. Reitman, D. 1991. "Endogenous quality differentiation in congested markets". *Journal of Industrial Economics*, 39(6), 621-647.

34. Rubinovitch, M. 1985. "The slow server problem". *Journal of Applied Probability*, 22(1), 205-213.

35. Singh, V. P. 1970. "Two-server Markovian queues with balking: heterogeneous vs. homogeneous servers". *Operations Research*, 18(1), 145-159.

36. So, K. C. 2000. "Price and time competition for service delivery". *Manufacturing & Service Operations Management*, 2(4), 392-409.

37. So, K. C. and J. S. Song. 1998. "Price, delivery time guarantees and capacity selection". *European Journal of Operations Research*, 111(1), 28-49

38. Stidham, S. 1992. "Pricing and capacity decisions for a service facility: stability and multiple local optima". *Management Science*, 38(8), 1121-1139.

39. Stidham, S. and C. Rump. 1998. "Stability and chaos in input pricing for a service facility with adaptive consumer response to congestion". *Management Science*, 44(2), 246-261.

40. Zhao, Y. and W. K. Grassmann. 1990. "The shortest queue model with jockeying". *Naval Research logistics*, 37(5), 773-787.