Product Differentiation and Operations Strategy for Price and Time Sensitive Markets

by

Sachin Jayaswal

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this dissertation, we study the interplay between a firm's operations strategy, with regard to its capacity management, and its marketing decision of product differentiation. For this, we study a market comprising heterogeneous customers who differ in their preferences for time and price. Time sensitive customers are willing to pay a price premium for a shorter delivery time, while price sensitive customers are willing to accept a longer delivery time in return for a lower price. Firms exploit this heterogeneity in customers' preferences, and offer a menu of products/services that differ only in their guaranteed delivery times and prices. From demand perspective, when customers are allowed to self-select according to their preferences, different products act as substitutes, affecting each other's demand. Customized product for each segment, on the other hand, results in independent demand for each product. On the supply side, a firm may either share the same processing capacity to serve the two market segments, or may dedicate capacity for each segment. Our objective is to understand the interaction between product substitution and the firm's operations strategy (dedicated versus shared capacity), and how they shape the optimal product differentiation strategy.

To address the above issue, we first study this problem for a single monopolist firm, which offers two versions of the same basic product: (i) regular product at a lower price but with a longer delivery time, and (ii) express product at a higher price but with a shorter delivery time. Demand for each product arrives according to a Poisson process with a rate that depends both on its price and delivery time. In addition, if the products are substitutable, each product's demand is also influenced by the price and delivery time of the other product. Demands within each category are served on a first-come-first-serve basis. However, customers for express product are always given priority over the other category when they are served using shared resources. There is a standard delivery time for the regular product, and the firm's objective is to appropriately price the two products and select the express delivery time so as to maximize its profit rate. The firm simultaneously

needs to decide its installed processing capacity so as to meet its promised delivery times with a high degree of reliability. While the problem in a dedicated capacity setting is solved analytically, the same becomes very challenging in a shared capacity setting, especially in the absence of an analytical characterization of the delivery time distribution of regular customers in a priority queue. We develop a solution algorithm, using matrix geometric method in a cutting plane framework, to solve the problem numerically in a shared capacity setting.

Our study shows that in a highly capacitated system, if the firm decides to move from a dedicated to a shared capacity setting, it will need to offer more differentiated products, whether the products are substitutable or not. In contrast, when customers are allowed to self-select, such that independent products become substitutable, a more homogeneous pricing scheme results. However, the effect of substitution on optimal delivery time differentiation depends on the firm's capacity strategy and cost, as well as market characteristics. The optimal response to any change in capacity cost also depends on the firm's operations strategy. In a dedicated capacity scenario, the optimal response to an increase in capacity cost is always to offer more homogeneous prices and delivery times. In a shared capacity setting, it is again optimal to quote more homogeneous delivery times, but increase or decrease the price differentiation depending on whether the status-quo capacity cost is high or low, respectively. We demonstrate that the above results are corroborated by real-life practices, and provide a number of managerial implications in terms of dealing with issues like volatile fuel prices.

We further extend our study to a competitive setting with two firms, each of which may either share its processing capacities for the two products, or may dedicate capacity for each product. The demand faced by each firm for a given product now also depends on the price and delivery time quoted for the same product by the other firm. We observe that the qualitative results of a monopolistic setting also extend to a competitive setting. Specifically, in a highly capacitated system, the equilibrium prices and delivery times are such that they result in more differentiated products when both the firms use shared capacities as compared to the scenario when both the firms use dedicated capacities. When the competing firms are asymmetric, they exploit their distinctive characteristics to differentiate their products. Further, the effects of these asymmetries also depend on the capacity strategy used by the competing firms. Our numerical results suggest that the firm with expensive capacity always offers more homogeneous delivery times. However, its decision on how to differentiate its prices depends on the capacity setting of the two firms as well as the actual level of their capacity costs. On the other hand, the firm with a larger market base always offers more differentiated prices as well as delivery times, irrespective of the capacity setting of the competing firms.

Acknowledgements

I would like to express my heartfelt thanks and gratitude to my advisor, Dr. Elizabeth Jewkes, for her continuous moral, intellectual and financial support during my Ph.D. study at the Department of Management Sciences, University of Waterloo. She has always been available for any help I needed. I would like to express my special thanks to my co-advisor, Dr. Saibal Ray, without whose guidance this thesis would not have been possible. I am very grateful to have had the chance to work with my advisors, who have encouraged me at every step of my Ph.D. study with great patience.

My gratitude also goes to the faculty of the Department of Management Sciences for providing me with the foundations for my Ph.D. research. I would like to thank Dr. James H. Bookbinder, Dr. Samir Elhedhli and Dr. Steve Drekic for their time and effort in serving on my Ph.D. committee. My special thanks to Dr. Miguel Anjos and Dr. Benny Martin who have always been available for their friendly advice.

I would like to thank all my friends in Waterloo-Kitchener who made my stay in Canada so memorable, especially my great house mates Sarvagya, Akhilesh, Aashish and Guru Prakash, who besides sharing the chores of my kitchen have also shared all my sorrows and happiness in last 4 years. I would also like to thank Emre Celebi, my first friend in the department, Meyappan, Edward, Joe, Bissan and Eman whose company I always enjoyed. My special thanks to my office mate, Navneet Vidyarthi, and his wife, Geetali Vidyarthi, for their love and affection.

Finally, I am forever indebted to my parents for their unconditional love and support, and for the sacrifices they have always made to see me realize my dreams in life. I am also indebted to my brother, Suman Jayaswal, my sisters, Sushma Jayaswal and Seema Jayaswal, for sharing my responsibilities towards my parents while I was away from home to pursue my studies. I am specially thankful to my sister, Seema Jayaswal, who trusts me more than I do myself, and has always reminded me of my responsibilities in life. My very special thanks goes to my

fiancee, Ritika, who has been solidly behind me during the good and bad phases of my life.

Dedication

To my parents

for the sacrifices they have always made to see me realize my dreams in life

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Time-Based Competition

Increasing market competition has forced modern businesses to introduce new products and ever greater variety at rapid rates, and speed has evolved as the competitive paradigm (Blackburn 1991, Hum and Sim 1996). As speed became a driver of business success, lead time reduction emerged as a dominant issue in manufacturing strategy (Van Beek and Van Putten 1987, Suri 1998, Hopp and Spearman 2000). In fact, the ability to offer customized products with short lead times is becoming a competitive advantage among suppliers (Andel 2002). This new shift in firms' focus is termed as time-based competition. Time-based competition mandates speed in every aspect of the business. Firms today compete primarily on three components of time: product development time, manufacturing lead time, customer response/delivery time (Kim and Tang 1997). Product development time is the time a firm needs to transform an idea into a product. Manufacturing lead time is the time to convert raw materials to finished goods. Response time is the time it takes to fulfill a customer's order. Competing in time gives a firm the advantage of increased market share, increased price premium, and reduced cost (Stalk and Hout 1990). Shorter product development time gives a firm an early entry into the market, enabling it to establish itself as a market leader. Shorter manufacturing lead time allows a manufacturer to provide the same level of customer service even with reduced finished goods and work-in-process inventories, which in turn helps mitigate the risk of obsolescence and cut inventory costs. Shorter response time increases customer satisfaction, which further helps repeat business.

In this thesis we focus on service and Make-to-Order (MTO) manufacturing industries where firms commonly use shorter response time as a competitive priority. As firms are moving from Make-to-Stock (MTS) to Make-to-Order (MTO) to reduce costs and increase market responsiveness, quoting effective prices and reliable lead times becomes especially important (Martin 2000, Vinas 2006). In service industries, customers regard total service time as a key concern (Stevenson 1999). The importance of a shorter response/delivery time has been highlighted in literature by several stories of their successful implementation. Progressive Insurance, an automobile insurance company based in Ohio, achieved a sevenfold growth of sales from $1.3 billion in 1991 to $9.5 billion in 2002 as a result of introducing an Immediate Response claims system, which dramatically reduced the claim handling time from 7-8 days to just nine hours (Hammer 2004). Shell Lubricants redesigned its order fulfillment process, thus reducing the cycle time by 75% and operating expenses by 45%, and boosting customer satisfaction 105% (Hammer 2004). Ray (2001) reports the case of an Electronics Manufacturing Service (EMS) company in Toronto, which specializes in supplying electronic components for a number of international Original Equipment Manufacturers (OEMs). The OEMs were ready to pay for many of the lead time reduction initiatives undertaken by the EMS since it helped them reduce the delivery time to their customers. Atlas Door, a leading supplier of customized industrial overhead doors in the United States, is able to fill an order for a door within four weeks, one third the industry average. Responsiveness has earned Atlas Door a large customer base, which is often willing to pay the premium price for quick delivery. Atlas is growing three times faster than the industry, and is five times more profitable than the industry average (Stalk and Hout 1990). Thomasville Furniture markets its quick-ship program, under which a customer order delivered within 30 days against the competitors' average response time of more than three months. Thomasville is growing four times faster than the industry, and the company is twice as profitable as the U.S. industry average (Stalk and Hout 1990). Other successful stories can be found in Charney (1991), Stalk and Hout (1990) and Blackburn (1991).

There are three basic response/delivery time based strategies that firms use to attract customers: (i) to serve customers as fast as possible, (ii) to encourage potential customers to get a delivery time quote before placing orders, and (ii) to guarantee a uniform delivery time to all potential customers (Ray and Jewkes 2004, So and Song 1998). The second strategy of encouraging customers to get a delivery time quote is more popular in make-to-order manufacturing industry where firms dynamically change the quoted delivery time based on the congestion in the system (Plambeck 2000). Our focus in this thesis is on the strategy of offering a uniform delivery time guarantee, which is also popular in make-to-order industries but has been more popularized by retail and service industries as it eliminates the uncertainty in receiving the service. Many firms today use their uniform delivery time guarantee in their promotion campaigns. For example, Cat Logistics, a subsidiary of Caterpillar, promises to ship service parts within 24 hours to its clients (Schmidt and Aschkenase 2004). Ameristock quotes maximum 10 seconds per internet equity trade (Boyaci and Ray 2003, Zhao et al. 2008). Tradewinds Coffee waives its shipping charges if the product is not delivered on time (Ho and Zheng 2004). Domino's Pizza advertises its 30 minutes delivery guarantee (Charney 1991). In freight services, Federal Express offers next day mail delivery by 11:00 a.m. (So and Song, 1998).

1.2 Market Segmentation and Product Differentiation

Shorter delivery time guarantee can have a major impact on both demand as well as price. Sterling and Lambert (1989), Blackburn et al. (1992), Maltz and Maltz

(1998) and Smith et al. (2000), besides others, have empirically shown the impact of delivery time on customer demand. In Industrial markets, a 5% increase in delivery time can lead to a loss of 24% of the demand from the existing customer base (Ballou 1998). The importance of speed or shorter lead times has also provided firms with new business opportunities. Firms try to exploit customers' sensitivity to speed/time to extract price premium for the same product by promising them a shorter delivery time. Amazon.com, for example, charges more than double the shipping costs to guarantee a delivery in two days against its normal delivery time of around a week (Ray and Jewkes, 2004). Amazon.com thus tries to serve both the market segments - one that is price sensitive and is willing to wait for a week for its delivery, and the other that is more time sensitive and is willing to pay a premium price for a faster delivery. Firms, like Amazon.com, that differentiate their products based on delivery times try to exploit heterogeneity in customers' preference for time and willingness to pay in order to create market segments that maximize the firm's revenue (Boyaci and Ray 2003).

Heterogeneity in customers' willingness to pay for delivery time guarantees may be inherent in their personalities. Some customers may be price sensitive and may not mind waiting to be served if that can reduce the price they need to pay. Others may be more impatient and thus willing to pay a price premium if that will shorten their waiting time. Heterogeneity may also be caused by situational factors. For example, whether or not a customer is willing to pay a premium may depend on the urgency of her need. Ultimately, a customer's choice from a menu of delivery times offered will depend on her perception of the price difference relative to the difference in the delivery times offered. If a particular customer feels the extra price she needs to pay is worth the improved service she is offered in return, she is likely to select an express delivery option; otherwise she will select a slower delivery. The services offered are thus often substitutable, and customers' decisions can be influenced by designing a menu that carefully discriminates between the services offered using differential pricing and delivery times. Market segmentation together with product differentiation can thus provide firms with a unique business opportunity to make greater revenues by influencing some customers to opt for express delivery at a higher price. For example, Plantgel, a firm selling nutrition gels for plants online, offers to process an order within a day for an extra $3, against a regular delivery of 10 days (Zhao et al. 2008). FedEx offers logistic services like "FedEx Next Flight", "FedEx First Overnight", "FedEx Priority Overnight" and "FedEx 2Day", each with a different price and delivery time guarantee to target different customer segments having different sensitivities to price and delivery time. Similarly, UPS offers "UPS Express Early A.M.", "UPS Express", "UPS Express Saver" and "UPS Expedited" for different categories of customers.

1.3 Product Differentiation and Operations Strategy

A firm's marketing decisions are often closely linked to its operations strategies. Different firms in an industry compete with each other by offering better deals, either in the form of a lower price, better service or both to their customers. In a make-to-stock industry, a higher service level (a faster delivery) translates into a better inventory management, whereas in a make-to-order or service industries, this usually translates into a better (server) capacity/queue management. One question that naturally arises in a firm's pursuit of market segmentation and product differentiation in a make-to-order or service industry is whether to pool or to differentiate the facilities used for each market segment. Examples from industry suggest the use of both, each having its own merits. FedEx, for example, uses separate facilities for its express and ground services. In the words of Frederick W. Smith, chairman, president and CEO of FedEx, "the optimal way to serve very distinct market segments, such as express and ground is to operate highly efficient, independent networks with different facilities, different cut-off times and different delivery commitments[1]". In contrast, UPS delivers express and ground

[1] http://www.fedex.com/us/about/express/pressreleases/pressrelease011900.html?link=4

services using one integrated network. According to UPS,“it is their integrated air and ground network that enhances pickup and delivery density and provides them with the flexibility to transport packages using the most efficient and cost-effective transportation mode or combination of modes[2]”.

Boyaci and Ray (2003) present other cases from industry that use dedicated capacities. For example, web hosting and content delivery firms maintain dedicated servers for customers like news sites whose content for online delivery is time sensitive. This makes possible real time update to their data. Other customers whose data do not require frequent updates are served using a different set of servers. Boyaci and Ray (2003) cite another example of a Southeast Asia based printing company, which uses separate facilities for time sensitive magazines like the Asian editions of Time, Neewsweek, etc., and separate facilities for books printed in mass scale. Further, the company uses a dedicated delivery system for time sensitive materials.

In contrast to the above examples, photo development stores offering one-hour express service and a cheaper three-day regular service share their capacities for the two different services. A mobile telephony service provider shares its facilities with other service providers to provide roaming services to their customers. The service provider in this case normally charges a higher price for roaming services compared to the basic service provided to its own customers. Similarly, a third party logistics service provider shares its fleet of vehicles to serve multiple firms with different delivery time guarantees (Sinha et al. 2008). Another interesting example provided by Sinha et al. (2008) is the possible sharing of rail-linked inland container depots (ICDs), required for inland rail container movement in India, which was until recently solely managed by Container Corporation of India Ltd. (Concor). Due to the recent opening up of inland rail container movement to private players, new firms may seek to lease these ICDs from Concor in the initial years due to their high infrastructural set up cost.

[2]http://sec.edgar-online.com/1999/10/20/11/0000940180-99-001230/Section2.asp

In the academic literature, authors advocate the use of dedicated channels when different customer segments have very different needs (Fuller et al. 1993, Farris 2002, Smith et al. 2000). In the transportation and logistics industry, for example, with a recent dramatic rise in "time-definite" premium transportation services, many carriers now maintain two totally separate capacities (trucks and information systems), one dedicated to time sensitive premium shipments and another dedicated to normal cost-effective shipments (Farris 2002). Whitt (1999) technically justifies this by arguing that when customer classes have different service time distributions, serving them using a common capacity increases the overall coefficient of variation of the service times, resulting in service quality for time sensitive customers being degraded by customers with longer service times. Thus there is a natural motivation for separate express checkout lines in supermarkets. However, if different classes of customers have the same service time distribution, it is known that shared capacity is more efficient as it exploits the benefits of pooling (Whitt 1999). In a call center, capacities (equipment and employees) are generally shared between normal and priority calls, with time sensitive priority customers served on a priority basis. Similarly, while boarding an aircraft, the same server generally serves economy as well as business/priority customers, although business class customers are prioritized. Offering different delivery time guarantees using a shared capacity (SC), however, creates supply-side interaction and requires mechanism for prioritizing orders. This creates operational complexities and potentially increasing costs. Providing different services using dedicated capacities (DC), on the other hand, requires additional capacity investment (Zhao et al. 2008).

1.4 Research Agenda and Organization of the Thesis

From the above discussion, we see that time has emerged as a key competitive priority in today's business, which is highlighted by several stories of successful implementation of time-based strategies. Firms that traditionally competed only based on costs have gradually shifted their focus to time. This shift in focus has come from a realization that customers do not only value money but also time, often to an extent that they are willing to pay a premium to get a faster service. Consequently, the last decade has witnessed vast literature on time-based competition. So and Song (1998), Palaka et al. (1998), Ray and Jewkes (2004), Hill and Khosla (1992), Tsay and Agarwal (2000), So (2000), Allon and Federgruen (2007), to name a few, have studied the problem of pricing and delivery time decisions for make-to-order/service industries in a time sensitive market. In this thesis, we aim to build upon some of the existing works to further contribute to this literature by identifying research avenues that are still unexplored. As discussed above, we find that often heterogeneity exists in customers' sensitivity to time (and their willingness to pay), which firms try to exploit by offering the same basic product or service with different delivery time guarantees and at different prices. However, we observe that different firms doing this may use very different operations strategies. For example, some firms use a dedicated set of resources for each market segment, while others pool their resources to serve the different market segments. This leads to the basic research question: how does the operations strategy, specifically capacity strategy, of a firm affect its pricing and delivery time differentiation strategy in a time sensitive market. This is the central issue we try to investigate in this thesis.

The rest of the thesis is organized as follows. In Chapter 2, we define our research problem in detail, and develop a basic modelling framework of a firm's pricing and delivery time decisions in a monopolistic setting. The mathematical models and the solution methods developed in Chapter 2 allow us to compare the price and delivery time differentiation strategies of a firm under dedicated and shared capacity settings. We study this problem extensively in chapter 3 to derive important managerial insights. In Chapter 4, we further extend our modelling framework to a competitive setting. We address the same research question to investigate if and how market competition affects our results. The competitive models developed in Chapter 4 are studied extensively in Chapter 5. Finally, we summarize our work and provide directions for future research in Chapter 6. We differ slightly from the usual convention in our presentation of this thesis in that we do not have a separate chapter for literature review. We instead dedicate a separate section on literature review in each of the Chapters 2 and 4. This allows us to position our work better with respect to the literature on monopolistic and competitive settings, which are reviewed separately in these two chapters.

1.5 Research Contribution

- We extend the existing literature on product differentiation in a segmented market by developing a modelling framework in a shared capacity setting. To the best of our knowledge, it is the first attempt to study product differentiations in a shared capacity framework. This allows us to study the effects of a firm's capacity strategy (dedicated versus shared) on its price and delivery time differentiation strategies. It also allows us to study the effect of substitution between different market segments on the product differentiation strategies, and how it interacts with the capacity strategy of a firm in shaping its optimal differentiation decisions.

- We further extend our modelling framework to study price and delivery time decisions in a competitive setting. This allows us to investigate if, and how, the capacity strategy of firms affect their product differentiation strategies in presence of market competition. This also allows us to better understand the effects of competition, per se, on price discrimination, given that research thus far has produced very contradictory results.

- Our study makes a significant technical contribution to the study of price and delivery time decisions by presenting a novel solution method, which links matrix geometric method to a cutting plane algorithm, to solve a complex mathematical model for the shared capacity setting.

- Our study provides several important insights of special interest to operations managers.

Some of the important managerial insights we generate from our study can be summarized as:

- A firm's capacity cost plays a major role in determining the relative product differentiation in the two capacity settings. Whereas the relative sensitivities of customers to price and time determines the effect of product substitution on product differentiation.

- In a high capacity cost environment, a firm with shared capacity should offer products with greater differentiation (both in terms of prices and delivery times) than a firm with dedicated capacities, irrespective of whether the products are substitutable or not.

- When a firm selling two non-substitutable products in independent markets decides to make both products available to all customers (thus introducing substitutability), it should reduce its price differentiation, irrespective of whether it operates under shared or dedicated capacity regime. However, as regards delivery times, whether the products should be more differentiated or more homogeneous depends on the firm's capacity strategy (as well as on its marginal capacity cost and market characteristics).

- The optimal response to any change in the capacity cost depends on the capacity strategy as well as the existing level of capacity cost. As the capacity cost increases, the optimal strategy for a firm with dedicated capacities is to offer a more homogeneous pricing and delivery time scheme for both substitutable and non-substitutable products. A shared capacity firm should also always offer more homogeneous delivery times, but needs to increase or decrease the price differentiation, depending on whether the status-quo capacity cost is high or low, respectively.

- Pure price competition decreases individual prices as well as price differentiation. Whereas when firms use delivery times, in addition to prices, as a strategic variable to compete, the effect of competition on product differentiation depends on customer behavior.

The second insight above has major implications for FedEx and UPS who use dedicated and shared capacity strategy, respectively. We show that the differentiation policies adopted by these firms indeed support our results. The third insight is important for firms thinking about modifying the customer access to their product offerings. The fourth insight is relevant in view of volatile fuel prices, which translates into fluctuations in capacity costs, and how firms like FedEx and UPS should change their product differentiation in order to adapt to this new reality.

Chapter 2

Monopolistic Market: Models & Solutions

2.1 Introduction

Progressive Insurance, an automobile insurance company based in Ohio, achieved a sevenfold growth of sales from \$1.3 billion in 1991 to \$9.5 billion in 2002 as a result of introducing an Immediate Response claims system, which dramatically reduced the claim handling time from 7-8 days to just nine hours (Hammer 2004). Shell Lubricants redesigned its order fulfillment process, thus reducing the cycle time by 75% and operating expenses by 45%, and boosting customer satisfaction 105% (Hammer 2004). The above examples highlight the importance of response/delivery time, in addition to pricing policy, to a firm's success. Firms, especially in service and make-to-order manufacturing sectors, are increasingly using explicit delivery time guarantees as a marketing strategy (Hammer 2004, Liu et al. 2007, Zhao et al. 2008). One form of delivery time guarantee, commonly used in retail and service industries, is to announce the delivery time in advance to all prospective customers[1]. For example, Cat Logistics, a subsidiary of Caterpillar, promises to

[1] Another form of time guarantee, popular in make-to-order manufacturing industry, is to dynamically change the quoted delivery time based on congestion in the system when a demand

ship service parts within 24 hours to its clients (Schmidt and Aschkenase 2004). Such guarantees are also used by firms like Ameristock, FedEx, UPS and Domino's Pizza (Zhao et al. 2008, Boyaci and Ray 2003).

Keeping the above discussion in mind, we study a setting where the end customer demand is sensitive to both the price charged and the delivery time guarantee offered. In that case, a firm needs to address two basic issues. The first is related to marketing, and involves determining whether to offer the same product to all its customers (i.e., guarantee the same delivery time at the same price for all), or to offer price-and-delivery-time differentiated products (different delivery times at different prices). Offering the same product/service with different different delivery time guarantees at different prices is popular when customers are heterogenous in their sensitivity to price and time. For example, Plantgel, a firm selling nutrition gels for plants online, offers to process an order within a day for extra \$3, against a regular delivery of 10 days (Zhao et al. 2008). FedEx offers logistic services like "FedEx Next Flight", "FedEx First Overnight", "FedEx Priority Overnight" and "FedEx 2Day", each with a different price and delivery time guarantee to target different customer segments having different sensitivities to price and delivery time. Similarly, UPS offers "UPS Express Early A.M.", "UPS Express", "UPS Express Saver" and "UPS Expedited" for different categories of customers. Obviously, managers also need to decide on the optimal prices and delivery times for whichever policy they choose. A firm's marketing decision cannot be decoupled from its operations strategy in a capacitated environment. So, if a firm decides to guarantee different delivery times to its different customer segments, the second choice to be made is whether to dedicate separate capacities for each market segment or to pool/share the capacities used for all segments, and what will be the corresponding optimal capacity level.

We study firms offering a menu of differentiated products/services to exploit heterogeneity in customers' preference for time and willingness to pay. Such firms

---

arrives (Plambeck 2004).

then need to deal with the issue of whether a given product be accessible to all customers, or need to be customized (and available) for only one segment. For example, the price and delivery time combinations that Dell quotes to government and health-care companies are very different from what they quote to individuals (MacWilliams 2001). In this case, Dell designs a specific service for each market segment, which is not available to the other. Dell's options are, therefore, *non-substitutable*, and the demand for each segment is independent of the other. Similarly, in the steel, chemical and consumer product industries, the price and delivery time quoted to a customer is tailored based on its geographical location and industry segment (Plambeck 2004 and references therein). On the other hand, FedEx and UPS offer logistic services like "FedEx First Overnight", "FedEx 2Day", "UPS Express", "UPS Express Saver", etc., each with a different guaranteed delivery time, to every customer willing to pay the corresponding price. In this case, customers self-select the (delivery time) option based on their preference for speed and willingness to pay. This allows them to switch their preferences, depending on the relative values of prices/delivery times for the products and/or their situational needs. For example, a customer that is otherwise price sensitive may opt for a faster delivery (and more expensive) option in case of an emergency. The menu of products offered are thus *substitutable*, creating a demand-side interaction between the different market segments.

Like the demand side, the supply side for different customer segments may also be independent or related to each other depending on the operations (capacity) strategy used by the firm. By operations (specifically capacity) strategy, we mean whether there are *dedicated capacities* (DC) for each customer segment or there is one *shared capacity* (SC) for all segments. Both strategies are used in practice. FedEx, for example, uses separate facilities for its express and ground services. In contrast, UPS delivers express and ground services using one integrated network. Photo development stores offering express and regular services also share capacity used for the two services. Note that offering different delivery time guarantees using a shared capacity creates a supply-side relationship between the different market segments a firm serves, and thus requires mechanisms for prioritizing orders. This creates operational complexities, potentially increasing costs. Providing different services using dedicated capacities implies that there is no such interaction, but requires additional capacity investment (Zhao et al. 2008).

Our primary objective in this thesis is to understand the interaction between (demand-side) product substitution and (supply-side) operations strategy in a capacitated environment, and how it affects a firm's optimal product differentiation policy. Specifically, we study the following issues.

- How does the operations strategy (dedicated or shared capacity) of a firm affect its optimal price and delivery time decisions for the two products, and hence its product differentiation policy? Are these effects impacted by whether there is a demand-side interaction or not (i.e., whether the products are substitutable or non-substitutable)?
- How does the substitutability between the products a firm offers shape its optimal differentiation decisions, and are these effects influenced by the firm's capacity strategy?
- How does the optimal product differentiation strategy of a firm change with increase in capacity cost under different demand and supply conditions?

In order to answer the above questions, we analyze and compare the four scenarios shown in Table 2.1. Comparison of the two scenarios under the dedicated and shared capacity columns demonstrates the effect of product substitution under two different capacity regimes. On the other hand, comparison of the two scenarios in the "without substitution" and "with substitution" rows shows the effect of the capacity strategy, depending on whether the products are substitutable or not. Note that, although our focus is on comparing different scenarios, our work also distinguishes itself by analyzing the problem of optimal product differentiation in a shared capacity setting, which has not been studied much in the literature (there are some studies in this setting but with very different objectives).

Table 2.1: Different scenarios

	Dedicated capacity ↓	*Shared capacity* ↓
Without substitution →	Non-substitutable products; dedicated capacity	Non-substitutable products; shared capacity
With substitution →	Substitutable products; dedicated capacity	Substitutable products; shared capacity

We first derive the optimal delivery times that the firm should guarantee and the optimal prices it should charge for the two products (consequently, the optimal level of product differentiation) as well as the optimal capacity level it should have/build[2] for each scenario. Note that while the dedicated capacity cases can be solved by functional optimization, for the shared capacity scenarios we utilize a novel methodology involving the matrix-geometric, the cutting plane and the golden section search methods.

Comparison of the results of the four scenarios for various levels of capacity cost enables us to illustrate the individual and joint effects of product substitution and operations strategy on the optimal product differentiation policy of the firm. The rest of the chapter is organized as follows. In §2.2, we briefly review the related literature. §2.3 defines the modelling framework, followed by a discussion on the solution methodology in §2.4. Analysis of results to draw important managerial insights is deferred to chapter 3.

2.2 Related Literature

The literature related to our study can be categorized into four groups, based on whether they consider demand-side and/or supply side interaction (like in Table 2.1).

[2]To satisfy the promised delivery times with a certain degree of reliability.

The papers in the first category study a single product, and hence product differentiation or substitutability is not an issue. These papers can thus be categorized as those that deal with "non-substitutable products; dedicated capacity" scenario. These include So and Song (1998), Palaka et al. (1998) and Ray and Jewkes (2004). All these papers study optimal pricing, delivery time and capacity decisions, while modelling the firm's operations as a single server queue. So and Song (1998) use an M/M/1 queueing model for a firm serving a stream of demands with a mean that has log-linear (Cobb-Douglas) relationship with the price charged and the delivery time guaranteed. They propose a mathematical framework to understand the interrelations among the pricing, delivery time and capacity decisions. They characterize the optimal decisions, and use numerical results to provide managerial insights into the effects of a firm's different operating characteristics on its optimal strategy.

Palaka et al. (1998) use a similar framework but with a linear relationship between the mean demand, price and the delivery time. They also take into account the work-in-process and lateness penalty costs. Ray and Jewkes (2004) further extend this line of work by explicitly modelling price as a function of delivery time, besides demand being a function of price and delivery time. Hill and Khosla (1992) also study a similar tradeoff between price and delivery time but in a deterministic framework.

Besides these, So (2000), Tsay and Agarwal (2000), Allon and Federgruen (2007) and Pekgun et al. (2006) also study similar problems but they are in a competitive setting, where two firms selling a common product compete on price and delivery time. Again, their models do not study product differentiation. We review these, and other papers using a competitive framework, in chapter 4.

The second category of papers takes into account product differentiation and substitution among multiple products, and assume that the products are processed using dedicated capacities. Boyaci and Ray (2003, 2006) are examples of such "substitutable products; dedicated capacity" scenario papers. Boyaci and Ray

(2003) study a firm selling the same product to two customer classes with different delivery time guarantees and at different prices. The firm uses dedicated facilities, each of which is modelled as an M/M/1 queueing system, to serve the two customer classes. The mean demand from each customer class is modelled as a linear function of its own price and delivery time as well as price and delivery time quoted to the other class. They develop a mathematical model that jointly determines the prices, delivery times and the capacity decisions, and study scenarios where the firm is constrained in capacity for none, one or both the customer classes.

Boyaci and Ray (2006) further extend this work to model the dependence of demand rates on delivery reliability guarantees (with which customers are served within their promised delivery times), in addition to prices and delivery times. Zhao et al. (2008) also use a similar modelling framework, but focus on comparing two different delivery time strategies - providing one uniform guaranteed delivery time (and charging one price) for all customers versus providing different guaranteed delivery times (and charging different prices) for different customer segments. Further, rather than explicitly modelling the demand rates for the two customer classes, they use an optimization model for the customers' product selection problem.

There is another stream of literature that models scenarios where capacities are shared for serving different customer segments. For example, Dewan and Mendelson (1990), Mendelson and Whang (1990), Stidham (1992), Afeche (2004), Afeche and Mendelson (2004) and Katta and Sethuraman (2005) study pricing and/or capacity selection issues for heterogeneous customers in a queuing context, wherein all customers are served by the same service facility. Since they do not deal with substitution issue, these papers fall under the "non-substitutable products; shared capacity" category. In general, the problem considered in these papers, except Afeche (2004), is to design an incentive compatible pricing and scheduling policy that maximizes the expected net value of the jobs processed by the system. In contrast, our model has the firm's profit maximization as its objective. Moreover, these models employ user delay costs, which is fundamentally different from our approach of using a delivery time guarantee. In further contrast, we do not explicitly model the customers' utility/value functions but assume that the demand function is an outcome of some underlying process whereby customers select the service class that maximizes their utility[3].

The work that is closest to ours in this category is by Sinha et al. (2008). They consider an operational setting in which a resource/server, which already serves an existing class (called primary class) of customers, is shared with a new class (called secondary class) of customers. The resource owner uses a delay dependent priority discipline (see Kleinrock 1964, Kanet 1982) to serve the two classes of customers. The problem is to determine the optimal price and the guaranteed delivery time (called quality of service) to the secondary customers, and the optimal parameter that specifies the relative delay dependent priority of one class of customers over the other. They use a linear demand function for the secondary class customers, which depends on its own price and delivery time, but is independent of the price and delivery time already being offered to the existing customers. In this sense, the services offered to the two customer classes are non-substitutable. They also assume that the resource owner has already entered into a long term agreement with the primary customers, and hence the price and delivery time offered to them are assumed as fixed. In contrast, we consider prices of both the classes as decision variables. Further, they consider service levels based on average delivery times of served customers, which is very different from our definition of service level based on probability distribution of the delivery times of served customers. This does not provide any bound on instances of unusually long delivery times. It is quite possible then that a large portion of the demands are actually not served within their promised delivery times, even if the promised delivery times are met on average. We, therefore, assume that firms select their capacity levels so as to fulfill their promised delivery times with a high level of reliability (generally 99%). This makes the delivery time guarantees more attractive, although it makes the problem a lot

[3]Liu et al. (2007) shows the equivalence between a utility function and the corresponding demand model.

more challenging to solve. In further contrast, they assume the installed capacity as fixed, whereas we assume that it is possible to alter the processing capacity in the short term by, for example, adding a shift or installing additional equipment.

Lastly, Ata and Van Mieghem (2008) study the conditions under which heterogeneous customers should be served by dedicated resources or by an integrated network through partial pooling of resources. In their setting, customer segments are served by capacities dedicated for each, but capacities can also be dynamically substituted. Their main goal is to understand the value of network integration. We can place this paper in the "substitutable products; shared capacity" cell since they consider resource substitution. However, note that they do not deal with product substitution or pricing/delivery time decisions, and so do not capture the interaction between product substitution and capacity strategy.

Our work also fits in the stream of price discrimination, extensively studied in the Economics (Industrial Organization) literature. The case of "without substitution" corresponds to third-degree price discrimination wherein different customers are quoted different prices based on their distinctive characteristics. Whereas the case of "with substitution" corresponds to second-degree price discrimination wherein customers are allowed to self-select from a given menu of options. More discussion on price discrimination can be found in Talluri and Van Ryzin (2004).

Our study complements the existing literature on pricing and delivery time decisions by delineating the individual and joint effects of supply and demand side interactions in a capacitated environment. This allows us to generate new managerial insights regarding how the optimal product differentiation strategy for firms should vary depending on their operations strategy, product offering portfolio, market characteristics and capacity costs.

2.3 Decision Models

2.3.1 Modelling Assumptions

We model a Make-to-Order (MTO) or a service firm, and hence delivery time is a key element to customer service. The firm offers a single product/service (henceforth called product) in a market comprising heterogenous customers that differ in their preferences for speed and willingness to pay. The firm exploits this heterogeneity in customers' preferences to create market segments in which customers are quoted a menu of different delivery times and corresponding prices for (otherwise) the same product. For simplicity, we assume the market is segmented into two customer classes, indexed by $k \in \{h, l\}$. Class h customers are high priority/express customers who are more time sensitive and are willing to pay a price premium for a shorter delivery time. Class l customers are low priority/regular customers who are more price sensitive and are willing to accept a longer delivery time for a price discount. p_k and L_k denote the price and delivery time offered by the firm to customer class $k \in \{h, l\}$.

Demand from customer class k arrives according to a Poisson process with rate $\lambda_k(p_k, L_k, k \in \{h, l\})$, which depends not only on its own absolute price and delivery time but also on its price and delivery time quoted relative to the other class. The firm can, therefore, attract new customers through price reductions and/or by offering shorter delivery times. Lowering the price and/or delivery time for one class can also induce customers to switch preferences. We assume that customers cannot observe the congestion levels of the firms, and their choices are only based on the prices and delivery times announced by the firms. The demand rates are modelled using the following linear functions, inspired by Tsay and Agrawal (2000) and Boyaci and Ray (2003):

$$\lambda_h = a - \beta_p^h p_h + \theta_p(p_l - p_h) - \beta_L^h L_h + \theta_L(L_l - L_h) \tag{2.1}$$

$$\lambda_l = a - \beta_p^l p_l + \theta_p(p_h - p_l) - \beta_L^l L_l + \theta_L(L_h - L_l) \tag{2.2}$$

where,

$2a$	:	market base
β_p^k	:	sensitivity of class k demand to its own price
β_L^k	:	sensitivity of class k demand to its own guaranteed delivery time
θ_p	:	sensitivity of demand to inter-class price difference
θ_L	:	sensitivity of demand to inter-class delivery time difference

$2a$ parameterizes the total market base. Mathematically, it is the total demand when price and delivery time offered to each customer class is zero. It captures the aggregate effect of all the factors other than price and delivery time on demand. For logistics service providers like FedEx and UPS, for example, these other factors may include factors like the convenience of pick-up, the ease with which deliveries can be tracked and the likelihood of the packages being damaged. For internet service providers, these may include factors like the frequency of service interruptions and the quality of the support staff (Allon and Federgruen 2008). Our demand model generalizes the one used by Tsay and Agrawal (2000) and Boyaci and Ray (2003) by using different sensitivities (to price and time) for regular and express customers. We feel it is necessary to use different sensitivities for the two customer classes as this is essentially what differentiates express customers from regular ones. However, the sensitivities of demand switchovers (θ_p and θ_L) are still the same across the two classes, as is required to make the total market size invariant to changes in these sensitivities. Our demand model differs from Pekgun et al. (2006) for similar reasons. We make the following assumptions regarding the market parameters:

Assumption 2.1. $\beta_p^k > 0$, $\beta_L^k > 0$, $\theta_p \geq 0$, $\theta_L \geq 0$, $\beta_p^h < \beta_p^l$ *and* $\beta_L^h > \beta_L^l$.

This is to ensure that demand from a market segment is decreasing in its own price and delivery time, and is increasing in price and delivery time offered to the other segment. $\beta_p^h < \beta_p^l$ and $\beta_L^h > \beta_L^l$ are required by definition of the two customer classes.

Assumption 2.2. *The market base a is sufficiently large.*

This assumption is required, as we will see in §2.4, to ensure that the optimal prices and demands are non-negative and that the optimal delivery time offered to time sensitive express customers is smaller than what is offered to price sensitive customers, i.e., $L_h < L_l$.

The exact behavior of the market depends on the specific combination of market parameter values. One extreme case is when $\theta_p = \theta_L = 0$, which models the "without substitution" scenario. In a "with substitution" scenario, two cases of special interest, as we see in chapter 3, are described as:

- *Time Difference Sensitive (TDS)*: We say the market is TDS type when the relative sensitivity of customers to the difference in delivery times (with respect to their own delivery time) is greater than their relative sensitivity to the price difference (with respect to their own price), such that $\theta_L/\beta_L^k > \theta_p/\beta_p^k$, $k \in \{h, l\}$.
- *Price Difference Sensitive (PDS)*: We say the market is PDS type when the relative sensitivity of customers to the price difference (with respect to their own price) is greater than their relative sensitivity to the difference in delivery times (with respect to their own delivery time), such that $\theta_p/\beta_p^k > \theta_L/\beta_L^k$, $k \in \{h, l\}$.

The choice of a linear demand function arises partly from its simplicity, which makes the model tractable, and allows us to obtain qualitative insights without much analytical complexity. Besides, it also possesses some desirable properties that are not exhibited even by the more popular Cobb-Douglas function (Palaka et al. 1998). This is clear from the expressions for the price and delivery time elasticities of demand:

$$\begin{aligned} E_{p_k} &= -\frac{d\lambda_k}{dp_k}\frac{p_k}{\lambda_k} \\ &= \frac{(\beta_p^k + \theta_p)p_k}{a - \beta_p^k p_k + \theta_p(p_j - p_k) - \beta_L^k L_k + \theta_L(L_j - L_k)}, \quad k, j \in \{h, l\}, j \neq k \end{aligned}$$

$$E_{L_k} = -\frac{d\lambda_k}{dL_k}\frac{L_k}{\lambda_k}$$
$$= \frac{(\beta_L^k + \theta_L)L_k}{a - \beta_p^k p_k + \theta_p(p_j - p_k) - \beta_L^k L_k + \theta_L(L_j - L_k)}, \quad k, j \in \{h, l\}, j \neq k$$

Clearly, the price elasticity of demand (E_{p_k}) for a given segment k is increasing in its own delivery time L_k. Similarly, the delivery time elasticity of demand (E_{L_k}) for a given segment k is increasing in its own price p_k. These properties are desirable since we expect customers to be more sensitive to their price when they have longer waiting times, and more sensitive to their delivery time when they are paying a higher price.

We assume the time it takes to serve a demand from class k is exponentially distributed with rate μ_k, $k \in \{h, l\}$. The service facility is thus modelled as an M/M/$\cdot$ queuing system. M/M/$\cdot$ queuing model is a traditional abstraction employed to make the problem tractable, especially when the emphasis is more on managerial insights than on accuracy (Palaka et al. 1998). Moreover, Lariviere and Van Mieghem (2004) have shown that self interested customers try to spread themselves out as much as possible and the arrival pattern this generates approaches a Poisson process as the number of customers and arrival points gets large. We further assume customers within each class are served on a first-come-first-serve (FCFS) basis. The firm can invest in its installed capacity to increase its processing rate μ_k. Since the pricing and delivery time decisions are generally short-term operating decisions, the capacity decision we consider here typically refers to expanding short term capacity in the existing facilities such as adding a shift or installing an additional equipment, rather than some long-term strategic decisions such as building new service facilities. We assume there is no economies of scale in investing in capacity. So a unit increment in μ_k per unit time always costs \$$A$. A may be different for different customer classes if they are served by different service capacities (e.g., express customers served by airplanes and regular customers served by trucks in a service logistics industry) or they may be equal if both the classes are served by the same service capacity. Using the same marginal capacity cost for the two customer classes, however, allows a meaningful comparison between the dedicated and the shared capacity settings. We also assume that the firm incurs the same operating cost of \$$m$ in serving a customer of either class.

The industry is assumed to have established a standard delivery time L_l for regular customers. The objective of the firm is to set the guaranteed delivery time L_h for express customers and the prices p_h and p_l for both classes, so as to maximize its profit per unit time. Obviously, a firm's pricing and delivery time decisions depend crucially on its capacity decision. Firms may charge premium prices by committing to shorter delivery times. This, however, puts pressure on the firm's available resources to reliably meet its promised delivery times. Failure to meet the guarantee may lead to penalties, either in the form of a discount, partial refund or an expedited delivery without additional charge to the customer (Liu et al. 2007). FedEx, for example, offers a money-back guarantee for every U.S. shipment that is even 1 minute late compared to its guaranteed delivery time [4]. Similarly, Black Angus Restaurants offer their customers free lunches if not served within 10 minutes (Charney 1991). A striking example is the case of seven online retailers, including Macys.com, Toysrus.com and CDNOW, that paid fines to the tune of \$1.5 million to settle a Federal Trade Commission lawsuit over late deliveries made in 1999 (Pekgun 2007). The firm, therefore, needs to simultaneously select the optimal service rates (i.e., capacities) μ_h and μ_l in order to meet the guaranteed delivery times with at least a minimum level of reliability α (called the target service level). The target service level α is set by the management as an internal performance measure, which is not quoted to the customers. Thus, we do not explicitly consider its impact on the mean demand in our demand model (2.1) and (2.2). However, since failure to honor its promised delivery time often leads to penalties for the firm, α is set to a high value, close to 1. This means that the chances of a customer of class k having to wait longer than L_k are very small. A schematic representation of the model is shown in Figure 2.1.

[4] http://www.fedex.com/us/services/options/mbg.html

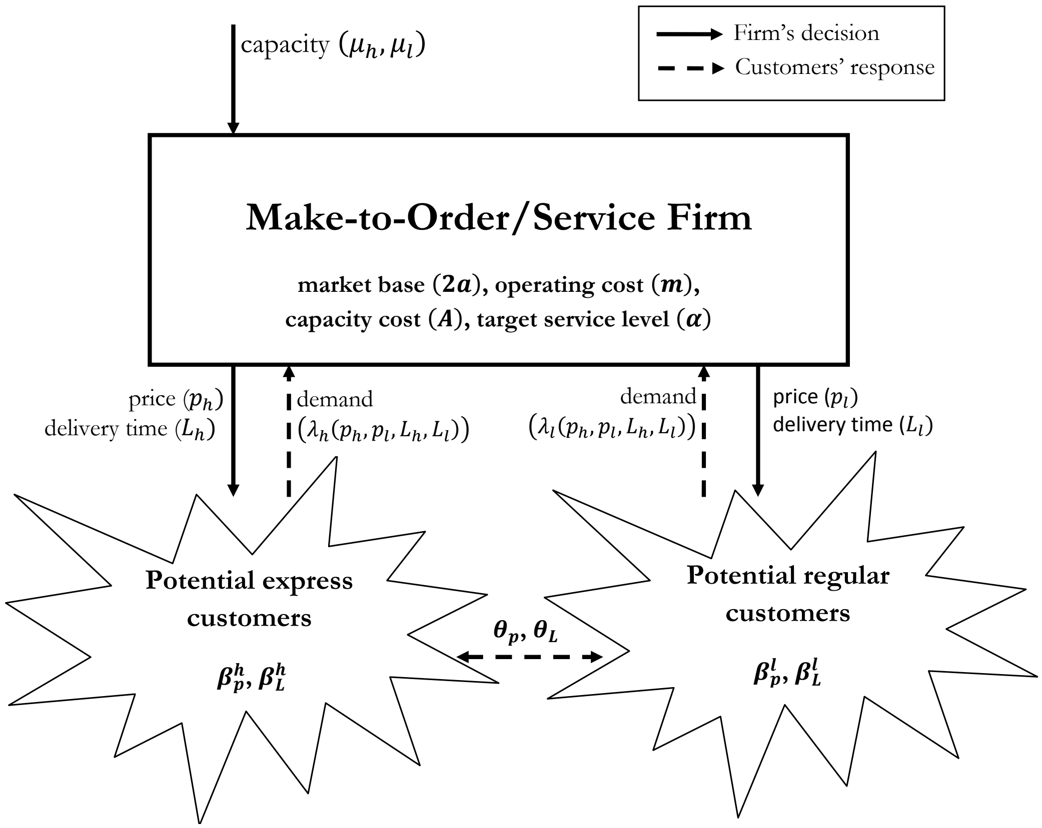


Figure 2.1: Schematic representation of a monopolistic model

Notation

k	:	index for customer class; $k \in \{h, l\}$
λ_k	:	mean demand rate from customer class k (units/unit time)
μ_k	:	mean processing rate for customer class k (units/unit time)
p_k	:	price charged to customer class k (\$/unit)
L_k	:	delivery time quoted to customer class k (time units)
W_k	:	steady state actual sojourn (waiting + service) time of customer class k (time units)
α	:	target service level (no unit)
$S_k(L_k)$	:	actual service level achieved for quoted delivery time L_k, i.e., $P(W_k \leq L_k)$ (no unit)
m	:	unit operating cost (\$/unit)
A	:	marginal capacity cost (\$/unit)

2.3.2 Mathematical Model

The firm's problem of determining the optimal prices, delivery times and processing rates can be mathematically stated as:

<u>**PDTDP**</u> :

$$\max_{p_h,p_l,L_h,\mu_h,\mu_l} \pi = (p_h - m)\lambda_h + (p_l - m)\lambda_l - A(\mu_h + \mu_l) \tag{2.3}$$

subject to:

$$L_h < L_l \tag{2.4}$$

$$p_h, p_l, \lambda_h, \lambda_l, \mu_h, \mu_l, L_h \geq 0 \tag{2.5}$$

$$\textit{Stability condition} \tag{2.6}$$

$$S_h(L_h) = P(W_h \leq L_h) \geq \alpha \tag{2.7}$$

$$S_l(L_l) = P(W_l \leq L_l) \geq \alpha \tag{2.8}$$

where λ_h and λ_l are given by (2.1) and (2.1) respectively. Constraint (2.4) requires that the guaranteed delivery time for high priority customers be shorter than that for the other class. Constraint set (2.5) is needed to define a realistic problem setting. Constraint (2.6) is the stability condition for the queuing system, which models the service facility at the firm. Later, we will see that, irrespective of the capacity setting, this condition is automatically satisfied by the remaining constraints, and hence excluding this constraint leaves the feasible region of the problem unchanged. Constraints (2.7) and (2.8) are delivery time reliability constraints (also called service level constraints), which say that the steady state actual delivery time W_h (resp., W_l) of a customer should not exceed the guaranteed delivery time L_h (resp., L_l) with a probability of at least α. We call the above mathematical model a Pricing and Delivery Time Decision Problem (PDTDP).

A special case of PDTDP is where the delivery times are fixed such that the prices and capacities are the only decisions made by the firm. This is relevant to situations where a firm may face a significantly higher stickiness for their delivery time decisions compared to their ability to vary prices (Allon and Federgruen 2007).

A relatively higher stickiness for delivery time decisions may arise, for example, when the services are partly outsourced to a third party. For example, a logistics firm like FedEx may maintain its own fleet of airplanes for international or inter-state deliveries but beyond that it may outsource its delivery services to a third party logistics service provider. In such a situation, its delivery times are dictated partly by its service level agreement with the third partly logistics service provider. Any change in its delivery time guarantee to its own customers thus requires a renegotiation of its service level agreements with the third party logistics service provider. In such a situation, the firm may not be able to revise its delivery time decisions as frequently as it can revise its prices.

Higher stickiness in delivery time decisions may also arise because of human resource practices or labor contracts that prohibit frequent changes to installed capacity via changes to workforce. Or it may arise due to long lead times for technology purchases (Allon and Federgruen 2007). In such a case, a firm may fix a part of its capacity investment, which varies with its guaranteed service level, by maintaining a fixed delivery time standard over a longer horizon. For example, airline call centers are designed to handle 80% of the economy class passengers within 20 seconds. Airlines have stuck for years to the same waiting time standard, while willing to change prices daily (Allon and Federgruen 2007). A significantly high stickiness in delivery time decisions may prevent a firm from frequently adjusting its delivery times in response to any change in market parameters (β's and θ's) or in its operating parameters (m, A). Under such situations, a firm optimizes its prices, treating its delivery times as fixed. We call this special case a Pricing Decision Problem (PDP).

Note that the above model ((2.1) - (2.8)) is a general one that is applicable to all the scenarios in Table 1. In what follows, we develop the exact framework for each of the four scenarios by specifying: i) the form of constraints (2.6)-(2.8) depending on the capacity strategy used (shared or dedicated), and ii) the form of the demand function that signifies absence or presence of product substitution.

Dedicated Capacity Setting

For a dedicated capacity setting, where each customer class is served by a separate M/M/1 server, the sojourn time distribution for either class of customers is known to be exponential. In this case, there is a separate stability condition for each of the queues. Hence, constraints (2.6), (2.7) and (2.8) can be expressed as:

$$\lambda_k - \mu_k < 0, \ k \in \{h, l\} \tag{2.6^{DC}}$$

$$S_h(L_h) = P(W_h \leq L_h) = 1 - e^{(\lambda_h - \mu_h)L_h} \geq \alpha \tag{2.7^{DC}}$$

$$S_l(L_l) = P(W_l \leq L_l) = 1 - e^{(\lambda_l - \mu_l)L_l} \geq \alpha \tag{2.8^{DC}}$$

The two demand scenarios, substitutable and non-substitutable products, can be obtained with $\theta_p > 0$, $\theta_L > 0$ and $\theta_p = \theta_L = 0$, respectively, in (2.1) and (2.2). We denote the resulting models of Pricing and Delivery Time Decision Problem in a Dedicated Capacity setting by $PDTDP_{DC}$.

As noted above, although we have explicitly included (2.6^{DC}) in $PDTDP_{DC}$ to ensure stability of the system, this is implicitly satisfied by the delivery time reliability constraints (2.7^{DC}) and (2.8^{DC})). This is clear from the following alternate representation of constraints (2.7^{DC}) and (2.8^{DC}):

$$\mu_k \geq -\frac{\ln(1-\alpha)}{L_k} + \lambda_k, \ k \in \{h, l\}$$

For a practical problem, L_l is finite, and L_h is also finite since $L_h < L_l$ for any feasible solution. For a finite L_k, $\frac{\ln(1-\alpha)}{L_k} < 0$ since $\alpha < 1$. This implies that any solution that satisfies (2.7^{DC}) and (2.8^{DC}) will automatically satisfy the stability condition: $\mu_k > \lambda_k$, $k \in \{h, l\}$.

Shared Capacity Setting

The firm's choice of shared capacity is modelled using a single server, which serves both customer classes employing a simple fixed priority scheme that always gives priority to time-sensitive customers. In other words, the firm reserves its capacity to serve high priority customers who pay a premium price, and uses the remaining capacity to serve low priority customers. This somewhat reflects the practice at UPS (Ata and Van Mieghem 2008). Dedicated capacities with partial pooling more accurately model the operational setting used by UPS where fast airplanes can serve both express and regular markets, while the slow trucks can serve only the regular market. We use shared capacity (with complete pooling) to study the extreme scenario and compare it with the dedicated capacity setting, typical of FedEx. The fixed priority scheme is also used by Plambeck (2004) where premium customers are given priority in scheduling. Customers within each class are served on a first-come-first-served (FCFS) basis. In this paper, we use a preemptive priority scheme, but the analysis can easily be extended to a non-preemptive priority discipline.

For a shared capacity setting, the sojourn time distribution $S_h(\cdot)$ for high priority customers in a preemptive priority queue is known to be exponential (So 2000). Hence, the delivery time reliability constraint (2.7) has the same analytical representation as that for the dedicated capacity setting. However, a closed form expression for the sojourn time distribution $S_l(\cdot)$ for low priority customers, appearing in equation (2.8) of PDTDP, is not known (Abate and Whitt 1997). We assume the single server serves customers of either class at the same rate $\mu_h = \mu_l = \mu$, which is a decision variable. Constraints (2.6) and (2.7) in a shared capacity setting can then be expressed as:

$$\lambda_h + \lambda_l - \mu < 0 \qquad (2.6^{SC})$$

$$S_h(L_h) = P(W_h \leq L_h) = 1 - e^{(\lambda_h - \mu)L_h} \geq \alpha \qquad (2.7^{SC})$$

We discuss how we tackle the issue of delivery reliability for regular customers (corresponding to Equation (2.8)) in the next section. Like before, the substitutable and non-substitutable demand cases can be obtained with $\theta_p > 0$, $\theta_L > 0$ and $\theta_p = \theta_L = 0$, respectively, in (2.1) and (2.2). We denote the resulting models of Pricing and Delivery Time Decision Problem in a Shared Capacity setting by $PDTDP_{SC}$ (including $S_l(L_l)$ constraint).

We now show that any solution that satisfies the two delivery time reliability constraints of $PDTDP_{SC}$ will automatically satisfy the stability condition (2.6^{SC}). We first note that the delivery time reliability constraint for express customers has the same form in both the dedicated and shared capacity settings. Therefore, as shown for dedicated case, any feasible solution will always satisfy: $\lambda_h < \mu$. However, if $\lambda_h < \mu$ but $\lambda_h + \lambda_l \rightarrow \mu$, then the queue of regular customers will grow infinitely long such that the probability of serving a regular customer within a finite time will approach 0. Thus, the delivery time reliability constraint for regular customers can never be satisfied. Therefore, any solution that satisfies the two delivery time reliability constraints of $PDTDP_{SC}$ will always satisfy the stability condition: $\lambda_h + \lambda_l < \mu$.

We summarize the above mathematical models in the following table, which corresponds to the four scenarios described in Table 2.1.

Table 2.2: Mathematical Models for the different scenarios as special cases of PDTDP

	Dedicated capacity ↓	*Shared capacity* ↓
Without substitution →	$PDTDP_{DC}$ with $\theta_p = \theta_L = 0$	$PDTDP_{SC}$ with $\theta_p = \theta_L = 0$
With substitution →	$PDTDP_{DC}$ with $\theta_p > 0$, $\theta_L > 0$	$PDTDP_{SC}$ with $\theta_p > 0$, $\theta_L > 0$

2.4 Solution Methodology

We now discuss the solution methodology for the models discussed in §2.3. The four different mathematical models for the corresponding scenarios described in §2.1 (Table 2.1) are shown in Table 2.2. We essentially have two different mathematical models: $PDTDP_{DC}$ and $PDTDP_{SC}$, corresponding to dedicated capacity and shared capacity settings. The corresponding mathematical models for "without substitution" and "with substitution" scenarios in each of the two capacity settings are obtained simply by substituting $\theta_p = 0$, $\theta_L = 0$ and $\theta_p > 0$, $\theta_L > 0$, respectively. We now discuss the solution to $PDTDP_{DC}$ and $PDTDP_{SC}$.

2.4.1 Dedicated Capacity Setting

We first state some propositions, which are used to arrive at the final results.

Proposition 2.1. *In a dedicated capacity setting, both the delivery time reliability constraints* (2.7^{DC}) *and* (2.8^{DC}) *in* $PDTDP_{DC}$ *are binding at optimality.*

Proof. Delivery time reliability constraints (2.7^{DC}) and (2.8^{DC}) can be restated as:

$$\mu_k \geq -\frac{\ln(1-\alpha)}{L_k} + \lambda_k \quad k \in \{l, h\}$$

The profit function π is decreasing in μ_k. Therefore, to maximize profit, the two service rates should be at their minimum levels that guarantee the desired service level α. This implies that at optimality, the two delivery time reliability constraints (2.7^{DC}) and (2.8^{DC}) must be binding, and the service rates are given by:

$$\mu_k = -\frac{\ln(1-\alpha)}{L_k} + \lambda_k, \; k \in \{h, l\}$$

□

Proposition 2.1 suggests that it is optimal for firms to stick to their minimum delivery time reliability (α) since a better service level to customers comes at an extra cost to the firm. As a result of Proposition 2.1, $PDTDP_{DC}$ reduces to maximizing (2.3) with μ_i as given above. Note that the stability conditions (2.6^{DC}) are automatically satisfied by the above equation. This allows us to reduce $PDTDP_{DC}$ to the following optimization problem:

<u>$\mathbf{PDTDP'_{DC}}$:</u>

$$\max_{p_h, p_l, L_h} \pi = (p_h - m - A)\lambda_h + (p_l - m - A)\lambda_l + A\frac{\ln(1-\alpha)}{L_h} + A\frac{\ln(1-\alpha)}{L_l} \tag{2.9}$$

subject to:

$L_h < L_l$

$p_h, p_l, \lambda_h, \lambda_l, L_h \geq 0$

Proposition 2.2. *For a fixed L_h, the objective function (2.9) of $PDTDP'_{DC}$ is strictly concave in p_h and p_l.*

Proof. The Hessian for (2.9), for a fixed L_h, is given by:

$$\begin{pmatrix} -2(\beta_p^h + \theta_p) & 2\theta_p \\ 2\theta_p & -2(\beta_p^l + \theta_p) \end{pmatrix}$$

Clearly, the first order leading principal minor of the Hessian is negative, while its determinant is positive. This proves that the objective function (2.9) in $PDTDP'_{DC}$ is strictly concave for a fixed L_h. □

Proposition 2.2 suggests that, for a fixed L_h, $PDTDP'_{DC}$ has a unique maximum, which can be obtained using functional optimization of its objective function (2.9), as long as p_h, p_l, λ_h and λ_l are non-negative and $L_h < L_l$. We ensure that these constraints are satisfied at optimality by imposing restrictions on our model parameter values.

Proposition 2.3. *For a fixed express delivery time L_h, the optimal prices in a dedicated capacity setting are given by:*

$$p_h^{DC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^l + 2\theta_p)a - (\beta_p^l\beta_L^h + \beta_p^l\theta_L + \beta_L^h\theta_p)L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{2.10}$$

$$p_l^{DC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^h + 2\theta_p)a + (\beta_p^h\theta_L - \beta_L^h\theta_p)L_h - (\beta_p^h\beta_L^l + \beta_p^h\theta_L + \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{2.11}$$

Proof. $p_h^{DC*}(L_h)$ and $p_l^{DC*}(L_h)$ are obtained by solving the following system of equations:

$$\frac{\partial \pi(L_h)}{\partial p_h} = 0$$

$$\frac{\partial \pi(L_h)}{\partial p_l} = 0$$

Since (2.9), for a fixed L_h, is strictly concave in p_h and p_l, solving the above system of equations gives a unique pair of prices that maximizes $\pi(L_h)$. □

The corresponding optimal price differentiation is then:

$$p_h^{DC*}(L_h) - p_l^{DC*}(L_h) = \frac{(\beta_p^l - \beta_p^h)a + \beta_p^h \beta_L^l L_l - \beta_p^l \beta_L^h L_h + (\beta_p^h + \beta_p^l)\theta_L(L_l - L_h)}{2(\beta_p^h \beta_p^l + \beta_p^h \theta_p + \beta_p^l \theta_p)} \tag{2.12}$$

Example 2.1: Assume the parameter values as shown in Table 2.3. The optimal prices obtained using Proposition 2.3, and other related variables, for $L_h = 0.50$ are shown in Table 2.4.

Table 2.3: Parameter values for Example 2.1

β_p^h	β_p^l	θ_p	β_L^h	β_L^l	θ_L	a	m	A	α	L_l
0.5	0.7	0.2	0.9	0.7	0.5	10	3	0.5	0.99	1

Table 2.4: Results for Example 2.1

$p_h^*(L_h)$	$p_l^*(L_h)$	$\mu_h^*(L_h)$	$\mu_l^*(L_h)$	$\lambda_h^*(L_h)$	$\lambda_l^*(L_h)$	$\pi^*(L_h)$
10.7585	8.7797	13.2353	7.9052	4.0250	3.3000	39.7305

Proposition 2.4. *The optimal express delivery time L_h^{DC*} in a dedicated capacity setting is given by the unique root of (2.13) in the interval $[0, L_l)$*

$$\frac{\partial \pi(L_h)}{\partial L_h} = -\left(\beta_L^h + \theta_L\right)\left(p_h^{DC*}(L_h) - m - A\right) + \theta_L\left(p_l^{DC*}(L_h) - m - A\right) - \frac{A\ln(1-\alpha)}{L_h^2} \tag{2.13}$$

where, $p_h^{DC*}(L_h)$ and $p_l^{DC*}(L_h)$ are given by (2.10) and (2.11).

Proof. Substituting the optimal prices, given by Proposition 2.3, into the objective function, and differentiating it with respect to L_h gives (2.13). Also,

$$\frac{\partial^2 \pi(L_h)}{\partial L_h^2} = \frac{(\beta_p^l + \theta_p)(\beta_L^h)^2 + (\beta_p^h + \beta_p^l)(\theta_L)^2 + 2\beta_p^l \beta_L^h \theta_L}{2(\beta_p^h \beta_p^l + \beta_p^h \theta_p + \beta_p^l \theta_p)} + \frac{2A \ln(1-\alpha)}{L_h^3} \tag{2.14}$$

$$\frac{\partial^3 \pi(L_h)}{\partial L_h^3} = -\frac{6A \ln(1-\alpha)}{L_h^4} \tag{2.15}$$

Let us understand the nature of the profit function $\pi(L_h)$ as we vary L_h. Since $L_h \in [0, L_l)$, we are interested in its behavior only for non-negative values of L_h.

Property 2.1. *As $L_h \to 0^+$, $\pi(L_h) \to -\infty$.*

This is obvious from the expression for $\pi(L_h)$ in (2.9).

Property 2.2. *$\pi(L_h)$ is increasing concave in L_h in the vicinity of $L_h = 0^+$.*

This can be easily verified by noting that as $L_h \to 0^+$, (2.13) $\to +\infty$ and (2.14) $\to -\infty$.

Property 2.3. *As L_h increases from 0, $\pi(L_h)$ changes from concave to convex for some $L_h \in (0, +\infty)$, and never becomes concave again.*

Since (2.15) is always positive for $L_h \in [0, +\infty) \Rightarrow$ (2.14) is monotonically increasing in $[0, +\infty)$. This implies that as L_h increases from 0, (2.14) changes sign from negative to positive, and hence $\pi(L_h)$ changes from concave to convex, for some $L_h \in (0, +\infty)$, and never changes to concave again. Using properties 2.1, 2.2 and 2.3, the nature of $\pi(L_h)$ in $[0, +\infty)$ can be summarized as shown as in Figure 2.2.

It is clear from the behavior of $\pi(L_h)$, as shown in Figure 2.2, that it has a unique maximum and at most one minimum in $[0, +\infty)$. The stationary points are given by the roots of (2.13) in $[0, +\infty)$, and the maximum is always the smaller of the two. Further, $\frac{\partial \pi(L_h)}{\partial L_h}\big|_{L_h = L_l} < 0$ is sufficient to guarantee that (2.13) has a unique root in the interval $[0, L_l)$, and that it is the point of maximum. The condition simplifies to:

$$\frac{-\left\{(\beta_p^l - \beta_p^h)\theta_L + \beta_p^l \beta_L^h + 2\beta_L^h \theta_p\right\} a + \left\{(\beta_p^l \beta_L^h + \beta_L^h \theta_p + \beta_L^l \theta_p)\beta_L^h + (\beta_p^l \beta_L^h - \beta_p^h \beta_L^l)\theta_L\right\} L_l}{2(\beta_p^h \beta_p^l + \beta_p^h \theta_p + \beta_p^l \theta_p)}$$

$$+ \frac{\beta_L^h (A + m)}{2} - \frac{A \ln(1-\alpha)}{(L_l)^2} < 0 \tag{2.16}$$

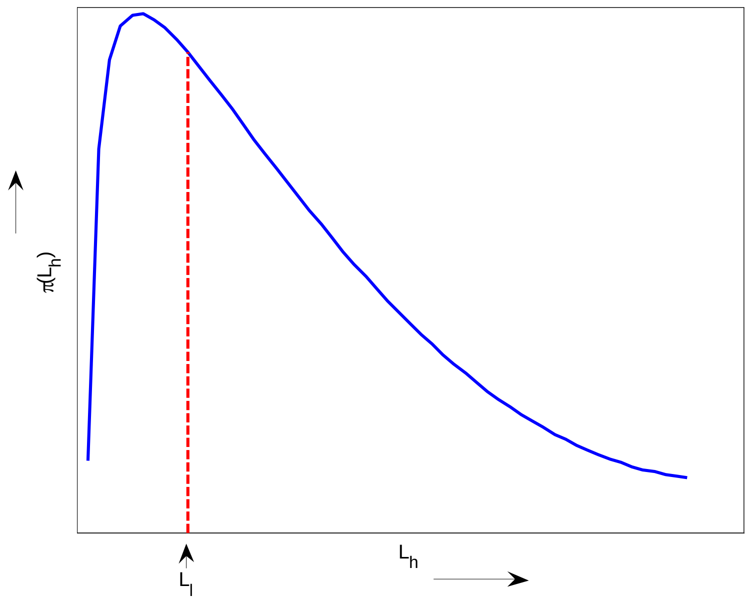


Figure 2.2: Behavior of the profit function for $L_h \in [0, +\infty)$

Since $\beta_p^h < \beta_p^l$ (Assumption 2.1), a necessary condition for (2.16) to hold is a to be high (Assumption 2.2). A sufficiently high value of a also guarantees $p_k > 0$, $p_h > p_l$ and $\lambda_k > 0$. □

Proposition 2.4 gives the optimal express delivery time L_h^{DC*} in a dedicated setting. L_h^{DC*} does not have a closed-form analytical solution. However, it can be obtained numerically using a simple bisection method (Burden and Faires 2000) since $\pi(L_h)$ is unimodal in $[0, L_l)$ (see proof of Proposition 2.4). The optimal prices can be obtained using Proposition 2.3 by substituting $L_h = L_h^{DC*}$. If $\theta_p > 0$, $\theta_L > 0$ in the above equations then we have the solution for the "substitutable products; dedicated capacity" case, while $\theta_p = \theta_L = 0$ in the above equations will generate the solution for the "non-substitutable products; dedicated capacity" scenario.

Example 2.2: Assume the parameter values as shown in Table 2.3. The optimal decisions, obtained using Proposition 2.4, demand and profit are shown in Table 2.5.

The above example suggests that if the firm had flexibility in selecting its express delivery time as well then $L_h = 0.5$ in Exampe 2.1 was not optimal. The firm should,

Table 2.5: Results for Example 2.2

L_h^*	p_h^*	p_l^*	μ_h^*	μ_l^*	λ_h^*	λ_l^*	π^*
0.5562	10.7032	8.7830	12.2653	7.9192	3.9857	3.3141	39.7753

in fact, increase its delivery time offered to express customers from 0.5 to 0.5562, and at the same time decrease its express price from 10.7585 to 10.7032 and increase its regular price from 8.7797 to 8.7830. This increases its profit from 39.7305 to 39.7753.

Example 2.3: Table 2.6 shows the optimal price, delivery time and capacity decisions, obtained using Proposition 2.4, for the parameter combinations shown in Table 2.3 and for various combinations of substitution parameters and capacity cost.

Table 2.6: Results for Example 2.3

	Without substitution ($\theta_p = \theta_L = 0$)		**With substitution** ($\theta_p = 0.2,\ \theta_L = 0.5$)		**With substitution** ($\theta_p = 0.4,\ \theta_L = 0.3$)	
	A=0.10	*A=1.0*	*A=0.10*	*A=1.0*	*A=0.10*	*A=1.0*
L_h^*	0.2494	0.8405	0.2389	0.8201	0.2569	0.8716
p_h^*	11.3255	11.2436	10.8152	10.6938	10.3582	10.3639
p_l^*	8.1929	8.6429	8.5642	9.0487	8.8789	9.2512
μ_h^*	22.5768	9.1010	23.5846	9.2912	22.1447	8.9103
μ_l^*	8.1702	7.8552	7.9799	7.8102	8.0587	7.8359

2.4.2 Shared Capacity Setting

The shared capacity model $PDTDP_{SC}$ is relatively more challenging to solve, especially in the absence of an analytical characterization of the delivery time reliability constraint (2.8) for regular customers. While the Laplace transform of the sojourn time distribution $S_l(\cdot)$, appearing in (2.8), and its first few moments are well known (see Stephan 1958, Cohen 1982, Heyman and Sobel 1982), the distribution itself is somewhat complicated and requires numerical computation for the inverse Laplace transform, thereby preventing its analytical characterization. There are approximations proposed in the literature for the sojourn time distribution. However, they are very complex and often not sufficiently accurate (Abate and Whitt 1997). Moreover, the appropriate form of approximation to use depends on the relative demand rates of the two customer classes, which can only be determined endogenously, and are not known in advance in our model. Further, even an analytical characterization of the sojourn time distribution or a good approximation will not produce an analytical solution similar to that for $PDTDP_{DC}$ since it cannot be guaranteed at the outset which of the constraints will be binding at optimality. So $PDTDP_{SC}$ does not lend itself to an easy solution using conventional optimization methods. We resolve this difficulty by solving it in two stages. We first solve $PDTDP_{SC}$ for a fixed L_h (we term it as Pricing Decision Problem (PDP_{SC})) numerically using the *matrix geometric method* in a *cutting plane* framework. We are able to obtain some analytical results for PDP_{SC} for the special case where L_h is sufficiently small. However, it is not possible to provide a rigorous mathematical proof for these results in absence of any analytical characterization of $S_l(\cdot)$. We, therefore, state these analytical results in a shared capacity case as observations rather than as propositions. Solution to PDP_{SC} is then used to solve $PDTDP_{SC}$ using the *golden section search* method. Again, some analytical results are possible for $PDTDP_{SC}$ for the special case where A is small, which we state as observations.

We now describe the matrix geometric method to numerically evaluate the sojourn time distribution, $S_l^k(\cdot)$, at a given point (p_h^k, p_l^k, μ^k) in the solution space of PDP_{SC}, which is used in the solution algorithm for PDP_{SC} and $PDTDP_{SC}$. We refer the reader to Neuts (1981) and Nelson (1991) for details of the matrix geometric method. The use of the matrix geometric method yields explicit recursive formulas for the joint stationary queue length distribution, which can provide significant computational improvements over the transform techniques also in use

(Miller 1981). Another recursive relation for the joint stationary queue length distribution has been obtained by White and Christie (1958). However, the use of the matrix geometric method also provides a recursive formula for $S_l^k(\cdot)$. Moreover, the matrix geometric method gives exact solutions, in contrast to simulation, which is another alternative method to evaluate $S_l^k(\cdot)$ that at best gives point estimates. The matrix geometric method is also computationally efficient compared to simulation. This is important in solving $PDTDP_{SC}$, which requires solving PDP_{SC} repeatedly for different values of L_h.

Matrix Geometric Method

Joint Stationary Queue Length Distribution: If we define $N_h(t)$ and $N_l(t)$ as state variables representing the number of high and low priority customers in the system at time t, then $\{\mathbf{N}(t)\} := \{N_l(t), N_h(t), t \geq 0\}$ is a continuous-time two-dimensional Markov chain with state space $\{\mathbf{n} = (n_l, n_h)\}$. The key idea we employ here is that $\{\mathbf{N}(t)\}$ is a *quasi-birth-and-death* (QBD) process, which allows us to develop a matrix geometric solution for the joint distribution of the number of customers of each class in the system. A simple implementation of the matrix geometric method, however, requires the number of states in the QBD process to be finite. For this, we treat the queue length of high priority customers (including the one in service) to be of finite size M, but of size large enough for the desired accuracy of our results. Since high priority customers are always served in priority over low priority customers, it is reasonable to assume that its queue size will always be bounded by some large number.

In the Markov process $\{\mathbf{N}(t)\}$, a transition can occur only if a customer of either class arrives or a customer of either class is served. The possible transitions are given in Table 2.7, where λ_h and λ_l at a given point (p_h^k, p_l^k, μ^k) are obtained using (2.1) and (2.2), respectively.

The infinitesimal generator Q associated with our system description is thus block-

Table 2.7: Transition rates for the priority queue

From	To	Rate	Condition
(n_l, n_h)	$(n_l, n_h + 1)$	λ_h	for $n_l \geq 0$, $n_h \geq 0$
(n_l, n_h)	$(n_l + 1, n_h)$	λ_l	for $n_l \geq 0$, $n_h \geq 0$
(n_l, n_h)	$(n_l, n_h - 1)$	μ	for $n_l \geq 0$, $n_h > 0$
(n_l, n_h)	$(n_l - 1, n_h)$	μ	for $n_l > 0$, $n_h = 0$

tridiagonal:

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

where B_0, A_0, A_1, A_2 are square matrices of order $M + 1$. These matrices can be easily constructed using the transition rates described above.

$$A_0 = \begin{pmatrix} \lambda_l & & & & \\ & \lambda_l & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_l \end{pmatrix};$$

$$A_2 = \begin{pmatrix} \mu & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix};$$

$$B_0 = \begin{pmatrix} * & \lambda_h & & & \\ \mu & * & \lambda_h & & \\ & \mu & * & \lambda_h & \\ & & \ddots & \ddots & \ddots \\ & & & \mu & * \end{pmatrix}$$

where $*$ is determined such that $A_0\mathbf{e} + B_0\mathbf{e} = \mathbf{0}$ is satisfied. $A_1 = B_0 - A_2$.

We denote $\mathbf{x}$ as the stationary probability vector of $\{\mathbf{N}(t)\}$:

$$\mathbf{x} = [x_{00}, x_{01}, \ldots, x_{0M}, x_{10}, x_{11}, \ldots, x_{1M}, \ldots, \ldots, x_{i0}, x_{i1}, \ldots, x_{iM}, \ldots, \ldots]$$

The vector $\mathbf{x}$ can be partitioned by levels into sub vectors $\mathbf{x}_i$, $i \geq 0$, where $\mathbf{x}_i = [x_{i0}, x_{i1}, \ldots, x_{iM}]$ is the stationary probability of states in level i ($n_l = i$). Thus, $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \ldots]$. $\mathbf{x}$ can be obtained using a set of balance equations, given in matrix form, by the following standard relations (Latouche and Ramaswami 1999, Nelson 1991, Neuts 1981):

$$\mathbf{x}Q = \mathbf{0}; \quad \mathbf{x}_{i+1} = \mathbf{x}_i R$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$A_0 + RA_1 + R^2 A_2 = \mathbf{0}$$

The matrix R can be computed using well known methods (Latouche and Ramaswami 1999). A simple iterative procedure often used is:

$$R(0) = 0 \ ; \quad R(n+1) = -\left[A_0 + R^2(n)A_2\right] A_1^{-1}$$

The probabilities $\mathbf{x}_0$ are determined from:

$$\mathbf{x}_0(B_0 + RA_2) = \mathbf{0}$$

subject to the normalization equation:

$$\sum_{i=0}^{\infty} \mathbf{x}_i \mathbf{e} = \mathbf{x}_0 (I - R)^{-1} \mathbf{e} = 1$$

where $\mathbf{e}$ is a column vector of ones of size $M + 1$.

We are aware that there are other recursive relations, not based on the matrix geometic method, in the literature to compute the joint stationary probabilities (see White and Christie 1958). However, the use of matrix geomatric method gives us the matrix R, required in the computation of $S_l(\cdot)$. We are also aware that there is a specialized method presented by Miller (1981) that exploits the special structure of the R matrix to compute the joint stationary probabilities for M/M/1 priority queues. However, we use the more general method since our main focus is on the managerial insights generated from our research, and not as much on the elegance of the solution algorithm.

Estimation of $S_l(\cdot)$: The delivery time W_l of a low priority customer is the time between its arrival to the system till it completes service. It may be preempted by one or more high priority customers for service. So it is difficult to characterize the distribution $S_l(\cdot)$. Ramaswami and Lucantoni (1985) present an efficient algorithm based on *uniformization* to derive the complimentary distribution function of the stationary waiting times in phase-type and QBD processes. The same approach is used by Leemans (2001) to derive the complimentary distribution of waiting times in a more complex queuing system. We adopt their algorithm to derive $S_l(\cdot)$, the distribution of the waiting time plus the time in service of low priority customers.

Consider a tagged low priority customer entering the system. The time spent by the tagged customer depends on the number of customers of either class already present in the system ahead of it, and also on the number of subsequent high priority arrivals before it completes its service. All subsequent low priority arrivals, however, have no influence on its time spent in the system. The tagged customer's time in the system is, therefore, simply the time until absorption in a modified Markov process $\{\tilde{\mathbf{N}}(t)\}$, obtained by setting $\lambda_l = 0$. Consequently, matrix $\tilde{A}_0$, representing transitions to a higher level, becomes a zero matrix. We define an *absorbing* state, call it state $0'$, as the state in which the tagged customer has finished its service. The infinitesimal generator for this process can be represented as:

$$\tilde{Q} = \left(\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & \cdots \\ \hline b_0 & \tilde{B}_0 & 0 & & & \\ 0 & A_2 & \tilde{A}_1 & 0 & & \\ 0 & & A_2 & \tilde{A}_1 & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\tilde{B}_0 = B_0 + A_0$; $\tilde{A}_1 = A_1 + A_0$; and $b_0 = [\mu \quad 0 \quad \cdots \quad 0]^T_{M+1}$. The first row and column in $\tilde{Q}$ corresponds to the absorbing state $\acute{0}$. The time spent in system by the tagged customer, which is the time until absorption in the modified Markov process with rate matrix $\tilde{Q}$, depends on the prices (p_h and p_l), through the arrival rates (λ_h and λ_l), and the service rate μ. For given prices (p_h^k, p_l^k) and service rate μ^k, the distribution of the time spent by a low priority customer in the system is $S_l^k(y) = 1 - \overline{S_l^k}(y)$, where $\overline{S_l^k}(y)$ is the stationary probability that a low priority customer spends more than y units of time in the system. Further, let $\overline{S_{li}^k}(y)$ denote the conditional probability that a tagged customer, who finds i low priority customers ahead of it, spends a time exceeding y in the system. The probability that a tagged customer finds i low priority customers is given, using the PASTA property (see Wolff 1982), by $\mathbf{x}_i = \mathbf{x}_0 R^i$. $\overline{S_l^k}(y)$ can be expressed as:

$$\overline{S_l^k}(y) = \sum_{i=0}^{\infty} \mathbf{x}_i \overline{S_{li}^k}(y)\mathbf{e} \qquad (2.17)$$

$\overline{S_{li}^k}(y)$ can be computed more conveniently by uniformizing the Markov process $\{\tilde{\mathbf{N}}(t)\}$ with a Poisson process with rate γ, where

$$\gamma = \max_{0 \le i \le M} (-\tilde{A}_1)_{ii} = \max_{0 \le i \le M} -(A_0 + A_1)_{ii}$$

so that the rate matrix $\tilde{Q}$ is transformed into the discrete-time probability matrix:

$$\hat{Q} = \frac{1}{\gamma}\tilde{Q} + I = \left(\begin{array}{c|ccccc} 1 & 0 & 0 & 0 & 0 & \cdots \\ \hline \hat{b_0} & \hat{B}_0 & 0 & & & \\ 0 & \hat{A}_2 & \hat{A}_1 & 0 & & \\ 0 & & \hat{A}_2 & \hat{A}_1 & 0 & \\ \vdots & & & \ddots & \ddots & \ddots \end{array}\right)$$

where $\hat{A}_2 = \frac{A_2}{\gamma}$, $\hat{A}_1 = \frac{\tilde{A}_1}{\gamma} + I$, $\hat{b_0} = \frac{b_0}{\gamma}$. In this uniformized process, points of a Poisson process are generated with a rate γ, and transitions occur at these epochs only. The probability that n Poisson events are generated in time y equals $e^{-\gamma y}\frac{(\gamma y)^n}{n!}$. Suppose the tagged customer finds i low priority customers ahead of it. Then, for its time in system to exceed y, at most i of the n Poisson points may correspond to transitions to lower levels (i.e., service completions of low priority customers). Therefore,

$$\overline{S^k_{li}}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \qquad i \geq 0 \tag{2.18}$$

where, $G_v^{(n)}$ is a matrix such that its entries are the conditional probabilities, given that the system has made n transitions in the discrete-time Markov process with rate matrix $\hat{Q}$, that v of those transitions correspond to lower levels (i.e., service completions of low priority customers). Substituting the expression for $\overline{S^k_{li}}(y)$ from (2.18) into (2.17), we obtain:

$$\overline{S^k_{l}}(y) = \sum_{n=0}^{\infty} d_n e^{-\gamma y}\frac{(\gamma y)^n}{n!} \tag{2.19}$$

where, d_n is given by:

$$d_n = \sum_{i=0}^{\infty} \mathbf{x}_0 R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \qquad n \geq 0 \tag{2.20}$$

Now,

$$
\begin{aligned}
&\sum_{i=0}^{\infty} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e} \\
&= \sum_{i=0}^{n+1} R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e} + \sum_{i=n+2}^{\infty} R^i \sum_{v=0}^{n} G_v^{(n)}\mathbf{e} && \big(\text{since } G_v^{(n)} = 0 \text{ for } v > n\big) \\
&= \sum_{v=0}^{n+1} \sum_{i=v}^{n+1} R^i G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+2}\mathbf{e} && \left(\text{since } \sum_{v=0}^{n} G_v^{(n)}\mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{n+1} (I-R)^{-1}(R^v - R^{n+2}) G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+2}\mathbf{e} \\
&= \sum_{v=0}^{n} (I-R)^{-1}R^v G_v^{(n)}\mathbf{e} + (I-R)^{-1}R^{n+1}G_{n+1}^{(n)}\mathbf{e} && \left(\text{since } \sum_{v=0}^{n+1} G_v^{(n)}\mathbf{e} = \mathbf{e}\right) \\
&= \sum_{v=0}^{n} (I-R)^{-1}R^v G_v^{(n)}\mathbf{e} && \big(\text{since } G_v^{(n)} = 0 \text{ for } v > n\big) \\
&= (I-R)^{-1}H_n\mathbf{e} && n \geq 0
\end{aligned}
$$

where, $H_n = \sum_{v=0}^{n} R^v G_v^{(n)}$. Therefore,

$$S_l^k(L_l) = 1 - \overline{S_l^k}(L_l) = \sum_{n=0}^{\infty} e^{-\gamma L_l} \frac{(\gamma L_l)^n}{n!} \mathbf{x}_0 (I-R)^{-1} H_n \mathbf{e} \tag{2.21}$$

H_n can be computed recursively as:

$$H_{n+1} = H_n \hat{A}_1 + R H_n \hat{A}_2; \quad H_0 = I$$

Therefore, for given prices (p_h^k, p_l^k) and service rate (μ^k), $S_l^k(\cdot)$ can be computed using (2.21).

Example 2.4: Assume the parameter values as shown in Table 2.3. Further, assume $L_h = 0.5$, $p_h = 10.7585$, $p_l = 8.5297$, $\mu = 13.1853$. Substituting these values in the demand model (2.1) - (2.2) gives $\lambda_h = 3.9750$ and $\lambda_l = 3.5250$. The service level $S_l(L_h = 0.5)$ for regular customers, obtained using the matrix geometric method, is 0.968901.

The above example shows that for the given prices and delivery times, the capacity level of $\mu = 13.1853$ is insufficient to provide the target service level of

0.99 to regular customers, and the firm either needs to invest more in its capacity level or alter its prices and/or delivery times to influence its regular demands in such a way that they can be served with a 99% reliability.

Example 2.5: Assume again the parameter values as shown in Table 2.3. Now assume a different set of values for the decision variables: $L_h = 0.5$, $p_h = 10.9061$, $p_l = 8.7012$, $\mu = 14.7843$. The service level $S_l(L_h = 0.5)$ now is 0.9900.

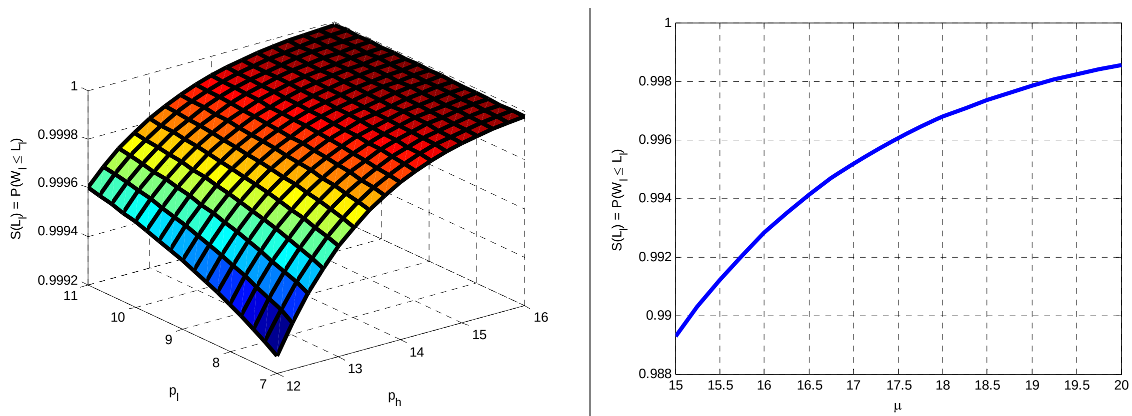
This example shows that for the given prices and delivery times, the capacity level of $\mu = 14.7843$ is just sufficient to provide the target service level of 0.99 to regular customers. Thus, the firm is able to satisfy the guaranteed service level by investing in its capacity level, which not only allows it to serve its regular customers with the desired reliability but also to charge higher prices to its regular as well as express customers. The net effect of this change in prices on customer demands is a drop in λ_h and λ_l from 3.9750 and 3.5250 to 3.9060 and 3.4001, respectively, allowing the firm to meet the new demand level λ_l reliably.

We now state an important property of $S_l(\cdot)$ based on our extensive numerical experiments, which provides the basis for the cutting plane method, described in the next section, to solve PDP_{SC}.

Property 2.4. *The sojourn time distribution of regular customers, $S_l(\cdot)$, in a shared capacity setting is:*

- *concave in (p_h, p_l)*
- *concave in μ.*

Figure 2.3 shows plots of $S_l(\cdot)$ vs. (p_h, p_l), and $S_l(\cdot)$ vs. μ, obtained using the matrix geometrix method described above. These plots suggest that $S_l(\cdot)$ is concave in (p_h, p_l) and separately in μ, although it is not possible to prove it mathematically in absence of an analytical characterization of $S_l(\cdot)$.



(a) $\boldsymbol{S_l}$ vs. $(\boldsymbol{p_h}, \boldsymbol{p_l})$ (b) $\boldsymbol{S_l}$ vs. μ

Figure 2.3: Service level vs. prices and capacity

Pricing Decision Problem (PDP_{SC})

We first solve the firm's optimization problem for a fixed L_h, which reduces it to a Pricing Decision Problem, which we denote as PDP_{SC}. On substituting (2.1) and (2.2) into (2.3), the objective function for PDP_{SC} is quadratic and concave. All constraints but (2.8), which does not have a closed form expression, are linear. Although the exact form of $S_l(\cdot)$ in constraint (2.8) is unknown, we exploit its special structure, determined numerically using the matrix geometric method. Property 2.4 suggests that for a fixed L_h, $S_l(\cdot)$ is concave in $(p_h,\, p_l)$ and separately in μ. However, this does not necessarily show the joint concavity of $S_l(\cdot)$ in $(p_h,\, p_l,\, \mu)$. We will, therefore, integrate into our solution method a mechanism to ensure that the concavity assumption is not violated.

Assuming $S_l(\cdot)$ is concave, it can be approximated by a set of tangent hyperplanes at various points $(p_h^k,\, p_l^k,\, \mu^k)$, $\forall\, k \in K$, as shown in Figure 2.4. That is:

$$S_l(\cdot) = \min_{k \in K} \left\{ S_l^k(\cdot) + (p_h - p_h^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) + (p_l - p_l^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) + (\mu - \mu^k)\left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \right\}$$

where $S_l^k(\cdot)$ denotes the value of $S_l(\cdot)$ at a fixed point (p_h^k, p_l^k, μ^k), and $\frac{\partial S_l^k(\cdot)}{\partial p_h}$, $\frac{\partial S_l^k(\cdot)}{\partial p_l}$ and $\frac{\partial S_l^k(\cdot)}{\partial \mu}$ are the partial gradients of $S_l(\cdot)$ at (p_h^k, p_l^k, μ^k). Constraint (2.8) can thus

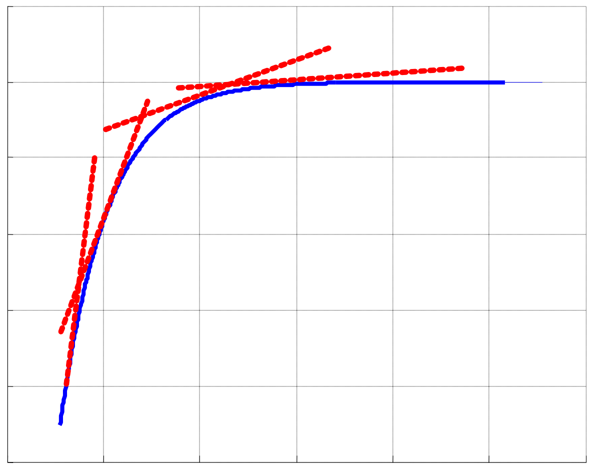


Figure 2.4: Piecewise linear approximation of $S_l(\cdot)$

be replaced by the following set of linear constraints:

$$S_l^k(\cdot)+(p_h-p_h^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right)+(p_l-p_l^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right)+(\mu-\mu^k)\left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \geq \alpha \quad \forall k \in K \quad (2.22)$$

Substituting the above set of constraints in place of (2.8), and the expressions (2.1) and (2.2) for λ_h and λ_l results in the following quadratic programming problem (QPP) with a finite but a large number of constraints, which makes it suitable for the cutting plane method (Kelley 1960).

PDP$_{(K)}$:

$$\max_{p_h, p_l, \mu} \pi = -(\beta_p^h + \theta_p)p_h^2 - (\beta_p^l + \theta_p)p_l^2 + 2\theta_p p_h p_l$$
$$+ \left\{-\beta_L^h L_h + \theta_L(L_l - L_h) + m\beta_p^h + a\right\} p_h$$
$$+ \left\{-\beta_L^l L_l + \theta_L(L_h - L_l) + m\beta_p^l + a\right\} p_l$$
$$- A\mu + (\beta_L^h L_h + \beta_L^l L_l)m - 2ma \tag{2.23}$$

subject to:

$$-(\beta_p^h + \theta_p)p_h + \theta_p p_l - \mu \leq \frac{\ln(1-\alpha)}{L_h} - a + (\beta_L^h + \theta_L)L_h - \theta_L L_l \tag{2.24}$$

$$\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) p_h + \left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) p_l + \left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \mu \geq \alpha - S_l^k(\cdot) +$$
$$\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) p_h^k + \left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) p_l^k + \left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \mu^k \qquad \forall k \in K \tag{2.25}$$

$$-\beta_p^h p_h - \beta_p^l p_l - \mu < \beta_L^h L_h + \beta_L^l L_l - 2a \tag{2.26}$$

$$-(\beta_p^h + \theta_p)p_h + \theta_p p_l \geq (\beta_L^h + \theta_L)L_h - \theta_L L_l - a \tag{2.27}$$

$$\theta_p p_h - (\beta_p^l + \theta_p)p_l \geq -\theta_L L_h + (\beta_L^l + \theta_L)L_l - a \tag{2.28}$$

$$p_h, p_l, \mu \geq 0 \tag{2.29}$$

Proposition 2.5. *The Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient for the global optimal solution of $PDP_{(K)}$.*

Proof. The Hessian of (2.23) is given by:

$$\begin{pmatrix} -2(\beta_p^h + \theta_p) & 2\theta_p & 0 \\ 2\theta_p & -2(\beta_p^l + \theta_p) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

This shows that the Hessian is negative semidefinite. Therefore, $PDP_{(K)}$ has a quadratic concave objective function. Moreover, all its constraints are linear. Hence, KKT conditions are both necessary and sufficient for its global optimal solution (Luenberger 1984). □

$PDP_{(K)}$ can be solved using any of the standard algorithms like Wolfe's Algorithm (Cooper 1974, Wolfe 1959). We use the matrix geometric method, described in §2.4.2, to numerically evaluate $S_l^k(\cdot)$ at a given point $(p_h^k,\, p_l^k,\, \mu^k)$. Once $S_l^k(\cdot)$ is evaluated at a point $(p_h^k,\, p_l^k,\, \mu^k)$, its gradients are obtained using the *finite difference method*, described below.

Estimation of the Gradient of $S_l(\cdot)$: There are several methods available in the literature to compute the gradients of $S_l(\cdot)$ (Carson and Maria 1997, Andradottir 1998). We use a *finite difference* method as it is probably the simplest and most intuitive, and can be easily explained (Atlason et al. 2004). Using the finite difference method, the gradients can be computed as:

$$\frac{\partial S_l^k(\cdot)}{\partial p_h} = \frac{S_l^{(p_h^k+dp_h,p_l,\mu)}(\cdot) - S_l^{(p_h^k-dp_h,p_l,\mu)}(\cdot)}{2dp_h}$$
$$\frac{\partial S_l^k(\cdot)}{\partial p_l} = \frac{S_l^{(p_h,p_l^k+dp_l,\mu)}(\cdot) - S_l^{(p_h,p_l^k-dp_l,\mu)}(\cdot)}{2dp_l}$$
$$\frac{\partial S_l^k(\cdot)}{\partial \mu} = \frac{S_l^{(p_h,p_l,\mu^k+d\mu)}(\cdot) - S_l^{(p_h,p_l,\mu^k-d\mu)}(\cdot)}{2d\mu}$$

where dp_h, dp_l and $d\mu$ (referred to as step sizes) are infinitesimal changes in the respective variables. These estimates of the gradients are used in the cutting plane algorithm to generate constraints/cuts of the form (2.25), which are added iteratively in the cutting plane algorithm, described next.

Cutting Plane Algorithm: We now describe the cutting plane algorithm to solve $PDP_{(K)}$. The algorithm fits the framework of Kelley's cutting plane method (Kelley 1960). It differs from the traditional description of the algorithm in that we use the matrix geometric method to generate the cuts and evaluate the function values instead of having an algebraic form for the function and using analytically determined gradients to generate the cuts. Figure 2.5 shows a flowchart of the cutting plane algorithm. The algorithm works as follows: We start with an empty constraint set (2.25), which results in a simple QPP, and obtain an initial solution $(p_h^0,\, p_l^0,\, \mu^0)$. We use the matrix geometric method to compute the distribution $S_l^{(p_h^0,p_l^0,\mu^0)}(\cdot)$ of W_l. If $S_l^{(p_h^0,p_l^0,\mu^0)}(\cdot)$ meets the delivery time reliability constraint α, we stop with an optimal solution to $PDP_{(K)}$. Otherwise we add to (2.25) a linear

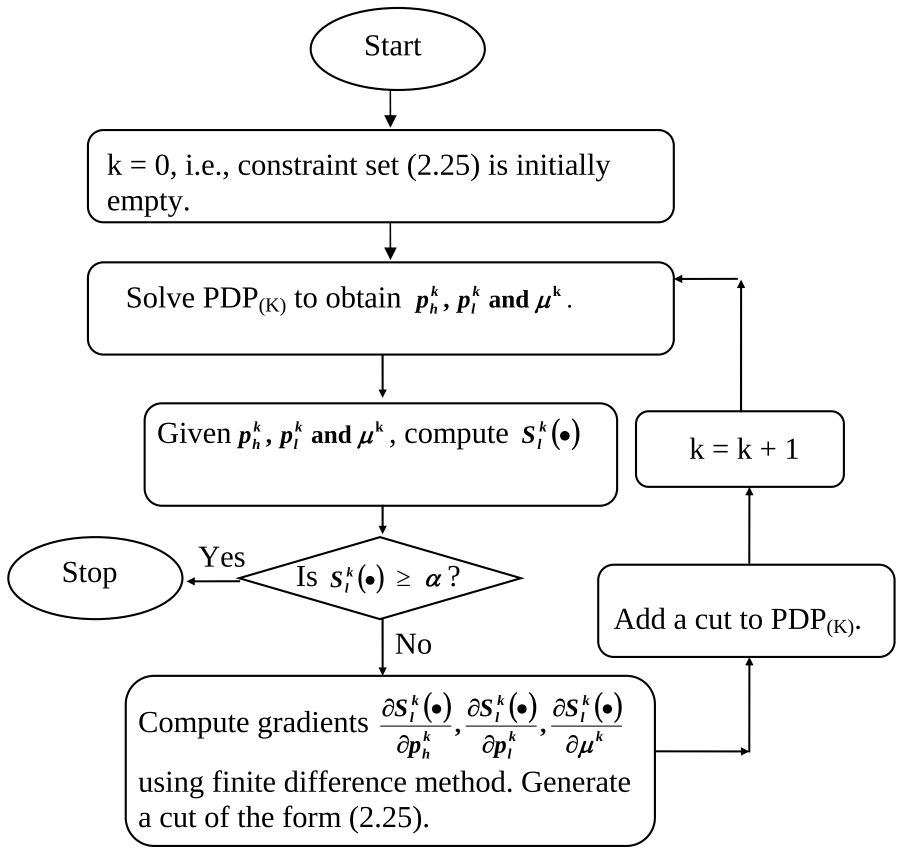


Figure 2.5: Cutting Plane Algorithm

constraint/cut generated using the finite difference method. The new cut eliminates the current solution but does not eliminate any feasible solution to $PDP_{(K)}$. This procedure repeats until the delivery time reliability constraint is satisfied within a sufficiently small tolerance limit ϵ such that $|S_l(\cdot) - \alpha| \leq \epsilon$. The method has been proved to converge (Atlason et al. 2004).

The success of the cutting plane algorithm relies on the concavity of $S_l(\cdot)$. We have already demonstrated, using computational results obtained by the matrix geometric method, that $S_l(\cdot)$ is concave in $(p_h,\ p_l)$ and separately concave in μ. However, it is difficult to establish the joint concavity of $S_l(\cdot)$ in $(p_h,\ p_l,\ \mu)$. If the concavity assumption is violated, then the algorithm may cut off parts of the feasible region and terminate with a solution that is suboptimal. We include a test to ensure the concavity assumption is not violated. This is done by ensuring that a new point, visited by the cutting plane algorithm after each iteration, lies below all the previously defined cuts, and that all previous points lie below the newly added cut. The test, however, cannot ensure that $S_l(\cdot)$ is concave unless it examines all the points in the feasible region. Still, it does help ensure that the concavity assumption is not violated at least in the region visited by the algorithm. Details of the test can be found in Atlason et al. (2004).

Example 2.6: Assume the parameter values as shown in Table 2.3. For $L_h = 0.5$, iteration 0 of $PDP_{(K)}$ corresponds to:

$\underline{\mathbf{PDP_{(k=0)}}}$:

$$\max_{p_h, p_l, \mu} \pi = -0.70p_h^2 - 0.90p_l^2 + 0.40p_h p_l + 11.30p_h + 11.15p_l - 0.50\mu - 56.55$$

subject to:

$$-0.70p_h + 0.20p_l - \mu \leq -19.0103$$

$$-0.50p_h - 0.70p_l - \mu < -18.85$$

$$-0.70p_h + 0.20p_l \geq -9.80$$

$$0.20p_h - 0.90p_l \geq -9.05$$

$$p_h, p_l, \mu \geq 0$$

For the solution algorithm, a bound (M) on the high priority queue size needs to be specified to facilitate use of the $m$atrix geometric method. Finding an appropriate value of M requires some experimentation. Computational experiments of a priority queue with a reasonable range of parameter values suggested $M = 100$ to be a good choice with little effect on the accuracy of results. For the cutting plane algorithm, we set the tolerance limit (ϵ) at 10^{-6}, and the step sizes ($dp_h, dp_l, d\mu$) for gradient estimation at 0.001. Table 2.8 shows the results of successive iterations. The optimal solution is (p_h = 10.9061, p_l = 8.7012, μ = 14.7843), which is obtained after 5 iterarions. Computational results, showing the number of cuts used and the time (in seconds) taken by the algorithm for a range of parameters values, are reported in Table 2.9. All computations are performed on a Pentium IV (3.06 GHz, 512 MB RAM) machine. The results suggest that the proposed algorithm is very efficient, taking only a few seconds.

Table 2.8: Iterations of the cutting plane algorithm

Iter.	(p_h, p_l, μ)	$(S_h(L_h), S_l(L_l))$	*Cut generated*
0	(10.7585, 8.5297, 13.1853)	(0.990000, 0.968901)	$0.0170p_h + 0.0074p_l + 0.0195\mu \geq 0.5243$
1	(10.8725, 8.7163, 14.0965)	(0.993793, 0.984251)	$0.0093p_h + 0.0033p_l + 0.0100\mu \geq 0.2760$
2	(10.8935, 8.7062, 14.6575)	(0.995350, 0.989084)	$0.0067p_h + 0.0021p_l + 0.0069\mu \geq 0.1945$
3	(10.9038, 8.7020, 14.7807)	(0.995645, 0.989964)	$0.0062p_h + 0.0019p_l + 0.0064\mu \geq 0.1794$
4	(10.9061, 8.7012, 14.7843)	(0.995657, 0.990000)	not needed

Table 2.9: Performance of the cutting plane algorithm

A →	**0.10**		**0.25**		**0.50**		**0.75**		**1.00**	
L_h ↓	*Cuts*	*Time*	*Cuts*	*Time*	*Cuts*	*Time*	*Cuts*	*Time*	*Cuts*	*Time*
0.10	0	0.08	0	0.09	0	0.08	0	0.05	0	0.09
0.20	0	0.08	0	0.11	0	0.08	0	0.09	0	0.08
0.30	0	0.09	0	0.08	0	0.08	0	0.09	0	0.09
0.40	0	0.13	0	0.17	0	0.16	0	0.13	0	0.16
0.50	4	3.81	4	3.16	4	2.42	4	2.39	4	2.38
0.60	5	3.17	5	3.16	5	3.14	5	3.13	5	3.14
0.70	6	4.02	6	4.03	6	4.02	6	3.95	6	3.94
0.80	6	4.48	6	4.47	6	4.39	6	4.42	6	4.41
0.90	7	5.7	7	5.7	7	5.67	7	5.63	7	5.59

An important observation to make from Table 2.9 is that the number of cuts generated by the algorithm is always 0 when L_h is small (≤ 0.40 in the above example). This means that the complicating delivery time reliability constraint for regular customers, which lacked any analytical characterization, is never binding at optimality for small L_h. This is depicted in Figure 2.6 in which the threshold value of L_h below which the delivery time reliability constraint for regular customers is never binding is indicated as L_h^T. Although this is depicted for a specific combination of parameter values, this is true in general. To give an intuitive explanation for this observation, we first state the following proposition.

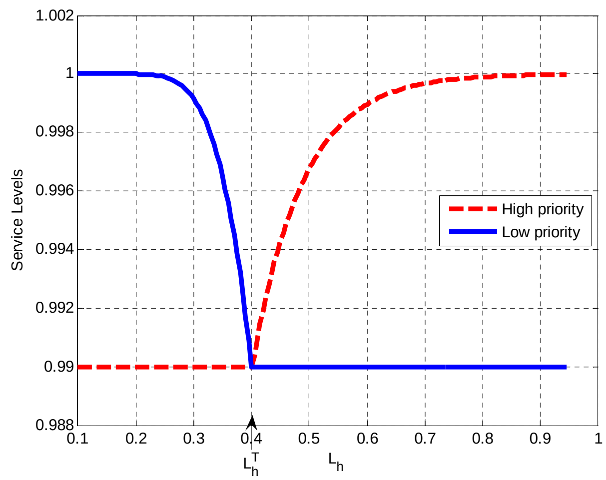


Figure 2.6: Service levels for high and low priority customers in a shared capacity setting

Proposition 2.6. *In a shared capacity setting, at least one of the delivery time reliability constraints is binding at optimality.*

Proof. Delivery time reliability constraint for express customers (2.7^{SC}) can be restated as:

$$\mu \geq \mu_h^\alpha = -\frac{\ln(1-\alpha)}{L_h} + \lambda_h \tag{2.30}$$

Although the actual analytical characterization of delivery time reliability constraint for regular customers is unknown, we still know that the service level always increases with the service rate, irrespective of the priority discipline. This implies that:

$$\mu \geq \mu_l^\alpha \tag{2.31}$$

where μ_l^α is the minimum service rate required to meet the desired service level α for low priority customers. Combining (2.30) and (2.31), we have:

$$\mu \geq \max\{\mu_h^\alpha, \mu_l^\alpha\} \tag{2.32}$$

Further, the profit function π is decreasing in μ. Therefore, to maximize profit, the service rate μ should be at its minimum level that guarantees the desired service level α for the two customer classes. This implies that at optimality, (2.32) should hold with equality, which means that at least one of the delivery time reliability constraints should be binding. Thus, the service rate at optimality is given by:

$$\mu = \max\{\mu_h^\alpha, \mu_l^\alpha\} \tag{2.33}$$

□

Clearly, μ_h^α increases as L_h decreases. Further, λ_l decreases as L_h decreases. This suggests that for L_h sufficiently small, $\mu = \max\{\mu_h^\alpha, \mu_l^\alpha\} = \mu_h^\alpha$ such that the capacity requirement is dictated solely by the demand from express customers. This allows us to solve PDP_{SC} analytically for small L_h. We state this result formally as an observation rather than a proposition since in the absence of an analytical expression for $S_l(\cdot)$, we are unable to provide a rigorous mathematical proof.

Observation 2.1. *When L_h is small, the optimal prices in a shared capacity setting are given by:*

$$p_h^{SC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^l + 2\theta_p)a - (\beta_p^l\beta_L^h + \beta_p^l\theta_L + \beta_L^h\theta_p)L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{2.34}$$

$$p_l^{SC*}(L_h) = \frac{m}{2} + \frac{(\beta_p^h + 2\theta_p)a + (\beta_p^h\theta_L - \beta_L^h\theta_p)L_h - (\beta_p^h\beta_L^l + \beta_p^h\theta_L + \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{2.35}$$

As noted earlier in §2.3.2, the stability condition (2.6^{SC}) is automatically satisfied by the two delivery time reliability constraints of $PDTDP_{SC}$ or PDP_{SC}. Further, we observe that for small L_h, the service rate μ is decided solely by the delivery time reliability constraint for express customers, such that:

$$\mu = \lambda_h - \frac{\ln(1-\alpha)}{L_h}$$

Substituting the expression for μ in the objective function, $p_h^{SC*}(L_h)$ and $p_l^{SC*}(L_h)$ can be obtained by solving simultaneously $\partial\pi/\partial p_h = 0$ and $\partial\pi/\partial p_l = 0$, in very much the same way as we did for the dedicated capacity case.

When the prices are described by the above relations (2.34) and (2.35), it can be shown that a small L_h, in fact, results in a relatively large express demand compared to regular demand, such that the capacity requirement μ is dictated only by the delivery time reliability constraint of express customers. Using (2.34) and (2.35), we obtain:

$$\begin{aligned}
\frac{d\lambda_h}{dL_h} &= -(\beta_p^h + \theta_p)\frac{\partial p_h}{\partial L_h} + \theta_p \frac{\partial p_l}{\partial L_h} - (\beta_L^h + \theta_L) \\
&= -\frac{\beta_L^h + \theta_L}{2} < 0 \\
\frac{d\lambda_l}{dL_h} &= -(\beta_p^l + \theta_p)\frac{\partial p_l}{\partial L_h} + \theta_p \frac{\partial p_h}{\partial L_h} + \theta_L \\
&= \frac{\theta_L}{2} > 0
\end{aligned}$$

This suggests that when L_h gets sufficiently small, express demand gets much larger compared to regular demand. Thus, the capacity requirement μ is dictated only by the demand from express customers, something we used to arrive at the results (2.34) and (2.35) at first place.

Observation 2.1 is important in that it provides us some basis to compare the pricing decisions in a shared versus dedicated capacity settings when the delivery times are fixed.

Pricing and Delivery Time Decision Problem

The Pricing and Delivery Time Decision Problem ($PDTDP_{SC}$) adds an additional dimension to PDP_{SC} by treating L_h as a decision variable, which the firm tries to jointly optimize along with p_h, p_l and μ. This makes constraint (2.7^{SC}) non-linear, and the model substantially more challenging to solve. We use the solution to PDP_{SC} and the golden section search method (Luenberger 1984, Winston and Venkataramanan 2003) to solve $PDTDP_{SC}$, which can be rewritten as:

$$\max_{L_h \in [0,\, L_l)} \quad f(L_h)$$

where $f(L_h)$ is a PDP_{SC} for a fixed L_h. We have shown in a dedicated capacity setting that $f(L_h)$ has a unique maximum when a is high. Our extensive numerical experiments with $f(L_h)$ suggests that a sufficiently large a guarantees that $f(L_h)$ has a unique maximum in a shared capacity setting as well, and hence $PDTDP_{SC}$ can be solved efficiently using the golden section search method. At each step, the algorithm solves PDP_{SC} to evaluate $f(L_h)$ for a given value of L_h.

Example 2.7: Assume the parameter values as shown in Table 2.3. The optimal decisions, demand and profit for a shared capacity setting are shown in Table 2.10.

Table 2.10: Results for Example 2.7

L_h^*	p_h^*	p_l^*	μ^*	λ_h^*	λ_l^*	π^*
0.5562	10.7032	8.7830	12.2653	3.9857	3.3141	39.7753

Example 2.8: Table 2.11 shows the optimal price, delivery time and capacity decisions for the parameter combinations shown in Table 2.3 and for various combinations of substitution parameters and capacity cost.

Table 2.11: Results for Example 2.8

	Without substitution ($\theta_p = \theta_L = 0$)		**With substitution** ($\theta_p = 0.2,\ \theta_L = 0.5$)		($\theta_p = 0.4,\ \theta_L = 0.3$)	
	A=0.10	*A=1.0*	*A=0.10*	*A=1.0*	*A=0.10*	*A=1.0*
L_h^*	0.2494	0.42755	0.2393	0.4277	0.2572	0.4276
p_h^*	11.3255	11.8355	10.8148	11.2960	10.3580	10.9126
p_l^*	8.1429	8.3985	8.5142	8.7753	8.8289	9.0739
μ^*	22.5774	14.4659	23.5440	14.5170	22.1047	14.3657

Observation 2.2. *When A is small, the optimal express delivery time L_h^{SC*} is given by the unique root of (2.36) in the interval [0, L_l).*

$$\frac{\partial \pi(L_h)}{\partial L_h} = -\left(\beta_L^h + \theta_L\right)\left(p_h^{SC*}(L_h) - m - A\right) + \theta_L\left(p_l^{SC*}(L_h) - m\right) - \frac{A\ln(1-\alpha)}{L_h^2} \tag{2.36}$$

where, $p_h^{SC*}(L_h)$ and $p_l^{SC*}(L_h)$ are given by (2.34) and (2.35).

We know from Observation 2.1 that when L_h is known to be small ($\leq L_h^T$), $p_h^{SC*}(L_h)$ and $p_l^{SC*}(L_h)$ can be obtained using (2.34) and (2.35). Substituting $p_h^{SC*}(L_h)$ and $p_l^{SC*}(L_h)$, given by (2.34) and (2.35), in the profit function π, and differentiating it with respect to L_h gives (2.36), while $\partial^2\pi(L_h)/\partial L_h^2$ and $\partial^3\pi(L_h)/\partial L_h^3$ are given by the same relations (2.14) and (2.15) as for the dedicated capacity case. Thus, for small A, the properties 2.1, 2.2 and 2.3 of π hold true in a shared capacity case as well. This implies that for a sufficiently high, π has a unique maximum, as shown in Figure 2.2, given by the root of (2.36).

We obtain the above result assuming that L_h is known to be small. We now show that when A is small, L_h^{SC*} is indeed small, such that the above result holds true. Using the Implicit Function Theorem, we get:

$$\frac{\partial L_h^{SC*}}{\partial A} = -\left(\frac{\partial^2\pi/\partial L_h \partial A}{\partial^2\pi/\partial L_h^2}\right)\Bigg|_{L_h = L_h^{SC*}},$$

where

$$\frac{\partial^2\pi}{\partial L_h \partial A}\Bigg|_{L_h = L_h^{SC*}} = \frac{\beta_L^h + \theta_L}{2} - \frac{ln(1-\alpha)}{(L_h^*)^2} > 0.$$

Since we know that

$$\frac{\partial^2\pi}{\partial L_h^2}\Bigg|_{L_h = L_h^{SC*}} < 0 \quad \Rightarrow \quad \frac{\partial L_h^{SC*}}{\partial A} > 0.$$

This implies that L_h^{SC*} is increasing in A. Therefore, a sufficiently small A guarantees that L_h^* is small, which we used at first place to arrive at the result.

2.5 Conclusions

In this chapter, we developed the modelling framework to study the optimal product differentiation strategy of a firm selling two "products", which are similar in all respect except in their prices and guaranteed delivery times, in a capacitated environment. Our primary objective was to understand how the demand-side product substitution and the supply-side operations strategy of the firm (using dedicated versus shared capacity) affect the optimal pricing and delivery time decisions as well as the optimal capacity level. For this, we developed a general mathematical model, special cases of which capture different scenarios depending on whether the products are substitutable or not, and whether the capacity strategy is shared or dedicated. The dedicated capacity setting allowed us to obtain analytical results. From a technical perspective, our methodology for dealing with the analytically-difficult shared capacity setting is somewhat novel. This involved linking a matrix geometric model for queuing performance analysis to a cutting plane algorithm for optimization.

In the following chapter, we use the solution methods developed in this chapter to extensively study and compare the results for the different scenarios to derive important managerial insights.

Chapter 3

Monopolistic Market: Analysis & Insights

In chapter 2, we developed a general mathematical model PDTDP for a firm's pricing and delivery time decision problem. We further studied how the same model can be adapted for the four different scenarios described in §2.1, and discussed the solution approach for each. We now study the different scenarios described in §2.1, and address the research issues posed therein. Specifically, we first study the individual and joint roles played by product substitution and a firm's operations (capacity) strategy in shaping its price and delivery time differentiation decisions. We then investigate how rising capacity costs affect product differentiation policy under different demand and supply conditions. Since the mathematical model for the shared capacity scenario does not, in general, have an analytical solution, we test our models numerically under different combinations of parameter values. Generalizations based on observable patterns that emerge from these numerical experiments are reported as observations. From these observations, we derive conclusions of managerial interest.

We first discuss the results for the pricing decision problem (PDP) for fixed delivery times. As discussed in chapter 2, PDP is appropriate when there is stickiness in a firm's delivery time decisions. Then we discuss the more general problem of pricing and delivery time decision (PDTDP).

3.1 Numerical Experiment Design

Our model setting described in chapter 2 involves the following parameters: a, m, α, L_l and A. Of these, we fix the value of $L_l = 1$ (so, delivery time differential $= 1 - L_h^*$). As regards the other parameters, we experiment with a large combination of their values as given in Table 3.1:

Table 3.1: Parameter settings for numerical experiments

Parameter	Number of Choices	Possible Values
a	5	{10, 50, 100, 200, 400}
m	6	{1, 2, 3, 4, 5, 6}
A	6	{0.1, 0.25, 0.5, 0.75, 1, 2}
α	5	{0.95, 0.96, 0.97, 0.98, 0.99}

Note that not all possible combination of values given in Table 3.1 are used in our experiment. An important assumption we have made throughout is that a is sufficiently high (see Assumption 2.1). This was required to guarantee a unique L_h^* that maximizes a firm's profit. We use only those combinations of parameter values from Table 3.1 that make a sufficiently large for our purpose. However, the figures that we present in this chapter use $a = 10$, $m = 3$, $\alpha = 0.99$, $A = 0.5$ (unless otherwise stated). Since the behavior of the prices depends on the market characteristics, we compare their optimal values under the different market settings. We use the following market parameter values:

- *Time Difference Sensitive* (TDS): $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$, $\theta_p = 0.2$, $\theta_L = 0.5$, such that $\theta_L/\beta_L^h > \theta_p/\beta_p^h$ and $\theta_L/\beta_L^l > \theta_p/\beta_p^l$.
- *Price Difference Sensitive* (PDS): $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$, $\theta_p = 0.4$, $\theta_L = 0.3$, such that $\theta_L/\beta_L^h < \theta_p/\beta_p^h$ and $\theta_L/\beta_L^l < \theta_p/\beta_p^l$.

3.2 Pricing Decision Problem

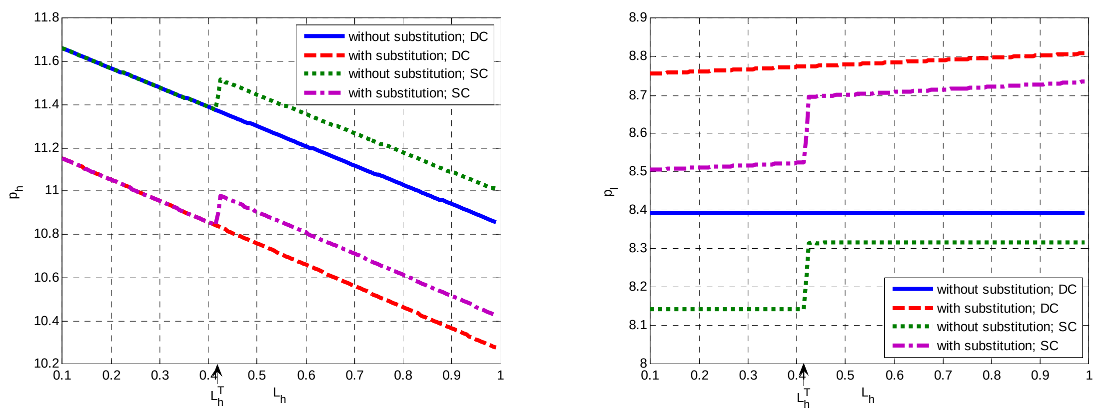
In this section, we assume the firm faces a significantly higher stickiness for their delivery time decisions compared to their ability to vary prices. Situations in which such a model will be more relevant are discussed in chapter 2. Under such situations, a firm optimizes its prices, treating its delivery times as fixed.

We start by studying the behavior of the optimal prices in response to a change in the guaranteed express delivery time L_h in each of the four scenarios. First of all, a change in operating philosophy from dedicated to shared capacity setting has *no effect* on the way the two prices behave with respect to L_h, except for a sudden jump in their values at a specific value of L_h, denoted as L_h^T, in a shared capacity setting. L_h^T is the value of L_h at which delivery time reliability constraint is binding for both the classes of customers (refer to Figure 2.6). Product substitution, on the other hand, affects the behavior of regular price only. Figure 3.1 shows the behavior of the two prices under different scenarios as we vary L_h, which is summarized in the following observation:

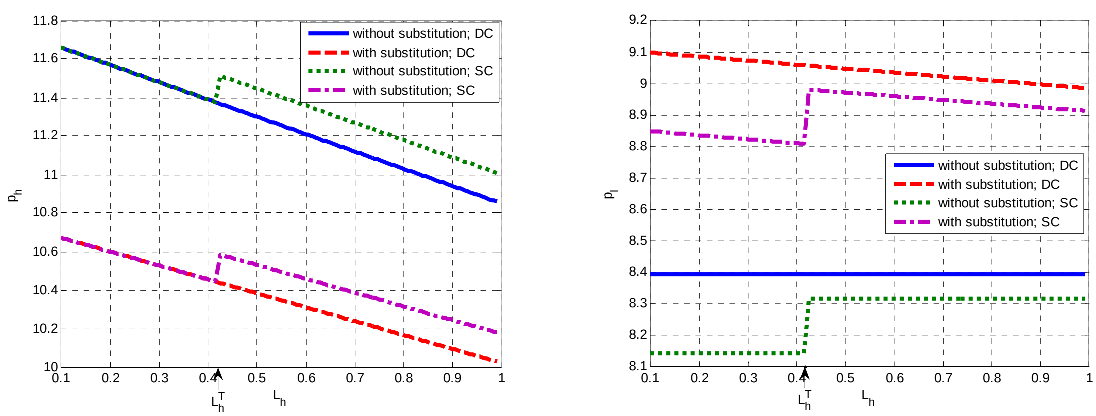
Observation 3.1. *In both the dedicated capacity (DC) and the shared capacity (SC) settings, a decrease in L_h results in (Refer to Figure 3.1): (a) an increase in p_h^* (b) a decrease in p_l^* if $\theta_L/\beta_L^h > \theta_p/\beta_p^h$; an increase p_l^* if $\theta_L/\beta_L^h < \theta_p/\beta_p^h$; and no change in p_l^* if $\theta_p = \theta_L = 0$.*

In a DC setting as well as for small A in an SC setting, the above observation follows directly from Proposition 2.2 and Observation 2.1. Figure 3.1 shows the behavior of the two prices in different scenarios as we vary L_h. This behavior is quite intuitive and is similar to what has been shown by Boyaci and Ray (2003) for the dedicated capacity case. Since express customers are time-sensitive, a firm can always charge them a higher price for a guaranteed shorter delivery time, as also evident from the expressions for the optimal prices ((2.10) and (2.11) for DC and (2.34) and (2.35) for SC). However, in absence of product substitution ($\theta_p = \theta_L = 0$), customers from a given class are not concerned about what is offered to the other class.

Thus, the price charged to regular customers is unaffected by any change in the delivery time guaranteed to the express ones (see (2.11) and (2.35)). With product substitution, the behavior of the optimal price for the regular class depends on the market conditions. In a TDS type market, the relative sensitivity of customers to the difference in delivery times (with respect to their own delivery times) is greater than their relative sensitivity to the price difference (with respect to their own price), such that $\theta_L/\beta_L^k > \theta_p/\beta_p^k$, $k \in \{l, h\}$. In such a market, a decrease in L_h results in a small gain in new express customers but a relatively larger number of regular customers switch to the express option. By increasing p_h and decreasing p_l simultaneously, the firm can attract new regular customers without causing a significant number of express customers to switch option, thus increasing the profit.



(a) TDS



(b) PDS

Figure 3.1: Comparison of prices in the four different scenarios

On the other hand, in a PDS type market, the relative sensitivity of customers to the difference in delivery times (with respect to their own delivery times) is less than their relative sensitivity to the price difference (with respect to their own prices), such that $\theta_L/\beta_L^k < \theta_p/\beta_p^k,\ k \in \{l, h\}$. In such a market, reducing L_h attracts a significant number of new customers who choose the express delivery option but relatively few of the regular customers switch to the express delivery option. Since customers preference between the two options is driven mainly by the price difference now, the firm cannot afford to increase the price difference, which will cause a significant number of express customers to switch options, causing a loss of profit. Thus the optimal strategy for the firm is to increase the price for the regular customers corresponding to an increase in price for the express customers.

We next do a comparison of the values of the optimal prices in the four scenarios to study the effects of product substitution and capacity strategy.

Observation 3.2. *If the express delivery time L_h is fixed and is sufficiently small, the relations shown in Table 3.2 hold between DC and SC.*

Table 3.2: Comparison between DC and SC when L_h is given and sufficiently small

Shared Capacity	**Relation**	**Dedicated Capacity**
p_h^{SC*}	$=$	p_h^{DC*}
p_l^{SC*}	$<$	p_l^{DC*}
λ_h^{SC*}	$\leq$	λ_h^{DC*}
λ_l^{SC*}	$>$	λ_l^{DC*}
$\lambda_h^{SC*} + \lambda_l^{SC*}$	$>$	$\lambda_h^{DC*} + \lambda_l^{DC*}$
μ^{SC*}	$<$	$\mu_h^{DC*} + \mu_l^{DC*}$

The relation between the prices in the two capacity settings follow directly by comparing (2.10) with (2.34) and (2.11) with (2.35). The relation between the demands can be explained by comparing the resulting demand functions obtained by substituting the prices (p_h^{DC*}, p_l^{DC*}) and (p_h^{SC*}, p_l^{SC*}) in the demand model (2.1) and (2.2). These relations suggest that when the express delivery time is sufficiently small, regular customers get a better price from the firm using shared capacities. For the relation between the capacity requirements in the two settings, we have:

$$\begin{aligned}
\mu^{SC*} &= \lambda_h^{SC*} - \frac{\ln(1-\alpha)}{L_h} \\
&\leq \lambda_h^{DC*} - \frac{\ln(1-\alpha)}{L_h} \\
&< \lambda_h^{DC*} - \frac{\ln(1-\alpha)}{L_h} + \lambda_l^{DC*} - \frac{\ln(1-\alpha)}{L_l} \\
&= \mu_h^{DC*} + \mu_l^{DC*}
\end{aligned}$$

This shows that when L_h is chosen to be small, a firm that shares its capacities achieves the benefits of capacity pooling. This is a well established fact and holds true even for larger L_h, as our numerical results suggest. However, the other relations between individual prices and demand may not hold true for larger L_h. The next observation shows a comparison between the two capacity settings that hold true irrespective of the value of the express delivery time chosen by a firm.

Observation 3.3. *For a given delivery time differentiation (Refer to Figure 3.1):*
- a change in capacity strategy from dedicated to shared results in (a) a generally higher p_h^*, (b) a lower p_l^*, and hence (c) a higher optimal price differentiation
- introduction of product substitutability results in (a) a lower p_h^*, (b) a higher p_l^*, and hence (c) a lower optimal price differentiation.

Managerially speaking, the above observation is significant. It shows that for a capacitated, pure-price competition environment, a firm's operations strategy (dedicated or shared capacity), as well as its marketing strategy (whether to make the products available for all market segments or to customize them for separate segments), affects both the absolute product prices as well as the optimal product differentiation. For small capacity costs, this observation follows directly from the comparison of the optimal prices in the two capacity settings, as given by Proposition 2.3 and Observation 2.1, respectively. Comparing the prices in the two settings, we see that the price for express customers remains the same, whereas that for regular customers in SC decreases by a constant amount $A/2$, thereby increasing the price differentiation.

As L_h increases, demand for express customers decreases, while that for regular customers increases. The supply system then faces increasing pressure to satisfy the demand from regular customers. Indeed, beyond L_h^T, the supply capacity in SC is dictated solely by the demand from regular customers. The problem is difficult to solve analytically in absence of a closed form expression for constraint (2.25). However, the numerical results suggest that as L_h increases to L_h^T, the firm needs to suddenly increase the prices for both the products. This further increases the price difference for express customers between the two capacity settings, and decreases it for regular customers. The price differentiation between the two customer classes is still higher in SC compared to DC.

The effect of product substitution in a dedicated capacity setting follows directly from (2.10) and (2.11).

$$p_h^{DC*}(L_h)\big|_{\theta_p,\theta_L>0} - p_h^{DC*}(L_h)\big|_{\theta_p,\theta_L=0}$$
$$= \frac{-(\beta_p^l - \beta_p^h)\theta_p a - (\beta_p^h\theta_L - \beta_L^h\theta_p)\beta_p^l L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)\beta_p^h L_l}{2\beta_p^h(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)}$$

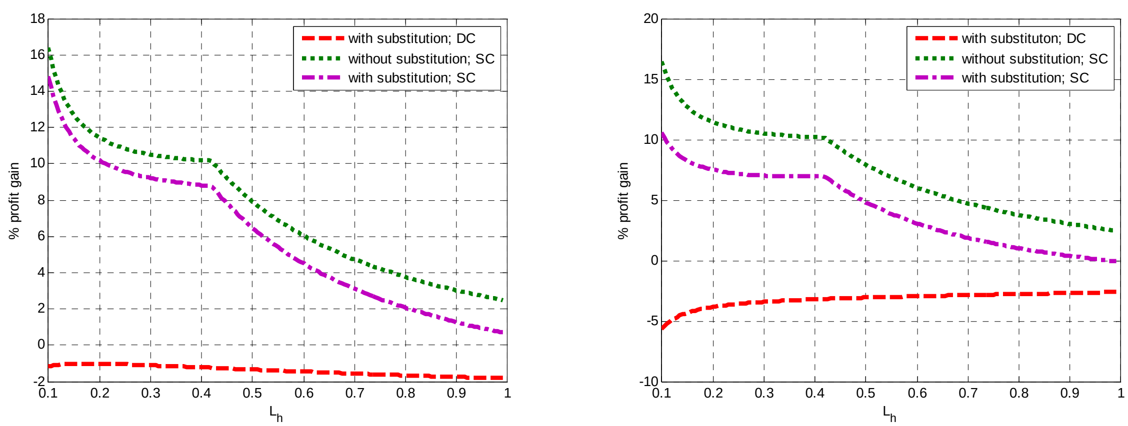
$$p_l^{DC*}(L_h)\big|_{\theta_p,\theta_L>0} - p_l^{DC*}(L_h)\big|_{\theta_p,\theta_L=0}$$
$$= \frac{(\beta_p^l - \beta_p^h)\theta_p a + (\beta_p^h\theta_L - \beta_L^h\theta_p)\beta_p^l L_h - (\beta_p^l\theta_L - \beta_L^l\theta_p)\beta_p^h L_l}{2\beta_p^h(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)}$$

The above relations show that for a sufficiently high a (Assumption 2.2), p_h decreases whereas p_l increases with substitution. The net result is a decrease in price differentiation. The effect is most pronounced when the market is simultaneously TDS for express customers and PDS for regular customers. The effect of product substitution in a shared capacity setting for small L_h can be explained similarly by substituting $\theta_p = \theta_L = 0$ in (2.34) and (2.35).

It is important to point out here that the effect of product substitution on the two prices for a given delivery time differentiation has been studied by Boyaci and

Ray (2003), albeit only in a dedicated capacity setting. However, their results differ significantly from ours. Their results suggest that product substitution may increase or decrease price differentiation, depending on the customers' behavior. Our results, in contrast, suggest that product substitution, for a given delivery time differentiation, always results in a lower price differentiation, irrespective of customers' behavior. This difference in the two results arises due to the difference in the modelling assumptions made. Boyaci and Ray (2003) use the same (price and delivery time) sensitivities ($\beta_p^h = \beta_p^l = \beta_p$, $\beta_L^h = \beta_L^l = \beta_L$) for the two customer classes, even though the customers are essentially categorized as price or time sensitive only based on the difference in their price and delivery time sensitivities.

Observation 3.4. *A change in capacity strategy from dedicated to shared results in higher profits, whereas introducing substitutability erodes profit. The effect, in general, is stronger at higher delivery time differentiation. (Refer to Figure 3.2).*



(a) TDS (b) PDS

Figure 3.2: % Profit gain for different scenarios over the "non-substitutable products; dedicated capacity" scenario

Figure 3.2 shows the % gain in profit in different scenarios over the "non-substitutable products; dedicated capacity" scenario. Regardless of the market characteristics, shared compared to dedicated capacity always leads to higher profits. The relative gain in profit in SC over DC increases with an increase in the delivery time differentiation (decrease in L_h). This gain in profit is attributed mostly to the savings in capacity related costs due to capacity pooling in SC. A unit decrease in the express delivery time L_h (corresponding to a unit increase in the delivery time differentiation) generates additional demand from express customers at a rate of $(\beta_L^h + \theta_L)$, out of which β_L^h are new customers and θ_L are regular customers who now switch to the express delivery option. The net result is an increase in the total demand at a rate β_L^h. A larger delivery time differentiation, therefore, leads to a larger capacity required to serve the increased demand, which increases the savings due to capacity pooling in SC. An increase in capacity cost will, therefore, increase such a gain in profit. Product substitution, on the other hand, results in lower profits. This is consistent with the revenue management theory, which suggests that a properly designed fence that prevents leakage of demand from high price segment to the low price segment enhances a firm's profit (Zhang 2007). Results obtained in this section are summarized in Table 3.3.

Table 3.3: Summary: Observations for the Pricing Decision Problem

Decision Variables	**Effect of SC compared to DC**	**Effect of product substitution**
$p_h^*(L_h)$	↑	↓
$p_l^*(L_h)$	↓	↑
$p_h^*(L_h) - p_l^*(L_h)$	↑	↓
$\pi^*(L_h)$	↑	↓

3.3 Pricing and Delivery Time Decision Problem

The last section focussed on the optimal pricing (and price differentiation) strategy, for a given delivery time differentiation. In this section we address the issue of overall product differentiation - both in terms of delivery time and price. So we now solve the pricing and delivery time decision problems $PDTDP_{DC}$ and $PDTDP_{SC}$. We first study the comparison of the optimal product differentiation under the four scenarios for a given marginal capacity cost A, and then study their behavior as A increases.

3.3.1 Optimal Product Differentiation for a Given Marginal Capacity Cost

We have performed extensive numerical experiments for our models for the four different scenarios to study the effects of capacity strategy and product substitution on a firm's price and delivery time decisions, and also on its optimal product differentiation. We present a small sample of these studies to illustrate our comparison of the optimal decisions in the four scenarios. We use the demand parameter values as defined in §3.1 for PDS and TDS type markets. Firm specific parameters are fixed at: $a = 10$, $m = 3$, $\alpha = 0.99$, $L_l = 1$. We use two different values for A to illustrate the difference in the behavior of optimal decisions in a shared capacity setting when capacity cost is high versus when it is small: (i) $A = 0.10$ (small capacity cost) (ii) $A = 1.0$ (high capacity cost). The results are presented in Table 3.4 for "without substitution" scenario and in Table 3.5 for "with substitution" scenario.

Table 3.4: Numerical Results: Without Product Substitution

	A = 0.10		**A = 1.0**	
	DC	**SC**	**DC**	**SC**
L_h^*	0.2494	0.2494	0.8405	0.42755
$L_l - L_h^*$	0.7506	0.7506	0.1595	0.57245
p_h^*	11.3255	11.3255	11.2436	11.8355
p_l^*	8.1929	8.1429	8.6429	8.3985
$p_h^* - p_l^*$	3.1326	3.1826	2.6007	3.4370

We now state some observations, based on the above numerical results, to illustrate the behavior of the optimal price and delivery time decisions of a firm,

Table 3.5: Numerical Results: With Product Substitution

	A = 0.10				**A = 1.0**			
	TDS		**PDS**		**TDS**		**PDS**	
	DC	**SC**	**DC**	**SC**	**DC**	**SC**	**DC**	**SC**
L_h^*	0.2389	0.2393	0.2569	0.2572	0.8201	0.4277	0.8716	0.4276
$L_l - L_h^*$	0.7611	0.7607	0.7431	0.7428	0.1799	0.5723	0.1284	0.5724
p_h^*	10.8152	10.8148	10.3582	10.358	10.6938	11.29595	10.3639	10.9126
p_l^*	8.5642	8.5142	8.8789	8.8289	9.0487	8.775273	9.2512	9.0739
$p_h^* - p_l^*$	2.251	2.3006	1.4793	1.5291	1.6451	2.520679	1.1127	1.8387

Table 3.6: Effect of Capacity Sharing

	in absence of product substitution		**in presence of product substitution**	
	small A	*large A*	*small A*	*large A*
L_h^*	–	↓	↑	↓
$L_l - L_h^*$	–	↑	↓	↑
p_h^*	–	↑	↓	↑
p_l^*	↓	↓	↓	↓
$p_h^* - p_l^*$	↑	↑	↑	↑

Table 3.7: Effect of Product Substitution

	DC		**SC**			
	TDS	*PDS*	*TDS*		*PDS*	
			small A	*large A*	*small A*	*large A*
L_h^*	↓	↑	↓	↑	↑	↑
$L_l - L_h^*$	↑	↓	↑	↓	↓	↓
p_h^*	↓	↓	↓	↓	↓	↓
p_l^*	↑	↑	↑	↑	↑	↑
$p_h^* - p_l^*$	↓	↓	↓	↓	↓	↓

and also its optimal product differentiation decisions, as a result of its operations strategy or product substitution. These observations are summarized in Tables 3.6 and 3.7. These observations hold true in general, independent of the system parameter values chosen even though we are unable to establish such results analytically, especially for large A for which we do not have analytical results in a shared capacity setting. Some of these observations can be explained analytically, especially when the capacity cost A is small (see Appendix A.1). These observations are summarized in the following:

Observation 3.5. - *If a firm decides to change its operations (capacity) strategy from dedicated to shared, then (whether the products are substitutable or not): (a) it should increase price differentiation, and (b) should also increase delivery time differentiation if capacity is expensive, but decrease it (or keep it at the same level) when capacity is cheap.*

- If a firm decides to make its market-customized products available to all customers (i.e, introduces substitutability), then: (a) it should decrease price differentiation irrespective of the capacity strategy, but (b) may need to increase or decrease delivery time differentiation, depending on the capacity strategy, market conditions and marginal capacity cost.

Managerial Implications: It is evident from Table 3.6 that when the marginal capacity cost is large, sharing capacities always increases both the optimal delivery time differentiation and price differentiation of a firm, regardless of the product/market characteristics. This is in contrast to the argument presented by Boyaci and Ray (2003) that sharing capacity will lead to *"averaging"* such that all customers are served at an average speed and charged an average price. This will happen only if the firm's operations department does not discriminate between the two market segments. However, as long as the firm has a mechanism to prioritize the orders from its time sensitive customers, it is always optimal for the marketing department to differentiate its product/service based on its price and delivery time guarantee for the different market segments. In fact, we find that such a priority mechanism in a shared capacity setting requires it to maintain even a higher level of product differentiation between the two customer classes compared to the dedicated capacity setting if the capacity cost is high.

We further look at the example of FedEx versus UPS to see if the industry practice corroborates our finding. As noted earlier, FedEx uses separate facilities for its express and ground services, whereas UPS delivers express and ground services using one integrated network. Table 3.8 shows two different price and delivery time combinations offered by FedEx[1] and UPS[2] for a normal package (within 1 lb) delivery between Waterloo and Toronto, Canada. Clearly, UPS, which uses a shared capacity policy, maintains a greater delivery time and price differentiation between the two options offered, compared to FedEx, which uses dedicated capacity. This seems to be in close agreement with our observation, assuming that the marginal capacity cost is sufficiently large.

Table 3.8: Price and delivery time differentiation by FedEx vs. UPS

FedEx			**UPS**		
Service	*Guaranteed Delivery*	*Rate*	*Service*	*Guaranteed Delivery*	*Rate*
FedEx First Overnight	by 9:00 AM next day	$ 33.40	UPS Express Early A.M.	by 8:00 AM next day	$ 42.2
FedEx Priority Overnight	by 12:00 PM next day	$ 18.84	UPS Express Saver	by 12:00 PM next day	$ 15.32

The above observation also has important implications for Dell or steel, chemical and consumer product industries, cited in §2.1, that quote a specific price and delivery time combination to one segment of customers, which is not available to the other segment. The products/services offered to different market segments

[1]http://www.fedex.com/ratefinder/home?cc=ca&language=en

[2]https://wwwapps.ups.com/ctc/request?loc=en_CA

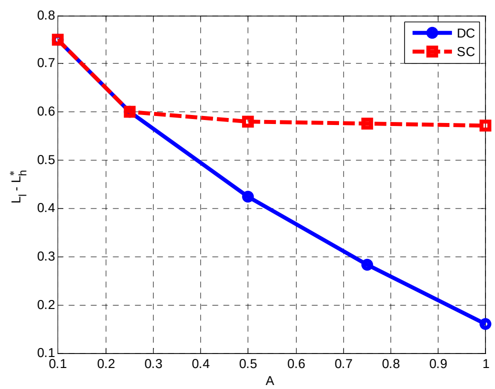
are thus non-substitutable. If these firms decide to make their products available across different market segments, allowing the customers to self-select their options, then this will require them to reduce price differentiation between the different products irrespective of the capacity strategy used. This is intuitive since the customers' preference for a given product are now affected not only by its absolute price, but also by its price compared to the other option. By keeping this price difference small, a firm can minimize the migration of customers to the lower price option, and hence its loss of revenue. The effect on the delivery time quotation will, however, depend on the market characteristics and capacity strategy used. In a dedicated capacity setting, the firm's optimal strategy will be to offer delivery time options with a greater differentiation if the market is TDS type. This is because increasing the delivery time difference in a TDS type market will induce more regular customers to switch to the express option than will the price difference cause express customers to switch to the regular option, thereby increasing its revenue. On the other hand, in a PDS type market, the firm should offer more homogeneous delivery time options. Since the customers are now more sensitive to the price difference, the firm can reduce the delivery time difference, which allows it to further decrease the price difference, thereby minimizing the migration of customers to the lower price option. In a shared capacity environment, optimal delivery time differentiation for a PDS type market will be the same as in DC, but for a TDS type, market the differentiation will further depend on the firm's marginal capacity cost.

3.3.2 Effects of Capacity Cost Increase

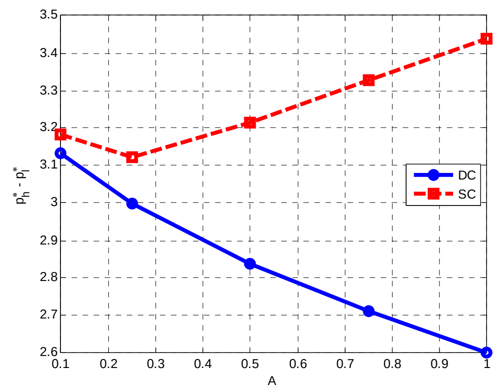
Another issue of potential managerial interest is how should the product differentiation strategy for a firm change as its marginal capacity cost A increases. Optimal delivery time and price differentiation decisions under various scenarios are shown in Figures 3.3 and 3.4 for a TDS type market and a PDS type market, respectively. The following observation summarizes our main finding in this context.

Observation 3.6. - *For a firm using a dedicated capacity strategy, its optimal response to any increase in marginal capacity cost is to decrease both the delivery time differentiation as well as the price differentiation.*

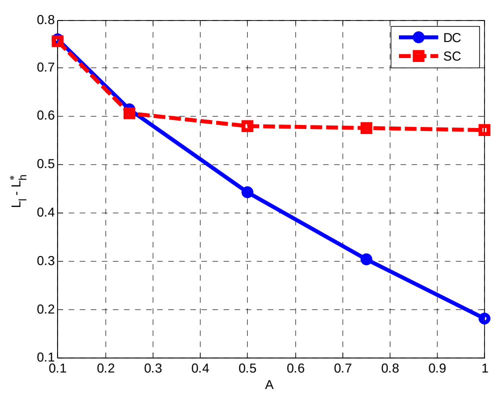
- For a firm using a shared capacity strategy, its optimal response to any increase in marginal capacity cost is to decrease the delivery time differentiation and may still need to increase the price differentiation, especially if the status-quo capacity cost is high. (Refer to Figures 3.3 and 3.4).



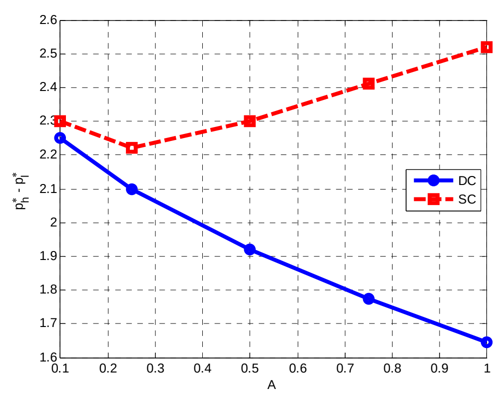
(a) Delivery time differentitaion versus marginal capacity cost in absence of product substitution



(b) Price differentitaion versus marginal capacity cost in absence of product substitution



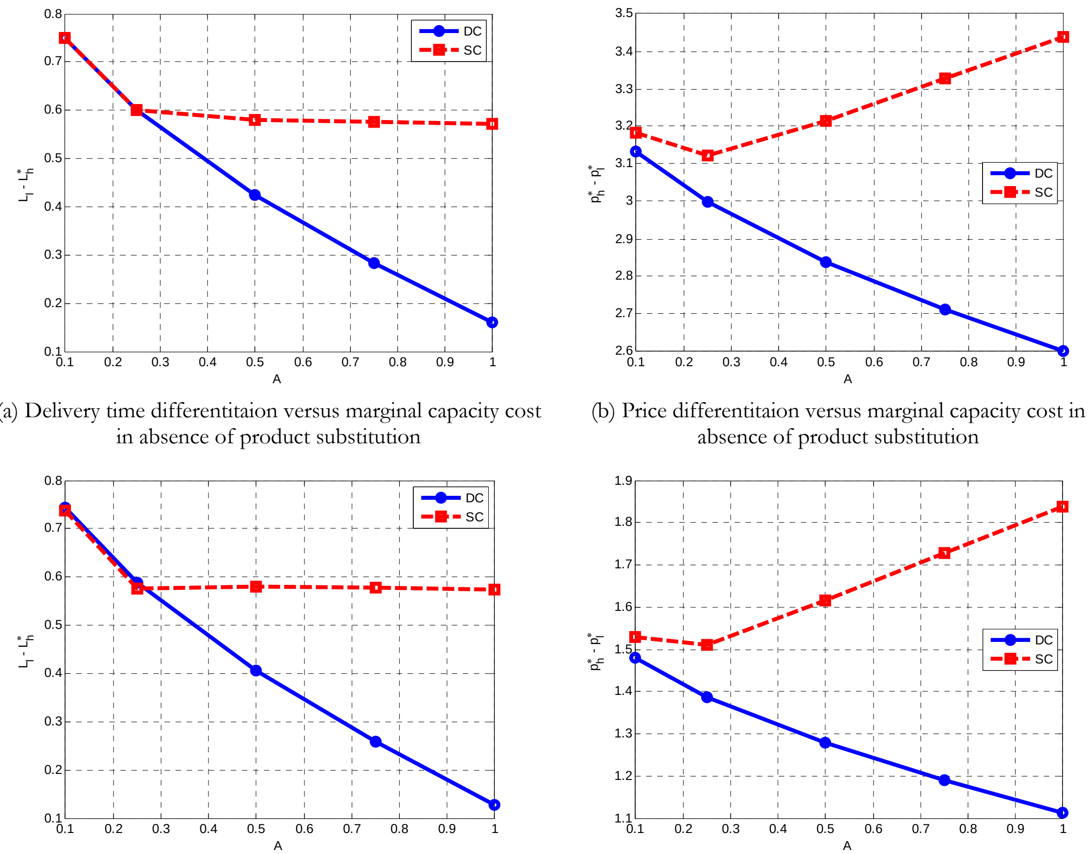
(c) Delivery time differentitaion versus marginal capacity cost in presence of product substitution



(d) Price differentitaion versus marginal capacity cost in presence of product substitution

Figure 3.3: Effects of operations strategy and product substitution in a TDS type market

One would expect increasing capacity cost to drive the prices higher. However, when the customers are also sensitive to the delivery time, the optimal strategy appears



(a) Delivery time differentitaion versus marginal capacity cost in absence of product substitution

(b) Price differentitaion versus marginal capacity cost in absence of product substitution

(c) Delivery time differentitaion versus marginal capacity cost in presence of product substitution

(d) Price differentitaion versus marginal capacity cost in presence of product substitution

Figure 3.4: Effects of operations strategy and product substitution in a PDS type market

to be otherwise. The optimal strategy is to react to an increase in capacity cost by offering longer express delivery time. Therefore, the delivery time differentiation always decreases with an increase in the marginal capacity cost, irrespective of the firm's operations strategy, absence or presence of product substitution or the market conditions. This can be intuitively explained as follows. With an increase in capacity cost, it becomes increasingly expensive for the firm to offer a shorter delivery time. Hence, for a given delivery time offered to the regular customers, the optimal strategy for the firm then is to offer a longer delivery time to the express customers, and hence reduce its delivery time differentiation. A longer express delivery time, in turn, decreases price differentiation in a dedicated capacity setting, as evident from (2.12).

An increase in the capacity cost in a shared capacity setting has similar effects if the status-quo capacity cost is small. However, the effect is more pronounced in a dedicated capacity setting. This is because an increase in capacity cost has a greater effect on the firm's profit in a dedicated capacity setting due to larger total capacity requirement. On the other hand, if the capacity cost is already high, any further increase in cost may require the firm to increase price differentiation in a shared capacity setting (although delivery time differentiation decreases).

Managerial Implications: Observation 3.6 suggests that rising fuel prices, which effectively increases capacity cost, has different implications for FedEx and UPS. The best strategy for FedEx, which operates in a dedicated capacity environment, in such a situation is to make its products more homogeneous, both in terms of delivery times and prices. Whereas UPS, which uses a shared capacity policy, needs to increase the price differentiation for its products, despite making them more homogeneous in terms of guaranteed delivery times, if the status-quo capacity cost is already high. This is likely to induce a portion of its express customers to switch to the regular delivery option. Note that the two firms should also be careful about altering their decisions properly as fuel prices start going down, as has been the case recently.

3.4 Conclusions & Future Research

In this chapter, we extensively studied the different mathematical models developed in chapter 2, which helped us generate important managerial insights. Our analytical/numerical study of the models clearly shows that the firms's operations strategy as well as its policy regarding whether to customize products for different markets or to make them available for all plays a major role in determining its optimal prices and delivery time. In a high capacity cost business environment, sharing the same capacity for processing the two products results in express (regular) customers being offered faster (slower) and more expensive (cheaper) products, compared to when there are dedicated capacities for each of them. This implies that the firm offers products with greater differentiation under a shared capacity setting. Interestingly, the above effect of the capacity strategy does not depend on any end customer characteristics or whether the products are substitutable or not. In contrast, the effects of substitutability of the products on delivery time decision do depend on the operations strategy used by the firm and the behavior of the end customers, in addition to the capacity cost. Specifically, the guaranteed delivery times for the two products may be more differentiated or more homogeneous when non-substitutable products become substitutable, depending on the values of the three factors. However, introduction of substitutability always results in cheaper express products and more expensive regular products, i.e., in a more homogeneous pricing scheme. We also demonstrated that as the capacity becomes more costly, the optimal response of the firm depends on its operations strategy, but not on demand characteristics. In that case, a dedicated capacity firm should always reduce (both price and delivery time) differentiation of its products, whereas a shared capacity firm should always offer more homogeneous delivery times, but needs to increase or decrease the price differentiation depending on whether the system is already highly capacitated or not, respectively.

The above results are managerially quite relevant. First of all, they show how capacitated firms should alter their product differentiation strategy when they make changes in their market coverage of product offerings and/or capacity strategy. We also show that managers need to pay close attention to two other factors - capacity cost of the business environment they are operating in and the behavior of their end customers - both of which play crucial role in many circumstances. As we demonstrated through the FedEx/UPS example, our results are corroborated by real-life practice. Our analysis regarding the effects of increasing capacity cost is especially relevant keeping in mind the volatility of fuel price, which directly impacts capacity cost in a number of sectors. We demonstrate how managers should optimally respond to these changing business environments in order to gain competitive advantage.

There are a number of directions in which this research can be extended. One possible extension would be to develop a good approximation for the sojourn time distribution $S_l(\cdot)$ of the low priority customers in a shared capacity setting, which can be used in the optimization model to simplify its analysis. Another possible extension may be to include the guaranteed delivery time for regular customers also as a decision variable. This will, however, bring in additional complexity in that determining the sufficiency condition for the optimal solution will be extremely challenging. In our present study, when L_l is treated as fixed, finding the sufficient condition required the profit function to be concave in L_h. This required imposing restrictions on the demand parameter a. Treating L_l also as a variable will require the profit function to be jointly concave in both L_h and L_l. This will require additional restrictions on the parameter values, which is extremely challenging to determine.

Further, in the shared capacity setting, we have assumed that the firm uses a static priority discipline that always prioritizes express customers. Occasionally, it may be prudent for the firm to give higher priority to regular customers when the queue of regular customers gets sufficiently long. As such, employing a delay dependent dynamic priority discipline (Kleinrock 1964, Kanet 1982) will be a better strategy. This will, however, make the problem extremely challenging, especially with service levels based on probability distribution of realized delivery times.

We have so far considered a monopolistic setting. It would be interesting to incorporate horizontal competition in the model as has been done by So (2000), Tsay and Agarwal (2000) and Pekgun et al. (2006). However, firms in all these studies compete for a single product, and hence product substitutability is not an issue. Modelling two competing firms each of which sells a menu of substitutable products (as is the case for FedEx and UPS) is another possible extension of our work, and is the focus of our study in the next chapter. In the following chapter, we answer the same research questions about the effects of a firm's capacity strategy on its product differentiation, but in a competitive setting, where the demand from a given customer segment is not only influenced by the service offered to other segments but also by what is offered by other competing firms.

Chapter 4

Competitive Market: Models & Solutions

4.1 Introduction

In chapters 2 and 3, we studied the effects of a firm's operations strategy, specifically sharing versus dedicating service capacities to different market segments, and product substitution on its price and delivery time differentiation strategy. Our study provided us with some useful insights. We found that sharing capacity results in a larger price differentiation between the different market segments and also a larger delivery time differentiation if the capacity cost is high. So, if FedEx, which uses dedicated capacities for different market segments, ever decides to pool its resources, it should make the two prices more different. However, the decision to make the delivery time more different or more similar will depend on its capacity cost. Further, we showed that product substitution always reduces price differentiation even if it makes the services more homogeneous in terms of their guaranteed delivery times. This has an important implication for Dell, used as an example in our earlier discussion that makes customized offers to a given customer segment, which are not available to other customer segments. If Dell decides to make a segment-specific offer available to the other segment as well, it should decrease the difference in the prices charged to the two segments. Although these are important managerial insights, they are based on the assumption that a firm makes these decisions in isolation, and the presence of other firms in the industry has no bearing on its price and delivery time decisions. In the FedEx versus UPS example, this is tantamount to saying that FedEx makes its price, delivery time and capacity decisions independent of the corresponding decisions by UPS, and vice-versa. That is to say, each firm operates as a monopolist.

Businesses in real world, however, rarely operate as monopolists. FedEx, for example, faces competition form UPS, DHL and others. Similarly, Dell competes with Hewlett Packard, IBM, Acer and others for its market share in the PC industry. In face of competition, firms make their decisions keeping in mind reactions from other firms. For example, in 2002, Sony announced to cut the price of its PlayStation 2 game console from \$299 to \$199. Micrsoft, in response, marked down the price of its Xbox console the very next day, which was followed by a similar price cut by Nintendo for its GameCube platform from \$149 to \$50 (Rudy 2002). Similary, Wal-Mart trumped price cuts by Netflix and Blockbuster by slashing the price of its standard DVD rentals-by-mail plan by 7.5% (Borland and Hansen 2004). UPS launched a new service that guarantees next day mail delivery by 8:30 a.m. in response to FedEX's "next day mail delivery by 11:00 a.m." (So and Song 1998). Other examples of price and/or lead-time competition can be found in Pekgun et al. (2006).

We know competition, in general, drives prices down. In the specific examples cited above, Sony, Microsoft, Nintendo, Wal-Mart, Netflix and Blockbuster all decided to cut their prices to compete with each other. However, whether competition increases or decreases price differentiation (price discrimination, as is popularly termed in Economics literature) is not clear, all the more so when the price discrimination is based on some endogenous category such as the delivery time guarantee (second degree discrimination (Talluri and Van Ryzin 2004)).

The textbook theory argues that competitive firms cannot price discriminate since they are price takers, while monopolists can price discriminate to the extent that there exists both homogeneity in consumers' demand elasticities and a useful sorting mechanism to distinguish between consumer types (Gerardi and Shapiro 2007). The textbook theory, therefore, predicts that competition should decrease price discrimination. This is further corroborated by the theoretical model of Rochet and Stole (2002) on second degree price discrimination. However, the theoretical models of Gale (1993) and Stole (1995) produce exactly the opposite results. As there is no overarching theoretical model, the relation between competition and price discrimination becomes an empirical question. However, different empirical studies have again produced very contrasting results. Borenstein (1989) and Borenstein and Rose (1994) found evidence of increasing price dispersion with competition in airline industry, thereby suggesting that competition increases price discrimination. However, a more detailed empirical study by Gerardi and Shapiro (2007) found a negative relation between market competition price dispersion, thereby suggesting that competition decreases price discrimination. Further, a more complex analysis is necessary when the firm must price discriminate on the basis of some endogenous category such as the delivery time preference (Varian 1989). Moreover, the effect of competition on delivery time differentiation between different customer segments itself is not clear.

We now extend the model for price and delivery time decisions, developed in chapter 2, to a competitive setting in which a firm's decisions are influenced by other firms in the industry. Our primary objective is to investigate if the managerial insights generated in the previous chapter for a monopolistic setting hold true, in general, for a competitive setting. Through our study, we try to shed more light into the effect of market competition, and also of firms capacity strategies, on price and delivery time differentiation/discrimination. Specifically, we try to address the following:

- How does the operation strategy (dedicated or shared capacity) of a firm affect its price and delivery time decisions for its substitutable products, and hence its product differentiation strategy, in presence of competition from other firms in the industry?

- How does asymmetry in firms' operating conditions (in terms of capacity settings or cost parameters) affect the equilibrium prices and delivery times?
- How does competition affect the price and delivery time differentiation decisions of a firm relative to a monopolistic setting?

To explore the impact of a firm's operations strategy (dedicated versus shared capacity), we compare the following three scenarios given in Table 4.1:

Table 4.1: Different capacity scenarios in a competitive market

Scenario	**Capacity Setting**	
	Firm 1	*Firm 2*
DD	Dedicated	Dedicated
SS	Shared	Shared
DS	Dedicated	Shared

Comparisons among the three scenarios brings out the effect of a firm's operations strategy on its price and delivery time differentiation in a competitive market. Comparing the results of a competitive market with those of a monopolistic market further brings out the effect of market competition.

Rest of the chapter is organized as follows. In §4.2, we provide a review of the related literature. We present our mathematical model and the underlying assumptions in §4.3. In §4.4, we describe the best response of a firm, given its competitor's price and delivery time decisions, for both the dedicated and shared capacity settings. Equilibrium solution for the duopoly problem is presented in §4.5. Discussion of results is deferred to chapter 5.

4.2 Literature Review

Literature on price and/or time competition can mostly be classified into (Cachon and Harker 2002) (i) papers on inventory games, and (ii) those on queueing games. Papers on inventory games are relevant in a make-to-stock setting where firms use inventory as their strategic tool to compete in the market. Parlar (1988), Li (1992), Ha et al. (2003), Dai (2003), Bernstein and Federgruen (2003), Bernstein and Federgruen (2004) are some of the papers that fall in this category. Papers on queueing games are pertinent to make-to-order or service industries, where firms use better (server) capacity/queue management to adjust their price and delivery time decisions, and thus compete in the market. Since our model is relevant to make-to-order or service industries, our focus is on the latter category.

One of the first papers on queueing games is by Levhari and Luski (1978), who consider two competing firms providing identical service to customers and having identical service time distribution. However, customers are heterogenous in their sensitivity to waiting times, which is captured using different costs that they attach to their delay in service. Customers decide whether or not to seek service from one of these firms. It is shown that at equilibrium, the firms charge different prices in general, and that the firm charging higher price specializes in serving more time sensitive customers. Kalai et al. (1992) consider competition between two firms with service speed as a strategic tool to capture a larger market share, but pricing is not a decision variable.

Literature on queueing games can further be categorized into (i) papers that aggregate price and waiting time cost into a single measure called "full price", and (ii) those that model price and delivery/waiting time as independent explanatory variables. Levhari and Luski (1978), Loch (1991), Lederer and Li (1997), Chen and Wan (2003) and Armony and Haviv (2003) belong to the former category. All these papers assume that customers associate a specific cost rate with their waiting time, and that they make their selection of a firm based only on their "full price", which is the sum of the actual price charged and the expected delay cost, disregarding any other service attributes. Further, they assume that all customers are in a position to assess the equilibrium steady state waiting times they will experience.

Chen and Wen (2003) study a duopoly market with two service providers, each modelled as an M/M/1 server, that compete for a single customer class on the basis of full price, but charge the same full price in the long run. Loch (1991) considers a duopoly market with service providers operating as M/G/1 servers with given service rates, and two customer classes, each with a given waiting cost rate and average service time. A customer selects the firm that offers the lowest full price. The author shows the existence of a Nash Equilibrium in which the customers are prioritized according to the "$c\mu$ rule" (see Mendelson and Whang 1990 for a discussion of the "$c\mu$ rule"). The author shows that the game has a unique pure-strategy Nash equilibrium when the two firms are symmetric, i.e., have the same service rates. However, pure-strategy equilibrium may not exist when the firms are asymmetric.

Armony and Haviv (2003) study a similar duopoly problem with two customer classes with each firm modelled as an M/M/1 server. Competition is modelled in two stages. In the first stage, firms compete on the basis of their prices. In the second stage, given the prices, the two customer classes decide whether or not to seek service or how to allocate their demand between the two firms.

Lederer and Li (1997) consider a more generalized model with an arbitrary number of competing firms, each modelled as an M/G/1 server, and an arbitrary number of customer classes. Firms compete in the market by selecting their prices, production rates and scheduling policies. The authors establish the existence of a Nash Equilibrium in which customers are prioritized according to the "$c\mu$ rule".

The second category of papers, which model price and delivery time as independent explanatory variables, include So (2000), Pekgun et al. (2006), Allon and Federgruen (2007, 2008). These papers model customers' aggregate demand for a firm's service as a function of its price, delivery time and/or other service attributes, each of which is explicitly advertised by the firm. So (2000) uses a multiplicative competitive interaction (MCI) model to represent the market shares of an arbitrary number of firms competing for the same product based on their prices and delivery time guarantees. Each firm is modelled as an M/M/1 server, which targets to meet its promised delivery time guarantee with at least a certain degree of reliability. The author shows how heterogeneous firms exploit their competitive advantage, in terms of a higher capacity or a lower operating cost, to differentiate their services.

Pekgun et al. (2006) study two firms competing in a common market based on their price and lead-time decisions, and explore the impact of centralization versus decentralization of these decisions, as quoted by the marketing and production departments, respectively. They model the competing firms as M/M/1 servers, and each firm's expected demand as a linear function of the prices and delivery times quoted by both the firms.

Allon and Federgruen (2007) study competition between an arbitrary number of firms, each modelled as an M/M/1 server, for a class of homogeneous customers. They model the expected demand for each firm as a separable function of all firms' prices and service levels, which is also linear in the prices. The service level is defined as the difference between an upper bound benchmark for waiting time and the firm's actual waiting time standard. They study three types of competition depending on the order in which the decision variables are selected: (i) service-level first, (ii) price first, and (iii) simultaneous game. Allon and Federgruen (2008) further extend this to a setting with multiple customer classes, all served by shared service facilities. Each firm competes by advertising its price and expected waiting time, and selects its optimal capacity level and a priority discipline to serve the customers. Their demand model is separable by customer class, i.e., the demand rate for a firm from a given customer class is independent of the prices and waiting time standards offered to other customer classes. They study the equilibrium behavior of the firms under three types of competition: (i) price competition, (ii) waiting time competition, and (iii) simultaneous competition.

It is worth mentioning the work of Tsay and Agrawal (2000) whose demand model, although in a deterministic setting, bears some similarity with Pekgun et al. (2006) and Allon and Federgruen (2007). They study a distribution system in a non-queueing framework in which a manufacturer supplies a common product to two retailers, who compete for end customers based on their retail prices and service.

There is another line of research on duopolies in price and time sensitive markets, where customers strategically choose the firm that maximizes their utility, which is generally a function of their price and other service attributes. Besbes and Zeevi (2005) model utility as a function of price and waiting time. Ho and Zheng (2004) model customers' utility as a function of their expected delivery time and service quality, whereas Li and Lee (1994) model it as a function of price, response time and quality. It is possible to derive an aggregate demand function from the underlying utility model (see Anderson et al. 1992, Farahat and Perakis 2008 Liu et al. 2007).

We position our work in the second category since we treat price and delivery times as independent variables announced by a firm. Although our demand model bears some similarity with those used by Tsay and Agrawal (2002), Pekgun et al. (2006) and Allon and Federgruen (2007), it is fundamentally different from them in that these models consider only a single customer class, and thus there is no market segmentation. To our knowledge, Loch (1991), Lederer and Li (1997), Armony and Haviv (2003) and Allon and Federgruen (2008) are the only works to have addressed the phenomenon of market segmentation. As noted earlier, these papers, except for Allon and Federgruen (2008), assume that customers aggregate the price and delivery time into a full price, and that they select the service provider on the basis of this full price only. In doing so, they assume that all customers are in a position to assess the equilibrium steady state waiting times they will experience, while in our model, waiting time standards are advertised to the different classes. Moreover, they consider the firms' capacity levels as exogenously given, and not a decision variable. Thus, Allon and Federgruen (2008) appears to be the closest to our work. However, they study completely segmented markets, which means that each customer is strictly assigned to a specific class, and she cannot switch between different classes. This is tantamount to saying that the specific service package (price and delivery time combination) offered to a given customer class is not available to any other class, and hence is non-substitutable. Our demand model is more generalized, which also captures product substitution. Moreover, Allon and Federgruen (2008) use a service level that is based on expected delivery times. In other words, they assume that firms select their capacity levels so that customers from each segment are served within their promised delivery times on average. As discussed in 2, this does not provide any bound on instances of unusually long delivery times.

4.3 Decision Models

4.3.1 Modelling Assumptions

We consider a service or a make-to-order manufacturing industry with two firms, indexed by $i \in \{1, 2\}$ and $j = 3 - i$, competing in a market that is segmented into 2 customer classes, indexed by $k \in \{h, l\}$. As described in chapter 2, class h customers are high priority/express customers who are more time sensitive and are willing to pay a price premium for a shorter delivery time. Class l customers are low priority/regular customers who are more price sensitive and are willing to accept a longer delivery time for a price discount. Firm i competes for its market share by selecting its prices p_k^i and guaranteed delivery times L_k^i offered to market segment $k \in \{h, l\}$. Firm i faces a demand from class k, generated according to a Poisson process with rate $\lambda_k^i \, (p_k^i, L_k^i, k \in \{h, l\}, i \in \{1, 2\})$, which depends on the decisions of both firms in the following way: each firm's expected demand from a given market segment is (i) decreasing in its price and delivery time offered to that segment, (ii) increasing in its price and delivery time offered to the other market segment, and (iii) increasing in the price and delivery time offered offered by the other firm to the same segment. We model this using the following system of linear equations:

$$\lambda_h^i = a^i - \beta_p^h p_h^i + \theta_p(p_l^i - p_h^i) - \beta_L^h L_h^i + \theta_L(L_l^i - L_h^i) + \gamma_p(p_h^j - p_h^i) + \gamma_L(L_h^j - L_h^i) \quad (4.1)$$

$$\lambda_l^i = a^i - \beta_p^l p_l^i + \theta_p(p_h^i - p_l^i) - \beta_L^l L_l^i + \theta_L(L_h^i - L_l^i) + \gamma_p(p_l^j - p_l^i) + \gamma_L(L_l^j - L_l^i) \quad (4.2)$$

where,

$2a^i$	:	market base of firm i
β_p^k	:	sensitivity of class k demand to its own price
β_L^k	:	sensitivity of class k demand to its own guaranteed delivery time
θ_p	:	sensitivity of demand to inter-class price difference
θ_L	:	sensitivity of demand to inter-class delivery time difference
γ_p	:	sensitivity of demand to inter-firm price difference
γ_L	:	sensitivity of demand to inter-firm delivery time difference

$2a^i$ parameterizes firm $i's$ market base. Mathematically, it is the demand faced by firm i when price and delivery time offered by each firm to each customer class is zero. It captures the aggregate effect of all the factors such as a firm's brand image, service quality, etc other than price and delivery time on demand. Hence, the firm offering the lowest price and the shortest delivery time to a market segment need not capture its entire demand. The relative values of a^i and a^j can be loosely used to describe comparative advantage in terms of a firm's market penetration. This may reflect the underlying preferences of customers for one firm over the other, which may be due to customers' appeal for a brand.

The behavior of the market depends on the relative sensitivities of customers to prices and delivery times, described through various market parameters. Two specific combinations of these parameters are of special interest, as we will see in our analysis in the next chapter. These combinations define specific market behavior, which are extensions of their counterparts described for the monopolistic setting in chapters 2 and 3.

- *Time Difference Sensitive (TDS)*: We say the market is TDS type when the relative sensitivity of customers to the inter-class delivery time difference (with respect to their own delivery time and inter-firm delivery time difference) is greater than their relative sensitivity to the inter-class price difference (with respect to their own price and inter-firm price difference), such that $\theta_L/(\beta_L^k + \gamma_L) > \theta_p/(\beta_p^k + \gamma_p)$, $k \in \{h, l\}$.

- *Price Difference Sensitive (PDS)*: We say the market is PDS type when the relative sensitivity of customers to the inter-class price difference (with respect to their own price and inter-firm price difference) is greater than their relative sensitivity to the inter-class delivery times difference (with respect to their own delivery time and inter-firm delivery time difference), such that $\theta_p/(\beta_p^k + \gamma_p) > \theta_L/(\beta_L^k + \gamma_L)$, $k \in \{h, l\}$.

The above demand model ((4.1) - (4.2)) is an extension of the monopolist demand model used in Chapters 2 and 3, and captures the cross-firm effect on a firm's demand. It also generalizes the demand model used by Boyaci and Ray (2003) to a competitive setting, and those used by Tsay and Agarwal (2000) and Pekgun et al. (2006) to segmented markets. Further, it generalizes the demand model used by Allon and Federgruen (2008) by taking into account product substitution (θ_p and θ_L). The total market size ($\sum_{i\in\{1,2\}}\sum_{k\in\{h,l\}}\lambda_k^i$) in our model is invariant to any changes in inter-firm or inter-class sensitivities, which only affects the distribution of the total demand among the firms and the customer classes. However, the pricing and delivery time decisions of the two firms affect both the total market size as well as the resulting demand for each firm and from each market segment. This is in sharp contrast to the multiplicative competitive interaction model used by So (2000), which assumes the total market size to be constant. We make the following assumptions regarding the market parameters:

Assumption 4.1. *$\beta_p^k > 0$, $\beta_L^k > 0$, $\theta_p \geq 0$, $\theta_L \geq 0$, $\gamma_p \geq 0$, $\gamma_L \geq 0$, $\beta_p^h < \beta_p^l$ and $\beta_L^h > \beta_L^l$.*

This is to ensure that demand from a market segment is decreasing in its own price and delivery time; increasing in price and delivery time offered by the same firm to the other segment; and increasing in price and delivery time offered by the other firm to the same segment. $\beta_p^h < \beta_p^l$ and $\beta_L^h > \beta_L^l$ are required by definition of the two customer classes. $\theta_p > 0$, $\theta_L > 0$ signifies product substitution, while $\gamma_p > 0$, $\gamma_L > 0$ signify the presence of price competition and delivery time competition in the market. $\gamma_p = \gamma_L = 0$ makes the demand of two firms independent, and hence decouples their decision making, resulting in a monopolistic setting.

Assumption 4.2. *The market base a^i is sufficiently large.*

This assumption is similar to Assumption 2.2 of chapter 2, and is required, as we will see in §4.4, to ensure that firm $i's$ best response always consists of non-negative prices and demands and a smaller delivery time to express customers compared to that offered to regular customers, i.e., $L_h^i < L_l^i$.

It is important to note that our demand model does not explicitly consider the impact of the reliability of delivery time guarantees. Firms that constantly miss on their promised delivery time will eventually lose their credibility with customers for future business, which defeats the very purpose of exploiting delivery time guarantees to attract customers. In fact, in the airline industry, independent government agencies (e.g., the Aviation Consumer Protection Division of the DOT) and Internet travel services (e.g., Expedia) report, on a flight-by-flight basis, the average delay and percentage of flights arriving within 15 minutes of their schedule (Allon and Federgruen 2007). Indeed, firms target to meet their guaranteed delivery times with at least a given level of reliability, and carefully monitor their delivery performance. To take this into consideration in our model, we use a reliability level α^i with which firm i tries to meet its delivery time guarantee. In our model, we restrict our discussions only to cases where the service reliability for each firm is the same, i.e., $\alpha^i = \alpha$. This is applicable in situations where there exists some industry standard and published reports (like those published by the Aviation Consumer Protection Division of the DOT and Expedia for airline industry) such that the delivery performance of each firm is readily available to customers. In this way, firms are discouraged from performing below the standard such that the market share is

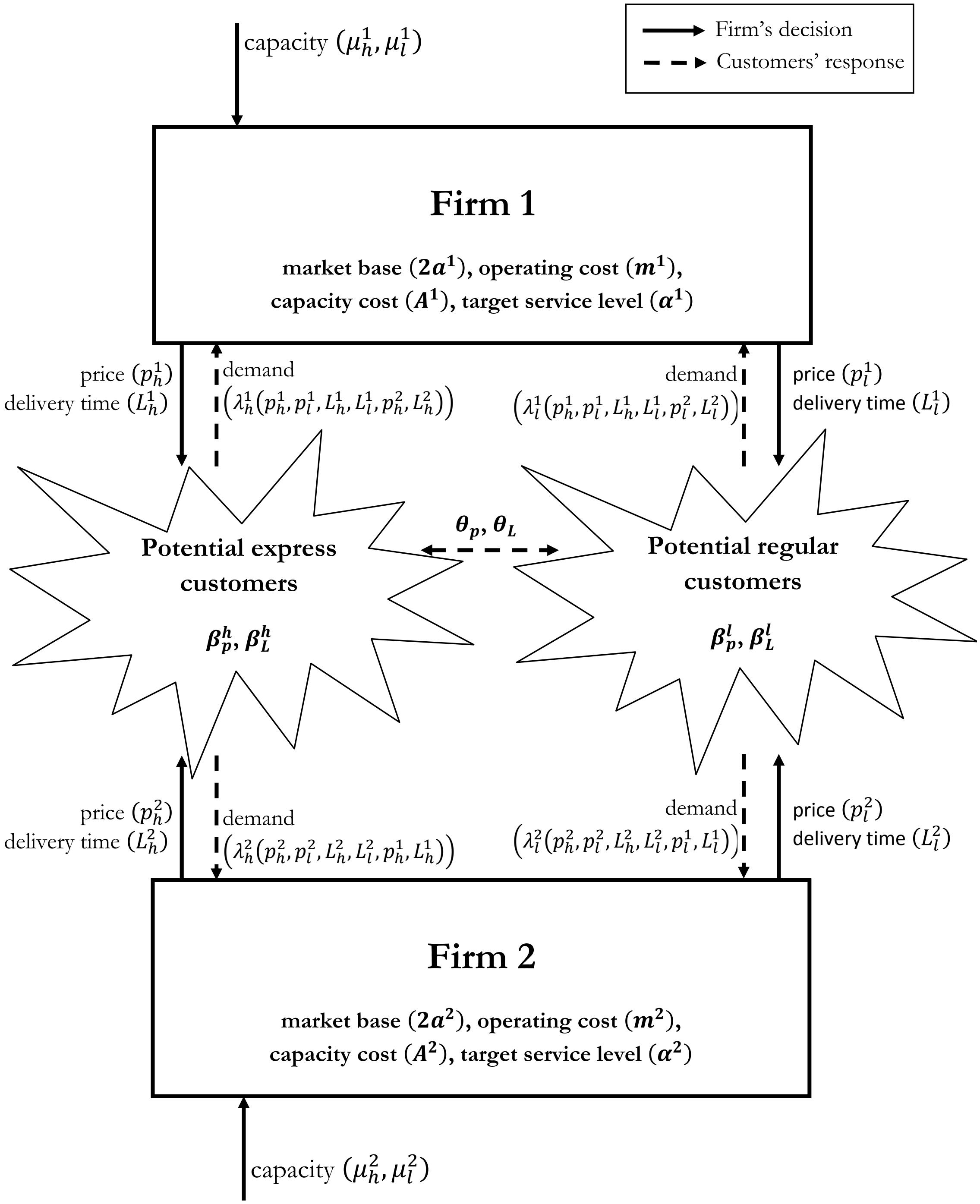


Figure 4.1: Schematic representation of a competitive model

then mainly affected by their promised times and prices, as depicted by our demand model. Of course, our model and analysis also allow for different service reliabilities for different firms.

We assume the time firm i takes to serve a customer from class k is exponentially distributed with a rate μ_k^i. Firm i, therefore, behaves like an M/M/$\cdot$ queuing system. Further, it serves its customers within a given class on a first-come-first-serve (FCFS) basis. We assume firm i incurs the same operating cost of $\$m^i$ and a marginal capacity cost of $\$A^i$ in serving a customer of either class.

The industry is assumed to have established a standard delivery time for regular customers, and hence $L_l^i = L_l^j = L_l$. Firm i selects and announces its express delivery time and the two prices $(L_h^i,\ p_h^i,\ p_l^i)$ so as to maximize its profit. Firm i does so taking into account the delivery time and prices $(L_h^j,\ p_h^j,\ p_l^j)$ selected by firm $j = 3 - i$ since they have an impact on firm $i's$ demands, and hence on its profit. It also needs to simultaneously select its optimal service rates (i.e., installed capacities) μ_k^i in order to meet the guaranteed delivery times with at least a minimum level of reliability α. A schematic representation of the model is shown in Figure 4.1.

Notation

i, j	:	indices for firm; $i \in \{1, 2\}$, $and\, j = 3 - i$
k	:	index for customer class; $k \in \{h, l\}$
λ_k^i	:	mean demand rate for firm i from customer class k (units/unit time)
μ_k^i	:	mean processing rate of firm i for customer class k (units/unit time)
p_k^i	:	price charged by firm i to customer class k (\$/unit)
L_k^i	:	delivery time quoted by firm i to customer class k (time units)
W_k^i	:	steady state actual sojourn (waiting + service) time of customer class k at firm i (time units)
α^i	:	target service level set by firm i (no unit)
$S_k^i(L_k^i)$	:	actual service level achieved by firm i for quoted delivery time L_k^i, i.e., $\mathrm{P}(W_k^i \le L_k^i)$ (no unit)
m^i	:	unit operating cost of firm i (\$/unit)
A^i	:	marginal capacity cost of firm i (\$/unit)

In the next section, we describe the best response of a firm, given its competitor's price, delivery time and capacity decisions.

4.4 The Best Response of a Firm

Given the price, delivery time and capacity decisions $(p_h^j, p_l^j, L_h^j, \mu_h^j, \mu_l^j)$ of firm $j = 3 - i$, firm $i \in \{1, 2\}$ selects its own corresponding decisions $(p_h^i, p_l^i, L_h^i, \mu_h^i, \mu_l^i)$ that maximize its profit and also ensure that its delivery time commitments are met reliably. As clear from the demand model ((4.1) - (4.2)), the demands for firm $i \in \{1, 2\}$, and its decisions in turn, depend on the price and delivery time decisions made by firm $j = 3 - i$. Firm $i's$ demand and its decisions, however, do not depend on the capacity level (μ_h^j, μ_l^j) selected by firm j. While competing with the other firm, each firm, therefore, possesses only two types of essential strategic instruments: prices and the delivery times. Firm $i's$ strategy can be defined as a vector of its strategic decision variables $s^i := (p_h^i, p_l^i, L_h^i)$, which it uses to compete against the other firm j. The best response of firm i to firm $j's$ strategy, $s^j := (p_h^j, p_l^j, L_h^j)$, is thus a strategy $s^{i*} := (p_h^{i*}, p_l^{i*}, L_h^{i*})$ such that $\pi^i(s^{i*}, s^j) = \max_{s^i} \pi^i(s^i, s^j)$, $i \in \{1, 2\}$ and $j = 3 - i$. Firm $i's$ best response is the solution to the following optimization problem:

<u>PDTDP$^{\mathbf{i}}$:</u>

$$\max_{p_h^i, p_l^i, L_h^i, \mu_h^i, \mu_l^i} \pi^i = (p_h^i - m^i)\lambda_h^i + (p_l^i - m^i)\lambda_l^i - A^i(\mu_h^i + \mu_l^i) \tag{4.3}$$

subject to:

$$L_h^i < L_l^i \tag{4.4}$$

$$p_h^i, p_l^i, \lambda_h^i, \lambda_l^i, \mu_h^i, \mu_l^i \geq 0 \tag{4.5}$$

$$\textit{Stability condition} \tag{4.6}$$

$$S_h^i(L_h^i) = P(W_h^i \leq L_h^i) \geq \alpha \tag{4.7}$$

$$S_l^i(L_l^i) = P(W_l^i \leq L_l^i) \geq \alpha \tag{4.8}$$

We call it the Pricing and Delivery Time Decision Problem for firm i ($PDTDP^i$). As noted in chapter 2, in certain situations delivery times may be relatively sticky. In such situations, firms use prices as the only strategic variables for competition. Firm $i's$ best response is then obtained by solving $PDTDP^i$ for fixed L_h^i and L_l^i. We call the resulting problem a Pricing Decision Problem, denoted by PDP^i. Notice that a firm's best response problem has a form very much similar to a firm's optimal decision problem in a monopolistic setting, described in chapter 2. Difference still arises between the two due to a different demand model for a firm in a competitive setting, which also takes into account the effect of the price and delivery time decisions of its competitor. Therefore, the best response of a firm can also be solved using similar solution methods as developed in chapter 2. We, therefore, keep our discussion of the solution method very brief, citing important results, which are later used to obtain the equilibrium decisions of competing firms.

$PDTDP^i$ is a generalized model of a firm's best response, irrespective of its capacity strategy. In what follows, we develop the specialized model of $PDTDP^i$

for firm i using dedicated or shared capacities by specifying the exact form of constraints (4.6)-(4.8), and discuss the solution method for each.

4.4.1 Dedicated Capacity Setting

As discussed in chapter 2, for a dedicated capacity setting, where each customer class is served by a separate M/M/1 server, the tail of the sojourn time distribution for either class of customers is known to be exponential. In this case, there is a separate stability condition for each of the queues. Hence, constraints (4.6), (4.7) and (4.8) can be expressed as:

$$\lambda_k^i - \mu_k^i < 0, \ k \in \{h, l\} \tag{4.6^{DC}}$$

$$S_h^i(L_h^i) = P(W_h^i \leq L_h^i) = 1 - e^{(\lambda_h^i - \mu_h^i)L_h^i} \geq \alpha \tag{4.7^{DC}}$$

$$S_l^i(L_l^i) = P(W_l^i \leq L_l^i) = 1 - e^{(\lambda_l^i - \mu_l^i)L_l^i} \geq \alpha \tag{4.8^{DC}}$$

We denote the resulting Pricing and Delivery Time Decision Problem for firm i in a dedicated capacity setting by $PDTDP_{DC}^i$.

Proposition 4.1. *In a dedicated capacity setting, both the delivery time reliability constraints* (4.7^{DC}) *and* (4.8^{DC}) *are binding at optimality.*

Proof. Delivery time reliability constraints (4.7^{DC}) and (4.8^{DC}) can be rewritten as:

$$\mu_k^i \geq -\frac{\ln(1-\alpha)}{L_k^i} + \lambda_k^i \quad k \in \{l, h\}$$

The profit function π^i is decreasing in μ_k^i. Therefore, to maximize profit, the two service rates should be at their minimum level that guarantees the desired service level α. This implies that at optimality, the two delivery time reliability constraints (4.7^{DC}) and (4.8^{DC}) must be binding, and the service rates are given by:

$$\mu_k^i = -\frac{\ln(1-\alpha)}{L_k^i} + \lambda_k^i, \ k \in \{h, l\}$$

□

Proposition 4.1 suggests that it is optimal for firms to stick to their minimum delivery time reliability (α) since a better service level to customers comes at an extra cost to the firm. As a result of proposition 4.1, $PDTDP^i_{DC}$ reduces to maximizing (4.3) with μ_i as given above. The stability conditions (4.6^{DC}) are automatically satisfied by the above equation. This allows us to reduce $PDTDP^i_{DC}$ to the following optimization problem:

$\underline{\mathbf{PDTDP^{i'}_{DC}:}}$

$$\max_{p^i_h, p^i_l, L^i_h} \pi^i = (p^i_h - m^i - A^i)\lambda^i_h + (p^i_l - m^i - A^i)\lambda^i_l + A^i \frac{\ln(1-\alpha)}{L^i_h} + A^i \frac{\ln(1-\alpha)}{L_l} \tag{4.9}$$

subject to:

$L^i_h < L_l$

$p^i_h, p^i_l, \lambda^i_h, \lambda^i_l, L^i_h \geq 0$

Proposition 4.2. *For a fixed L^i_h, the objective function (4.9) of $PDTDP^{i'}_{DC}$ is strictly concave in p^i_h and p^i_l.*

Proof. The Hessian for (4.9), for a fixed L^i_h, is given by:

$$\begin{pmatrix} -2(\beta^h_p + \theta_p + \gamma_p) & 2\theta_p \\ 2\theta_p & -2(\beta^l_p + \theta_p + \gamma_p) \end{pmatrix}$$

Clearly, the first order leading principal minor of the Hessian is negative, while its determinant is positive. This proves that the objective function (4.9) in $PDTDP^{i'}_{DC}$ is strictly concave for a fixed L^i_h. □

Proposition 4.2 suggests that, for a fixed L^i_h, $PDTDP^{i'}_{DC}$ has a unique maximum, which can be obtained using functional optimization of its objective function (4.9), as long as p^i_h, p^i_l, λ^i_h and λ^i_l are non-negative. Proposition 4.3 gives the best response prices of firm i in a dedicated capacity setting to a given strategy used by firm j.

Proposition 4.3. *If firm i operates in a dedicated capacity setting, then for a given strategy $s^j := (p_h^j, p_l^j, L_h^j)$ by firm $j = 3-i$, the best response $s^{i*} := (p_h^{i*}(L_h^i), p_l^{i*}(L_h^i))$ for a fixed L_h^i by firm $i \in \{1,2\}$ is given by:*

$$\begin{aligned} p_h^{i*}(L_h^i) =& \frac{(\beta_p^l + 2\theta_p + \gamma_p)a^i - \{\beta_p^l(\theta_L + \gamma_L) + \beta_L^h(\beta_p^l + \theta_p + \gamma_p) + \theta_p\gamma_L + \gamma_L\gamma_p + \theta_L\gamma_p\}L_h^i}{2D} \\ & \frac{+\{(\beta_p^l + \gamma_p)\theta_L - (\beta_L^l + \gamma_L)\theta_p\}L_l^i + (\beta_p^l\gamma_p + \gamma_p\theta_p + \gamma_p^2)p_h^j + (\theta_p\gamma_p)p_l^j}{2D} \\ & \frac{+(\beta_p^l\gamma_L + \gamma_L\gamma_p + \theta_p\gamma_L)L_h^j + (\theta_p\gamma_L)L_l^j}{2D} + \frac{A^i + m^i}{2} \end{aligned} \tag{4.10}$$

$$\begin{aligned} p_l^{i*}(L_h^i) =& \frac{(\beta_p^h + 2\theta_p + \gamma_p)a^i - \{\beta_p^h(\theta_L + \gamma_L) + \beta_L^l(\beta_p^h + \theta_p + \gamma_p) + \theta_p\gamma_L + \gamma_L\gamma_p + \theta_L\gamma_p\}L_l^i}{2D} \\ & \frac{+\{(\beta_p^h + \gamma_p)\theta_L - (\beta_L^h + \gamma_L)\theta_p\}L_h^i + (\theta_p\gamma_p)p_h^j + (\beta_p^h\gamma_p + \gamma_p\theta_p + \gamma_p^2)p_l^j}{2D} \\ & \frac{+(\theta_p\gamma_L)L_h^j + (\beta_p^h\gamma_L + \gamma_L\gamma_p + \theta_p\gamma_L)L_l^j}{2D} + \frac{A^i + m^i}{2} \end{aligned} \tag{4.11}$$

where $D = \beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p + \beta_p^h\gamma_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2$.

Proof. $p_h^{i*}(L_h)$ and $p_l^{i*}(L_h)$ are obtained by solving the following system of equations:

$$\frac{\partial \pi(L_h^i)}{\partial p_h^i} = 0$$

$$\frac{\partial \pi(L_h^i)}{\partial p_l^i} = 0$$

Since (4.9), for a fixed L_h^i, is strictly concave in p_h^i and p_l^i, solving the above system of equations gives a unique pair of prices that maximizes $\pi^i(L_h^i)$. □

The corresponding optimal price differentiation is then:

$$\begin{aligned} p_h^{i*}(L_h^i) - p_l^{i*}(L_h^i) =& \frac{\{(\beta_p^l - \beta_p^h)a^i + (\beta_p^h + \beta_p^l)\theta_L + (\gamma_L + 2\theta_L)\gamma_p\}(L_l^i - L_h^i)}{2D} \\ & \frac{-(\beta_p^l\beta_L^h + \beta_p^l\gamma_L + \beta_L^h\gamma_p)L_h^i + (\beta_p^h\beta_L^l + \beta_p^h\gamma_L + \beta_L^l\gamma_p)L_l^i + (\beta_p^l + \gamma_p)\gamma_p p_h^j}{2D} \\ & \frac{-(\beta_p^h + \gamma_p)\gamma_p p_l^j + (\beta_p^l + \gamma_p)\gamma_L L_h^j - (\beta_p^h + \gamma_p)\gamma_L L_l^j}{2D} \end{aligned} \tag{4.12}$$

Example 4.1: Assume the customer specific and firm specific parameter values as shown in Table 4.2. Given firm $1's$ strategy $s^1 := (p_h^1 = 10,\ p_l^1 = 8,\ L_h^1 = 0.5)$, firm $2's$ best response prices and other related variables for $L_h^2 = 0.5$ are given in Table 4.3.

Table 4.2: Parameter values for Example 4.1

β_p^h	β_p^l	θ_p	β_L^h	β_L^l	θ_L	γ_p	γ_L	a^2	m^2	A^2	α	L_l
0.5	0.7	0.4	0.9	0.7	0.1	0.4	0.4	10	3	1	0.99	1

Table 4.3: Results for Example 4.1

$p_h^{2*}(L_h)$	$p_l^{2*}(L_h)$	$\mu_h^{2*}(L_h)$	$\mu_l^{2*}(L_h)$	$\lambda_h^{2*}(L_h)$	$\lambda_l^{2*}(L_h)$	$\pi^{2*}(L_h)$
9.0894	8.0405	14.2103	8.6302	5.0	4.025	27.8944

(4.10) and (4.11) in proposition 4.3 suggest that in pricing its product for a given customer segment, a firm should take into account the price quoted by the other firm not only to the same customer segment but also to the other customer segment. This, at first thought, sounds surprising. This is because our demand functions (4.1) and (4.2) suggest that a firm's demand from a given segment is not influenced by what is offered to the other segment by the other firm. To make things clear, our demand function (4.1), for example, suggests that the demand faced by firm 1 from the express segment depends on the price charged by firm 2 to the express segment, but is not influenced by what firm 2 charges to the regular customers. However, (4.10) suggests that in pricing its product for express customers, firm 1 should keep in mind not only the price charged by firm 2 to the express customers but also the price charged by firm 2 to the regular customers. This is because firm $2's$ price to regular customers influences its demand from express customers as well. So, in pricing its product for express customers, firm 1 should take into account the other factors that influence express customers' decision, which includes the price charged by firm 2 to regular customers.

Proposition 4.4. *If firm i operates in a dedicated capacity setting, then for a given strategy $s^j := (p_h^j, p_l^j, L_h^j)$ by firm $j = 3 - i$, the best response express delivery time L_h^{i*} of firm $i \in \{1, 2\}$ is given by the unique root of (4.13) in the interval [0, L_l)*

$$\frac{\partial \pi^i(L_h^i)}{\partial L_h^i} = -\left(\beta_L^h + \theta_L + \gamma_L\right)\left(p_h^{i*}(L_h^i) - m^i - A^i\right) + \theta_L\left(p_l^{i*}(L_h^i) - m^i - A^i\right) - \frac{A\ln(1-\alpha)}{(L_h^i)^2} \tag{4.13}$$

where, $p_h^{i*}(L_h^i)$ and $p_l^{i*}(L_h^i)$ are given by (4.10) and (4.11).

Proof. Substituting the optimal prices, given by 4.3, into the objective function, and differentiating it with respect to L_h gives (4.13). Also,

$$\frac{\partial^2 \pi^i(L_h^i)}{\partial (L_h^i)^2} = -\left(\beta_L^h + \theta_L + \gamma_L\right)\left(\frac{\partial p_h^i(L_h^i)}{\partial L_h^i}\right) + \theta_L\left(\frac{\partial p_l^i(L_h^i)}{\partial L_h^i}\right) + \frac{2A^i ln(1-\alpha)}{(L_h^i)^3} \tag{4.14}$$

$$\frac{\partial^3 \pi^i(L_h^i)}{\partial (L_h^i)^3} = -\frac{6A^i\ln(1-\alpha)}{(L_h^i)^4} \tag{4.15}$$

Let us understand the nature of the profit function $\pi^i(L_h^i)$ as we vary L_h^i. Since $L_h^i \in [0, L_l)$, we are interested in its behavior only for non-negative values of L_h. Note the similarity of (4.13), (4.14) and (4.15) to the corresponding expressions (2.13), (2.14) and (2.15) in chapter 2. Therefore, we expect the profit function $\pi(L_h^i)$ to hold similar properties.

Property 4.1. *As $L_h^i \to 0^+$, $\pi^i(L_h^i) \to -\infty$.*

Property 4.2. *$\pi^i(L_h^i)$ is increasing concave in L_h^i in the vicinity of $L_h^i = 0^+$.*

Property 4.3. *As L_h^i increases from 0, $\pi^i(L_h^i)$ changes from concave to convex for some $L_h^i \in (0, +\infty)$, and never becomes concave again.*

Using properties 4.1, 4.2 and 4.3, the nature of $\pi^i(L_h^i)$ in $[0, +\infty)$ can be summarized as shown as in Figure 4.2.

It is clear from the behavior of $\pi^i(L_h^i)$, as shown in Figure 4.2, that it has a unique maximum and at most one minimum in $[0, +\infty)$. The stationary points are given by the roots of (4.13) in $[0, +\infty)$, and the maximum is always the smaller of the two. Further, $\frac{\partial \pi^i(L_h^i)}{\partial L_h^i}\Big|_{L_h^i = L_l} < 0$ is sufficient to guarantee that (4.13) has only

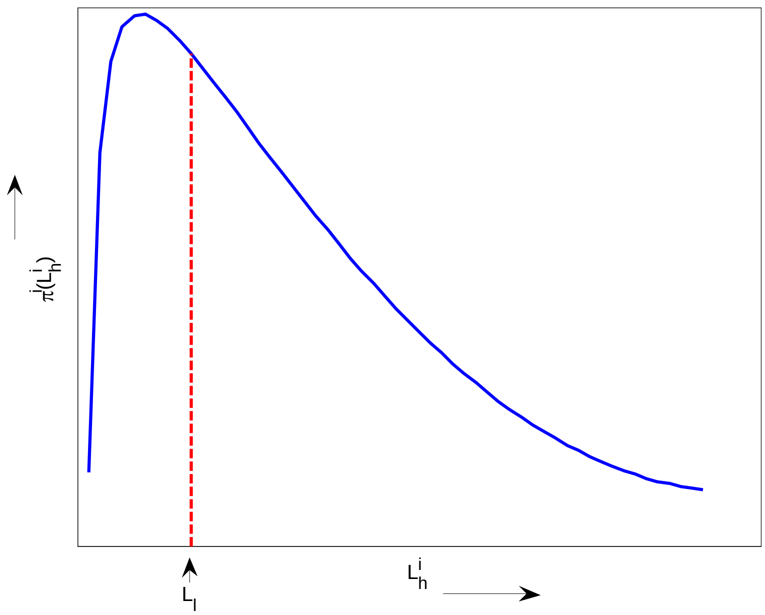


Figure 4.2: Behavior of the profit function for $L_h^i \in [0, +\infty)$

one root in the interval $[0, L_l)$, and that it is the point of maximum. The condition simplifies to:

$$\frac{K_1 a^i + K_2 L_l^i + K_3 A^i + K_4 m^i}{2(\beta_p^h \beta_p^l + \beta_p^h \theta_p + \beta_p^l \theta_p + \beta_p^h \gamma_p + \beta_p^l \gamma_p + 2\theta_p \gamma_p + \gamma_p^2)} - \frac{A^i \ln(1-\alpha)}{(L_l^i)^2} < 0 \tag{4.16}$$

where K_1, K_2, K_3, K_4 are functions only of the market parameters (β_p^k, β_L^k, θ_p, θ_L, γ_p, γ_L), and hence are constants. Further,

$$K_1 = -\left\{(\beta_p^l - \beta_p^h)\theta_L + (\beta_L^h + \gamma_L)(\beta_p^l + 2\theta_p + \gamma_p)\right\}$$

Since $\beta_p^h < \beta_p^l$ (Assumption 4.1), a necessary condition for (4.16) to hold is a^i to be high (Assumption 4.2). A sufficiently high value of a^i also guarantees $p_k^i > 0$, $p_h^i > p_l^i$ and $\lambda_k^i > 0$. □

Proposition 4.4 gives the best response express delivery time L_h^{i*} in a dedicated setting. L_h^{i*}, like L_h^* for the monopolistic market in chapter 2, does not have a closed-form analytical solution. However, it can be obtained numerically using a simple bisection method since $\pi^i(L_h^i)$ is unimodal in $[0, L_l)$. The best response prices can be obtained using proposition 4.3 by substituting $L_h^i = L_h^{i*}$.

Example 4.2: Consider again the customer specific and firm specific parameter values as shown in Table 4.2. Given firm $1's$ strategy $s^1 := (p_h^1 = 10,\ p_l^1 = 8,\ L_h^1 = 0.5)$, firm $2's$ best response response, $s^{2*} := (p_h^{2*},\ p_l^{2*},\ L_h^{2*})$ and other related variables are shown in Table 4.4.

Table 4.4: Results for Example 4.2

L_h^{2*}	p_h^{2*}	p_l^{2*}	μ_h^{2*}	μ_l^{2*}	λ_h^{2*}	λ_l^{2*}	π^{2*}
0.8452	8.8908	7.9990	10.2073	8.6474	4.7584	4.0423	29.3833

4.4.2 Shared Capacity Setting

The firm's choice of shared capacity is modelled using a single server, which serves both customer classes employing a simple fixed priority scheme that always gives priority to time-sensitive customers. As discussed in chapter 2, for a shared setting the delivery time reliability constraint (4.7) has an analytical closed-form representation, similar to that for the dedicated capacity setting. However, a closed form expression for the sojourn time distribution $S_l(\cdot)$ for low priority customers, appearing in (4.8), is not known. We assume the single server serves customers of either class at the same rate $\mu_h^i = \mu_l^i = \mu^i$. Constraints (4.6) and (4.7) in a shared capacity setting can then be expressed as:

$$\lambda_h^i + \lambda_l^i - \mu^i < 0 \tag{4.6^{SC}}$$

$$S_h^i(L_h^i) = P(W_h^i \leq L_h^i) = 1 - e^{(\lambda_h^i - \mu^i)L_h^i} \geq \alpha \tag{4.7^{SC}}$$

In absence of a closed-form analytical expression for (4.8) in a shared capacity setting, we approximate it by the following set of linear constraints, as described in §2.4.2:

$$S_l^{ik}(\cdot) + (p_h^i - p_h^{ik})\left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_h^i}\right) + (p_l^i - p_l^{ik})\left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_l^i}\right) + (\mu^i - \mu^{ik})\left(\frac{\partial S_l^{ik}(\cdot)}{\partial \mu^i}\right) \geq \alpha \quad \forall k \in K \tag{4.17}$$

where $S_l^{ik}(\cdot)$ denotes the value of $S_l^i(\cdot)$ at a fixed point $(p_h^{ik}, p_l^{ik}, \mu^{ik})$, which can be obtained numerically using the matrix geometric method described in §2.4.2. $\frac{\partial S_l^{ik}(\cdot)}{\partial p_h^i}$, $\frac{\partial S_l^{ik}(\cdot)}{\partial p_l^i}$ and $\frac{\partial S_l^{ik}(\cdot)}{\partial \mu^i}$ are the partial gradients of S_l^i at $(p_h^{ik}, p_l^{ik}, \mu^{ik})$, which can be obtaned using the finite difference method described in §2.4.2. Thus, for a given strategy $s^j := (p_h^j, p_l^j, L_h^j)$, of firm j, firm $i's$ best response prices (p_h^{i*}, p_l^{i*}) for a fixed L_h^i in a shared capacity setting are given by the solution of the following optimization problem:

$\underline{\mathbf{PDP^i_{(K)}}:}$

$$\begin{aligned}
\max_{p_h^i, p_l^i, \mu^i} \; & \pi^i = -(\beta_p^h + \theta_p + \gamma_p)(p_h^i)^2 - (\beta_p^l + \theta_p + \gamma_p)(p_l^i)^2 + 2\theta_p p_h^i p_l^i \\
& + \left\{ -\beta_L^h L_h^i + \theta_L(L_l^i - L_h^i) + \gamma_L(L_h^j - L_l^j) + \gamma_p p_h^j + m^i(\beta_p^h + \gamma_p) + a^i \right\} p_h^i \\
& + \left\{ -\beta_L^l L_l^i + \theta_L(L_h^i - L_l^i) + \gamma_L(L_l^j - L_l^i) + \gamma_p p_l^j + m^i(\beta_p^l + \gamma_p) + a \right\} p_l - A^i \mu \\
& + (\beta_L^h + \gamma_L) m^i L_h^i + (\beta_L^l + \gamma_L) L_l^i m^i - \gamma_L(L_h^j + L_l^j) m^i - \gamma_p(p_h^j + p_l^j)^i - 2m^i a^i \qquad (4.18)
\end{aligned}$$

subject to:

$$\begin{aligned}
& -(\beta_p^h + \theta_p + \gamma_p) p_h^i + \theta_p p_l^i - \mu^i \\
& \le \frac{\ln(1-\alpha)}{L_h^i} - a^i + (\beta_L^h + \theta_L + \gamma_L) L_h^i - \theta_L L_l^i - \gamma_p p_h^j - \gamma_L L_h^j \qquad (4.19)
\end{aligned}$$

$$\begin{aligned}
& \left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_h^i}\right) p_h^i + \left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_l^i}\right) p_l^i + \left(\frac{\partial S_l^{ik}(\cdot)}{\partial \mu^i}\right) \mu^i \ge \alpha - S_l^{ik}(\cdot) + \\
& \left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_h^i}\right) p_h^{ik} + \left(\frac{\partial S_l^{ik}(\cdot)}{\partial p_l^i}\right) p_l^{ik} + \left(\frac{\partial S_l^{ik}(\cdot)}{\partial \mu^i}\right) \mu^{ik} \qquad \forall k \in K \qquad (4.20)
\end{aligned}$$

$$\begin{aligned}
& -(\beta_p^h + \gamma_p) p_h^i - (\beta_p^l + \gamma_p) p_l^i - \mu^i \\
& < (\beta_L^h + \gamma_L) L_h^i + (\beta_L^l + \gamma_L) L_l^i - \gamma_p(p_h^j + p_l^j) - \gamma_L(L_h^j + L_l^j) - 2a^i \qquad (4.21)
\end{aligned}$$

$$-(\beta_p^h + \theta_p + \gamma_p) p_h^i + \theta_p p_l^i \ge (\beta_L^h + \theta_L + \gamma_L) L_h^i - \theta_L L_l^i - \gamma_p p_h^j - \gamma_L L_h^j - a^i \qquad (4.22)$$

$$\theta_p p_h^i - (\beta_p^l + \theta_p + \gamma_p) p_l^i \ge -\theta_L L_h^i + (\beta_L^l + \theta_L + \gamma_L) L_l^i - \gamma_p p_l^j - \gamma_L L_l^j - a^i \qquad (4.23)$$

$$p_h^i, p_l^i, \mu^i \ge 0 \qquad (4.24)$$

The best response prices of firm i to a given strategy of firm j can thus be obtained by solving $PDP^i_{(K)}$ using the cutting plane algorithm described in 2.4.2.

Example 4.3: Consider the same problem setting as described in Example 4.1. Assume now firm 2 uses shared capacities. Firm $2's$ best response response prices and other related variables for $L_h = 0.5$ are shown in Table 4.5.

Table 4.5: Results for Example 4.3

$p_h^{2*}(L_h)$	$p_l^{2*}(L_h)$	$\mu^{2*}(L_h)$	$\lambda_h^{2*}(L_h)$	$\lambda_l^{2*}(L_h)$	$\pi^{2*}(L_h)$
9.3545	7.8915	16.5226	4.5958	4.3545	33.9814

In chapter 2, we argued that when a firm's express delivery time is sufficiently small, its capacity requirement in a shared capacity setting is dictated solely by the demand from express customers. Using a similar argument, when L_h^i is small, we have:

$$\mu^i = \lambda_h^i - \frac{\ln(1-\alpha)}{L_h^i}$$

Substituting the expression for μ^i in the objective function, $p_h^{i*}(L_h)$ and $p_l^{i*}(L_h)$ can be obtained by solving simultaneously $\partial\pi^i/\partial p_h^i = 0$ and $\partial\pi^i/\partial p_l^i = 0$, in very much the same way as we did for the dedicated case.

Observation 4.1. *When L_h^i is small, the best response prices of firm $i \in \{1,2\}$ in a shared capacity setting to a given strategy $s^j := (p_h^j, p_l^j, L_h^j)$ by firm $j = 3-i$ are given by:*

$$\begin{aligned} p_h^{i*}(L_h^i) =& \frac{(\beta_p^l + 2\theta_p + \gamma_p)a^i - \{\beta_p^l(\theta_L + \gamma_L) + \beta_L^h(\beta_p^l + \theta_p + \gamma_p) + \theta_p\gamma_L + \gamma_L\gamma_p + \theta_L\gamma_p\}L_h^i}{2D} \\ & \frac{+\{(\beta_p^l + \gamma_p)\theta_L - (\beta_L^l + \gamma_L)\theta_p\}L_l^i + (\beta_p^l\gamma_p + \gamma_p\theta_p + \gamma_p^2)p_h^j + (\theta_p\gamma_p)p_l^j}{2D} \\ & \frac{+(\beta_p^l\gamma_L + \gamma_L\gamma_p + \theta_p\gamma_L)L_h^j + (\theta_p\gamma_L)L_l^j}{2D} + \frac{A^i + m^i}{2} \end{aligned} \tag{4.25}$$

$$\begin{aligned} p_l^{i*}(L_h^i) =& \frac{(\beta_p^h + 2\theta_p + \gamma_p)a^i - \{\beta_p^h(\theta_L + \gamma_L) + \beta_L^l(\beta_p^h + \theta_p + \gamma_p) + \theta_p\gamma_L + \gamma_L\gamma_p + \theta_L\gamma_p\}L_l^i}{2D} \\ & \frac{+\{(\beta_p^h + \gamma_p)\theta_L - (\beta_L^h + \gamma_L)\theta_p\}L_h^i + (\theta_p\gamma_p)p_h^j + (\beta_p^h\gamma_p + \gamma_p\theta_p + \gamma_p^2)p_l^j}{2D} \\ & \frac{+(\theta_p\gamma_L)L_h^j + (\beta_p^h\gamma_L + \gamma_L\gamma_p + \theta_p\gamma_L)L_l^j}{2D} + \frac{m^i}{2} \end{aligned} \tag{4.26}$$

where $D = \beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p + \beta_p^h\gamma_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2$.

When the prices are described by the above relations (4.25) and (4.26), it can be shown that a small L_h^i will, in fact, result in a relatively large express demand compared to regular demand for firm i. Using (4.25) and (4.26), we obtain:

$$
\begin{aligned}
\frac{d\lambda_h^i}{dL_h^i} &= -(\beta_p^h + \theta_p + \gamma_p)\frac{\partial p_h^i}{\partial L_h^i} + \theta_p \frac{\partial p_l^i}{\partial L_h^i} - (\beta_L^h + \theta_L + \gamma_L) \\
&= -\frac{\beta_L^h + \theta_L + \gamma_L}{2} < 0 \\
\frac{d\lambda_l^i}{dL_h^i} &= -(\beta_p^l + \theta_p + \gamma_p)\frac{\partial p_l^i}{\partial L_h^i} + \theta_p \frac{\partial p_h^i}{\partial L_h^i} + \theta_L \\
&= \frac{\theta_L}{2} > 0
\end{aligned}
$$

This suggests that when L_h^i gets sufficiently small, λ_h^i gets much larger compared to λ_l^i. Thus, the capacity requirement μ^i is dictated only by the demand from express customers, something we used to arrive at the results (4.25) and (4.26) at first place.

The best response delivery time for express customers L_h^i is given by:

$$\arg \max_{L_h^i \in [0, L_l)} f(L_h^i)$$

where $f(L_h^i)$ is a $PDP_{(K)}^i$ for a given L_h^i and given price and delivery time decisions $(L_h^j,\ p_h^j,\ p_l^j)$ of firm j. The best response L_h^i can be obtained using the golden section search method as we did for the monopolistic setting in chapter 2.

Example 4.4: The best response delivery time L_h^{2*} and the corresponding prices $(p_h^{2*},\ p_l^{2*})$ of firm 2 for the same problem described in Example 4.3 are given in Table 4.6.

Table 4.6: Results for Example 4.4

L_h^{2*}	p_h^{2*}	p_l^{2*}	μ^{2*}	λ_h^{2*}	λ_l^{2*}	π^{2*}
0.3827	9.3371	7.8661	16.8042	4.7723	4.3740	34.7228

Observation 4.2. *When the capacity cost A^i of firm i is small, its best response express delivery time L_h^{i*} in a shared capacity setting is given by the unique root of (4.27) in the interval [0, L_l).*

$$\frac{\partial \pi^i(L_h^i)}{\partial L_h^i} = -\left(\beta_L^h + \theta_L + \gamma_L\right)\left(p_h^{i*}(L_h^i) - m^i - A^i\right) + \theta_L\left(p_l^{i*}(L_h^i) - m^i\right) - \frac{A^i \ln(1-\alpha)}{(L_h^i)^2} \tag{4.27}$$

where, $p_h^{i*}(L_h)$ and $p_l^{i*}(L_h)$ are given by (4.25) and (4.26).

We know from Observation 4.1 that when L_h^i is known to be small, $p_h^{i*}(L_h^i)$ and $p_l^{i*}(L_h^i)$ can be obtained using (4.25) and (4.26). Substituting $p_h^{i*}(L_h^i)$ and $p_l^{i*}(L_h^i)$, given by (4.25) and (4.26), in the profit function π^i, and differentiating it with respect to L_h^i gives (4.27), while $\partial^2\pi^i(L_h^i)/\partial(L_h^i)^2$ and $\partial^3\pi^i(L_h^i)/\partial(L_h^i)^3$ are given by the same relations (4.14) and (4.15) as for the dedicated capacity case. Thus, the properties 4.1, 4.2 and 4.3 of π^i hold true in a shared capacity case as well. This implies that for a^i sufficiently high, π^i has a unique maximum, as shown in Figure 4.2, given by the root of (4.27).

We obtain the above result assuming that L_h^i is known to be small. We now show that when A^i is small, L_h^{i*} is indeed small, such that the above result holds true.

$$\frac{\partial L_h^{i*}}{\partial A^i} = -\left(\frac{\partial^2\pi^i/\partial L_h^i \partial A^i}{\partial^2\pi^i/\partial (L_h^i)^2}\right)\Bigg|_{L_h^i = L_h^{i*}},$$

where

$$\left.\frac{\partial^2\pi^i}{\partial L_h^i \partial A^i}\right|_{L_h^i = L_h^{i*}} = \frac{\beta_L^h + \theta_L + \gamma_L}{2} - \frac{ln(1-\alpha)}{(L_h^{i*})^2} > 0.$$

Since we know that

$$\left.\frac{\partial^2\pi^i}{\partial (L_h^i)^2}\right|_{L_h^i = L_h^{i*}} < 0 \quad \Rightarrow \quad \frac{\partial L_h^{i*}}{\partial A^i} > 0.$$

This implies that L_h^{i*} is increasing in A^i. Therefore, a small A^i guarantees that L_h^{i*} is small, which we used at first place to arrive at the result.

4.5 Duopoly Problem

We now study the price and delivery time decisions for a duopoly problem. One basic question is to investigate whether an equilibrium exists, and if so, how will the equilibrium change under different operational settings and market characteristics. To study the impact of a firm's operations strategy (dedicated versus shared capacity), we study and compare the three scenarios shown in Table 4.1. The optimization problem that each firm solves for its best response in each of these scenarios is given in Table 4.7.

Table 4.7: Mathematical models for a firm's best response in different capacity scenarios

Scenario	**Capacity Setting**	
	Firm 1	*Firm 2*
DD	$PDTDP^1_{DC}$	$PDTDP^2_{DC}$
SS	$PDTDP^1_{SC}$	$PDTDP^2_{SC}$
DS	$PDTDP^1_{DC}$	$PDTDP^2_{SC}$

Under competition, both firms simultaneously announce their price and delivery time decisions. We assume that firm $i \in \{1,2\}$ has full knowledge of the operational setting of firm $j = 3-i$, including its capacity strategy and also its parameters A, m and a. Firm i can thus correctly anticipate the best response of firm j to its own moves, and can hence strategically plan its own strategy. Equilibrium is reached when none of the firms can do better by unilaterally deviating from its decisions. A Nash equilibrium is thus a vector of strategies (s^{i*}, s^{j*}) such that for each firm i, $\pi^i(s^{i*}, s^{j*}) = \max_{s^i} \pi^i(s^i, s^{j*})$, $i \in \{1,2\}$ and $j = 3-i$. In other words, the strategy used by either firm is the best response to the strategy chosen by the other.

The equilibrium solution can be obtained by the simultaneous solution of the best responses for firms $i = 1, 2$. Proposition 4.3 gives the best response prices of a firm using dedicated capacities. Thus, when the delivery time decisions are fixed, such that firms compete purely using prices, equilibrium prices in a DD setting can be obtained in closed-form by the simultaneous solution of the 4 linear equations given by (4.10) and (4.11) (2 equations corresponding to each $i \in \{1, 2\}$). Since the equilibrium prices in a DD setting have closed-form solution, we first study the DD setting in a greater detail.

4.5.1 Competition in a DD Setting

We first study the Equilibrium results under pure price competition wherein the firms face a significantly higher stickiness for their delivery time decisions compared to their ability to vary prices. Situations in which such a model will be more relevant are discussed in chapter 2. In such situations, firms compete using only prices as their strategic variables, treating their delivery times as fixed.

Pure Price Competition

Proposition 4.5. *Pure price competition in a DD setting always results in a unique equilibrium. Further, if the firms are identical, then the equilibrium prices are symmetric, given by:*

$$\begin{aligned} p_h^*(L_h) = &\frac{(2\beta_p^l + 4\theta_p + \gamma_p)a - \{\beta_L^h(2\beta_p^l + 2\theta_p + \gamma_p) + (2\beta_p^l + \gamma_p)\theta_L\}L_h}{D_1} \\ &\frac{+\{(2\beta_p^l + \gamma_p)\theta_L - 2\beta_L^l\theta_p\}L_l}{D_1} \\ &\frac{+(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + \beta_p^h\gamma_p + 2\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(A+m)}{D_1} \end{aligned} \tag{4.28}$$

$$\begin{aligned} p_l^*(L_h) = &\frac{(2\beta_p^h + 4\theta_p + \gamma_p)a + \{(2\beta_p^h + \gamma_p)\theta_L - 2\beta_L^h\theta_p\}L_h}{D_1} \\ &\frac{-\{\beta_L^l(2\beta_p^h + 2\theta_p + \gamma_p) + (2\beta_p^h + \gamma_p)\theta_L\}L_l}{D_1} \\ &\frac{+(2\beta_p^l\beta_p^h + 2\beta_p^l\theta_p + \beta_p^l\gamma_p + 2\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(A+m)}{D_1} \end{aligned} \tag{4.29}$$

where $D_1 = 4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2.$

Proof. The equilibrium prices in a DD setting are given by the simultaneous solution of the 4 linear equations given by (4.10) and (4.11) for $i \in \{1, 2\}$. The system of equations in matrix notation is given by $\mathbf{Ax} = \mathbf{b}$.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \frac{-(\beta_p^l+\theta_p+\gamma_p)\gamma_p}{2D} & \frac{-\theta_p\gamma_p}{2D} \\ 0 & 1 & \frac{-\theta_p\gamma_p}{2D} & \frac{-(\beta_p^h+\theta_p+\gamma_p)\gamma_p}{2D} \\ \frac{-(\beta_p^l+\theta_p+\gamma_p)\gamma_p}{2D} & \frac{-\theta_p\gamma_p}{2D} & 1 & 0 \\ \frac{-\theta_p\gamma_p}{2D} & \frac{-(\beta_p^h+\theta_p+\gamma_p)\gamma_p}{2D} & 0 & 1 \end{pmatrix} \tag{4.30}$$

where $D = \beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p + \beta_p^h\gamma_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2$

$$\mathbf{x} = \begin{pmatrix} p_h^{1*} & p_l^{1*} & p_h^{2*} & p_l^{2*} \end{pmatrix}^T$$

and $\mathbf{b}$ is a 4x1 matrix of constants. $\mathbf{A}$ is symmetric and strictly diagonally dominant since we have $A_{ij} = A_{ji}\ \forall i, j$ and $\sum_{j\neq i} |A_{ij}| < A_{ii}\ \forall i$. Hence, $\mathbf{A}$ is symmetric positive definite (Horn and Johnson 1985). This implies that $\mathbf{A}$ is full-rank, and hence the system of linear equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution. This proves the uniqueness of the equilibrium.

Further, when the firms are identical, they have the same operating parameter values ($a^1 = a^2$; $m^1 = m^2$; $A^1 = A^2$; $\alpha^1 = \alpha^2$; $L_l^1 = L_l^2$; $L_h^1 = L_h^2$). Denote the equilibrium solution by the 2-tuple $(s^{1*}(L_h), s^{2*}(L_h))$, where $s^{i*}(L_h) := (p_h^{i*}(L_h), p_l^{i*}(L_h))$. Assume the contrary that the equilibrium solution is not symmetric, i.e., $s^{1*}(L_h) \neq s^{2*}(L_h)$. Since the two firms are identical, this implies that (s^{2*}, s^{1*}) must also be a Nash Equilibrium, which contradicts the uniqueness of the Nash Equilibrium. Hence, $s^{1*}(L_h) = s^{2*}(L_h)$. Substituting $p_h^{1*}(L_h) = p_h^{2*}(L_h) = p_h^*(L_h)$ and $p_l^{1*}(L_h) = p_l^{2*}(L_h) = p_l^*(L_h)$ in the expressions for the best response prices, given by (4.10) and (4.11), and solving the resulting system of 2 equations in 2 unknown gives (4.28) and (4.29). □

The corresponding price differentiation for a given L_h is then:

$$\begin{aligned} p_h^*(L_h) - p_l^*(L_h) =& \frac{2(\beta_p^l - \beta_p^h)a + 2(\beta_p^l + \beta_p^h + \gamma_p)\theta_L(L_l - L_h) + \beta_L^l(2\beta_p^h + \gamma_p)L_l}{D_1} \\ & \frac{-(2\beta_p^l + \gamma_p)\beta_L^h L_h + (\beta_p^l - \beta_p^h)\gamma_p(A + m)}{D_1} \end{aligned} \tag{4.31}$$

Example 4.5: Assume the customer specific and firm specific parameter values as shown in Table 4.8. The equilibrium prices for $L_h^1 = L_h^2 = 0.50$ in a DD scenario are shown in Table 4.9.

Table 4.8: Parameter values for Example 4.5

β_p^h	β_p^l	θ_p	β_L^h	β_L^l	θ_L	γ_p	γ_L
0.55	0.75	0.15	0.9	0.7	0.5	0.4	0.4
a^1	m^1	A^1	a^2	m^2	A^2	α	L_l
10	3	0.01	10	3	0.01	0.99	1

Table 4.9: Results for Example 4.5

Firm 1		**Firm 2**	
p_h^{1*}	p_l^{1*}	p_h^{2*}	p_l^{2*}
8.166499	6.800659	8.166499	6.800659

Price and Delivery Time Competition

We now study the equilibrium solution under both price and delivery time competition. The equlibrium express delivery times in a DD setting are given by the simultaneous solution of the system of 2 non-linear equations, given by (4.13) = 0 for $i = 1, 2$. In absence of a closed-form solution for this system of non-linear equations, we design an iterative procedure, described in Figure 4.3, that always converges to the equilibrium solution. We solve for an equilibrium solution assuming the game is played dynamically, starting at an initial solution, until none of the firms has an incentive to deviate from its decision unilaterally.

Proposition 4.6. *The iterative algorithm given in Figure 4.3 converges to a unique Nash Equilibrium in a DD setting.*

Proof. The proof of convergence of the equilibrium delivery times is based on an important result that for a given set of parameter values, each firm monotonically increases or decreases its express delivery time in response to an increase in the correspoding express delivery time by its competitor. If the two firms play the game iteratively, any one of the following will happen:

- Both L_h^1 and L_h^2 increase monotonically or both L_h^1 and L_h^2 decrease monotonically
- L_h^1 increases monotonically and L_h^2 decreases monotonically or vice-versa

Since the express delivery times are bounded above and below ($L_h^i \in [0, L_l^i)$), they will converge ultimately. The equilibrium prices are then given by the unique solution to the system of linear equations given by the best response prices of the two firms at the equilibrium express delivery times. Details of the proof are given in Appendix B.1. □

1. *Initialization*: For each firm i, set $p_h^i = p_l^i = m^i$, $L_h^i = 0$ or $L_h^i = L_l$.
2. *Iterative step*: Start with $i = 1$. Use the best response obtained for Firm i problem. Repeat this for $i = 2$.
3. *Convergence criteria*: Repeat step 2 until each firm's decision values differ from their previous values by less than some predetermined tolerance level ϵ.

Figure 4.3: Iterative Algorithm for Nash Equilibrium

4.5.2 Competition in an SS or a DS Setting

When one of the competing firms uses shared capacities, we do not have a closed-form analytical characterization of its best response prices and delivery time. In such a situation, equlibrium prices and delivery times are obtained using the iterative procedure, described in Figure 4.3. Our extensive numerical experiments confirm its convergence to a unique Nash Equilibrium in all operational settings. Further, in a pure price competition, the equlibrium prices can be obtained using the same iterative procedure by fixing L_h^i.

Observation 4.3. *The iterative procedure given in Figure 4.3 always converges to a unique Nash Equilibrium under all capacity settings.*

Example 4.6: Assume the customer specific and firm specific parameter values as shown in Table 4.8. The equilibrium result for each of the three operational settings: (i) DD, (ii) SS and (iii) DS is shown in Table 4.10

Table 4.10: Results for Example 4.6

	DD		**SS**		**DS**	
	Firm 1	*Firm 2*	*Firm 1*	*Firm 2*	*Firm 1*	*Firm 2*
L_h^*	0.075916	0.075916	0.075927	0.075927	0.075915	0.075928
p_h^*	8.487575	8.487575	8.487409	8.487409	8.48746	8.487527
p_l^*	6.748059	6.748059	6.742129	6.742129	6.747257	6.742933

In the above example, we obtain a unique Nash Equilibrium for each of the operational settings in the above example. Further, the equilibrium is symmetric when both the firms are identical (i.e., have the same firm-specific parameters: a^i, A^i, m^i, and use the same capacity strategy).

Observation 4.4. *If the firms are identical, then pure price competition in an SS setting results in symmetric equilibrium prices, which for small L_h are given by:*

$$p_h^*(L_h) = \frac{(2\beta_p^l + 4\theta_p + \gamma_p)a - \{\beta_L^h(2\beta_p^l + 2\theta_p + \gamma_p) + (2\beta_p^l + \gamma_p)\theta_L\}L_h}{D_1}$$
$$\frac{+\{(2\beta_p^l + \gamma_p)\theta_L - 2\beta_L^l\theta_p\}L_l}{D_1}$$
$$\frac{+(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + \beta_p^h\gamma_p + 2\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)m}{D_1}$$
$$\frac{+(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + \beta_p^h\gamma_p + 2\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 3\theta_p\gamma_p + \gamma_p^2)A}{D_1} \tag{4.32}$$

$$p_l^*(L_h) = \frac{(2\beta_p^h + 4\theta_p + \gamma_p)a + \{(2\beta_p^h + \gamma_p)\theta_L - 2\beta_L^h\theta_p\}L_h}{D_1}$$
$$\frac{-\{\beta_L^l(2\beta_p^h + 2\theta_p + \gamma_p) + (2\beta_p^h + \gamma_p)\theta_L\}L_l}{D_1}$$
$$\frac{+(2\beta_p^l\beta_p^h + 2\beta_p^l\theta_p + \beta_p^l\gamma_p + 2\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)m}{D_1} + \frac{\theta_p\gamma_p A}{D_1} \tag{4.33}$$

where $D_1 = 4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2$.

A mathematical justification for the above observation is given in Appendix B.2. The corresponding price differentiation is given by:

$$p_h^*(L_h) - p_l^*(L_h) = \frac{2(\beta_p^l - \beta_p^h)a + 2(\beta_p^l + \beta_p^h + \gamma_p)\theta_L(L_l - L_h) + \beta_L^l(2\beta_p^h + \gamma_p)L_l}{D_1}$$
$$\frac{-(2\beta_p^l + \gamma_p)\beta_L^h L_h + (\beta_p^l - \beta_p^h)\gamma_p m}{D_1}$$
$$\frac{+(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + \beta_p^h\gamma_p + 2\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 2\theta_p\gamma_p + \gamma_p^2)A}{D_1} \tag{4.34}$$

Example 4.7: Assume the same parameter values as described in Table 4.8. Assume that the firms' delivery time decisions are sticky such that they compete only in their prices. Further, assume that both the firms have a small express delivery time, $L_h^1 = L_h^2 = L_h = 0.10$. The equilibrium prices in an SS scenario are shown in Table 4.11.

4.6 Conclusions

In this chapter, we extended our models developed in chapter 2 for optimal product differentiation strategy to a competitive setting. Our primary objective was to understand how the capacity strategies used by competing firms affect their price and

Table 4.11: Results for Example 4.7

Firm 1		Firm 2	
p_h^{1*}	p_l^{1*}	p_h^{2*}	p_l^{2*}
8.469186	6.745116	8.469186	6.745116

delivery time differentiation decisions, and if the qualitative results of a monopolistic setting also extend to a competitive setting. For this we developed a general mathematical model, special cases of which capture three different scenarios, depending on the capacity strategy used by either firm. We extended the solution methods developed for the monopolistic setting to obtain the best response of a firm when competing with another firm. We finally designed an iterative algorithm to obtain the price and delivery time decisions at equilibrium of competing firms.

In the following chapter, we study the models developed in this chapter to understand how equilibrium decisions of competing firms are shaped by the capacity strategy they choose. We also use the results of the competitive setting to investigate how competition, per se, affects the product differentiation of a firm.

Chapter 5

Competitive Market: Analysis & Insights

In chapter 4, we extended the models developed in chapter 2 to a competitive framework, and adapted them for the different scenarios depending on the capacity strategy used by the competing firms. In this chapter, we study in detail the different models developed in chapter 4. Based on our numerical study, we draw important insights into how the capacity strategies of competing firms influence their product differentiation strategies. In §5.1, we describe the experimental setup for our numerical study of various scenarios. §5.2 presents some general observations on the best response of a firm. Comparisons of equilibrium results under pure price competition and more general price and delivery time competition are presented in §5.3 and §5.4, respectively. We conclude the chapter with a summary of main results and directions for future research in §5.5.

5.1 Numerical Experiment Design

Our model setting described in chapter 4 involves the following parameters: a^i, m^i, A^i, α and L_l. Of these, we fix the value of $L_l = 1$ and $\alpha = 0.99$. As regards the other parameters, we experiment with a large combination of their values as given in Table 5.1:

Table 5.1: Parameter settings for numerical experiments

Parameter	**Number of Choices**	**Possible Values**
a^i	6	{10, 15, 20, 25, 50, 100}
m^i	6	{1, 2, 3, 4, 5, 6}
A^i	2X4	{0.01, 0.025, 0.05, 0.1} (small A)
		{0.25, 0.50, 0.75, 1.0} (large A)

Of these $(6X6X2X4)^2$ possible choices, we select those combinations for our experiments that satisfy the condition for unique L_h^{i*}, as given in the proof of proposition 4.4. For the market parameters, we use the following combinations:

- *Time Difference Sensitive* (TDS): $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$, $\theta_p = 0.2$, $\theta_L = 0.5$, $(\gamma_p, \gamma_L) \in \{(0,0),(0,0.4),(0.4,0),(0.4,0.4),(0.1,0.95),(0.95,0.1)\}$, such that $\theta_L/(\beta_L^k + \gamma_L) > \theta_p/(\beta_p^k + \gamma_p)$, $k \in \{h, l\}$.
- *Price Difference Sensitive* (PDS): $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$, $\theta_p = 0.4$, $\theta_L = 0.1$, $(\gamma_p, \gamma_L) \in \{(0,0),(0,0.4),(0.4,0),(0.4,0.4),(0.1,0.95),(0.95,0.1)\}$, such that $\theta_p/(\beta_p^k + \gamma_p) > \theta_L/(\beta_L^k + \gamma_L)$, $k \in \{h, l\}$.

For illustration, we use the parameter setting shown in Tables 5.2 and 5.3, unless stated otherwise.

Table 5.2: Market parameters used in illustrative examples

Market Type ↓	β_p^h	β_p^l	θ_p	β_L^h	β_L^l	θ_L	γ_p	γ_L
TDS	0.55	0.75	0.15	0.9	0.7	0.5	0.4	0.4
PDS	0.5	0.7	0.4	0.9	0.7	0.1	0.4	0.4

Table 5.3: Firm-specific parameters used in illustrative examples

Firm 1					**Firm 2**				
a^1	A^1	m^1	α^1	L_l^1	a^2	A^2	m^2	α^2	L_l^2
10	0.25	3	0.99	1	10	0.25	3	0.99	1

We first make some general observations on the best response of firm i to a given strategy of firm j under different capacity settings.

5.2 Best Response of a Firm

Observation 5.1. *Given a strategy $s^j := (p_h^j, p_l^j, L_h^j)$ of firm $j = 3 - i$, a decrease in the express delivery time L_h^i by firm $i \in \{1, 2\}$ results in: (a) an increase in p_h^{i*} (b) a decrease in p_l^{i*} if $\theta_L/(\beta_L^h + \gamma_L) > \theta_p/(\beta_p^h + \gamma_p)$; and an increase in p_l^{i*} if $\theta_L/(\beta_L^h + \gamma_L) < \theta_p/(\beta_p^h + \gamma_p)$.*

For the best response of a firm using dedicated capacities, the above observation follows directly from Proposition 4.3. In a shared capacity setting, the above observation follows from Observation 4.1 for small L_h^i. Observation 5.1 suggests that given the decisions of the other firm, a firm's best response express price always decreases with an increase in its express delivery time. The effect of any change in its express delivery time on the regular price, however, depends on whether the market is TDS ($\theta_L/(\beta_L^h + \gamma_L) > \theta_p/(\beta_p^h + \gamma_p)$)or PDS ($\theta_L/(\beta_L^h + \gamma_L) < \theta_p/(\beta_p^h + \gamma_p)$).

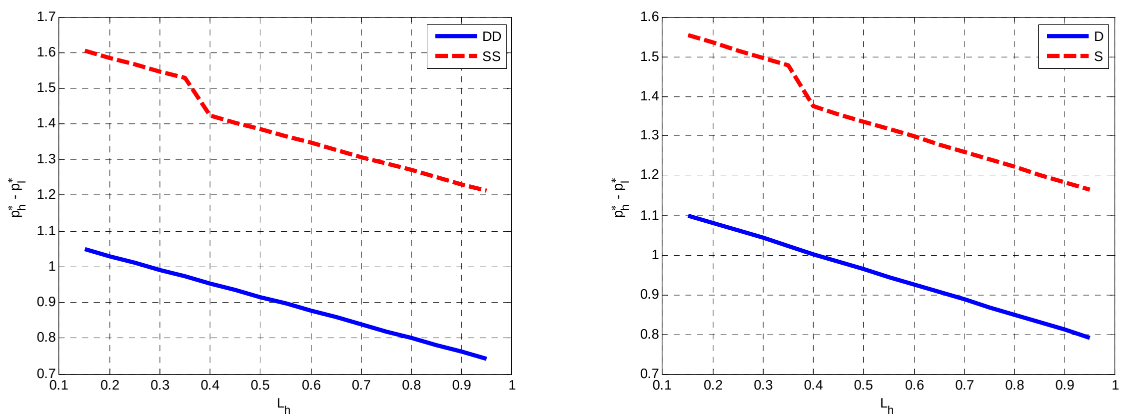
In the remainder of this chapter, we compare the equilibrium results of competing firms under different capacity settings. We first study the Equilibrium results under pure price competition wherein the firms face a significantly higher stickiness for their delivery time decisions compared to their ability to vary prices.

5.3 Pure Price Competition

5.3.1 Effect of Capacity Strategy

Observation 5.2. - *Pure price competition under SS, compared to DD, results in: (a) a larger price differentiation at equilibrium.*

- Pure price competition under DS results in: (a) a larger price differentiation at equilibrium for the firm using shared capacities compared to the other firm using dedicated capacities.



(a) price differentiation of symmetric firms in DD versus SS

(b) price differentiation of firms in DS

Figure 5.1: Price differentiation in different capacity settings under pure price competition

The above observation, for small L_h, follows directly by comparing (4.31) with (4.34). This is shown in Figure 5.1(a) for a PDS type market. Although the behavior of the prices of each firm may change in a different market type, the above observation still holds true, irrespective of the market behavior. Further, comparing (4.28) with (4.32) and (4.29) with (4.33) suggests that when two symmetric firms compete using shared capacities, their equilibrium prices for both the market segments are smaller than when they both compete using dedicated capacities. This suggests that all customers are better off when the competing firms use shared capacities. Note that when the two firms competing in DD or SS setting are identical, their prices at equilibrium are symmetric so that the differentiation at equilibrium of both the firms coincide. Our numerical results suggest that when the two firms compete in a DS setting, their price decisions at equilibrium are not symmetric even if the firms are otherwise identical. The asymmetry in their equilibrium prices results from the asymmetry in the capacity strategies. The firm using shared capacities always maintains a higher price differentiation compared to the other firm that uses dedicated capacities as shown in Figure 5.1(b).

5.3.2 Effect of Price Competition

We have so far analyzed our models in a competitive setting to study the effects of firms' capacity strategies on their price differentiation strategies. We now study the effect of price competition on a firm's price and delivery time decisions in a given capacity setting. We know competition generally drives prices down. But how does competition affect price differentiation? To answer this, we compare the optimal prices of a monopolist with its equilibrium prices when it faces price competition from an identical firm. A monopolist setting can be represented using the mathematical model of chapter 2 for a single firm with a market base $2a$. Alternatively, it can be represented using a competitive model of chapter 4 for two identical firms, each with a market base $2a$, but with $\gamma_p = \gamma_L = 0$. The later case represents two identical firms operating in geographically different markets such that they do not poach each other's market share. From the firms' point of view, there is no difference between the two scenarios as they both result in the same monopolist prices. In contrast, a competitive setting represents a situation in which two firms operate in the same geographical market, each with a market base $2a$, such that each firm's demand is affected by the relative prices of the two firms. Mathematically, this corresponds to $\gamma_p > 0$, $\gamma_L > 0$ in our competitive model.

The way competition affects price differentiation may also be influenced by the operations (capacity) strategy of the competing firms. However, in absence of an analytical characterization of optimal prices when one of the firms uses shared capacities, we restrict our study only to a setting where the firms use dedicated capacities. The effect of price competition is summarized in the following proposition.

Proposition 5.1. *Pure price competition in a dedicated capacity setting always results in: (a) a lower express price p_h^*, (a) a lower regular price p_l^*, and (c) a lower price differentiation ($p_h^* - p_l^*$). Further, the effects are more pronounced in presence of product substitution.*

Proof. See Appendix C.1 □

The effect of competition on individual prices is not surprising. In fact, it is well established in theory that competition always decreases prices (Varian 1989). This is observed in practice as well as highlighted by various real-world examples in §4.1. However, researchers seem to be divided in their understanding of the effect of competition on price differentiation. Our model, with an important linkage between marketing decision of price discrimination and operation's capacity related decisions, provides results that concur with the traditional theory on price discrimination that predicts that market competition decreases a firms ability to use price discrimination. Our results are also consistent with the findings of Gerardi and Shapiro (2007), which suggest that price discrimination has decreased with increase in competition in the airline industry. Our result, however, is in contrast with empirical results of (Borenstein and Rose (1994), which conclude otherwise for the airline industry. Further, our results suggest that the effects of competition on individual prices as well as price discrimination are more pronounced in presence of product substitution. This suggests that the degree of price discrimination (second degree in presence of product substitution, and third degree in absence of product substitution) further plays a role in deciding the intensity of the effect of price competition.

5.4 Price and Delivery Time Competition

We now consider a more general situation where firms have flexibility in quoting the delivery times to their express customers. We still assume there is a standard delivery time for regular customers established by the industry.[1] In such a situation, firms compete by strategically selecting both the express delivery time and the two prices. We first study the effect of firms' capacity strategies on product differentiation at equilibrium, and then we study the effect of competition on the firms' decisions. Finally, we study the effect of asymmetry, in terms of firms' operating parameters, on their price and delivery time decisions.

5.4.1 Effect of Capacity Strategy

Observation 5.3. - *Price and delivery time competition under SS, compared to DD, results in: (a) a larger price differentiation at equilibrium, and (b) a larger delivery time at equilibrium if capacity cost is high, but a smaller delivery time differentiation when capacity cost is small.*
- Price and delivery time competition under DS results in: (a) a larger price differentiation at equilibrium for the firm using shared capacities, and (b) a larger delivery time differentiation at equilibrium for the firm using shared capacities if capacity cost is high, but a smaller delivery time differentiation for the firm using shared capacities when capacity cost is small.

The above observation is an extension of the results obtained in chapter 3 to a competitive setting. This observation suggests that the capacity strategies of firms have the same influence on their product differentiation decisions in a competitive market as they have in a monopolist setting. The first part of the observation (for DD versus SS) can be shown analytically for the case when the capacity cost A is

[1] In a more general situation, firms may optimally select delivery times for both the customer segments. However, our assumption of a standard delivery time for regular customers is made mainly for the tractability of the model.

Table 5.4: Numerical Results for DD and SS settings

	A = 0.01				**A = 0.25**			
	PDS		*TDS*		*PDS*		*TDS*	
	DD	*SS*	*DD*	*SS*	*DD*	*SS*	*DD*	*SS*
L_h^*	0.079835	0.07984	0.075916	0.075927	0.40887	0.375944	0.393785	0.382392
$L_l - L_h^*$	0.920165	0.92016	0.924084	0.924073	0.59113	0.624056	0.606215	0.617608
p_h^*	8.48642	8.4861	8.487575	8.487409	8.475324	8.496719	8.397922	8.407862
p_l^*	7.4262	7.42033	6.748059	6.742129	7.536979	7.410724	6.933531	6.790338
$p_h^* - p_l^*$	1.06022	1.06577	1.739516	1.74528	0.938345	1.085995	1.464391	1.617524

Table 5.5: Numerical Results for DS setting

	A = 0.01				**A = 0.25**			
	PDS		*TDS*		*PDS*		*TDS*	
	D	*S*	*D*	*S*	*D*	*S*	*D*	*S*
L_h^*	0.079836	0.079838	0.075915	0.075928	0.409076	0.375427	0.393641	0.381862
$L_l - L_h^*$	0.920164	0.920162	0.924085	0.924072	0.590924	0.624573	0.606359	0.618138
p_h^*	8.486179	8.486345	8.48746	8.487527	8.469203	8.502327	8.395587	8.410847
p_l^*	7.425453	7.42108	6.747257	6.742933	7.520666	7.426818	6.914207	6.809853
$p_h - p_l^*$	1.060726	1.065265	1.740203	1.744595	0.948537	1.075509	1.481381	1.600994

small (see Appendix C.2). We illustrate this using numerical results obtained for the parameter setting described in Tables 5.2 and 5.3 for two levels of capacity cost: (i) $A = 0.01$ (for small capacity cost) and (ii) $A = 0.25$ (for large capacity cost). A comparison of the equilibrium prices and delivery times in an SS versus a DD setting is shown in Table 5.4, and for a DS setting is shown in Table 5.5.

5.4.2 Effect of Price and Delivery Time Competition

When firms use delivery time, in addition to price, as a strategic tool to attract demand and compete in the market, this leads to another question of interest: how does competition affect both price and delivery time differentiation? To answer this, we compare the equilibrium prices and delivery time decisions in a competitive setting with that under a monopolistic setting, discussed in chapter 2. Although the effect may depend on the capacity strategy used by the firms, we restrict our study to the case where both firms use dedicated capacities since that leads to some analytical results.

The effect of competition on price and delivery time differentiation, in general, depends on the relative intensities of price competition (γ_p) and delivery time competition (γ_L), as well as other demand parameters. The following proposition summarizes the effect of competition for the following special cases: (i) $\gamma_p = 0$, $\gamma_L > 0$ (ii) $\gamma_p > 0$, $\gamma_L = 0$.

Proposition 5.2. *Price and delivery time competition in a dedicated capacity setting:*
- decreases both delivery time differentiation and price differentiation when $\gamma_L = 0$.
- increases both delivery time differentiation and price differentiation when $\gamma_p = 0$.

Proof. See Appendix C.3 □

The above proposition suggests that price and delivery time competition may increase or decrease price and delivery time differentiation, depending on customers' behavior. This is intuitive. When $\gamma_L = 0$, customers' choice of a firm is not influenced by the relative delivery times but by the relative prices offered by the two firms. In such a situation, firms tend to cut prices to attract customers. At the same time, they increase their express delivery time, and hence decrease their delivery time differentiation, in order to cut their capacity cost and maintain their profit. It further follows from (4.31) that a smaller delivery time differentiation in a DD setting also results in a smaller price differentiation. On the other hand, when $\gamma_p = 0$, customers' choice of a firm is not influenced by the relative prices but by the relative delivery times offered by the two firms. In such a situation, firms try to cut their delivery times to attract customers. This results in a smaller express delivery time, and hence a larger delivery time differentiation. Again, it follows from (4.31) that a larger delivery time differentiation also allows the firms to maintain a larger price differentiation.

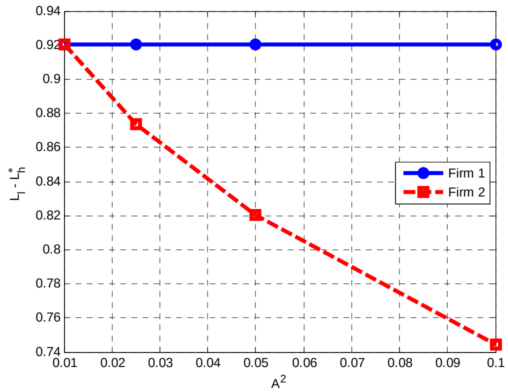
5.4.3 Effect of Asymmetry Between Firms

We have thus far studied firms that are symmetric with respect to their market base a, capacity cost A and operating cost m, although they may use different capacity strategies. For symmetric firms, we have studied the effects of firms' capacity strategies on their product differentiation strategies. However, competing firms, in reality, may be asymmetric with respect to one or more of these parameters. In such a scenario, a firm will try to exploit its competitive advantage of a lower capacity cost A, or a higher market base a due to its better brand appeal. We study the effects of such asymmetry on the equilibrium decisions of the competing firms. We study such asymmetric competition in both DD as well as SS settings to see if, and how, these effects vary with the capacity settings.

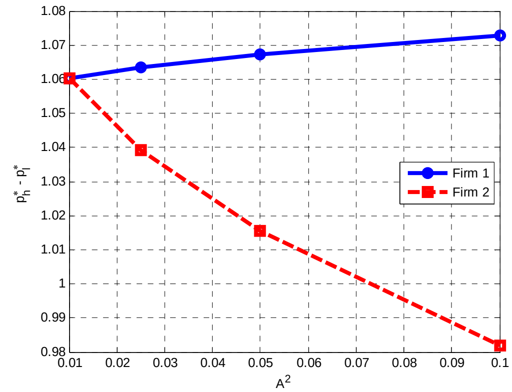
Asymmetry in Capacity Cost

Observation 5.4. *If one of the firms, which are otherwise identical, has a higher capacity cost, then compared to the other firm at equilibrium:*
- in a DD setting, it has (a) a smaller delivery time differentiation, and (b) a smaller price differentiation (Refer to Figure 5.2).
- in an SS seting, it has (a) a smaller delivery time differentiation, and (b) a smaller price differentiation if the status-quo capacity cost is small, but a larger price differentiation if the status-quo capacity cost is high (Refer to 5.3).

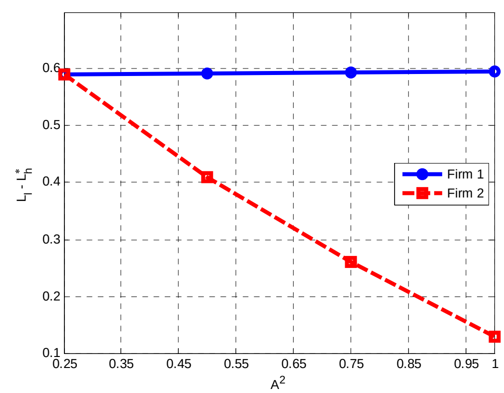
Figure 5.2 shows the equilibrium price and delivery time differentiations of the two firms in a DD setting that differ in their capacity costs but are otherwise identical. Figure 5.3 shows similar plots for an SS setting. We show these plots for a PDS type market (parameter values shown in Table 5.2), although the qualitative results are independent of the specific market parameters. Firm-specific parameters are as shown in Table 5.3. In one set of experiments, we fix the capacity cost of firm 1, A^1, at 0.01 and vary that for firm 2, A^2, from 0.01 to 0.10. In another set of experiments, we fix A^1 at 0.25 and vary A^2 from 0.25 to 1.0. This helps



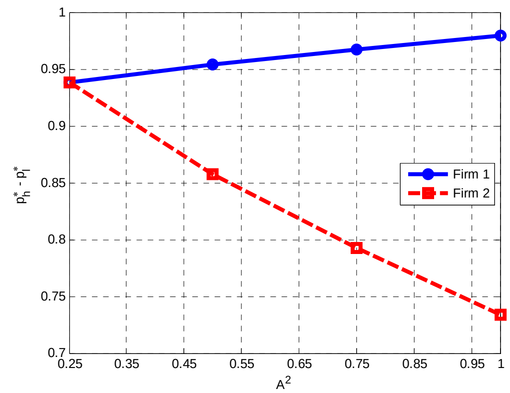
(a) Delivery time differentiation versus marginal capacity cost when capacity cost is low



(b) Price differentiation versus marginal capacity cost when capacity cost is low



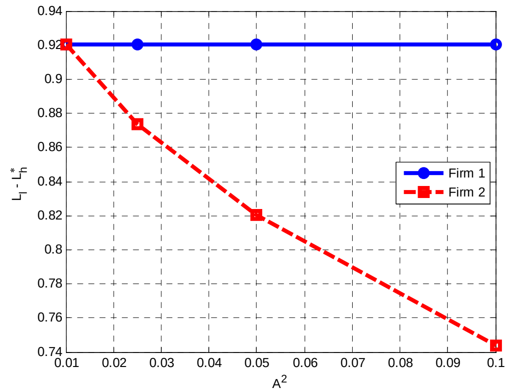
(c) Delivery time differentiation versus marginal capacity cost when capacity cost is high



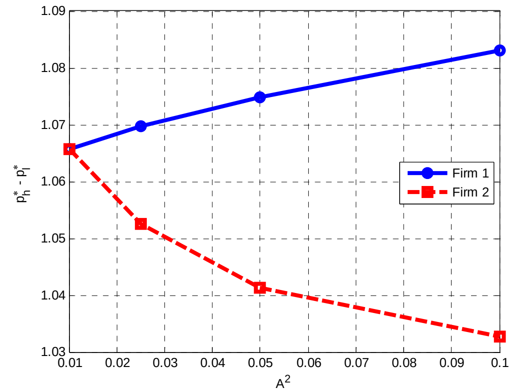
(d) Price differentiation versus marginal capacity cost when capacity cost is high

Figure 5.2: Effects of capacity cost asymmetry on product differentiation decisions in a DD setting

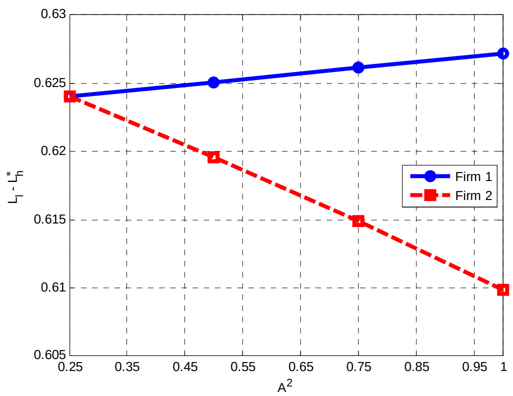
us capture the effect of a larger capacity cost incurred by firm 2 on the decisions of the two firms at equilibrium. As evident from the plots, when the firms are symmetric ($A^2 = A^1$), the delivery time and price differentiations of both firms coincide. Any increase in firm $2's$ capacity cost (A^2) always decreases its delivery time differentiation at equilibrium, irrespective of the capacity settings used by the two firms. An increase in A^2 also decreases firm $2's$ price differentiation at equilibrium in a DD setting. In an SS setting, an increase in A^2 decreases firm $2's$ price differentiation only for small capacity costs; for larger capacity costs, it increases its price differentiation. Thus, our results for the monopolistic setting,



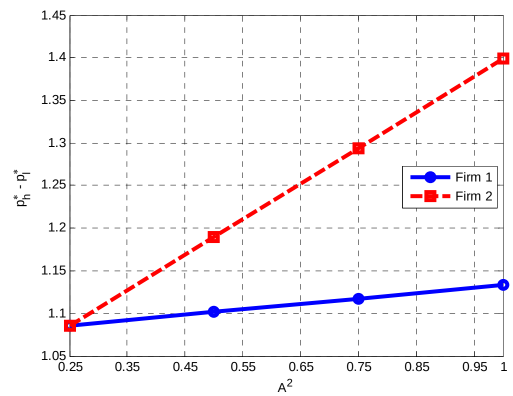
(a) Delivery time differentiation versus marginal capacity cost when capacity cost is low



(b) Price differentiation versus marginal capacity cost when capacity cost is low



(c) Delivery time differentiation versus marginal capacity cost when capacity cost is high



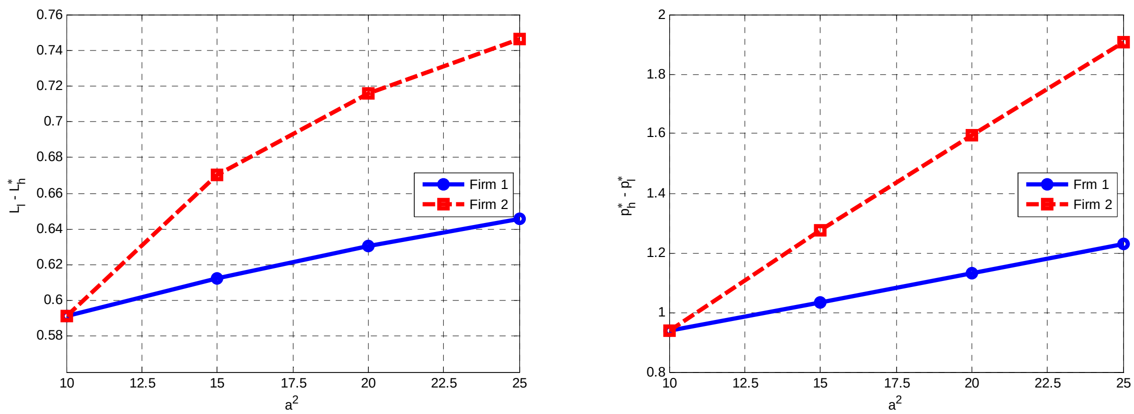
(d) Price differentiation versus marginal capacity cost when capacity cost is high

Figure 5.3: Effects of capacity cost asymmetry on product differentiation decisions in an SS setting

described in chapter 3, also extend to a competitive setting. However, the effect of an increase in firm $2's$ capacity cost may have a similar or contrasting effect on firm 1, depending on the market parameters and the level of the capacity cost. Whatever be the effects on individual firms, when $A^2 > A^1$, firm 2 always has a smaller delivery time differentiation, irrespective of the capacity settings. It also has a smaller price differentiation in a DD setting, but a higher price differentiation for larger capacity costs in an SS setting.

Asymmetry in Market Base

Observation 5.5. *If one of the firms, which are otherwise identical, has a larger market base, then compared to the other firm at equilibrium:*
- it always has (a) a larger delivery time differentiation, and (b) a larger price differentiation, irrespective of the capacity strategy of either firm (Refer to Figures 5.4 and 5.5).

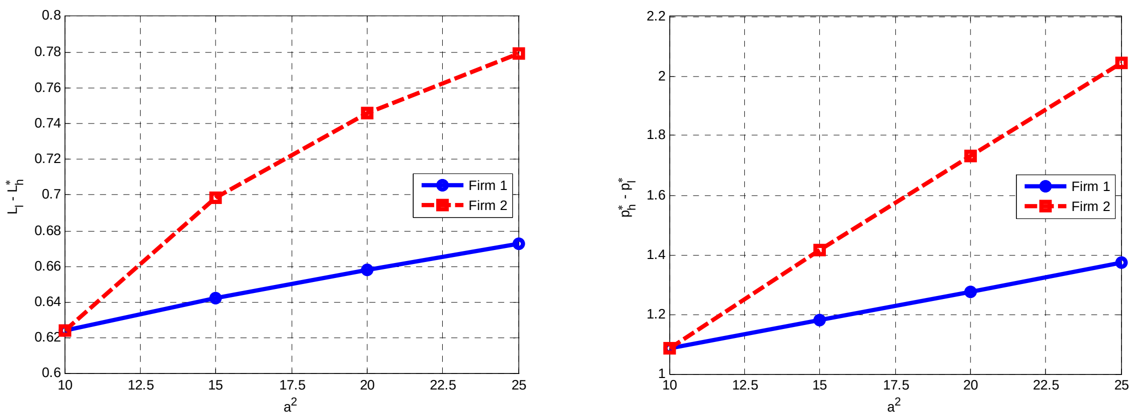


(a) Delivery time differentiation versus market base

(b) Price differentiation versus market base

Figure 5.4: Effects of asymmetry in market base on product differentiation decisions in a DD setting

We illustrate this result using a sample from our numerical experiments. We consider two firms that have different market bases ($a^1 \neq a^2$), but are otherwise identical. Difference in the market bases of the two firms means that one firm always has a higher mean demand even if they both offer the same delivery times at the same prices. This may be the result of a difference in their brand appeal to the customers or due to a more convenient locations or a better customer experience at one of the firms. We assume the market is PDS type (parameter values shown in Table 5.2), although the generalizations drawn are independent of the specific market parameters. Firm specific parameters are as shown in Table 5.3. The market base a^1 for firm 1 is now fixed at 10, while that for firm 2 (a^2) is varied. Figures 5.4 and 5.5 show the equilibrium price and delivery time differentiations



(a) Delivery time differentiation versus market base

(b) Price differentiation versus market base

Figure 5.5: Effects of asymmetry in market base on product differentiation decisions in an SS setting

of the two firms in a DD and an SS setting, respectively. This helps us capture the effect of a larger market base of firm 2 on the decisions of the two firms at equilibrium. As evident from the plots, when the firms are symmetric ($a^2 = a^1$), the delivery time and price differentiations of both firms coincide. Any increase in firm $2's$ market base (a^2) increases its delivery time differentiation as well as the price differentiation at equilibrium, irrespective of the capacity settings used by the two firms. Although firm $1's$ price and delivery time differentiation decisions also increase with a^2 in this case, this is specific only to this set of market parameters. In general, the behavior of firm $1's$ decisions depends on the market parameters. Whatever be the effects on individual firms, when $a^2 > a^1$, firm 2 always has a larger delivery time differentiation and a larger price differentiation, irrespective of the capacity settings and market parameters.

5.5 Conclusions & Future Research

In this chapter, we extensively studied the different mathematical models developed in chapter 4, which helped us generate important managerial insights. Our analytical/numerical study of the models clearly shows that the qualitative results of a monopolistic setting regarding the effect of firms' operations strategy on their price and delivery time differentiation decisions also hold true in a competitive environment. Specifically, when processing capacities are expensive, the firm with shared capacities offers faster and more expensive product to time sensitive customers and slower and cheaper product to price sensitive customers compared to the firm using dedicated capacities. This implies that the firm with shared capacities offers products with greater differentiation. Further, the above effect of the capacity strategy does not depend on any end customer characteristics or whether the products are substitutable or not.

We also demonstrated that when firms are asymmetric with respect to their capacity related costs or their market bases, each firm tries to use its distinctive advantage to uniquely differentiate its products. The way a firm exploits its distinctive advantage of lower capacity cost further depends on its own capacity strategy and also of its competitor. Specifically, the firm with cheaper capacities makes its products more differentiated if both firms use dedicated capacities. If both firms use shared capacities, then the firm with cheaper capacities again makes its delivery times more differentiated, but may offer more homogeneous or more differentiated prices depending on the level of capacity cost. Whereas the firm with a larger market base always offers more differentiated products, irrespective of the capacity strategy of either firm.

Our study provides further insight into the effect of competition on price discrimination. We showed that when firms use dedicated capacities, pure price competition always reduces individual prices as well as price discrimination. However, when firms use delivery times, in addition to prices, as strategic variables to compete in the market, the effect of competition on product differentiation further depends on customers' behavior.

The above results are managerially quite relevant. First of all, they show how managers should anticipate the outcome of competition, in terms of product differentiation, given its own capacity strategy as well as the capacity strategy of its competitor. It also enables managers to anticipate how its competitor will use its distinctive advantage to differentiate its products, and hence how to respond optimally.

The competitive framework studied in chapters 4 and 5 can also be extended along the same directions discussed in chapter 3. As discussed in chapter 3, one possible extension would be to develop a good approximation for the sojourn time distribution $S_l(\cdot)$ of the low priority customers in a shared capacity setting, which will allow for a closed-form solution to the best response problem. This will also allow for a proof of convergence and uniqueness of the Nash Equilibrium when one of the firms uses shared capacities. Further, the mathematical model for the best response in a shared capacity setting can be extended to include delay dependent dynamic priority discipline. Another possible extension may be to include the guaranteed delivery time for regular customers also as a decision variable.

Chapter 6

Summary & Future Research

6.1 Summary

In this dissertation, we focused on firms that exploit heterogeneity in customers' preferences by offering a menu of products/services that differ only in their prices and guaranteed delivery times. For such firms, we looked at two very different operations strategies (sharing service/production capacities among different market segments versus using dedicated facilities for each segment), and the way each affects their marketing decision of differentiating their products for different customer segments. We also looked at how a firm's optimal product differentiation is further affected by the degree of discrimination (second degree discrimination versus third degree discrimination) it employs. From demand perspective, when customers are allowed to self-select from the menu (second degree discrimination), different products act as substitutes, affecting each other's demand. Customized product for each segment (third degree discrimination), on the other hand, results in independent demand for each product. We tried to understand the interaction between a firm's operations strategy (dedicated versus shared capacities) and its marketing strategy (second versus third degree discrimination), and how they shape the optimal product differentiation decisions.

In chapters 2 and 3, we studied the above issues for a single profit maximizing monopolist firm, which offers two versions of the same basic product: (i) regular product at a lower price but a longer delivery time, and (ii) express product with a shorter delivery time but for a higher price. Demand was assumed to be uncertain, modelled using a Poisson process. We modelled the price and time sensitivity of customers using a mean of the Poisson demand as a deterministic function of its price and delivery time, and used different price and time sensitivities for the two customer segments. Further, we modelled product substitution by making the mean demand for each product to also depend on the price and delivery time of the other product. We modelled these dependencies of the mean demand using linear functions for analytical tractability. However, we saw linear demand models exhibit some important desirable properties, which are not exhibited by other more complicated functions. We modelled the dedicated capacity strategy of a firm using separate servers catering to each market segment. The shared capacity strategy, on the other hand, was modelled using a single server, which serves both the market segments, giving preemptive priority to customers for its express product. We looked at how to optimally price the two products and select their guaranteed delivery times so as to maximize the firm's profit rate. For analytical tractability, we assumed that the delivery time for the regular product is already established. The firm simultaneously needs to decide its optimal processing capacity, represented by its exponential processing rate, so as to meet its promised delivery times with a predetermined minimum level of reliability.

Different combinations of operations strategy and marketing strategy of the firm resulted in four possible scenarios: (i) Non-substitutable products; dedicated capacity (ii) Non-substitutable products; shared capacity (iii) Substitutable products; dedicated capacity (iv) Substitutable products; shared capacity. Comparison of the four scenarios allowed us to study the interactions between a firm's operations and marketing strategies on optimal product differentiation. On the technical side, the problem in a shared capacity setting became very challenging, especially in the absence of an analytical characterization of sojourn time distribution of regular customers in a priority queue. We resolved this difficulty by developing a solution algorithm, using matrix geometric method in a cutting plane framework, to solve the problem numerically. Numerical solution of the problem in a shared capacity setting prevented analytical comparisons of the different scenarios, and we derived most of our insights from a numerical study, giving intuitive explanations and mathematical justification wherever possible.

Our study showed that in a highly capacitated system, the firm should offer products with greater differentiation if they use shared capacities compared to if they use dedicated capacities, whether the products are substitutable or not. In contrast, when customers are allowed to self-select, such that independent products become substitutable, a more homogeneous pricing scheme results. However, the effect of substitution on optimal delivery time differentiation depends on the firm's capacity strategy and cost, as well as market characteristics. The optimal response to any change in capacity cost also depends on the firm's operations strategy. In a dedicated capacity scenario, the optimal response to an increase in capacity cost is always to offer more homogeneous prices and delivery times. In a shared capacity setting, it is again optimal to quote more homogeneous delivery times, but increase or decrease the price differentiation depending on whether the status-quo capacity cost is high or low, respectively. We also demonstrated that the above results are corroborated by real-life practices, and provided a number of managerial implications in terms of dealing with issues like volatile fuel prices.

In chapters 4 and 5, we extended our analysis to a competitive setting with two firms, each of which may either share its processing capacities for the two products, or may dedicate capacity for each product. The demand faced by each firm for a given product now also depends on the price and delivery time quoted for the same product by the other firm. We first studied the best response of each firm, given the price and delivery time decisions of the other firm. We developed a solution algorithm, which always converges to a unique solution, to determine the decisions of the competing firms at equilibrium. We used the solution algorithm to study the equilibrium decisions in three different scenarios (i) both firms use dedicated capacities (DD); (ii) both firms use shared capacities (SS); (iii) one firm uses dedicated while the other uses shared capacities (DS). Comparing the solutions of different scenarios, we derived generalizations on the effects of capacity strategy on the product differentiation decisions of competing firms.

From our study, we observed that the qualitative results of a monopolistic setting also extend to a competitive setting. Specifically, in a highly capacitated system, the equilibrium prices and delivery times are such that they result in products with greater differentiation when both the firms use shared capacities as compared to the scenario when both the firms use dedicated capacities. When the competing firms are asymmetric, they exploit their distinctive characteristics to differentiate their products. Further, the effects of these asymmetries also depend on the capacity strategy used by the competing firms. Our numerical results suggested that the firm with expensive capacity always offers more homogeneous delivery times. However, its decision on how to differentiate its prices depends on the capacity setting of the two firms as well as the actual level of their capacity costs. On the other hand, the firm with a larger market base always offers more differentiated prices as well as delivery times, irrespective of the capacity setting of the competing firms. Comparing the equilibrium solutions of our competitive setting with the optimal solution of the monopolistic setting, we observed that competition may increase or decrease product differentiation depending on the market structure.

The insights generated from our study are managerially quite relevant. First of all, they show how capacitated firms should alter their product differentiation strategy when they make changes in their market coverage of product offerings and/or capacity strategy. We also showed that managers need to pay close attention to two other factors - capacity cost of the business environment they are operating in and the behavior of their end customers - both of which play crucial role in many circumstances. Our analysis regarding the effects of any change in capacity cost is especially relevant keeping in mind the volatility of fuel price, which directly impacts capacity cost in a number of sectors. We demonstrated how managers should optimally respond to these changing business environments in order to gain competitive advantage. Our insights from the competitive model demonstrate how managers should anticipate the outcome of competition, in terms of product differentiation, given its own capacity strategy as well as the capacity strategy of its competitor. It also enables managers to anticipate how its competitor will use its distinctive advantage to differentiate its products, and hence how to respond optimally.

6.2 Directions for Future Research

The summary presented in the previous section shows our understanding of the effects of operations strategy and product substitution on the product differentiation strategy of a firm in a service or make-to-order (MTO) industry. The models developed in this dissertation have potential for being extended and further evolved. We have already discussed in chapters 3 and 5 the directions along which the current models can be extended. The models developed in this dissertation can further be extended for Make-to-Stock (MTS) and Assemble-to-Order (ATO) manufacturing environment. A key feature distinguishing MTS and ATO from our models for MTO is the issue of inventory management. Models for MTS/ATO will thus require integrating inventory decisions in the modelling framework. While MTS produces to stock end-products, ATO pools the inventories of different products by producing components that can be quickly assembled into different end-products. It will be worthwhile to study how inventory pooling in ATO affects the product differentiation decision of a firm.

The models studied in this thesis include static decisions. We see pricing and lead time setting for segmented markets in a dynamic setting as another possible extension. Further, there is also a scope for empirical research to understand how firms actually manage their capacities to serve different market segments, and also to estimate the different demand parameters.

APPENDICES

Appendix A

Mathematical Appendices for Chapter 3

A.1 Explanations for Observation 3.5

For small A:

$$\left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{SC} - \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{DC} = \frac{\theta_L A}{2} \geq 0 \tag{A.1}$$

Absence of product substitution ($\theta_p = \theta_L = 0$) implies (A.1) = 0. This suggests that sharing capacity, when it is relatively inexpensive, has no effect on the optimal express delivery time, and hence on delivery time differentiation, in absence of product substitution. Presence of product substitution ($\theta_p > 0$, $\theta_L > 0$), on the other hand, implies (A.1) > 0. Further, $\pi(L_h)$ is increasing concave in L_h for $L_h \leq L_h^{DC*}$. Similarly, $\pi(L_h)$ is increasing concave in L_h for $L_h \leq L_h^{SC*}$. This, together with (A.1) > 0, implies that:

$$L_h^{SC*} := \{L_h^{SC} : \partial\pi/\partial L_h^{SC} = 0\} > L_h^{DC*} := \{L_h^{DC} : \partial\pi/\partial L_h^{DC} = 0\} \; for \; \theta_L > 0$$

This implies that when A is small, sharing capacity in presence of product substitution increases optimal L_h, and hence decreases delivery time differentiation. This, together with Proposition 2.3 and Observation 2.1, explains the effect of capacity sharing on optimal p_h.

The effect of product substitution in a dedicated capacity setting follows from the following expression:

$$\left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{\theta_p,\theta_L>0} - \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{\theta_p,\theta_L=0} = \frac{-\beta_p^l(\beta_p^l-\beta_p^h)(\beta_p^h\theta_L-\beta_L^h\theta_p)a}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)}$$
$$+\frac{\{\beta_p^h\beta_p^l(\beta_p^h+\beta_p^l)(\theta_L)^2+\beta_p^h(\beta_p^l)^2\beta_L^h\theta_L+(\beta_p^l)^2\beta_L^h(\beta_p^h\theta_L-\beta_L^h\theta_p)\}L_h}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)}$$
$$-\frac{\{\beta_L^h(\beta_p^l\theta_L-\beta_L^l\theta_p)+(\beta_p^h+\beta_p^l)(\theta_L)^2+\beta_p^h\beta_L^l\theta_L\}L_l}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)} \tag{A.2}$$

A high value of a (Assumption 2.2) makes (A.2) negative (resp., positive) if $\beta_p^h\theta_L - \beta_L^h\theta_p > 0$ (resp., < 0). Also, the profit function is increasing concave in L_h. Therefore, (A.2) < 0 (resp., > 0) implies that optimal L_h^* decreases (resp., increases) with substitution. This implies that product substitution decreases (resp., increases) $L_h^* := \{L_h : \partial\pi/\partial L_h = 0\}$, and hence increases (resp., decreases) the delivery time differentiation for a TDS (resp., PDS) type market. The effect of product substitution in a shared capacity setting for small A can be similarly explained.

A.2 Explanations for Observation 3.6

For a dedicated capacity setting, from Implicit Function Theorem, we get:

$$\frac{\partial L_h^*}{\partial A} = -\left.\left(\frac{\partial^2\pi/\partial L_h\partial A}{\partial^2\pi/\partial L_h^2}\right)\right|_{L_h=L_h^*}, \quad \text{where} \quad \left.\frac{\partial^2\pi}{\partial L_h\partial A}\right|_{L_h=L_h^*} = \frac{\beta_L^h}{2} - \frac{ln(1-\alpha)}{(L_h^*)^2} > 0$$

Since we know that

$$\left.\frac{\partial^2\pi}{\partial L_h^2}\right|_{L_h=L_h^*} < 0 \quad \Rightarrow \quad \frac{\partial L_h^*}{\partial A} > 0$$

This implies that the optimal delivery time differentiation decreases with an increase in the marginal capacity cost A.

The effect of an increase in A on price differentiation in DC follows from:

$$\frac{d(p_h^*-p_l^*)}{dA} = \frac{\partial(p_h^*-p_l^*)}{\partial A} + \frac{\partial(p_h^*-p_l^*)}{\partial L_h^*}\frac{\partial(L_h^*)}{\partial A}$$
$$= -\frac{\beta_p^l\beta_L^h+\beta_p^h\theta_L+\beta_p^l\theta_L}{2(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)}\frac{\partial(L_h^*)}{\partial A} < 0$$

Appendix B

Mathematical Appendices for Chapter 4

B.1 Proof of Proposition 4.6

Proof. Given the strategy of firm $j \in \{1, 2\}$, the best response express delivery time of firm $i = 3 - j$ satisfies:

$$\frac{\partial \pi^i}{\partial L_h^i} = 0$$

Taking the total derivative of the above relation with respect to the express delivery time L_h^j of firm j, we get:

$$\frac{d}{dL_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right) = \frac{\partial}{\partial L_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right) + \frac{\partial}{\partial p_h^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{\partial p_h^j}{\partial L_h^j} + \frac{\partial}{\partial p_l^j}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{\partial p_l^j}{\partial L_h^j} + \frac{\partial}{\partial L_h^i}\left(\frac{\partial \pi^i}{\partial L_h^i}\right)\frac{dL_h^i}{dL_h^j} = 0$$

$$\Rightarrow \frac{dL_h^i}{dL_h^j} = \frac{-\left[\frac{\partial^2 \pi^i}{\partial L_h^j \partial L_h^i} + \frac{\partial^2 \pi^i}{\partial p_h^j \partial L_h^i}\frac{\partial p_h^j}{\partial L_h^j} + \frac{\partial^2 \pi^i}{\partial p_l^j \partial L_h^i}\frac{\partial p_l^j}{\partial L_h^j}\right]}{\frac{\partial^2 \pi^i}{\partial (L_h^i)^2}}$$

For a DD setting, the above relation simplifies to:

$$\frac{dL_h^i}{dL_h^j} = \frac{-\left[\gamma_p\left\{\left(\frac{\partial p_h^j}{\partial L_h^j}\right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j}\right)^2\right\} + \gamma_L\left(\frac{\partial p_h^j}{\partial L_h^j}\right)\right]}{\frac{\partial^2 \pi^i}{\partial (L_h^i)^2}} \tag{B.1}$$

We know that for $L_h \leq L_h^*$:

$$\frac{\partial^2 \pi^i}{\partial (L_h^i)^2} < 0$$

The numerator in RHS of (B.1) consists of terms that are functions only of the market parameters, and hence is a constant for a given parameter setting. Further,

$$\gamma_p \left\{ \left(\frac{\partial p_h^j}{\partial L_h^j} \right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j} \right)^2 \right\} > 0 \, and \, \gamma_L \left(\frac{\partial p_h^j}{\partial L_h^j} \right) < 0$$

Therefore, we have:

$$\frac{dL_h^i}{dL_h^j} \geq 0 \, if \, \gamma_p \left\{ \left(\frac{\partial p_h^j}{\partial L_h^j} \right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j} \right)^2 \right\} \geq \gamma_L \left(\frac{\partial p_h^j}{\partial L_h^j} \right) \tag{B.2}$$

$$\frac{dL_h^i}{dL_h^j} < 0 \, if \, \gamma_p \left\{ \left(\frac{\partial p_h^j}{\partial L_h^j} \right)^2 + \left(\frac{\partial p_l^j}{\partial L_h^j} \right)^2 \right\} < \gamma_L \left(\frac{\partial p_h^j}{\partial L_h^j} \right) \tag{B.3}$$

This suggests that if the market parameters are such that (B.2) holds, firm i always increases (decreases) its express delivery time L_h^i in response to a corresponding increase (decrease) in firm $j's$ express delivery time L_h^j. We let $p_h^i(n)$, $p_l^i(n)$ and $L_h^i(n)$ be the best response decisions of firm i at the n^{th} iteration of the procedure. If $L_h^i(0) = 0$, then $L_h^i(n) \geq L_h^i(0)$ for all n. We will show that if (B.2) holds, $L_h^i(n)$ is increasing in n for $i \in \{1, 2\}$. As L_h^i is bounded above ($L_h^i < L_l$), for $i \in \{1, 2\}$, this will establish that the iterative procedure converges. We prove the convergence by induction as follows:

1. (Step $n = 1$): We know that $L_h^i(1) \geq L_h^i(0)$ for $i \in \{1, 2\}$.

2. (Step $n - 1$): Assume that $L_h^i(n-1) \geq L_h^i(n-2)$ for $i \in \{1, 2\}$.

3. (Step n): Given the inductive assumption from Step $n-1$, (B.2) implies that $L_h^i(n) \geq L_h^i(n-1)$ for $i \in \{1, 2\}$.

This completes our induction. In case (B.3) holds, convergence of the algorithm can proved similarly by letting $L_h^1(0) = L_l$ and $L_h^2(0) = 0$ and by showing that $L_h^1(n)$ is decreasing in n while $L_h^2(n)$ is increasing in n.

We show the uniqueness of the Nash Equilibrium by contradiction. For any given (L_h^1, L_h^2), the equilibrium prices are uniquely determined by the simultaneous solution of the 4 linear equations given by (4.10) and (4.11) for $i \in \{1, 2\}$. Therefore, an Equilibrium solution is completely specified by the pair of express delivery times (L_h^1, L_h^2). Suppose there exist two different equilibrium solutions $\Phi = (L_h^1, L_h^2)$ and $\Phi' = (L_h^{1'}, L_h^{2'})$. By numbering the firms and the two solutions appropriately, we can assume that $L_h^{1'} > L_h^1$, which results in $L_h^{2'} > L_h^2$ if (B.2) holds, else $L_h^{2'} < L_h^2$ if (B.3) holds. We will show that in either condition, such two equilibrium solutions cannot both satisfy the optimality equation (4.13) $= 0$ for $i \in \{1, 2\}$.

For any given (L_h^1, L_h^2), the equilibrium prices, obtained by the simultaneous solution of the 4 linear equations given by (4.10) and (4.11) for $i \in \{1, 2\}$, can be expressed as:

$$p_h^1 = K_h^1 + K_h^{11} L_h^1 + K_h^{12} L_h^2$$
$$p_h^2 = K_h^2 + K_h^{12} L_h^1 + K_h^{11} L_h^2$$
$$p_l^1 = K_l^1 + K_l^{11} L_h^1 + K_l^{12} L_h^2$$
$$p_l^2 = K_l^2 + K_l^{12} L_h^1 + K_l^{11} L_h^2$$

where K_h^1, K_h^2, K_h^{11}, K_h^{12}, K_l^1, K_l^2, K_l^{11}, K_l^{12} are functions of market and firm specific parameters, and hence are constants for a given problem setting. Substituting the equilibrium prices in the optimality equation (4.13) for $i \in \{1, 2\}$, we get:

$$f^1(L_h^1, L_h^2) = K_1 + K_{11} L_h^1 + K_{12} L_h^2 - A^1 \ln(1-\alpha)/(L_h^1)^2 = 0 \qquad \text{(B.4)}$$
$$f^2(L_h^1, L_h^2) = K_2 + K_{12} L_h^1 + K_{11} L_h^2 - A^2 \ln(1-\alpha)/(L_h^2)^2 = 0 \qquad \text{(B.5)}$$

For $(L_h^{1'}, L_h^{2'})$ to also be an equilibrium, it should hold that $f^1(L_h^{1'}, L_h^{2'}) = f^2(L_h^{1'}, L_h^{2'})$ $= 0$. It should also hold that $f^{1\Delta} = f^1(L_h^1, L_h^2) - f^1(L_h^{1'}, L_h^{2'}) = 0$ and $f^{2\Delta} = f^2(L_h^1, L_h^2) - f^2(L_h^{1'}, L_h^{2'}) = 0$, and therefore, $f^{\Delta} = f^{1\Delta} + f^{2\Delta} = 0$. However, given (B.2) holds and $\partial^2\pi^i/\partial(L_h^i)^2 < 0$, $\partial^2\pi^i/\partial(L_h^{i'})^2 < 0$, $f^{\Delta} \neq 0$ for $L_h^{1'} > L_h^1$ and $L_h^{2'} > L_h^2$. Similarly, given (B.3) holds and $\partial^2\pi^i/\partial(L_h^i)^2 < 0$, $\partial^2\pi^i/\partial(L_h^{i'})^2 < 0$, $f^{\Delta} \neq 0$ for $L_h^{1'} > L_h^1$ and $L_h^{2'} < L_h^2$. Thus, we conclude that there is a unique equilibrium solution in a DD setting. □

B.2 Explanation for Observation 4.4

When both the firms use shared capacities and both have small express delivery times, the equilibrium prices can be obtained from the simultaneous solution of the 4 linear equations given by (4.25) and (4.26) for $i \in \{1, 2\}$. The system of equations in matrix notation is given by $\mathbf{Ax} = \mathbf{b}$, where:

$$\mathbf{x} = \begin{pmatrix} p_h^{1*} & p_l^{1*} & p_h^{2*} & p_l^{2*} \end{pmatrix}^T$$

and $\mathbf{A}$ is the same matrix as that for the DD setting, given by (4.30). The only difference between the system of equations between a DD and an SS setting is in the 4x1 matrix of constants, $\mathbf{b}$. We have already shown (see proof of Proposition 4.5) that $\mathbf{A}$ is a full-rank matrix, and hence the system of linear equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution. Proof for symmetry of the equilibrium solution is the same as for the DD setting. Hence, $s^{1*}(L_h) = s^{2*}(L_h)$. Substituting $p_h^{1*}(L_h) = p_h^{2*}(L_h) = p_h^*(L_h)$ and $p_l^{1*}(L_h) = p_l^{2*}(L_h) = p_l^*(L_h)$ in the expressions for the best response prices, given by (4.25) and (4.26), and solving the resulting system of 2 equations in 2 unknown gives (4.32) and (4.33).

Appendix C

Mathematical Appendices for Chapter 5

C.1 Proof of Proposition 5.1

Comparing the monopolist prices, given by Proposition 2.3 with the equilibrium prices, given by Proposition 4.3, we get:

$$p_h^{DD*}(L_h)\Big|_{duopoly} - p_h^{DC*}(L_h)\Big|_{monopoly}$$
$$= \frac{-\gamma_p\left\{K_1^h a + K_2^h L_h + K_3^h L_l + K_4^h(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{C.1}$$

$$p_l^{DD*}(L_h)\Big|_{duopoly} - p_l^{DC*}(L_h)\Big|_{monopoly}$$
$$= \frac{-\gamma_p\left\{K_1^l a + K_2^l L_h + K_3^l L_l + K_4^l(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{C.2}$$

$$\left(p_h^{DD*}(L_h) - p_l^{DD*}(L_h)\right)\Big|_{duopoly} - \left(p_h^{DC*}(L_h) - p_l^{DC*}(L_h)\right)\Big|_{monopoly}$$
$$= \frac{-\gamma_p\left\{K_1^d a + K_2^d L_h + K_3^d L_l + K_4^d(A+m)\right\}}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \tag{C.3}$$

where, K_1^h, K_2^h, K_3^h, K_4^h, K_1^l, K_2^l, K_3^l, K_4^l, K_1^d, K_2^d, K_3^d, K_4^d are some functions only of the system parameters, and hence are constants. Clearly, when $\gamma_p = 0$,

$p_h^{DD*}(L_h)\Big|_{duopoly} = p_h^{DC*}(L_h)\Big|_{monopoly}$ and $p_l^{DD*}(L_h)\Big|_{duopoly} = p_l^{DC*}(L_h)\Big|_{monopoly}$. For, $\gamma_p > 0$, (C.1), (C.2) and (C.3) are dictated mainly by K_1^h and K_1^l and K_1^d, respectively since a is assumed to be large (see Assumptions 2.2 and 4.2). Further,

$$K_1^h = 2(\beta_p^l)^2 + 2\beta_p^h\theta_p + 6\beta_p^l\theta_p + 8\theta_p^2 + \beta_p^l\gamma_p + 2\theta_p\gamma_p > 0$$

$$K_1^l = 2(\beta_p^h)^2 + 6\beta_p^h\theta_p + 2\beta_p^l\theta_p + 8\theta_p^2 + \beta_p^h\gamma_p + 2\theta_p\gamma_p > 0$$

$$K_1^d = (\beta_p^l - \beta_p^h)\gamma_p + 2\{(\beta_p^l)^2 - (\beta_p^h)^2\} + 4(\beta_p^l - \beta_p^h)\theta_p > 0$$

Therefore, $K_1^h > 0$, $K_1^l > 0$ and $K_1^d > 0 \Rightarrow$ (C.1) < 0, (C.2) < 0 and (C.3) < 0, respectively if $\gamma_p > 0$. This shows that pure price competition decreases both the express and regular prices as well as the price differentiation.

C.2 Explanations for Observation 5.3

$$\frac{\partial\pi(L_h)}{\partial L_h}\Big|_{SS} - \frac{\partial\pi(L_h)}{\partial L_h}\Big|_{DD}$$
$$= \frac{\left\{(2\beta_p^h\beta_p^l + 2\beta_p^h\theta_p + 2\beta_p^l\theta_p + \beta_p^l\gamma_p + 2\theta_p\gamma_p)\theta_L + (\beta_L^h + \gamma_L)\theta_p\gamma_p\right\} A_h}{4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 2\beta_p^h\gamma_p + 4\beta_p^l\theta_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2} \geq 0 \quad \text{(C.4)}$$

Further, for $L_h \leq L_h^{DC*}$ in DC, profit function is increasing concave in L_h. Also, for $L_h \leq L_h^{SC*}$ in SC, profit function is increasing concave in L_h. This, together with (C.4), shows that $L_h^{SC*} := \{L_h^{SC} : \partial\pi/\partial L_h^{SC} = 0\} \geq L_h^{DC*} := \{L_h^{DC} : \partial\pi/\partial L_h^{DC} = 0\}$. Thus, SS setting results in a larger L_h^* and hence a smaller delivery time differentiation if the products are substitutable (i.e., $\theta_p > 0$ and $\theta_L > 0$).

Further, comparison of equilibrium prices (comparison of (4.28) with (4.32) and (4.29) with (4.33)) suggests that both express and regular prices are smaller under SS setting compared to DD setting for a given L_h. Whereas, comparison of (4.31) with (4.34) shows that the price differentiation increases in an SS setting for a given L_h. A larger L_h in SS setting compared to DD setting partly offsets the difference in the price differentiation in the two settings, but all our numerical results suggest that the price differentiation is still higher in SS setting.

C.3 Proof of Proposition 5.2

The effect of competition on the express delivery time when firms use dedicated capacities is given by:

$$\left.\frac{\partial\pi(L_h)}{\partial L_h}\right|_{duopoly} - \left.\frac{\partial\pi(L_h)}{\partial L_h}\right|_{monopoly}$$
$$= \frac{-\{K_1 a + K_2 L_h + K_3 L_l + K_4(A+m)\}}{2(4\beta_p^h\beta_p^l + 4\beta_p^h\theta_p + 4\beta_p^l\theta_p + 2\beta_p^h\gamma_p + 2\beta_p^l\gamma_p + 4\theta_p\gamma_p + \gamma_p^2)(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{C.5}$$

where, K_1, K_2, K_3 and K_4 are some functions only of the system parameters, and hence are constants. For large a, (Assumptions 2.2 and 4.2), (C.5) is dictated mainly by K_1, which is given by:

$$\begin{aligned}
K_1 = &\left\{4\beta_p^h(\beta_p^l)^2 + 4(\beta_p^l)^2\theta_p + 8\beta_p^h\theta_p^2 + 8\beta_p^l\theta_p^2 + 12\beta_p^h\beta_p^l\theta_p\right\}\gamma_L \\
&- \left\{\beta_L^h\beta_p^l + 2\beta_L^h\theta_p + [(\beta_p^l)^2 - (\beta_p^h)^2]\theta_L\right\}\gamma_p^2 \\
&- \left\{2\beta_p^h\beta_L^h\theta_p + 2(\beta_p^l)^2\beta_L^h + 4\beta_p^l\theta_L + 8\beta_L^h\theta_p^2\right\}\gamma_p \\
&- \left\{[6\beta_p^l\beta_L^h - 4\beta_p^h\theta_L]\theta_p + 2[(\beta_p^l)^2 - (\beta_p^h)^2]\theta_L\right\}\gamma_p \\
&+2\left\{\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p\right\}\gamma_L\gamma_p
\end{aligned} \tag{C.6}$$

Clearly, the effect of competition on L_h, and hence on delivery time differentiation, depends on the relative intensities of price competition (γ_p) and delivery time competition (γ_L), as well as other demand parameters. $\gamma_p = 0$ and $\gamma_L > 0$ results in C.6 > 0, and hence C.5 < 0. Thus, L_h is smaller under competition when $\gamma_p = 0$. Further, (C.1), (C.2) and (C.3) suggest that for a given L_h, the equilibrium prices as well as the price differentiation under DD coincide with the monopolist prices and price differentiation under DC for $\gamma_p = 0$. However, a smaller L_h under DD compared to DC results in a larger price differentiation.

$\gamma_p > 0$ and $\gamma_L = 0$, on the other hand, results in C.6 < 0, and hence C.5 > 0. Thus, L_h is larger under competition. A larger L_h results in a smaller price differentiation.

References

[1] J. Abate and W. Whitt. Low-priority waiting-time tail probabilities. *Queueing Systems*, 25:173–233, 1997.

[2] P. Afeche. Incentive compatible revenue managemet in queueing systems: optimal strategic delay and other delaying tactics. Working paper, Kellog School of Management, Northwestern University, Evasnton, IL, USA, 2004.

[3] P. Afeche and H. Mendelson. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science*, 50(7):869–882, 2004.

[4] G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 53(1):37–55, 2007.

[5] G. Allon and A. Federgruen. Competition in service industries with segmented markets. Working paper, Graduate School of Business, Columbia University, NY, USA, 2008.

[6] T. Andel. From common to custom: The case for make-to-order. *Material Handling Management*, November 2000.

[7] S. P. Anderson, A. de Palma, and J. F. Thisse. *Discrete Choice Theory and Product Differentiation.* MIT Press, Cambridge, MA, USA, 1992.

[8] S. Andradottir. A review of simulation optimization techniques. In J. S. Carson D. J. Medeiros, E. F. Watson and M. S. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 151–158, 1998.

[9] H. Armony and M. Haviv. Price and delay competition between two service providers. *European Journal of Operational Research*, 147(1):32–50, 2003.

[10] B. Ata and J. A. Van Mieghem. The value of dynamic resource pooling: Should a service network be integrated or product-focused? Working paper, Kellog School of Management, Northwestern University, IL, USA, 2008.

[11] J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127(1–4):333–358, 2004.

[12] R. H. Ballou. *Business Logistics Management.* Prentice Hall, Upper Saddle River, NJ, USA, 1998.

[13] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithm.* John Wiley & Sons Publication, NY, USA, 1993.

[14] F. Bernstein and A. Federgruen. Pricing and replenishment strategies in a distribution system with competing retailers. *Operations Research*, 51(3):409–426, 2003.

[15] F. Bernstein and A. Federgruen. A general equilibrium model for industries with price and service competition. *Operations Research*, 52(6):868–886, 2004.

[16] O. Besbes and A. Zeevi. Implications of choice model parameter estimation in a competitive environment. Working paper, Graduate School of Business, Columbia University, NY, USA, 2005.

[17] J. D. Blackburn. *Time-Based Competition: The Next Battleground in American Manufacturing.* Business One Irwin, Homewood, IL, USA, 1991.

[18] A. S. Blinder, E. Canetti, D. Lebow, and J. Rudd. *Asking About Prices: A New Approach to Understanding Price Stickiness.* Russell Sage Foundation Publications, NY, USA, 1998.

[19] S. Borenstein and N. Rose. Competition and price dispersion in the U.S. airline industry. *The Journal of Political Economy*, 102(4):653–683, 1994.

[20] J. Borland and E. Hansen. DVD price wars: How long can they go? CNET News, URL `http://news.cnet.com/2100-1026_3-5439388.html`, November 05 2004.

[21] T. Boyaci and S. Ray. Product differentiation and capacity selection cost interaction in time and price sensitive markets. *Manufacturing and Service Operations Management*, 5(1):18–36, 2003.

[22] T. Boyaci and S. Ray. The impact of capacity costs on product differentiation in delivery time, delivery reliability and price. *Production and Operations Management*, 15(2):179–197, 2006.

[23] R. L. Burden and J. D. Faires. *Numerical Analysis*. Brooks Cole, CA, USA, 2000.

[24] M. Busse and M. Rysman. Competition and price discrimination in yellow pages advertising. *The RAND Journal of Economics*, 36(2):378–390, 2005.

[25] G. P. Cachon and P. T. Harker. Competition and outsourcing with scale economies. *Management Science*, 48(10):1314–1333, 2002.

[26] Y. Carson and A. Maria. Simulation optimization: Methods and applications. In D. H. Withers S. Andradottir, K. J. Healy and B. L. Nelson, editors, *Proceedings of the 1997 Winter Simulation Conference*, pages 118–126, 1997.

[27] C. Charney. *Time to Market: Reducing Product Lead Time*. Society of Manufacturing Engineers, Dearborn MI, USA, 1991.

[28] H. Chen and Y. Wan. Price competition of make-to-order firms. *IIE Transactions*, 35:817–832, 2003.

[29] S. K. Clerides. Price discrimination with differentiated products: Definition and identification. *Economic Enquiry*, 42(3):402–412, July 2004.

[30] J. W. Cohen. *The Single Server Queue.* North-Holland Publishing Company, The Netherlands, 1982.

[31] L. Cooper. *Applied Nonlinear Programming.* Aloray Publisher, Englewood, NJ, USA, 1974.

[32] Y. Dai. *Game theoretic approach to supply chain management.* PhD thesis, Industrial Engineering, North Carolina State University, Raleigh, North Carolina, USA, 2003.

[33] S. Dewan and H. Mendelson. User delay costs and internal pricing for a service facility. *Management Science*, 36(12):1502–1507, 1990.

[34] A. Farahat and G. Perakis. Price competition among multiproduct firms. Working paper, Johnson Graduate School of Management, Cornell University, Ithaca, NY, USA, 2008.

[35] M. T. Farris. The rise of premium transportation through technology. Working paper, Center for logistics education and research, University of North Texas, TX, USA, 2002.

[36] J. B. Fuller, J. O'Conor, and R. Rawlinson. Tailored logistics: The next advantage. *Harvard Business Review*, 71(3):87–98, 1993.

[37] I. Gale. Price dispersion in a market with advanced-purchases. *Review of Industrial Organization*, 8:451–464, 1993.

[38] K. S. Gerardi and A. H. Shapiro. Does competition reduce price discrimination? new evidence from the airline industry. Working paper No. 07-7, Federal Reserve Bank of Boston, 2007.

[39] A. Y. Ha, L. Li, and S. Ng. Price and delivery logistics competition in a supply chain. *Management Science*, 49(9):1139–1153, 2003.

[40] M. Hammer. Deep change: How operational innovation can transform your company. *Harvard Business Review*, 82(4):85–93, 2004.

[41] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research*, volume 1. Dover Publications, Mineola, NY, USA, 2004.

[42] A. V. Hill and I. S. Khosla. Models for optimal lead time reduction. *Producton and Operations Management*, 1(2):185–197, 1992.

[43] T. H. Ho and Y. S. Zheng. Setting customer expectation in service delivery: an integrated marketing-operations perspective. *Management Science*, 50(4):479–488, 2004.

[44] W. J. Hopp and M. L. Spearman. *Factory Physics*. McGraw-Hill Higher Education, USA, 2nd edition, 2000.

[45] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

[46] S. H. Hum and H. H. Sim. Time-based competition: Literature review and implications for modeling. *International Journal of Operations & Production Management*, 16(1):75–90, 1996.

[47] E. Kalai, M. Kamien, and M. Rubinovitch. Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163, 1992.

[48] J. J. Kanet. A mixed delay dependent queue discipline. *Operations Research*, 30(1):93–96, 1981.

[49] A. K. Katta and J. Sethuraman. Pricing strategies and service differentiation - a profit maximization perspective. Working paper, Department of Industrial Engineering and Operations Research, Columbia University, NY, USA, 2005.

[50] J. E. Kelley. The cutting plane method for solving convex programs. *SIAM Journal on Applied Mathematics*, 8(4):703–712, 1960.

[51] I. Kim and Tang C. S. Lead time and response time in a pull production control system. *European Journal of Operational Research*, 101:474–485, 1997.

[52] L. Kleinrock. A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11:329–341, 1964.

[53] M. A. Lariviere and J. A. Van Mieghem. Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing and Service Operations Management*, 6(1):23–40, 2004.

[54] G. Latouche and V. Ramaswami. *An Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[55] P. J. Lederer and L. Li. Pricing, production, scheduling and delivery-time competition. *Operations Research*, 45(3):407–420, 1997.

[56] H. T. Leemans. Waiting time distribution in a two-class two-server heterogeneous priority queue. *Performance Evaluation*, 43(2–3):133–150, 2001.

[57] D. Levhari and I. Luski. Duopoly pricing and waiting lines. *European Economic Review*, 11(1):17–35, 1978.

[58] L. Li. The role of inventory in delivery time competition. *Management Science*, 38(2):182–197, 1992.

[59] L. Li and Y. S. Lee. Pricing and delivery-time performance in a competitive environment. *Management Science*, 40(5):633–646, 1994.

[60] L. Liu, M. Parlar, and S. X. Zhang. Pricing and lead time decisions in decentralized supply chains. *Management Science*, 53(5):713–725, 2007.

[61] C. Loch. *Pricing in markets sensitive to delay*. PhD thesis, Stanford University, Stanford, CA, USA, 1991.

[62] D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wiley Publishing Company, CA, USA, 1984.

[63] A. Maltz and E. Maltz. Customer service in the distrbution channel: empirical findings. *Journal of Business Logistics*, 19(2):103–120, 1998.

[64] R. Martin. Kpi: Customer service rules. Technical report, June 10 2000.

[65] G. McWilliams. Lean machine: How Dell fine-tunes its pc pricing to gain edge in a slow market. *Wall Street Journal*, June 8 2001.

[66] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research*, 38(5):870–883, 1990.

[67] D. R. Miller. Computation of steady-state probabilities for m/m/1 priority queues. *Operations Research*, 29(5):945–958, 1981.

[68] R. Nelson. Matrix geometric solutions in markov models: A mathematical tutorial. Technical report, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA, 1991.

[69] M. F. Neuts. *Matrix Geometric Solutions in Stochastic Models*. Dover Publications, Mineola, NY, USA, 1981.

[70] K. S. Palaka, S. Erlebacher, and D. H. Kropp. Lead time setting, capacity utilization, and pricing decisions under lead time demand. *IIE Transactions*, 30(2):151–163, 1998.

[71] M. Parlar. Game theoretic analysis of the substitutable product inventory problem with random demands. *Naval Research Logistics Quarterly*, 35(3):397–409, 1988.

[72] P. Pekgun. *An analysis of pricing and leadtime policies within the marketing/operatons interface*. PhD thesis, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA, 2007.

[73] P. Pekgun, P. M. Griffin, and P. Keskinocak. Centralized vs. decentralized competition for price and lead-time sensitive demand. Working paper, School of Systems and Industrial Engineering, Georgia Institute of Technology, Atlanta, GA, USA, 2006.

[74] E. L. Plambeck. Optimal leadtime differentiation via diffusion approximations. *Operations Research*, 52(2):213–228, 2004.

[75] V. Ramaswami and D. M. Lucantoni. Stationary waiting time distribution in queues with phase type service and quasi-birth-and death processes. *Communication in Statistics - Stochastic Models*, 1(2):125–136, 1985.

[76] S. Ray. *Lead time management in supply chains*. PhD thesis, Department of Management Science, University of Waterloo, Waterloo, ON, Canada, 2001.

[77] S. Ray and E. M. Jewkes. Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research*, 153(3):769–781, 2004.

[78] J. Rochet and L. A. Stole. Nonlinear pricing with random participation. *The Review of Economic Studies*, 69(1):277–311, 1997.

[79] J. Rudy. The video-game price war winners. Business Week, May 24 2002.

[80] M. S. Schmidt and S. Aschkenase. The building blocks of service. *Supply Chain Management Review*, 8:34–40, 2004.

[81] S. K. Sinha, N. Rangaraj, and N. Hemachandra. Pricing surplus server capacity for mean waiting time sensitive customers. Technical Report, Industrial Engineering and Operations Research, Indian Institute of Technology, Bombay, URL `http://www.ieor.iitb.ac.in/~nh/sl-pricing-TR.pdf`, 2008.

[82] M. Smith, J. Bailey, and E. Brynjolfsson. *Understanding the Digital Economy*, chapter Understanding digital markets: Review and assessment. MIT Press, Boston, MA, USA, 2000.

[83] K. C. So. Price and time competition for service delivery. *Manufacturing and Service Operations Management*, 2(4):392–409, 2000.

[84] K. C. So and J. S. Song. Price, delivery time guarantees and capacity selection. *European Journal of Operational Research*, 111(1):28–49, 1998.

[85] G. Stalk and T. M. Hout. *Competing Against Time*. The Free Press, NY, USA, 1990.

[86] F. F. Stephan. Two queues under preemptive priority with poisson arrival and service rates. *Operations Research*, 6(3):399–418, 1958.

[87] J. U. Sterling and D. M. Lambert. Customer service research: past, present and future. *International Journal of Physical Distribution and Materials Management*, 19(2):2–23, 1989.

[88] W. J. Stevenson. *Production and Operations Management*. Irwin McGraw-Hill, USA, 1999.

[89] S. Stidham. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science*, 38(8):1121–1139, 1992.

[90] L. A. Stole. Nonlinear pricing and oligopoly. *Journal of Economics & Management Strategy*, 44(4):529–562, 1995.

[91] L. A. Stole. Price discrimination and competition. In M. Armstrong and R. H. Porter, editors, *HandBook of Industrial Organization*, volume 3. North-Holland, Amsterdam, 2007.

[92] R. Suri. *Quick Response Manufacturing: A Companywide Approach to Reducing Lead Times*. Productivity Press, USA, 1998.

[93] K. Talluri and G. J. Van Ryzin. *The Theory and Practice of Revenue Management*. Kluwer Acedemic Publishers, Boston, USA, 2004.

[94] A. Tsay and A. N. Agrawal. Channel dynamics under price and service competition. *Manufacturing and Service Operations Management*, 2(4):372–391, 2000.

[95] P. Van Beek and C. Van Putter. OR contributions to flexibility improvement in production/inventory systems. *European Journal of Operational Research*, 31:52–60, 1987.

[96] H. R. Varian. Price discrimination. In R. Schmalensee and R. Willig, editors, *Handbook of Industrial Organization*, volume 1. North-Holland, Amsterdam, 1989.

[97] H. R. Varian. *Intermediate Microeconomics: A Modern Apprach*. W. W. Norton & Company, NY, USA, 1990.

[98] T. Vinas. Swagelok CEO Arthur Anton on moving beyond 'make-to-stock'. Industry Week, 2006.

[99] H. White and L. G. Christie. Queuing with preemptive priority or with breakdown. *Operations Research*, 6(1):79–96, 1958.

[100] W. Whitt. Partitioning customers into service groups. *Management Science*, 45(11):1579–1592, 1999.

[101] W. L. Winston and M. Venkataramanan. *Introduction to Mathematical Programming*. Thomson Learning, CA, USA, 2003.

[102] P. Wolfe. The simplex method for quadratic programming. *Econometrica*, 27:382–398, 1959.

[103] R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[104] M. Zhang. *Fencing in a revenue management context*. PhD thesis, Richard Ivey School of Business, University of Western Ontario, ON, Canada, 2007.

[105] X. Zhao, K. E. Stecke, and A. Prasad. Lead time and price quotation mode selection: Uniform or differentiated? Working paper, The School of Management, The University of Texas at Dallas, TX, USA, 2008.