Product Differentiation and Operations Strategy in a Capacitated Environment

Sachin Jayaswal[a,*], Elizabeth Jewkes[b], Saibal Ray[c]

[a]*Indian Institute of Management, Vastrapur, Ahmedabad, Gujarat 380 015, India. Ph: +91-79-6632-4877, Fax: +91-79-6632-6896, E-mail: sachin@iimahd.ernet.in*

[b]*Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. Ph: +1-519-888-4567-x32475, Fax: +1-519-746-7252, E-mail: emjewkes@engmail.uwaterloo.ca*

[c]*Desautels Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, Quebec H3A 1G5, Canada. Ph: +1-514-398-3270, Fax: +1-514-398-3876, E-mail: saibal.ray@mcgill.ca*

Abstract

We study a firm selling two products/services, which are differentiated solely in their prices and delivery times, to two different customer segments in a capacitated environment. From a demand perspective, when both products are available to all customers, they act as substitutes, affecting each other's demand. Customized products for each segment, on the other hand, result in independent demand for each product. From a supply perspective, the firm may either share the same capacity or may dedicate a different 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 this interaction shapes the optimal product differentiation strategy. We show 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 independent products become substitutable, it results in a more homogeneous pricing scheme. Moreover, the optimal response to an increase in capacity cost also depends on the firm's operations strategy. In a dedicated capacity scenario, the optimal response is always to offer more homogeneous prices and delivery times. In a shared capacity setting, it is always optimal to quote more homogeneous delivery times, but to increase or decrease the price differentiation depending on whether the status-quo capacity cost is high or low, respectively.

Keywords: pricing, product differentiation, delivery time guarantee, product substitution, capacity sharing

*Corresponding author

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, thereby reducing the cycle time by 75% and operating expenses by 45%, and boosting customer satisfaction by 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; 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 ship service parts within 24 hours to its clients (Schmidt and Aschkenase, 2004). Similar 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, this paper deals with a setting where the end customer demand is sensitive to both the price charged as well as the delivery time guarantee offered. In such a 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., to 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). Obviously, managers need to decide on the optimal price and delivery time levels for whichever policy they choose. If a firm decides to guarantee different delivery times to its different customer segments, then the second issue 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 for each option.

We study firms offering a menu of differentiated products/services to exploit heterogeneity in customers' preferences for time and their willingness to pay. Such firms then need to decide whether to make a given product accessible to all customers or to customize it (and make it available) for only one segment. For example, the price and delivery time combina-

[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 arrives (Plambeck, 2004).

tions that Dell quotes to government and health-care companies are very different from what it quotes to individuals (McWilliams, 2001). In this case, Dell designs an exclusive 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 are tailored based on its geographical location and industry segment (Plambeck, 2004). On the other hand, FedEx and UPS offer logistics 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 preferences for speed and their 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. The menu of products offered are thus *substitutable*, creating a demand-side interaction between the different market segments.

Just like the demand side, the supply side for different customer segments may either 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.[2] In contrast, UPS delivers express and ground services using one integrated network.[3] Photo development stores offering both express and regular services also share capacity used for the two services. It is important to note that offering different delivery time guarantees using a shared capacity creates a supply-side relationship between the different market segments a firm serves. Thus, it 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 paper 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: (1) *How does the operations strategy (dedicated or shared capacity) of a*

[2]http://www.fedex.com/us/about/express/pressreleases/pressrelease011900.html?link=4
[3]http://sec.edgar-online.com/1999/10/20/11/0000940180-99-001230/Section2.asp

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)? (2) *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?* (3) *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 1. A 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, a 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 1: Scenarios studied in this paper

	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 solve for the optimal delivery times that a 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 (to satisfy the promised delivery times with a certain degree of reliability) for each scenario. While the dedicated capacity cases can be solved by functional optimization, for the shared capacity scenarios, we utilize a novel methodology integrating 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 allows us to illustrate both the individual and the joint effects of product substitution and operations strategy on the optimal product differentiation policy of a firm. Some of the important

managerial insights we generate from our analysis are: (1) *In a high capacity cost environment, a firm with shared capacity should offer more differentiated products (both in terms of prices and delivery times) than a firm with dedicated capacities, irrespective of whether the products are substitutable or not.* (2) *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 a shared or a dedicated capacity regime. However, with regard to 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).* (3) *As the capacity cost increases, the optimal strategy for a firm with dedicated capacities is to offer more homogeneous pricing and delivery time schemes for both substitutable and non-substitutable products. A shared capacity firm should also always offer more homogeneous delivery times, but will need to increase or decrease the price differentiation level depending on whether the status-quo capacity cost is high or low, respectively.*

The rest of the paper is organized as follows. In §2, we briefly review the related literature. §3 defines the modelling framework, followed by a discussion on the solution methodology in §4. §5 presents a discussion on the managerial insights highlighted above. The paper concludes with a summary of results and a discussion about future research in §6.

2. Related Literature

The literature related to our paper can be categorized into four groups based on whether they consider demand-side and/or supply-side interaction (similar to Table 1).

The papers in the first category deal with 'non-substitutable products, dedicated capacity' scenario and 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 one single server queue. These models also consider a single product, and hence product differentiation or substitutability is not an issue. So (2000), Tsay and Agrawal (2000), and Pekgun et al. (2006) study similar problems but in a competitive setting, where two firms selling a common product compete on price and delivery time. Again, their models do not study differentiation or capacity sharing since customers are assumed to be homogeneous.

The second category of papers takes into account product differentiation and substitution among multiple products, but assumes that the products are processed using dedicated

capacities. Boyaci and Ray (2003, 2006) are examples of such 'substitutable products, dedicated capacity' scenario papers. 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.

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 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 the substitutability issue, these papers fall under the 'non-substitutable products, shared capacity' category. In general, the problem considered in these papers is to design a pricing scheme 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. Finally, 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' category since they consider resource substitution. However, they do not deal with product substitution or pricing/delivery time decisions, and, therefore, do not capture the interaction between product substitution and capacity strategy.

Our study further extends the research 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.

3. Decision Models

We model a firm that offers a single basic product/service (henceforth called product) in a market comprising heterogeneous 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 $i \in \{h, l\}$. 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_i and L_i denote the price and delivery time offered by the firm to customer class $i \in \{h, l\}$.

Demand from customer class i arrives according to a Poisson process with rate λ_i, 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[4], inspired by Tsay and Agrawal (2000) and Boyaci and Ray (2003):

$$\lambda_h = a - \beta_p^h p_h + \theta_p(p_l - p_h) - \beta_L^h L_h + \theta_L(L_l - L_h) \tag{1}$$

$$\lambda_l = a - \beta_p^l p_l + \theta_p(p_h - p_l) - \beta_L^l L_l + \theta_L(L_h - L_l) \tag{2}$$

where,

- $2a$: market base
- β_p^i : sensitivity of class i demand to its own price
- β_L^i : sensitivity of class i 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

[4]This demand model also establishes the effects of price and time differentiation on the demand rates: one extra unit of price differentiation reduces the demand rate from express customer and increases that from regular customers by the same amount, while one extra unit of time differentiation increases the demand rate from express customers and reduces that from regular customers by the same amount.

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. 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.[5] By definition, $\beta_p^i > 0$, $\beta_L^i > 0$, $\theta_p \geq 0$, $\theta_L \geq 0$, $\beta_p^h < \beta_p^l$ and $\beta_L^h > \beta_L^l$.

We assume the time it takes to serve a demand from class i is exponentially distributed with rate μ_i, $i \in \{h, l\}$. The service facility is thus modelled as an M/M/$\cdot$ queuing system.[6] Customers within each class are served on a first-come-first-served (FCFS) basis. The firm can invest in its installed capacity to increase its processing rate μ_i. We assume there are no economies of scale in investing in capacity. So, a unit increment in μ_i always costs $\$A$.[7] 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.[8] 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 (to avoid any further delay) without additional charge to the customer. FedEx also offers a money-back guarantee for every U.S. shipment that is even 1 minute late compared to its guaranteed delivery time.[9] The firm, therefore, needs to simultaneously select the optimal service rates (i.e., capacities)

[5]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.

[6]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).

[7]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 logistics services 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 for a meaningful comparison between the dedicated and the shared capacity settings.

[8]Although we make this assumption mainly for the tractability of the model, this is still a reasonably realistic representation of certain business settings. For example, in most of the online retail web hosting services, any updating of content, if not done in express fashion, must be done within one day (Boyaci and Ray, 2003).

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

μ_h and μ_l in order to meet the guaranteed delivery times with at least a minimum level of reliability α (referred to as the target service level). We refer to this problem as the Pricing and Delivery Time Decision Problem [$PDTDP$]. Mathematically, it can be stated as:

[**PDTDP**] :

$$\max_{p_h,p_l,L_h,\mu_h,\mu_l} \pi = (p_h - m)\lambda_h + (p_l - m)\lambda_l - A(\mu_h + \mu_l) \tag{3}$$

$$\text{s.t.} \quad L_h < L_l \tag{4}$$

$$p_h, p_l, \lambda_h, \lambda_l, \mu_h, \mu_l, L_h \geq 0 \tag{5}$$

$$\textit{Stability condition} \tag{6}$$

$$S_h(L_h) = P(W_h \leq L_h) \geq \alpha \tag{7}$$

$$S_l(L_l) = P(W_l \leq L_l) \geq \alpha \tag{8}$$

where λ_h and λ_l are given by (1) and (2) respectively. Constraint (4) requires that the guaranteed delivery time for high priority customers be shorter than that for the other class. Constraint set (5) is required to define a realistic problem setting. Constraint (6) is the stability condition for the queuing system, which models the service facility at the firm. Constraints (7) and (8) are delivery time reliability constraints (also called service level constraints), which require 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. 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 (1) and (2).

Note that the above formulation 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 (6)-(8) depending on the capacity strategy used (shared or dedicated), and ii) the form of the demand function that signifies the absence or presence of product substitution.

Dedicated Capacity Setting: For a dedicated capacity setting, wherein 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 (Gross and Harris, 1998; Ross, 2003; So and Song, 1998). In this case, there is a separate stability condition for each of the queues. Hence,

constraints (6), (7), and (8) can be expressed as:

$$\lambda_i - \mu_i < 0,\ i \in \{h, l\} \tag{6^{DC}}$$

$$S_h(L_h) = P(W_h \leq L_h) = 1 - e^{(\lambda_h - \mu_h)L_h} \geq \alpha \tag{7^{DC}}$$

$$S_l(L_l) = P(W_l \leq L_l) = 1 - e^{(\lambda_l - \mu_l)L_l} \geq \alpha \tag{8^{DC}}$$

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 (1) and (2). We denote the resulting models of Pricing and Delivery Time Decision Problem in a Dedicated Capacity setting by $[PDTDP_{DC}]$.

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 allocates its resources to first serve high priority customers who pay a premium price, and then uses any remaining capacity to serve low priority customers. This practice somewhat reflects the one at UPS (Ata and Van Mieghem, 2008). Dedicated capacities with partial pooling more accurately model the operational setting used by UPS wherein fast airplanes can serve both express and regular markets, while the slow trucks serve only the regular market. We use shared capacity to study the extreme scenario (with complete pooling) and compare it with the dedicated capacity setting, typical of FedEx. 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 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 (Chang, 1965). Hence, the delivery time reliability constraint (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 constraint (8) of $[PDTDP]$ problem, is not known (Abate and Whitt, 1997). We assume that the single server serves customers of either class at the same rate $\mu_h = \mu_l = \mu$. Constraints (6) and (7) in a shared capacity setting can then be expressed as:

$$\lambda_h + \lambda_l - \mu < 0 \tag{6^{SC}}$$

$$S_h(L_h) = P(W_h \leq L_h) = 1 - e^{(\lambda_h - \mu)L_h} \geq \alpha \tag{7^{SC}}$$

We discuss how we tackle the issue of delivery reliability for regular customers (corresponding to Equation (8)) in the next section. As in the dedicated capacity setting, 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 (1) and (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).

4. Solution Methodology

We now discuss the solution methodology for the models discussed in §3.

4.1. Dedicated Capacity Setting

The dedicated capacity model $[PDTDP_{DC}]$ has been studied by Boyaci and Ray (2003) (both substitutable and non-substitutable products) for the special case $\beta_p^h = \beta_p^l = \beta_p$ and $\beta_L^h = \beta_L^l = \beta_L$. We briefly present the solution for our more general case for the sake of completeness.

Proposition 1. *For a fixed express delivery time L_h, the optimal prices in a dedicated capacity setting are given by:*

$$p_h^{DC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^l + 2\theta_p)a - (\beta_p^l\beta_L^h + \beta_p^l\theta_L + \beta_L^h\theta_p)L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{9}$$

$$p_l^{DC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^h + 2\theta_p)a + (\beta_p^h\theta_L - \beta_L^h\theta_p)L_h - (\beta_p^h\beta_L^l + \beta_p^h\theta_L + \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{10}$$

The optimal guaranteed delivery time for the express customers L_h^{DC*} is given by the unique root of (11) in the interval [0, L_l).

$$\frac{\partial\pi(L_h)}{\partial L_h} = -\left(\beta_L^h + \theta_L\right)\left(p_h^{DC*}(L_h) - m - A\right) + \theta_L\left(p_l^{DC*}(L_h) - m - A\right) - \frac{A\ln(1-\alpha)}{L_h^2} \tag{11}$$

Proof. See Appendix A. □

The corresponding optimal price differentiation is then:

$$p_h^{DC*}(L_h) - p_l^{DC*}(L_h) = \frac{(\beta_p^l - \beta_p^h)a + \beta_p^h\beta_L^l L_l - \beta_p^l\beta_L^h L_h + (\beta_p^h + \beta_p^l)\theta_L(L_l - L_h)}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{12}$$

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.

4.2. Shared Capacity Setting

The shared capacity model $[PDTDP_{SC}]$ is relatively more challenging to solve for, especially in the absence of an analytical characterization of the delivery time reliability con-

straint (8) for regular customers. While the Laplace transform of the sojourn time distribution $S_l(\cdot)$, appearing in (8), and its first few moments are well known (Stephan, 1958), 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 choice of appropriate approximation to be used depends on the demand rates of the two customer classes, which can only be determined endogenously, and is 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 $[PDP]$) using the *matrix geometric method* in a *cutting plane* framework. Solution to $[PDP]$ is then used to solve $[PDTDP_{SC}]$ using the *golden section search* method.

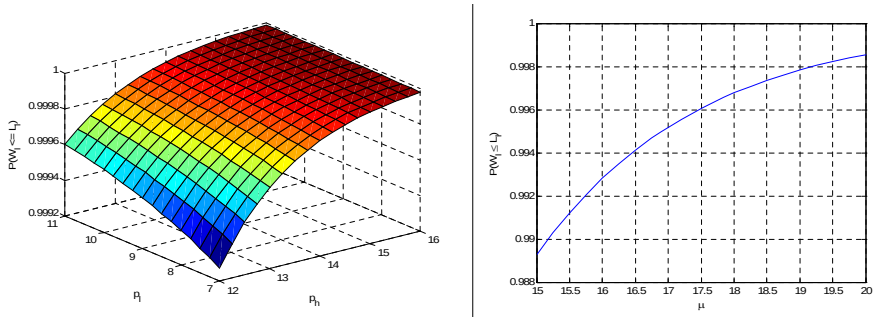
4.2.1. The Pricing Decision Problem [PDP]

On substituting (1) and (2) into (3), the objective function for $[PDP]$ is quadratic. All constraints are linear, except for (8), which does not have a closed form expression. Although the exact form of $S_l(\cdot)$ in constraint (8) is unknown, we exploit its special structure, determined numerically using the matrix geometric method. Plots of $S_l(\cdot)$ vs. (p_h, p_l), and $S_l(\cdot)$ vs. μ are shown in Figure 1. These plots suggest that $S_l(\cdot)$ 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, μ). 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, μ^k), $\forall\ k \in K$:

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

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 (8) can thus be replaced



(a) S_l vs. (p_h, p_l) (b) S_l vs. μ

Figure 1: Service level vs. prices and capacity

by the following set of linear constraints:

$$S_l^k(\cdot) + (p_h - p_h^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) + (p_l - p_l^k)\left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) + (\mu - \mu^k)\left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \geq \alpha \quad \forall k \in K \tag{13}$$

Substituting the above set of constraints in place of (8), and the expressions (1) and (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 is suitable for the use of cutting plane method (Kelley, 1960).

$\underline{[\mathbf{PDP_{(K)}}]:}$

$$\max_{p_h,p_l,\mu} \pi = -(\beta_p^h + \theta_p)p_h^2 - (\beta_p^l + \theta_p)p_l^2 + 2\theta_p p_h p_l +$$
$$\left\{-\beta_L^h L_h + \theta_L(L_l - L_h) + m\beta_p^h + a\right\} p_h +$$
$$\left\{-\beta_L^l L_l + \theta_L(L_h - L_l) + m\beta_p^l + a\right\} p_l - A\mu + (\beta_L^h L_h + \beta_L^l L_l)m - 2ma \tag{14}$$

$$\text{s.t.} \quad p_h, p_l, \mu \geq 0 \tag{15}$$

$$-(\beta_p^h + \theta_p)p_h + \theta_p p_l \geq (\beta_L^h + \theta_L)L_h - \theta_L L_l - a \tag{16}$$

$$\theta_p p_h - (\beta_p^l + \theta_p)p_l \geq -\theta_L L_h + (\beta_L^l + \theta_L)L_l - a \tag{17}$$

$$-\beta_p^h p_h - \beta_p^l p_l - \mu < \beta_L^h L_h + \beta_L^l L_l - 2a \tag{18}$$

$$-(\beta_p^h + \theta_p)p_h + \theta_p p_l - \mu \leq \frac{\ln(1-\alpha)}{L_h} - a + (\beta_L^h + \theta_L)L_h - \theta_L L_l \tag{19}$$

$$\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) p_h + \left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) p_l + \left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \mu \geq \alpha - S_l^k(\cdot) +$$
$$\left(\frac{\partial S_l^k(\cdot)}{\partial p_h}\right) p_h^k + \left(\frac{\partial S_l^k(\cdot)}{\partial p_l}\right) p_l^k + \left(\frac{\partial S_l^k(\cdot)}{\partial \mu}\right) \mu^k \qquad \forall k \in K \tag{20}$$

It is easy to verify that the Hessian of (14) is negative semidefinite. Therefore, $[PDP_{(K)}]$ has a quadratic concave objective function. Moreover, all its constraints are linear. Hence,

Karush-Kuhn-Tucker (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 such as Wolfe's Algorithm (Wolfe, 1959).

We use the matrix geometric method to numerically evaluate $S_l^k(\cdot)$ at a given point (p_h^k, p_l^k, μ^k) (see Appendix B). We refer our readers to Neuts (1981) for details of the matrix geometric method. The use of the matrix geometric method yields explicit recursive formulas for the joint stationary probabilities, which can provide significant computational improvements over the transform techniques (Miller, 1981). Moreover, it gives exact solutions, in contrast to simulation, which is another alternative method to evaluate $S_l(\cdot)$ 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 $[PDP]$ repeatedly for different values of L_h. Once $S_l(\cdot)$ is evaluated at a point (p_h^k, p_l^k, μ^k), its gradients are obtained using the *finite difference method* (described in Appendix C). The gradients are used to generate cuts of the form (20), which are added iteratively in the cutting plane algorithm. The details of the cutting plane algorithm along with its computational performance are presented in Appendix D.

4.2.2. The Pricing and Delivery Time Decision Problem $[PDTDP_{SC}]$

The Pricing and Delivery Time Decision Problem $[PDTDP_{SC}]$ adds an additional dimension to $[PDP]$ 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 (7^{SC}) non-linear, and the model substantially more challenging to solve. We use the solution to $[PDP_{(K)}]$ and the golden section search method (Luenberger, 1984) 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 $[PDP]$ for a fixed L_h. We have shown that in a dedicated capacity setting $f(L_h)$ has a unique maximum when a is high. Our extensive numerical experiments with $f(L_h)$ suggest 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 a $[PDP_{(K)}]$ to evaluate $f(L_h)$ for a given value of L_h.

5. Analysis and Managerial Insights

We study the four scenarios described in §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 policy. We then investigate how rising capacity costs affect product differentiation policy under different demand and supply conditions. Since our mathematical model for shared capacity setting lacks a closed-form 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, followed by their mathematical justification and intuitive explanations wherever possible. From these observations, we derive conclusions of managerial interest. We first discuss the observations for the pricing decision problem $[PDP]$ for a fixed delivery time differentiation. We then discuss the more general problem of pricing and delivery time decision $[PDTDP]$.

Our model setting 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 $= 1 - L_h^*$). With regard to the other parameters, we experiment with a large combination of their values (see Appendix E for the experiment design). However, the figures that we present in this paper use $a = 10$, $m = 3$, $\alpha = 0.99$, $A = 0.5$ (unless otherwise stated). For the solution algorithm, a bound (M) on the high priority queue size needs to be specified to facilitate use of the matrix geometric method. Computational experiments of priority queues with an appropriate range of parameter values suggested $M = 100$ to be a good choice with little effect on the accuracy of the waiting time distributions. 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, which are good enough for the accuracy of the results up to 3 decimal places.

5.1. Pricing Decision Problem [PDP]

[PDP] is relevant to situations where a firm may face a significantly higher stickiness for its delivery time decisions compared to its ability to vary prices. A relatively higher stickiness for delivery time decisions may arise, for example, when the services are partly outsourced to a third party. In such a situation, the firm may not be able to revise its delivery time decisions as frequently as it can revise its prices, and may optimize its prices, treating its delivery times as fixed. Other factors contributing to stickiness in delivery time decisions can be found in Allon and Federgruen (2007).

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. Our results show that 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). Product substitution, on the other hand, affects the behavior of regular price only. Figure 3 shows the behavior of the two prices under different scenarios as we vary L_h, which is summarized in the following observation:

Observation 1: *In both the dedicated capacity (DC) and the shared capacity (SC) settings, a decrease in L_h results in: (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 in p_l^* if $\theta_L/\beta_L^h < \theta_p/\beta_p^h$; and no change in p_l^* if $\theta_p = \theta_L = 0$.*

This behavior is quite intuitive and is similar to what has been shown by Boyaci and Ray (2003) for the dedicated capacity case (also refer to the expressions of the two prices, given by (9) and (10)). Since express customers are time-sensitive, a firm can always charge them a higher price for a guaranteed shorter delivery time. However, in the 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 customers. With product substitution, the behavior of the optimal price for the regular class depends on the market conditions. In a market with $\theta_L/\beta_L^h > \theta_p/\beta_p^h$, henceforth referred to as a "*Time Difference Sensitive*" (TDS) market, the relative sensitivity of express 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). 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 options, thereby increasing its profit. The effect of the market with $\theta_L/\beta_L^h < \theta_p/\beta_p^h$, henceforth referred to as a "*Price Difference Sensitive*" (PDS) market, can also be similarly explained (refer to Boyaci and Ray (2003) for details).

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 policy. Since the behavior of the prices depends on the market characteristics, we compare their optimal values under the

different market settings (TDS and PDS), described earlier. We use the following market parameter values for the two market types, which are defined with respect to both the customer classes:

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

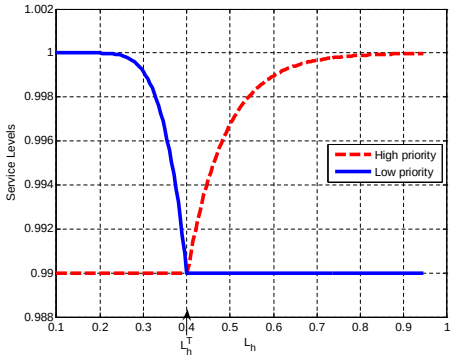
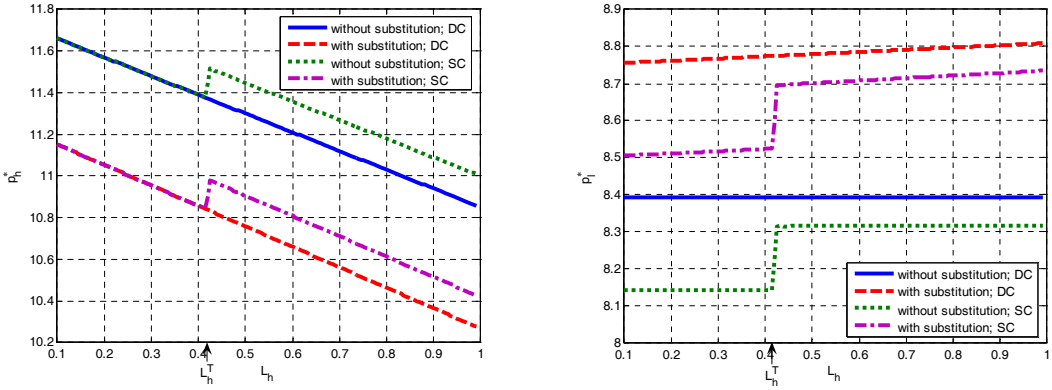


Figure 2: Service levels for high and low priority customers in a shared capacity setting

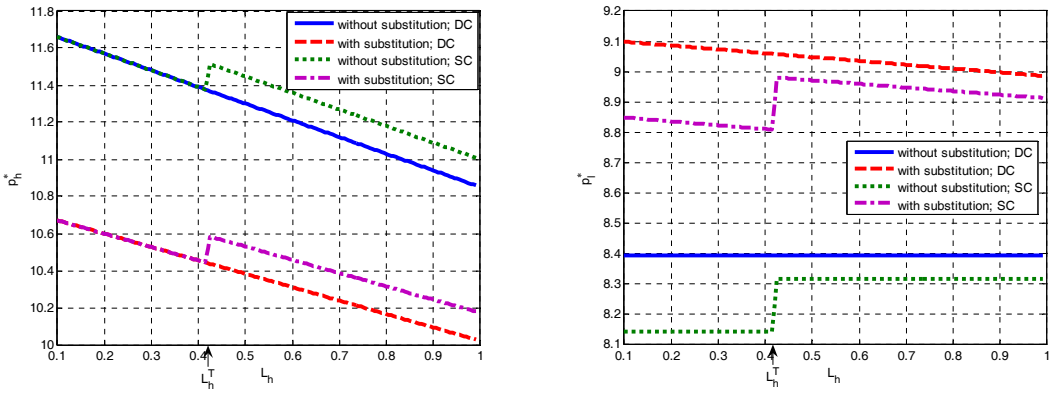
Observation 2: *For a given delivery time differentiation* (refer to Figure 3):
- a change in capacity strategy from dedicated to shared results in (a) a generally higher p_h^*, (b) a lower p_l^*, and hence (c) a higher optimal price differentiation.
- introduction of product substitutability results 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 policy (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 strategy. The underlying reason for this observation is described below.

We first note that as in DC, the stability condition in SC is automatically satisfied by the respective two delivery time reliability constraints. This is because as $\lambda_h + \lambda_l \to \mu$, the queue of waiting customers grows infinitely long such that the probability of serving a customer from at least one of the classes within a finite time will approach 0. Further, for



(a) TDS



(b) PDS

Figure 3: Comparison of prices in the four different scenarios

values of L_h smaller than L_h^T, it is relatively more expensive for firms in SC to meet the target delivery time reliability α for express customers than for regular customers. Thus, the capacity requirement is dictated by the demand from express customers such that the optimal combination of prices and capacity level that satisfies the delivery time reliability constraint (19) for the express class automatically satisfies the delivery time reliability constraint (20) for the regular class as well as the stability condition (18). This is an important observation, which suggests that for values of L_h below L_h^T, relaxing both the complicating constraint (20) and the stability condition (18) will still give the same optimal solution to $[PDP]$, and the resulting problem can be solved analytically (see Appendix F for details). The optimal prices are then given by:

$$p_h^{SC*}(L_h) = \frac{A+m}{2} + \frac{(\beta_p^l + 2\theta_p)a - (\beta_p^l\beta_L^h + \beta_p^l\theta_L + \beta_L^h\theta_p)L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{21}$$

$$p_l^{SC*}(L_h) = \frac{m}{2} + \frac{(\beta_p^h + 2\theta_p)a + (\beta_p^h\theta_L - \beta_L^h\theta_p)L_h - (\beta_p^h\beta_L^l + \beta_p^h\theta_L + \beta_L^l\theta_p)L_l}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \tag{22}$$

and

$$p_h^{SC*}(L_h) - p_l^{SC*}(L_h) = \frac{A}{2} + \frac{(\beta_p^l - \beta_p^h)a + \beta_p^h\beta_L^l L_l - \beta_p^l\beta_L^h L_h + (\beta_p^h + \beta_p^l)\theta_L(L_l - L_h)}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} \quad (23)$$

Comparing the above prices with those for the dedicated capacity setting (Proposition 1), we see that for values of L_h smaller than L_h^T, the price for express customers remains the same, whereas that for regular customers decreases by a constant amount $A/2$, thereby increasing the price differentiation.

As L_h increases, the 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 (20). 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 (9) and (10).

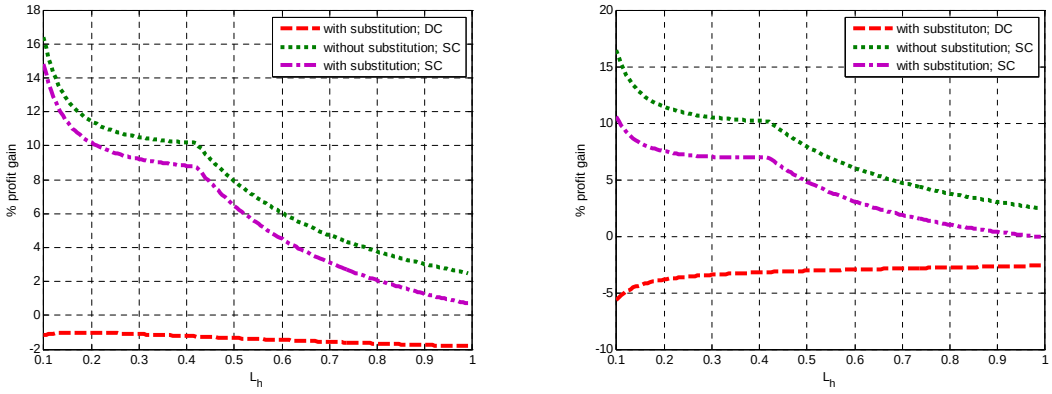
$$\begin{aligned}
&p_h^{DC*}(L_h)\big|_{\theta_p,\theta_L>0} - p_h^{DC*}(L_h)\big|_{\theta_p,\theta_L=0} \\
&= \frac{-(\beta_p^l - \beta_p^h)\theta_p a - (\beta_p^h\theta_L - \beta_L^h\theta_p)\beta_p^l L_h + (\beta_p^l\theta_L - \beta_L^l\theta_p)\beta_p^h L_l}{2\beta_p^h(\beta_p^h\beta_p^l + \beta_p^h\theta_p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

The above equations show that for a sufficiently high a, which also guarantees $L_h < L_l$ and $p_h > p_l$ (see Appendix A), 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 A can be explained similarly by substituting $\theta_p = \theta_L = 0$ in (21) and (22).

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: *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 4).



(a) TDS (b) PDS

Figure 4: % Profit gain for different scenarios over the 'non-substitutable products, dedicated capacity' scenario

Figure 4 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. This can be shown analytically for small L_h. For a small L_h, we first note that the following relations hold between

capacity requirements of the two settings:

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

The first $\leq$ relation above can be explained by comparing the resulting express demand functions obtained by substituting the prices (p_h^{DC*}, p_l^{DC*}) and (p_h^{SC*}, p_l^{SC*}) in the demand model (1) and (2), and using the relations: $p_h^{SC*} = p_h^{DC*}$ and $p_l^{SC*} < p_l^{DC*}$ (from Observation 1). The above derived relationship shows that a firm in a shared capacity setting benefits from capacity pooling. The gain in profit in SC over DC is attributed predominantly to the savings in capacity related costs due to capacity pooling.

We further observe that the relative gain in profit in SC over DC increases with an increase in the delivery time differentiation (decrease in L_h). 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 the principles of revenue management, which suggest that a properly designed fence that prevents leakage of demand from the high price segment to the low price segment enhances a firm's profit (Zhang, 2007).

5.2. The Pricing and Delivery Time Decision Problem [PDTDP]

The last section focussed on the optimal pricing (and the resulting 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, namely, $[PDTDP_{DC}]$ and $[PDTDP_{SC}]$. We first do a comparison of the optimal product differentiation under the four scenarios for a given marginal capacity cost A, and then study its behavior in each scenario as A increases.

Table 2: 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

Table 3: 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

We present a small sample from our extensive numerical experiments to illustrate our comparison of the optimal decisions in the four scenarios. We use the following demand parameters: $a = 10$, $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$. For product substitution, the following two parameter combinations correspond to: (i) TDS: $\theta_p = 0.2$ and $\theta_L = 0.5$ (ii) PDS: $\theta_p = 0.4$ and $\theta_L = 0.3$. Other parameters are fixed at: $m = 3$, $\alpha = 0.99$, $L_l = 1$. 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 2 for 'without substitution' scenario and in Table 3 for 'with substitution' scenario. General observations arising from our numerical results are summarized in Observation 4. These observations hold true in general, independent of the system parameter values chosen even though we are unable to establish these results analytically, especially for large A for which we do not have analytical results in a shared capacity setting. Some of these observations, especially when the capacity cost A is small, are explained analytically in Appendix F.

Observation 4: - *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 inexpensive.
- 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 its capacity strategy, market conditions, and marginal capacity cost.

Managerial Implications: Clearly, when the marginal capacity cost is large, sharing capacities always increases both the optimal delivery time and price differentiation of a firm, regardless of the product/market characteristics. This result 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.

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 4 shows two different price and delivery time combinations offered by FedEx[10] and UPS[11] 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 appears to be in close agreement with our observation, assuming that the marginal capacity cost is sufficiently large.

The above observation also has important implications for Dell or steel, chemical, and consumer product industries, cited in §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 are thus non-substitutable. If these

[10]http://www.fedex.com/ratefinder/home?cc=ca&language=en
[11]https://wwwapps.ups.com/ctc/request?loc=en_CA

Table 4: 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 a.m. next day	33.40 CAD	UPS Express Early A.M.	by 8:00 a.m. next day	42.2 CAD
FedEx Priority Overnight	by 12:00 p.m. next day	18.84 CAD	UPS Express Saver	by 12:00 p.m. next day	15.32 CAD

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 result is intuitive since the customers' preferences 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 more differentiated delivery time options 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 the firm's 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.

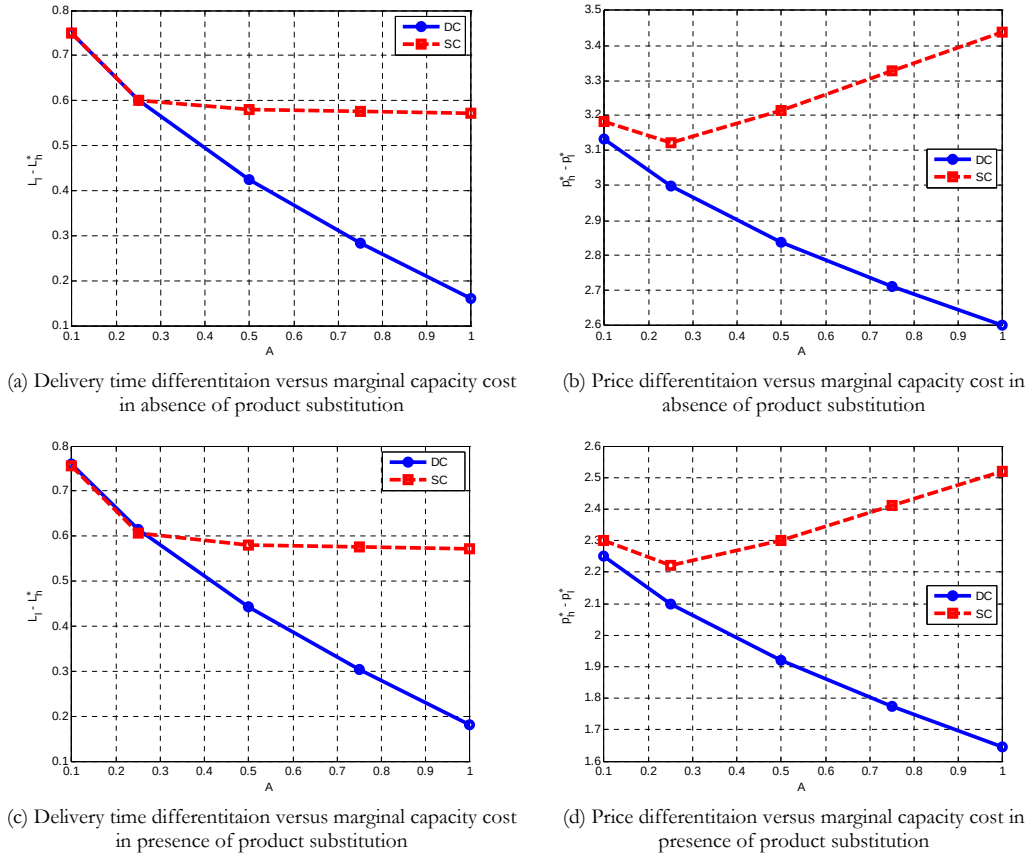
Effects of Capacity Cost Increase

Another issue of potential managerial interest is how the product differentiation strategy for a firm should change as its marginal capacity cost A changes. The details of the effects of A on the optimal decision variable values are shown in Appendix H. In Figure 5, we show the optimal delivery time and price differentiation decisions under various scenarios only for a TDS type market, although the nature of the results remain the same even for a PDS type

market. The following observation summarizes our main finding in this context.

Observation 5: - *For a firm using a dedicated capacity strategy, its optimal response to any increase in marginal capacity cost is to: (a) decrease the delivery time differentiation, and (b) decrease the price differentiation.*

- For a firm using a shared capacity strategy, its optimal response to any increase in marginal capacity cost is to: (a) decrease the delivery time differentiation, and (b) also decrease the price differentiation when the status-quo capacity cost is low, but to increase it when the status-quo capacity cost is high. (Refer to Figure 5).



(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 5: Effects capacity cost

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 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 (TDS/PDS). 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 thereby reduce its delivery time differentiation. A longer express delivery time, in turn, decreases price differentiation in a dedicated setting, as evident from (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 a 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). Mathematical justifications for these results are given in Appendix H.

Managerial Implications: Observation 5 suggests that an increase in capacity costs 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.

It is important to note that we have so far in our analysis assumed the same marginal capacity cost A for the two customer segments, which allows for a meaningful comparison between the dedicated and shared capacity settings. However, in reality, when different customer segments are served by dedicated capacities, they may incur different marginal capacity costs. In the logistics services industry, for example, express customers served by airplanes will have a marginal capacity cost different from regular customers served by trucks. Let A_h and A_l ($A_l \leq A_h$) represent the marginal capacity costs for the express and regular customer segments, respectively of a firm operating with dedicated capacities. It is important in such a setting to understand the optimal response of the firm separately to an increase in A_h and A_l. We only briefly state our results in this regard, which concur with Boyaci and Ray (2003) (the mathematical details are provided in Appendix H): *For a firm using a dedicated capacity strategy, its optimal response is to: (a) decrease its delivery time differentiation with any increase in A_h, and (b) increase its delivery time differentiation*

with any increase in A_l. Further, if both the marginal capacity costs increase by the same amount (such that A_h - A_l remains constant), then the optimal response is to decrease both its delivery time differentiation as well as its price differentiation.

6. Conclusions and Future Research

In this paper we studied the optimal product differentiation strategy of a firm selling two 'products', which are similar in all respects except in their prices and guaranteed delivery times, in a capacitated environment. Our primary objective was to understand how the demand-side product substitution interacts with the firm's supply-side operations policy (using dedicated versus shared capacity) in shaping 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 captures the four scenarios of our interest depending on whether the products are substitutable or not, and whether the capacity strategy is shared or dedicated. 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.

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 (less expensive) products, compared to when there are dedicated capacities for each of them. This implies that the firm offers more differentiated products under a shared capacity setting. Interestingly, the above effect of the capacity strategy does not depend either 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 (operations strategy and capacity cost of the firm and the behavior of the end customers). However, introduction of substitutability always results in less expensive express products and more expensive regular products, i.e., a more homogeneous pricing scheme.

We also demonstrated that as the capacity becomes more expensive, the optimal response of the firm depends on its operations strategy but not on the demand characteristics. In such a 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. However, a shared capacity firm needs to increase or decrease the price differentiation depending on whether the system is already highly capacitated or not, respectively. Obviously, in reality, some of our assumptions required for tractability purposes indeed may not hold true. Moreover, it may be difficult to implement some of our suggestions (e.g., budget constraints may force firms to opt for a shared capacity strategy even when a dedicated capacity strategy is optimal, or external competitive forces may require firms to select a certain non-optimal pricing and/or delivery time strategy). Consequently, the above insights should be viewed more as recommendations for managers rather than definite methodologies for decision-making.

There are a number of directions in which this research can be extended. One possible extension will 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 also include the guaranteed delivery time for the regular customers (L_l) 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. Finally, it will be interesting to incorporate horizontal competition in the model as has been done by So (2000) and Tsay and Agarwal (2000). However, firms in these studies compete for a single product, and hence product substitutability is not an issue. We see modelling two competing firms, each of which sells two substitutable products (as is the case for FedEx and UPS), as another possible extension of our work. This may, however, make the model extremely challenging, and developing a good approximation is imperative in that case.

References

Abate J, Whitt W. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. Queueing Systems 1997;25; 173–233.

Afeche P, Mendelson H. Pricing and priority auctions in queueing systems with a generalized delay cost structure. Management Science 2004;50; 869–882.

Allon G, Federgruen A. Competition in service industries. Operations Research 2007;53;

37–55.

Ata B, Van Mieghem JA . The Value of Partial Resource Pooling: Should a Service Network be Integrated or Product-Focused? Management Science 2009;55; 115–131.

Atlason J, Epelman MA, Henderson SG. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 2004;127; 333–358.

Boyaci T, Ray S. Product differentiation and capacity cost interaction in time and price sensitive markets. Manufacturing and Service Operations Management 2003;5; 18–36.

Boyaci T, Ray S. The impact of capacity costs on product differentiation in delivery time, delivery reliability and price. Production and Operations Management 2006;15; 179–197.

Chang W. Preemptive priority queues. Operations Research 1965;13; 820–827.

Dewan S, Mendelson H. User delay costs and internal pricing for a service facility. Management Science 1990;36; 1502–1517.

Gross D, Harris CM. Fundamentals of queueing theory (4th ed.). John Wiley & Sons, Inc.: New York, USA; 2008.

Hammer M. Deep Change: How operational innovation can transform your company. Harvard Business Review 2004;82; 84–93.

Katta AK, Sethuraman J. Pricing strategies and service differentiation in an M/M/1 Queue - A profit maximization perspective. Working paper 2005; Department of Industrial Engineering and Operations Research, Columbia University, NY, USA.

Kelley JE Jr. The cutting-plane method for solving convex programs. Journal of the Society for Industrial and Applied Mathematics 1960;8; 703–712.

Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. The American Statistical Association and the Society for Industrial and Applied Mathematics: Philadelphia, USA; 1999.

Luenberger DG. Linear and Nonlinear Programming. Addison-Wiley: CA, USA; 1984.

McWilliams G. Lean machine: How Dell fine-tunes its PC pricing to gain edge in a slow market. Wall Street Journal 2001; June 8.

Mendelson H, Whang S. Optimal incentive-compatible priority pricing for the M/M/1 queue. Operations Research 1990;38; 870–883.

Miller DR. Computation of steady-state probabilities for M/M/1 priority queues. Operations Research 1981;29; 945–958.

Neuts MF. Matrix-geometric solutions in stochastic models: An algorithmic approach. Dover Publications: Mineola, USA; 1981.

Palaka KS, Erlebacher D, Kropp H. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions 1998;30; 151–163.

Pekgun P, Griffin PM, Keskinocak P. Centralized vs. decentralized competition for price and lead-time sensitive demand. Working paper 2006; H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Plambeck EL. Optimal leadtime differentiation via diffusion approximations. Operations Research 2004;52; 213–228.

Ramaswami V, Lucantoni DM. Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes. Stochastic Models 1985;1; 125–136.

Ray S, Jewkes EM. Customer lead time management when both demand and price are lead time sensitive. European Journal of Operational Research 2004;153; 769–781.

Ross SM. Introduction to probability models (10th ed.). Academic Press: San Diego, USA; 2010.

Schmidt M, Aschkenase S. The building blocks of service. Supply Chain Management Review 2004;8; 34–40.

So KC. Price and time competition for service delivery. Manufacturing and Service Operations Management 2000;2; 392–409.

So KC, Song JS. Price, delivery time guarantees and capacity selection. European Journal of Operational Research 1998;111; 28–49.

Stephan FF. Two queues under preemptive priority with poisson arrival and service rates. Operations Research 1958;6; 399–418.

Stidham S. Pricing and capacity decisions for a service facility: Stability and multiple local optima. Management Science 1992;38; 1121–1139.

Tsay AA, Agrawal N. Channel dynamics under price and service competition. Manufacturing and Service Operations Management 2000;2; 372–391.

Wolfe P. The simplex method for quadratic programming. Econometrica 1959;27; 382–398.

Zhang M. Fencing in a revenue management context. Ph.D. Thesis 2007; Richard Ivey School of Business, University of Western Ontario, ON, Canada.

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

Appendix

Appendix A: Solution to [$PDTDP_{DC}$]

It is well known that at optimality, the two delivery time reliability constraints (7^{DC}) and (8^{DC}) must be binding (Palaka et al., 1998; So and Song, 1998; Boyaci and Ray, 2003). This implies that the two service rates will be given by:

$$\mu_i = -\frac{\ln(1-\alpha)}{L_i} + \lambda_i,\ i \in \{h, l\}$$

As a result, [$PDTDP_{DC}$] reduces to maximizing (3) with μ_i as given above. The system stability conditions (6^{DC}) are automatically satisfied by the expressions for μ_i. Upon substituting the expressions for μ_i into (3), and taking it partial derivatives with respect to p_h and p_l gives the following Hessian for a fixed L_h:

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

Clearly, the Hessian is negative definite. This shows that the objective function $\pi(L_h)$ is strictly concave for a fixed L_h, and, therefore, has a unique pair of optimal prices $p_h^{DC*}(L_h)$ and $p_l^{DC*}(L_h)$, which can be obtained by solving the following system of equations:

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

Substituting the optimal prices given by (9) and (10) into the objective function, and differentiating it with respect to L_h gives:

$$\frac{\partial \pi(L_h)}{\partial L_h} = -\left(\beta_L^h + \theta_L\right)\left(p_h^{DC*}(L_h) - m - A\right) + \theta_L\left(p_l^{DC*}(L_h) - m - A\right) - \frac{A\ln(1-\alpha)}{L_h^2} \quad \text{(A1)}$$

$$\frac{\partial^2 \pi(L_h)}{\partial L_h^2} = \frac{(\beta_p^l + \theta_p)(\beta_L^h)^2 + (\beta_p^h + \beta_p^l)(\theta_L)^2 + 2\beta_p^l\beta_L^h\theta_L}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)} + \frac{2A\ln(1-\alpha)}{L_h^3} \quad \text{(A2)}$$

$$\frac{\partial^3 \pi(L_h)}{\partial L_h^3} = -\frac{6A\ln(1-\alpha)}{L_h^4} \quad \text{(A3)}$$

The first three derivatives of $\pi(L_h)$ suggest that it has the following properties: (i) As $L_h \to 0^+$, $\pi(L_h) \to -\infty$. (ii) $\pi(L_h)$ is increasing concave in L_h in the vicinity of $L_h = 0^+$. (iii) As L_h increases from 0, $\pi(L_h)$ changes from concave to convex for some $L_h \in (0, +\infty)$, and never becomes convex again. It is clear from the above properties of $\pi(L_h)$ that it has a unique maximum and at most one minimum in $[0, +\infty)$. The stationary points are given by the roots of (A1) 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 (A1) has only one root in the interval $[0, L_l)$, and that it is the point of maximum. The condition

simplifies to:

$$\frac{-\left\{(\beta_p^l-\beta_p^h)\theta_L+\beta_p^l\beta_L^h+2\beta_L^h\theta_p\right\}a+\left\{(\beta_p^l\beta_L^h+\beta_L^h\theta_p+\beta_L^l\theta_p)\beta_L^h+(\beta_p^l\beta_L^h-\beta_p^h\beta_L^l)\theta_L\right\}L_l}{2(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)}$$
$$+\frac{\beta_L^h(A+m)}{2}-\frac{A\ln(1-\alpha)}{(L_l)^2}<0 \qquad \text{(A4)}$$

Since $\beta_p^h<\beta_p^l$, a necessary condition for (A4) to hold is a to be high. A sufficiently high value of a also guarantees $\lambda_i>0$, $p_i>0$ and $p_h>p_l$.

Appendix B: The Matrix Geometric Method

The 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:

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 $n_l>0$, $n_h=0$

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

$$A_0 = \begin{pmatrix} \lambda_l & & & & \\ & \lambda_l & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_l \end{pmatrix}; \quad A_2 = \begin{pmatrix} \mu & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}; \quad B_0 = \begin{pmatrix} * & \lambda_h & & & \\ \mu & * & \lambda_h & & \\ & \mu & * & \lambda_h & \\ & & \ddots & \ddots & \ddots \\ & & & \mu & * \end{pmatrix}$$

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

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; 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 $M + 1$.

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 of waiting times in phase-type and QBD processes. 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, by $\mathbf{x}_i = \mathbf{x}_0 R^i$. $\overline{S_l^k}(y)$ can be expressed as:

$$\overline{S_l^k}(y) = \sum_{i=0}^{\infty} \mathbf{x}_i \overline{S_{li}^k}(y)\mathbf{e} \tag{B1}$$

$\overline{S_{li}^k}(y)$ 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:

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

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 the 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,

$$\overline{S_{li}^k}(y) = \sum_{n=0}^{\infty} e^{-\gamma y}\frac{(\gamma y)^n}{n!}\sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \qquad i \ge 0 \tag{B2}$$

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 $\overline{S_{li}^k}(y)$ from (B2) into (B1), we obtain:

$$\overline{S_l^k}(y) = \sum_{n=0}^{\infty} d_n e^{-\gamma y}\frac{(\gamma y)^n}{n!} \tag{B3}$$

where, d_n is given by:

$$d_n = \sum_{i=0}^{\infty} \mathbf{x}_0 R^i \sum_{v=0}^{i} G_v^{(n)}\mathbf{e}, \qquad n \ge 0 \tag{B4}$$

Now,

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

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

$$S_l^k(L_l) = 1 - \overline{S_l^k}(L_l) = \sum_{n=0}^{\infty} e^{-\gamma L_l}\frac{(\gamma L_l)^n}{n!}\mathbf{x}_0(I-R)^{-1}H_n\mathbf{e} \tag{B5}$$

H_n can be computed recursively as:

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

Therefore, for given prices (p_h^k, p_l^k) and service rate (μ^k), $S_l^k(\cdot)$ in (16) can be computed using (B5).

Appendix C: Estimation of the Gradient of $S_l(\cdot)$

There are several methods available in the literature to compute the gradients of $S_l(\cdot)$. 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:

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

where dp_h, dp_l and $d\mu$ (referred to as step sizes) are infinitesimal changes in the respective variables.

Appendix D: The 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 1 shows a flowchart of the cutting plane algorithm. The algorithm works as follows: We start with an empty constraint set (20), which

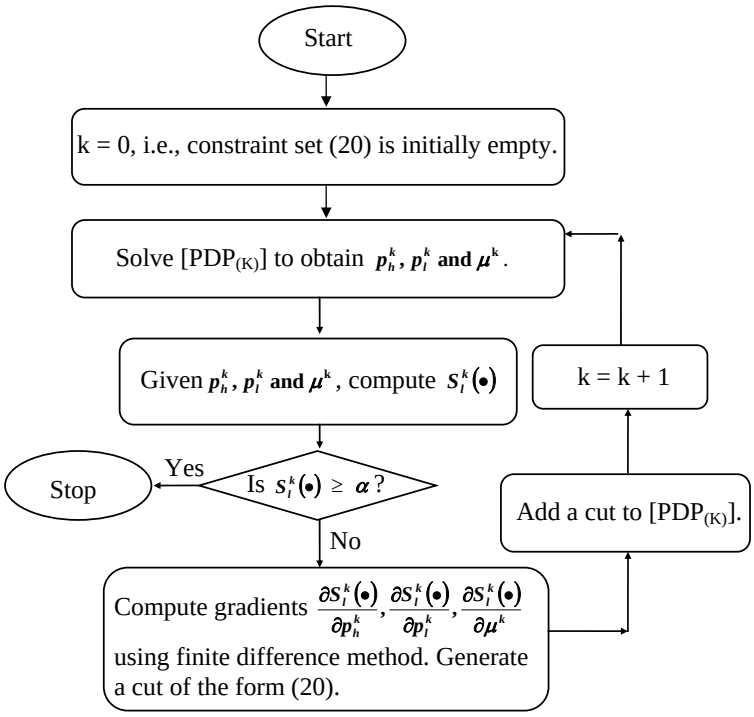


Figure 1: Cutting Plane Algorithm

results in a simple QPP, and obtain an initial solution (p_h^0, p_l^0, μ^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 $[PDP_{(K)}]$, else we add to (16) a linear 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, μ). 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. We used this test in our numerical experiments, which did ensure that the concavity assumption was not violated for any of the cases studied, at least in the region visited by the algorithm. Details of the test can be found in Atlason et al. (2004).

The table below presents the computational performance, in terms of the number of iterations and the CPU time (in seconds), of the algorithm for different combinations of values of L_h and A for $a = 10$, $m = 3$, $\alpha = 0.99$, $L_l = 1$, $\beta_p^h = 0.5$, $\beta_p^l = 0.7$, $\theta_p = 0.2$, $\beta_L^h = 0.9$, $\beta_L^l = 0.7$, $\theta_L = 0.5$, $M = 100$, $\epsilon = 10^{-6}$ and $dp_h = dp_l = d\mu = 0.001$. All computations are performed on a Pentium IV (3.06 GHz, 512 MB RAM) machine.

	A=0.10		A=0.25		A=0.50		A=0.75		A=1.00	
	Cuts	Time	Cuts	Time	Cuts	Time	Cuts	Time	Cuts	Time
L_h=0.10	0	0.25	0	0.2	0	0.23	0	0.22	0	0.25
L_h=0.20	0	0.23	0	0.25	0	0.27	0	0.39	0	0.37
L_h=0.30	0	0.27	0	0.34	0	0.34	0	0.28	0	0.25
L_h=0.40	0	0.37	0	0.31	0	0.31	0	0.27	0	0.31
L_h=0.50	4	7.33	4	7.13	4	7.39	4	7.14	4	7.46
L_h=0.60	5	10.06	5	10.17	5	9.66	5	9.66	5	9.48
L_h=0.70	6	11.9	6	12.21	6	13.38	6	12.98	6	12.04
L_h=0.80	6	12.9	6	13.28	6	12.75	6	13.21	6	12.98
L_h=0.90	7	17.41	7	16.07	7	15.46	7	15.49	7	15.85

Appendix E: Numerical Experiment Design

The different parameter values used in the numerical experiment are:

For a given parameter combination, we repeat the experiment for both the TDS and PDS market

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}

types, as defined in §5.1.

Appendix F: PDP in SC for L_h smaller than L_h^T

$[PDP_{(K)}]$ in absence of constraints (16) is similar in form to the pricing problem with a single class of customers (see Palaka et al. 1998; So and Song 1998; Ray and Jewkes, 2004). For the

resulting problem, it is well known that at optimality, the delivery time reliability constraint (15) for the express class must be binding, which implies:

$$\mu = -\frac{\log(1-\alpha)}{L_h} + a - (\beta_p^h + \theta_p)p_h + \theta_p p_l - (\beta_L^h + \theta_L)L_h + \theta_L L_l$$

As a result, [PDP] reduces to maximizing (14) with μ as given above. This allows the resulting problem to be solved analytically using the same tricks used in Appendix A.

Appendix G: Effects of Product Substitution and Capacity Sharing

The effect of sharing capacity can be explained mathematically for small A. When L_h is small, the optimal prices in SC are given by (21) and (22). Substituting (21) and (22) in $\pi^{SC}(L_h)$, it is easy to verify that $\pi^{SC}(L_h)$ has the same properties as does $\pi^{DC}(L_h)$, given in Appendix A. Hence, for a sufficiently high value of a (see Appendix A), $\pi^{SC}(L_h)$ has a unique maximizer, given by the root of (G1) in the interval $[0, L_l)$:

$$\frac{\partial \pi^{SC}(L_h)}{\partial L_h} = -\left(\beta_L^h + \theta_L\right)\left(p_h^{DC*}(L_h) - m - A\right) + \theta_L\left(p_l^{DC*}(L_h) - m\right) - \frac{A\ln(1-\alpha)}{L_h^2} \qquad \text{(G1)}$$

For small A, the effect of sharing capacity on delivery time differentiation can be explained using the following relation:

$$\left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{SC} - \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{DC} = \frac{\theta_L A}{2} \geq 0 \qquad \text{(G2)}$$

Absence of product substitution ($\theta_p = \theta_L = 0$) implies (G2) $= 0$. 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. Presence of product substitution ($\theta_p > 0$, $\theta_L > 0$), on the other hand, implies (G2) > 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 (G1) > 0, implies that:

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

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 (9) and (21), 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:

$$\begin{aligned}
&\left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{\theta_p,\theta_L>0} - \left.\frac{\partial \pi(L_h)}{\partial L_h}\right|_{\theta_p,\theta_L=0} \\
&= \frac{-\beta_p^l(\beta_p^l-\beta_p^h)(\beta_p^h\theta_L-\beta_L^h\theta_p)a}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)} \\
&+ \frac{\{\beta_p^h\beta_p^l(\beta_p^h+\beta_p^l)(\theta_L)^2+\beta_p^h(\beta_p^l)^2\beta_L^h\theta_L+(\beta_p^l)^2\beta_L^h(\beta_p^h\theta_L-\beta_L^h\theta_p)\}L_h}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)} \\
&- \frac{\{\beta_L^h(\beta_p^l\theta_L-\beta_L^l\theta_p)+(\beta_p^h+\beta_p^l)(\theta_L)^2+\beta_p^h\beta_L^l\theta_L\}L_l}{2\beta_p^h\beta_p^l(\beta_p^h\beta_p^l+\beta_p^h\theta_p+\beta_p^l\theta_p)}
\end{aligned} \tag{G3}$$

A high value of a, which is required for $L_h < L_l$ and $p_h > p_l$, makes (G3) negative (resp., positive) if $\beta_p^h\theta_L - \beta_L^h\theta_p > 0$ (resp., < 0). We have already shown that the profit function is increasing concave until it attains its maximum. Therefore, (G3) < 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.

Appendix H: Effect of Capacity Cost

For a dedicated capacity setting:

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

Since we know that

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

The effect of an increase in the marginal capacity cost on the price differentiation in a dedicated capacity setting is evident from the following equation:

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

For a shared capacity setting, the effect of the marginal capacity cost can be similarly shown analytically for small A.

Further, if the marginal capacity costs are different for the two customer segments in a dedicated

capacity setting ($A_l \leq A_h$), then:

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

Similarly,

$$\frac{\partial L_h^*}{\partial A_l} = -\left(\frac{\partial^2\pi/\partial L_h \partial A_l}{\partial^2\pi/\partial L_h^2}\right)\Big|_{L_h=L_h^*} \quad , \quad \text{where} \quad \frac{\partial^2\pi}{\partial L_h \partial A_l}\Big|_{L_h=L_h^*} = \quad -\frac{\theta_L}{2} < 0.$$

Since we know that

$$\frac{\partial^2\pi}{\partial L_h^2}\Big|_{L_h=L_h^*} < 0 \quad \Rightarrow \quad \frac{\partial L_h^*}{\partial A_h} > 0 \quad \text{and} \quad \frac{\partial L_h^*}{\partial A_l} < 0.$$

If A_h - A_l is constant ($= d$), then:

$$\frac{dL_h^*}{dA_h} = \frac{\partial L_h^*}{\partial A_h} + \frac{\partial L_h^*}{\partial A_l}\frac{\partial A_l}{\partial A_h} = \frac{\partial L_h^*}{\partial A_h} + \frac{\partial L_h^*}{\partial A_l} = -\left(\frac{\beta_L^h/2 - ln(1-\alpha)/(L_h^*)^2}{\partial^2\pi/\partial L_h^2}\right)\Big|_{L_h=L_h^*}$$

Since we know that

$$\frac{\partial^2\pi}{\partial L_h^2}\Big|_{L_h=L_h^*} < 0 \quad \Rightarrow \quad \frac{dL_h^*}{dA_h} > 0.$$

Further,

$$\begin{aligned}\frac{d(p_h^* - p_l^*)}{dA_h} &= \frac{\partial(p_h^* - p_l^*)}{\partial A_h} + \frac{\partial(p_h^* - p_l^*)}{\partial L_h^*}\frac{\partial L_h^*}{\partial A_h} + \frac{\partial(p_h^* - p_l^*)}{\partial A_l}\frac{\partial A_l}{\partial A_h} + \frac{\partial(p_h^* - p_l^*)}{\partial L_h^*}\frac{\partial L_h^*}{\partial A_l}\frac{\partial A_l}{\partial A_h} \\ &= -\frac{\beta_p^l\beta_L^h + \beta_p^h\theta_L + \beta_p^l\theta_L}{2(\beta_p^h\beta_p^l + \beta_p^h\theta_p + \beta_p^l\theta_p)}\frac{dL_h^*}{dA_h} < 0.\end{aligned}$$