This article was downloaded by: [128.100.40.87] On: 30 January 2014, At: 08:25
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Pricing Time-Sensitive Services Based on Realized Performance

Philipp Afèche, Opher Baron, Yoav Kerner,

To cite this article:
Philipp Afèche, Opher Baron, Yoav Kerner, (2013) Pricing Time-Sensitive Services Based on Realized Performance. Manufacturing & Service Operations Management 15(3):492-506. http://dx.doi.org/10.1287/msom.2013.0434

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2013, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

RIGHTSLINK

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 3, Summer 2013, pp. 492–506
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)
http://dx.doi.org/10.1287/msom.2013.0434
© 2013 INFORMS

Pricing Time-Sensitive Services Based on Realized Performance

Philipp Afèche, Opher Baron
Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada
{afeche@rotman.utoronto.ca, opher.baron@rotman.utoronto.ca}

Yoav Kerner
Department of Industrial Engineering and Management, Ben-Gurion University of the Negev,
Beer Sheva 84105, Israel, kerneryo@bgu.ac.il

Services such as FedEx charge up-front fees but reimburse customers for delays. However, lead-time pricing studies ignore such delay refunds. This paper contributes to filling this gap. It studies revenue-maximizing tariffs that depend on realized lead times for a provider serving multiple time-sensitive customer types. We relax two key assumptions of the standard model in the lead-time pricing literature. First, customers may be *risk averse* (RA) with respect to payoff uncertainty, where payoff equals valuation, minus delay cost, minus payment. Second, tariffs may be *arbitrary* functions of realized lead times. The standard model assumes risk-neutral (RN) customers and restricts attention to flat rates. We report three main findings: (1) With RN customers, flat-rate pricing maximizes revenues but leaves customers exposed to payoff variability. (2) With RA customers, flat-rate pricing is suboptimal. If types are distinguishable, the optimal lead-time-dependent tariffs fully insure delay cost risk and yield the same revenue as under optimal flat rates for RN customers. With indistinguishable RA types, the differentiated first-best tariffs may be incentive-compatible even for uniform service, yielding higher revenues than with RN customers. (3) Under price and capacity optimization, lead-time-dependent pricing yields higher profits with less capacity compared to flat-rate pricing.

Key words: delay; lead times; pricing; queueing systems; revenue management; risk aversion
History: Received: December 20, 2011; accepted: January 1, 2013. Published online in *Articles in Advance* May 3, 2013.

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

1. Introduction

Firms in a range of industries sell services and products with an inherent lead time between order placement and delivery. Their customers face *lead-time uncertainty*. Several firms offer tariffs that depend on *realized* lead times: They charge up-front fees but issue refunds for delivery delays. For example, such tariffs are commonly used by transportation carriers and by make-to-order suppliers of critical components. FedEx offers delay refunds, as does Beta LAYOUT, a custom printed-circuit-board supplier with headquarters in Germany. Delay penalty clauses are also common in contracts for construction projects (Friedlander 2001).

Although delay refunds are important in practice, they have so far been ignored in the lead-time pricing literature. This seems to be the first paper to study the rationale for and the design of tariffs that charge based on realized lead times. We also refer to such tariffs as *lead-time-dependent*. We address three fundamental questions for a *revenue-maximizing* service provider that serves time-sensitive customers:

1. Under what conditions can charging based on realized lead time increase revenue?

2. What are the properties of the optimal lead-time-dependent pricing scheme?

3. What is the value of optimal lead-time-dependent pricing?

1.1. Analytical Framework, Main Results, and Contributions

We model the provider as a queueing system. The customer population may comprise multiple types. Customers of the same type may differ in their valuations for instant delivery but have the same delay cost and utility functions. The provider is informed about aggregate demand statistics and designs a static (menu of) price-lead-time tariff(s) to maximize her revenue rate. We consider these tariff design decisions taking the number of price-service classes and the scheduling policy as given. Customers cannot observe the queue length and base their purchase decisions on the posted tariff(s).

This paper reports the following main contributions to the lead-time pricing literature:

1. *Modeling.* We relax two assumptions in the standard model since Naor (1969). (i) The key novelty of our model is that customers may be *risk averse* to delay cost and payment variability. That is, we allow a customer's utility to be concave in her service payoff, which equals her valuation, minus delay cost, minus payment. The standard model assumes risk-neutral customers; that is, they evaluate the cost of service based on the sum of expected delay cost plus payment. (ii) We allow general tariffs whereby a customer's total payment can be an *arbitrary function* of

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

her realized lead time. The standard model restricts attention to schemes that charge one or more *flat rates*; that is, a customer's payment is determined ex ante, at the instant of her purchase decision.

2. *Results.* We obtain novel results on the value and structure of optimal lead-time-dependent pricing: (i) If customers are risk neutral, as in the standard model, charging based on realized lead times has zero value, and flat-rate pricing is optimal. However, flat-rate pricing leaves customers exposed to delay cost risk. In reality, such delay cost risk may concern customers. (ii) If customers are risk averse, flat-rate pricing reduces the system utilization and revenue. If the provider can distinguish among types, then optimal lead-time-dependent tariffs fully protect against delay cost risk and yield the same revenue as under optimal flat rates for risk-neutral customers. (iii) Pricing based on realized lead times can also be an attractive tool for price discrimination. If the provider serves indistinguishable customer types with uniform service, e.g., first-in-first-out (FIFO), a menu of differentiated tariffs can have a positive value. Even the differentiated first-best tariff set may be incentive-compatible, yielding higher revenues than with risk-neutral customers. In contrast, with risk-neutral customers, a differentiated menu of tariffs has no value under uniform service. (iv) The simplest practical refund policy, which issues a full refund for late delivery, performs well relative to the optimal lead-time-dependent tariff. (v) Under joint pricing and capacity optimization, optimal pricing based on the realized lead time yields higher profits with less capacity compared to flat-rate pricing. The profit gain can be significant, particularly if the capacity cost is significant.

Our model and results provide some theoretical support for customer risk aversion as one reason for the use of lead-time-dependent pricing in practice. These findings also suggest that it is critical for providers to understand customer preferences with respect to delay cost and payment risk.

1.2. Literature and Positioning

We categorize the lead-time management literature into three streams—*operations*, *information*, and *pricing*—based on the levers used for managing lead times.

Operations levers focus on managing lead times through capacity, admission control, routing, sequencing, and expediting. This stream includes Baker (1984), Wein (1991), Duenyas (1995), Duenyas and Hopp (1995), Spearman and Zhang (1999), Plambeck et al. (2001), Harrison (2003), Plambeck (2004), Ho and Zheng (2004), Keskinocak and Tayur (2004), and Shang and Liu (2011).

Information levers focus on managing customer expectations and behavior by quoting lead times or waiting times, and by releasing information on factors like queue lengths that affect lead times. This stream includes Hassin (1986), Whitt (1999), Armony and Maglaras (2004), Dobson and Pinker (2006), Guo and Zipkin (2007), Armony et al. (2009), and Allon et al. (2011).

Pricing levers focus on regulating the total demand rate and customers' service class choices. Papers in this stream are closest to ours. See Hassin and Haviv (2003) for an excellent survey. As noted above, our model has two distinctive features: it captures customer risk aversion with respect to delay costs and payments, and it allows general price tariffs. Some papers assume customers with nonlinear delay cost functions in the standard model (e.g., Dewan and Mendelson 1990, Van Mieghem 2000, Kittsteiner and Moldovanu 2005, Ata and Olsen 2009, Bansal and Maglaras 2009b, Kumar and Randhawa 2010). In these cases, customers are not risk neutral with respect to lead-time uncertainty, but still risk neutral with respect to the resulting delay cost variability. Risk considerations are absent in aggregate demand models that capture arrival rates as decreasing functions of prices and lead times, where each price is a flat rate (e.g., So and Song 1998, Boyaci and Ray 2003, Charnsirisakskul et al. 2006, Allon and Federgruen 2007, Çelik and Maglaras 2008). In some models, purchase decisions do not explicitly depend on lead-time variability, only on the quoted lead times and flat rates, but the provider has a strong incentive to keep lead-time variability small and incurs the cost of managing the system accordingly. In So and Song (1998), the provider has to build enough safety capacity to meet an exogenous lead-time reliability constraint. In Charnsirisakskul et al. (2006), Çelik and Maglaras (2008), and Feng et al. (2011), the provider incurs early/late delivery or expediting costs when actual lead times deviate from quoted ones; these costs are exogenous and do not affect customers' purchase decisions, unlike in our setting where tariffs may specify delay discounts that customers consider in their purchase decisions.

The price flexibility in tariffs that depend on realized lead times is also different from that in price-service differentiation and in dynamic (state-dependent) pricing. In price-service differentiation, the provider offers a menu of flat rates, each for a different service class and based on some lead-time statistic for that class (e.g., Mendelson and Whang 1990, Maglaras and Zeevi 2005). In dynamic pricing, the flat rate fluctuates over time, for example, based on the queue length (e.g., Low 1974, Chen and Frank 2001, Çelik and Maglaras 2008, Ata and Olsen 2009, Feng et al. 2011). In these settings, different flat rates reflect performance fluctuations across service classes or across consecutive customers, but unlike in ours,

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

each customer knows her payment exactly when she makes her purchase decision.

A few studies show that it can be optimal to depart from flat-rate pricing by charging customers based on their realized processing (or service) time. Doing so either allows the provider to manipulate customers' service class or service rate choices (Mendelson and Whang 1990; Hassin 1995; Ha 1998, 2001; Kittsteiner and Moldovanu 2005), or to benefit from spending more time with customers than necessary (Debo et al. 2008). In these papers, departing from flat-rate pricing is optimal only because one party is ex ante better informed about, and/or has control over, customers' processing times; see §3 for details. In contrast, in our setting all parties are equally informed about processing times, which are exogenous, and charges are based on the entire realized lead times in response to customers' delay cost risk considerations.

This paper is also related to pricing studies without queueing in which customers are uncertain about a component of their utility when they make their purchase decisions. A number of papers consider the design of pricing contracts with refunds in advance-purchase situations where customers learn their valuations over time, for example, Courty and Hao (2000), Gallego and Sahin (2010), and Akan et al. (2013). Liu and van Ryzin (2008) and Bansal and Maglaras (2009a) consider risk-averse customers in settings with uncertain product availability.

Delay refund contracts can be viewed as a form of insurance. As such this paper is also related to the economics literature on insurance for risk-averse agents. Rothschild and Stiglitz (1976) and Stiglitz (1977) are seminal studies for competitive and monopoly markets, respectively. See Landsberger and Meilijson (1999) for a general model of insurance under adverse selection.

1.3. Plan of the Paper

In §2 we specify and discuss the model. In §3 we study under what conditions charging based on realized lead time can increase revenue and the properties of the optimal lead-time-dependent pricing schemes. In §4 we study the value of optimal lead-time-dependent pricing, first for fixed capacity and then under joint price and capacity optimization. In §5 we provide concluding remarks. Proofs are in the online supplement, available at http://dx.doi.org/10.1287/msom.2013.0434.

2. Model

We model a capacitated provider that serves delay-sensitive customers as a queueing system with well-defined moments of the steady-state lead-time distributions. We use the terms "lead time" and "delay" interchangeably; both refer to the entire time interval between order placement and delivery, that is, the system sojourn time including waiting and time in service. Except in §4.2, we study a system with fixed processing capacity. When considering the capacity explicitly, we denote it by μ. Potential customers have unit demand and arrive according to an exogenous stationary stochastic process with a finite rate Λ. The provider is risk neutral and makes static price and lead-time decisions to maximize her long-run average revenue rate. We say "optimal" to mean revenue-maximizing (or profit-maximizing when capacity is a decision variable). In contrast, we consider customers that are risk averse and maximize their expected utility given the posted information. It is standard to assume that both the provider and the customers are risk neutral. That providers are risk neutral seems plausible because they typically serve a significant volume of customers. However, customers may not be risk neutral, as noted in §1 and further discussed below.

Customers have independent and identically distributed (i.i.d.) processing requirements, unless specified otherwise. Service time realizations become known only once processing is completed. We normalize the marginal cost of serving a customer to zero. The population of potential customers may consist of one or more types. Each customer is characterized by three attributes: a *valuation*, a *delay cost* function, and a *utility* function. Customers of the same type may differ in their valuations, but they have the same delay cost and utility functions. The provider may offer one or more *price-service classes*. We use "type" and "class" in reference to a customer group and a price-service option, respectively. Each class has two attributes: a *price function* and a *lead-time distribution*.

Below we first formalize the problem for the basic single-type, single-class model and then outline how it extends to multiple types and/or classes. We next discuss key features of our model and how it relates to the standard model and to pricing schemes in practice.

2.1. One Type, One Class

In this case customers only differ in their valuations, which are continuous, nonnegative i.i.d. random variables with a continuous, strictly positive probability density function f. Let F denote the cumulative distribution function (c.d.f.), $\bar{F} = 1 - F$, and $\bar{F}^{-1}$ be its inverse. If all customers with valuation higher than the marginal valuation $\underline{v}$ decide to buy, then their arrival rate is $\lambda(\underline{v}) := \Lambda\bar{F}(\underline{v})$. We also call λ the demand rate. Conversely, the marginal value function $\underline{v}(\lambda) := \bar{F}^{-1}(\lambda/\Lambda)$ maps arrival rates to marginal valuations. A customer with valuation v who experiences lead time w has *net valuation* $v - C(w)$, where

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

the delay cost function $C\colon \mathbb{R}_+ \to \mathbb{R}$ is increasing with $C(0) = 0$. It captures the opportunity cost and/or the diminished value due to delay. The *payoff* from service for a customer with valuation v and lead time w is $v - C(w) - P(w)$, where $P\colon \mathbb{R}_+ \to \mathbb{R}$ is an arbitrary price function or *tariff* chosen by the provider and $P(w)$ is the customer's *payment*. The *full price* equals delay cost plus payment, so payoff equals valuation minus full price. Customers base their decisions on the utility of their payoff. A customer with payoff $v - C(w) - P(w)$ has utility

$$U(v - C(w) - P(w)),$$

where U is an increasing and (weakly) concave utility function with $U(0) = 0$. We call customers with $U(X) = X$ risk neutral (RN) and those with strictly concave utility U risk averse (RA).

Because of lead-time variability, a customer's full price is uncertain at the instant of her purchase decision. Let W denote the steady-state lead time. Given the capacity, the scheduling policy, and the statistical properties of the arrival and service processes, the distribution of W only depends on the arrival rate λ. We write $W(\lambda)$ when making this dependence explicit. For example, in a FIFO $M/M/1$ queue with service rate μ the distribution of $W(\lambda)$ is exponential with parameter $\mu - \lambda$.

The provider does not know individual customers' valuations but is informed about aggregate demand characteristics, that is, the valuation distribution F, the delay cost function C, the utility function U, the rate Λ, and the statistical properties of the arrival and service processes. Based on this information and the relationship between λ and $W(\lambda)$, the provider chooses and announces a price function P and a distribution of W, taking into account the resulting purchase decisions and arrival rate λ. Customers cannot observe the queue length and evaluate their payoff distribution based on the announced tariff P and distribution of W. Customers with valuation v buy if and only if their expected utility $\mathrm{E}[U(v - C(W) - P(W))]$ is nonnegative. Purchase decisions are irrevocable; that is, we assume no reneging or retrials. We require that the announced distribution of W matches the distribution of $W(\lambda)$, that is, the actual steady-state lead-time distribution given the resulting arrival rate. This requirement captures the notion that reputation effects and third-party auditors commit the provider to perform in line with her announcements.

The provider solves the revenue-maximization problem

$$\max_{\lambda, P} \ \lambda\, \mathrm{E}\big[P(W(\lambda))\big] \tag{1}$$

$$\text{s.t.} \ \ \mathrm{E}\big[U\big(\underline{v}(\lambda) - C(W(\lambda)) - P(W(\lambda))\big)\big] = 0. \tag{2}$$

The demand relationship (2) requires that for any price function P and corresponding equilibrium arrival rate λ, customers with marginal valuation $\underline{v}(\lambda)$ have zero expected utility. For a given $W(\lambda)$, the expected utility $\mathrm{E}[U(v - C(W(\lambda)) - P(W(\lambda)))]$ strictly increases in v, so (2) ensures that customers buy if and only if their valuation exceeds $\underline{v}(\lambda)$, and it rules out suboptimal pricing that leaves all customers with strictly positive expected utility if $\lambda = \Lambda$.

Constant Absolute Risk Aversion (CARA) and Linear Delay Costs. Our fundamental structural results hold for any RA customers. For more specific results, in §§3.3 and 4, we assume exponential (CARA) utility functions, given by $U(X) = 1 - \exp(-rX)$, $r > 0$, and linear delay costs $C(W) = cW$. In this case, (2) yields

$$1 - \exp(-r\,\underline{v}(\lambda))\, \mathrm{E}\big[\exp r\big(cW(\lambda) + P(W(\lambda))\big)\big] = 0. \tag{3}$$

Consider the linear tariff $P(W) = \alpha - \beta W$, with α, $\beta \geq 0$ constants. Let $\tilde{W}(\lambda, s) := \mathrm{E}[\exp(sW(\lambda))]$ and $L(\lambda, s) := \ln \tilde{W}(\lambda, s)$ denote, respectively, the moment-generating function (MGF) and the semi-invariant MGF of the random variable $W(\lambda)$, evaluated at s. By (3), the equilibrium arrival rate λ satisfies

$$\begin{aligned} \alpha &= \underline{v}(\lambda) - \frac{\ln(\mathrm{E}[\exp(r(c-\beta)W(\lambda))])}{r} \\ &= \underline{v}(\lambda) - \frac{L(\lambda, r(c-\beta))}{r}. \end{aligned} \tag{4}$$

That is, if the provider announces the tariff $P(W) = \alpha - \beta W$ and the distribution of $W(\lambda)$, then λ satisfies (4). Note that $\alpha + L(\lambda, r(c-\beta))/r$ is the *certainty equivalent* (CE) of the full price, that is, customers are indifferent between paying this certain amount and the random full price $\alpha + (c - \beta)W(\lambda)$. Similarly, $L(\lambda, r(c-\beta))/r$ is the CE of the net delay cost $(c-\beta)W(\lambda)$. For a FIFO $M/M/1$ queue with service rate μ,

$$\frac{L(\lambda, r(c-\beta))}{r} = \ln\left(\frac{\mu - \lambda}{\mu - \lambda - r(c-\beta)}\right).$$

2.2. Multiple Types and/or Classes

In cases with more than one type and/or more than one price-service class, we specify whether the provider can distinguish among types and how problem (1)–(2) generalizes. We index customer types by $i \in \{1, 2, \ldots, N\}$. The functions $\underline{v}_i$, C_i, and U_i specify the type i attributes as explained above. An arrival is of type i with probability Λ_i/Λ, where Λ_i is the arrival rate of potential type i customers. Let λ_i denote the type i arrival rate, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $\lambda = \sum_{i=1}^N \lambda_i$. We index price-service classes by $k \in \{1, 2, \ldots, K\}$, where W_k denotes the steady-state lead time of class k and $P_k(W_k)$ its price function.

2.3. Discussion

2.3.1. Risk-Averse Customers. The key modeling novelty is that we let demand be sensitive to both

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

delay cost and payment variability by modeling customers as risk averse with respect to their payoff.

The standard model assumes that customers are risk neutral with respect to their payoff, so $U(X) = X$, and pay a flat rate p, so $P(W) = p$. In evaluating the cost of service, they only consider its expected full price: $\mathrm{E}[C(W)] + p$. A customer with valuation v makes a purchase if her expected payoff is nonnegative, i.e., $v \geq \mathrm{E}[C(W)] + p$. However, because of lead-time variability, customers face delay cost risk. A customer's ex post payoff may be lower than her ex ante expectation and even negative. Specifically, for the marginal customer whose valuation equals the expected full price, the ex post payoff is negative whenever the realized delay cost exceeds its mean. For example, under linear delay costs, this event occurs whenever the realized lead time exceeds its mean; in an $M/M/1$ queue with FIFO service, this probability equals $1/e \approx 0.37$. Obviously, customers who end up with lower than expected, or even negative, payoff are less satisfied. In reality, such delay cost risk may concern customers and motivate providers to compensate them based on their actual delays, particularly if losses due to delay costs can be significant. (See Holt and Laury 2002 for experimental evidence of risk aversion even with low stakes.) For example, in commercial shipments, construction projects, and the procurement of critical components, delays can translate into considerable financial losses for customers. Providers in these industries commonly offer contracts that specify compensation payments for delays. However, by assuming that customers are indifferent between any two tariffs with the same expected full price, the standard model ignores these delay cost risk concerns. This limitation calls for a model that fits settings where customers are sensitive to full price risk, that is, to delay cost variability and to how much they pay as a function of their ex post delay cost. Our model with risk-averse customers provides a natural framework for such settings and subsumes the standard model as a special case.

2.3.2. Delay Cost Structure and Risk Neutrality in the Standard Model. Some papers study the standard model with delay cost functions that are convex (Dewan and Mendelson 1990, Van Mieghem 2000, Kittsteiner and Moldovanu 2005, Kumar and Randhawa 2010) or with convex-concave delay costs that capture sensitivity to deadlines (Ata and Olsen 2009, Bansal and Maglaras 2009b). In the standard model, customers with nonlinear delay costs are not risk neutral with respect to *lead-time uncertainty* but are still risk neutral "with respect to money," that is, the resulting delay *cost* variability; they are indifferent between any two delay cost distributions that have the same mean. For illustration, consider the deadline delay cost structure $C(W) = c \cdot I\{W > \bar{w}\}$, where I denotes the indicator function; that is, the delay cost $c > 0$ is incurred only if the lead time exceeds the deadline $\bar{w} > 0$. In this sense, customers are *satisficers* with respect to their delay (Bansal and Maglaras 2009b). In this case $\mathrm{E}[C(W)] = c\,\Pr\{W > \bar{w}\}$. In the standard model, customers are indifferent between any two tariffs that yield the same mean payment, for example, a flat rate p and the lead-time-sensitive tariff $P(W) = p + c \cdot \Pr\{W > \bar{w}\} - c \cdot I\{W > \bar{w}\}$ that charges more than p if delivery is on time and less than p if it is late. The mean payment is p and the mean full price is $p + c\,\Pr\{W > \bar{w}\}$ for both tariffs, but only the lead-time-sensitive tariff eliminates full price variability.

2.3.3. Price and Lead-Time Quotation. The assumption that the provider announces an arbitrary price function and a lead-time distribution describes a generalized price and lead-time quotation model. There are clear parallels between our model, the standard model, and current schemes in practice. In general, customers may not need the entire lead-time distribution to evaluate their expected utility; the information they require depends on the tariff and on their delay cost structure and risk preferences. In the standard model with RN customers, flat-rate pricing, and linear delay costs, customers only need to know the expected lead time. Lead-time-sensitive pricing schemes in practice typically specify a target lead time, a regular price for "on-time" delivery/completion, and a schedule of refund payments as a function of the delay. For example, a number of transportation providers offer such contracts. The simplest contracts specify a full refund for late delivery. Package carriers like FedEx and UPS offer such contracts for their express delivery services, as do certain less-than-truckload (LTL) carriers (Bohman 2003). In such cases, customers with a corresponding deadline delay cost structure only need to know the *on-time probability*. More sophisticated delay refund schedules, for example, for ocean freight or construction projects, specify two or more refund levels as a function of the delay, on a time scale of days, hours, or even minutes (Friedlander 2001). In such cases, customers may require more information on the lead-time distribution to forecast their expected utility. On-time performance and related lead-time statistics are typically published by the carriers themselves but are also increasingly available from third-party providers of information, auditing, and/or refund claim processing services. For example, the company PackageFox (packagefox.com) sells such services to FedEx and UPS express delivery customers and releases on-time performance statistics. With growing competitive pressures and the proliferation of sophisticated IT solutions, customers are gaining access to increasingly

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

detailed and up-to-date information on lead-time distributions. For example, the maritime shipping analyst SeaIntel and the ocean cargo technology provider INTTRA recently launched a monthly schedule reliability report that provides detailed container delivery time statistics for each major carrier and port–port combination (Burnson 2012). The increasing availability of detailed lead-time forecasts and the proliferation of third-party tracking/auditing services make it increasingly manageable for customers to evaluate their expected utility before a purchase, verify their actual lead times ex post, and enforce contracts by collecting delay refunds. To summarize, our model framework is general enough to accommodate a range of delay refund schemes that are found in practice.

3. The Optimal Lead-Time-Dependent Pricing

In this section, we address the first two questions posed in the introduction: Under what conditions can charging based on realized lead time increase revenue? What are the properties of the optimal lead-time-dependent pricing scheme? We start with RN customers in §3.1. We then consider distinguishable and indistinguishable RA customer types, in §§3.2 and 3.3, respectively.

3.1. Standard Model: RN Customers

We have the following revenue equivalence result:

Proposition 1. *Suppose that customers are RN. Then the following holds for any given number of price-service classes and any scheduling policy*:

1. *If customers have i.i.d. service requirements, the maximum expected revenue rate over all price functions can be attained by charging for each price-service class k a flat rate* $P_k(W_k) = p_k$.

2. *If customer types differ in their service requirements, then part* 1 *holds under the restriction that the provider can distinguish among types.*

By Proposition 1, pricing independently of the realized lead times entails no revenue loss in the standard model with RN customers. However, charging flat rates in the presence of lead-time variability exposes customers to full price risk—their delay cost is ex ante uncertain and their ex post payoff may be negative. The optimal flat rates moderate full price variability by controlling the utilization and the resulting lead-time variability. Still, whatever the variability level and the delay cost structure, with flat rates, the probability and/or magnitude of negative payoff realizations may be significant for some customers. The provider could reimburse customers for long lead times to eliminate their full price risk, but by Proposition 1, it has no incentive to do so if they are RN.

The distinction between parts 1 and 2 of Proposition 1 is important. If types have different service requirements, the lead-time distribution of a given service class may vary by type. If types are indistinguishable, the maximum attainable revenue over all tariffs may not be attainable by charging a flat rate for each class. In particular, ensuring incentive-compatibility at the optimal arrival rates may only be feasible by charging based on the realized *processing* (i.e., service) time.

Part 2 of Proposition 1 is related to a few exceptions in the literature when it is optimal for the provider to charge based on the realized processing time, in contrast to our tariffs that depend on the entire lead time, including the time in queue. In these papers, unlike in ours, the rationale for departing from flat-rate pricing is that one party is ex ante better informed about and/or has control over processing times. Mendelson and Whang (1990) characterize the welfare-maximizing priority pricing mechanism for a multiclass $M/M/1$ queue serving multiple indistinguishable RN types with type-dependent service time distributions. In their setting, service-time-dependent tariffs may be optimal to deter customers with long jobs from buying a class targeted to customers with shorter jobs. Similar results are given in Hassin (1995) and Kittsteiner and Moldovanu (2005) for priority auctions in queues with privately informed customers that have heterogeneous service requirements. In Ha (1998, 2001), customers choose their service requirements. The optimal tariffs include a component that depends on the realized processing time. In Debo et al. (2008), customers arrive to a visible FIFO queue of an expert who controls the service time. Under certain conditions, the expert may benefit from increasing the service time and charge customers per hour.

3.2. Distinguishable RA Customer Types

In this section, we show that charging based on realized lead times has positive value in the case of RA customers, unlike in the RN case. We also address the second question posed in the introduction: What are the properties of the optimal lead-time-dependent pricing scheme? We consider the case of distinguishable customer types in this section and that of indistinguishable types in §3.3. Whether the provider can distinguish among types depends on the characteristics of its customer base and its services/products. For example, a firm may be able to distinguish among types based on their location or if they are identifiable as residential versus business versus government customers. Firms may also be able to distinguish among types if their preferences are correlated with product attributes; for example, a firm that sells lower- and higher-value products may know that customers

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

who buy higher-value products are more time sensitive and risk averse with respect to shipping delays.

Proposition 2 characterizes the optimal lead-time-dependent pricing scheme for distinguishable customer types. We also say *first-best* to mean optimal in this case. The case of distinguishable types subsumes the special case of a single type and serves as a benchmark for that of indistinguishable types. From the provider's perspective, the key benefit of being able to distinguish among types is that she can limit each type to a single, targeted price-service class. Proposition 2 therefore identifies the tariff structure that is optimal in the absence of customer choice among classes.

Proposition 2. *Suppose that the provider sells service to N distinguishable customer types, serving all customers of the same type with the same price-service class. Given the provider's scheduling policy, let $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*)$ be a revenue-maximizing demand vector if all customer types were RN and distinguishable, i.e., $\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}} \sum_{i=1}^N \lambda_i \cdot (\underline{v}_i(\lambda_i) - \mathrm{E}[C_i(W_i(\boldsymbol{\lambda}))])$.*

1. *The optimal type i tariff P_i^* charges a fixed (lead-time-independent) up-front fee, equal to that type's marginal valuation, minus a lead-time-dependent discount, equal to its delay cost*:

$$P_i^*(W_i) = \underline{v}_i(\lambda_i^*) - C_i(W_i). \tag{5}$$

If type i is RA, then this optimal price function is unique.

2. *Under the optimal price functions* (5)*, all types' arrival rates and expected payments, and the provider's maximum expected revenue rate, are the same as if all types were RN.*

When customers are risk averse, it is optimal for the firm to charge based on realized performance. Under the optimal tariff in (5), the provider internalizes the delay cost of customers, which eliminates their payoff risk. The first-best tariffs for RA customers are therefore independent of risk-aversion levels and yield the same optimal arrival rates and revenue as if customers were RN.

The optimal tariff structure in (5), with a fixed up-front fee and a lead-time-sensitive refund schedule that matches the structure of customer delay costs, is invariant to the number of types and the operational characteristics of the system. These properties only affect the revenue-maximizing demand vector $\boldsymbol{\lambda}^*$, and the resulting up-front fees and lead-time distributions. Since delay costs increase in lead times, a customer's payment under the optimal tariff decreases in her realized lead time. For example, for the deadline delay cost structure $C(W) = c \cdot I\{W > \bar{w}\}$, the optimal tariff refunds the amount c for late delivery, similar to the simplest delay refund policies in practice.

Table 1 compares, for a single type, the mean and standard deviation of payments and full prices under optimal flat-rate pricing for RN customers and optimal lead-time-dependent pricing for RA customers. The delay cost variability is borne by customers when they are RN but by the provider when customers are RA. As a result, under the optimal tariff in (5), all customers have nonnegative ex post utility, but the provider has (potentially unlimited) liability for delays.

Table 1 Optimal Payments and Full Prices: RN vs. RA Customers (One Type, $\lambda^* = \boldsymbol{\lambda}^*$, $\underline{v}^* := \underline{v}^*(\lambda^*)$, $W^* := W(\lambda^*)$)

	Payment $P(W)$		Full price $P(W) + C(W)$	
Customers	RN	RA	RN	RA
Amount	$\underline{v}^* - \mathrm{E}[C(W^*)]$	$\underline{v}^* - C(W^*)$	$\underline{v}^* - \mathrm{E}[C(W^*)] + C(W^*)$	$\underline{v}^*$
Mean	$\underline{v}^* - \mathrm{E}[C(W^*)]$	$\underline{v}^* - \mathrm{E}[C(W^*)]$	$\underline{v}^*$	$\underline{v}^*$
Std. dev.	0	stdev$[C(W^*)]$	stdev$[C(W^*)]$	0

3.3. Two Indistinguishable RA Customer Types

The model with distinguishable RN types is a natural benchmark since the first-best revenue is independent of risk aversion. The first-best revenue is generally not attainable if the provider cannot distinguish types. In this section, we study the effect of this information constraint on the optimal lead-time-dependent pricing scheme. We focus on a system with uniform service (e.g., FIFO) and consider two indistinguishable types with CARA utility and linear delay costs, assuming without loss of generality that $c_1 > c_2$; that is, type 1 customers are more impatient. We start with *uniform pricing*, that is, a single tariff, and then consider *differentiated pricing* through a menu of tariffs.

3.3.1. Uniform Pricing: One Linear Tariff. For simplicity, the provider may offer a single lead-time-dependent tariff. Proposition 3 characterizes the optimal tariff with a linear delay refund.

Proposition 3. *Consider a system with uniform service for two indistinguishable RA customer types, with linear delay costs $c_1 > c_2$ and CARA utility functions. Suppose the provider offers a single linear tariff $P(W) = \alpha - \beta W$ and it is optimal to serve some, but not all, customers of each type. Let α^* and β^* be the optimal tariff parameters and $\underline{v}_i^*$ the resulting marginal type i valuation.*

1. *If $r_1, r_2 > 0$, the up-front fee exceeds the marginal valuation of the patient type and is lower than that of the impatient type, $\underline{v}_2^* < \alpha^* < \underline{v}_1^*$, and the delay discount rate exceeds the delay cost rate of the patient type and is lower than that of the impatient type, $c_2 < \beta^* < c_1$.*

2. *If $r_i > r_j = 0$, it is optimal to eliminate the RA type's delay cost risk: $\alpha^* = \underline{v}_i^*$ and $\beta^* = c_i$.*

Uniform pricing leaves customers with some delay cost risk. If both types are RA (part 1 of Proposition 3), the presence of type j with different delay sensitivity $c_j \neq c_i$ pulls β closer to c_j. Setting β outside the

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

interval (c_2, c_1) is suboptimal because it reduces the expected payment that both types are willing to pay. Under the optimal tariff with $\beta^* \in (c_2, c_1)$, the impatient type 1 customers have positive utility for instant service, but their full price increases and their utility decreases in lead time. By contrast, every patient customer's utility *increases* and her full price decreases in lead time; those with valuation larger than the upfront fee α^* have positive utility for every lead time, whereas the utility of those with lower valuation is negative for instant service and positive only at sufficiently long lead times. If one type is RN, as in part 2 of Proposition 3, its expected payment is invariant to β, so it is optimal to eliminate the delay cost risk of the other, RA type.

3.3.2. Differentiated Pricing: Incentive-Compatibility of the First-Best Tariff Set. Charging a single linear tariff is appealing for simplicity, but differentiated pricing through a menu of tariffs may generate more revenue. The first-best tariff set generates the maximum revenue among all tariff menus, which raises the question: Can the first-best tariff set be incentive-compatible (IC), and if so, under what conditions? It is well known that under uniform service the answer is generally negative for RN customers; that is, all tariffs that are selected by some customers must have the same expected payment. We show that the answer is positive for RA customers, and we shed light on the relationship to the RN case. We first characterize the optimality conditions of the first-best problem and then the conditions for the first-best solution to be IC.

Optimality Conditions of the First-Best Solution. By Proposition 2, for distinguishable customer types with linear delay costs, the optimal tariff set satisfies $P_i^*(W) = \alpha_i^* - \beta_i^* W$ for $i = 1, 2$, where $\beta_i^* = c_i$, $\alpha_i^* = \underline{v}_i(\lambda_i^*)$, and λ_i^* is the first-best arrival rate of type i customers. The first-best demand vector $\boldsymbol{\lambda}^*$ is the solution of

$$\max_{\boldsymbol{\lambda}} \sum_{i=1,2} \lambda_i \big(\underline{v}_i(\lambda_i) - c_i \,\mathrm{E}[W(\lambda)]\big) \tag{6}$$

$$\text{s.t. } 0 \le \lambda_i \le \Lambda_i, \quad i = 1, 2, \tag{7}$$

$$\lambda_1 + \lambda_2 < \mu, \tag{8}$$

where $\lambda = \lambda_1 + \lambda_2$.

For simplicity, we assume that it is optimal to serve some, but not all, customers of each type. Since it cannot be optimal to operate at capacity and the objective function is strictly concave, the following first-order conditions (FOC) for the first-best demand vector are necessary and sufficient:

$$\begin{aligned} &\underline{v}_1(\lambda_1^*) - c_1 \,\mathrm{E}[W(\lambda^*)] + \lambda_1^* \underline{v}_1'(\lambda_1^*) \\ &\quad = \underline{v}_2(\lambda_2^*) - c_2 \,\mathrm{E}[W(\lambda^*)] + \lambda_2^* \underline{v}_2'(\lambda_2^*) \\ &\quad = \sum_{i=1,2} \lambda_i^* c_i \frac{d\,\mathrm{E}[W(\lambda^*)]}{d\lambda}. \end{aligned} \tag{9}$$

Incentive-Compatibility of the First-Best Tariff Set. Suppose that the provider announces a menu of two linear tariffs $P_i(W) = \alpha_i - \beta_i W$ for $i = 1, 2$, and a distribution of W that is consistent with the lead-time distribution given the resulting arrival rates. Then, customers of each type purchase the service at the tariff targeted to their type, if and only if

$$\alpha_i = \underline{v}_i(\lambda_i) - \frac{L(\lambda, r_i(c_i - \beta_i))}{r_i}, \quad i = 1, 2; \tag{10}$$

that is, the demand relationship (4) holds for each tariff (which ensures individual rationality), and

$$\alpha_i + \frac{L(\lambda, r_i(c_i - \beta_i))}{r_i} \le \alpha_j + \frac{L(\lambda, r_i(c_i - \beta_j))}{r_i}, \quad i \ne j \in \{1, 2\}, \tag{11}$$

which ensure incentive-compatibility. By (11), a type i customer prefers the type i tariff (α_i, β_i) only if her CE of the full price under this tariff (the left-hand side) does not exceed her CE of the full price for the type j tariff (the right-hand side). Substituting for α_1 and α_2 from (10) into (11) yields the incentive-compatibility constraints in a form that depends only on $\boldsymbol{\lambda}$ and the delay discount rates:

$$\begin{aligned} &\frac{L(\lambda, r_1(c_1 - \beta_1))}{r_1} - \frac{L(\lambda, r_2(c_2 - \beta_1))}{r_2} \\ &\quad \le \underline{v}_1(\lambda_1) - \underline{v}_2(\lambda_2) \\ &\quad \le \frac{L(\lambda, r_1(c_1 - \beta_2))}{r_1} - \frac{L(\lambda, r_2(c_2 - \beta_2))}{r_2}. \end{aligned} \tag{12}$$

The types' marginal valuation difference is bounded by the difference in the CEs of their net delay costs under the type 1 tariff (the left-hand side) and under the type 2 tariff (the right-hand side). The first-best tariff set is IC if and only if the demand vector $\boldsymbol{\lambda}^*$ that solves (9) satisfies (12) for $\beta_1 = c_1$ and $\beta_2 = c_2$.

In the standard model with RN customers, differentiated pricing has no value under uniform service because any two IC price functions yield the same expected payment. For RN customers, the CE of the net delay cost of any tariff equals its mean:

$$\lim_{r_i \to 0} \frac{L(\lambda, r_i(c_i - \beta_j))}{r_i} = (c_i - \beta_j)\,\mathrm{E}[W(\lambda)], \quad \text{for all } i, j$$

(see part 3 of Lemma 1 in the online supplement). This implies from (10) that the expected payment of the tariff chosen by type i must equal its expected marginal net value

$$\alpha_i - \beta_i \,\mathrm{E}[W(\lambda)] = \underline{v}_i(\lambda_i) - c_i \,\mathrm{E}[W(\lambda)],$$

and from (12), the types' marginal valuation difference equals their expected delay cost difference:

$$\underline{v}_1(\lambda_1) - \underline{v}_2(\lambda_2) = (c_1 - c_2)\,\mathrm{E}[W(\lambda)].$$

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

As a result, both tariffs must yield the same mean payment: $\alpha_1 - \beta_1 E(W(\lambda)) = \alpha_2 - \beta_2 E(W(\lambda))$.

Proposition 4 proves that, unlike in the standard model, with RA customers differentiated pricing can have positive value even under uniform service. In fact, the provider may be able to implement the differentiated first-best tariff set. Furthermore, when customer types are indistinguishable, risk aversion allows the provider to extract more revenue than with RN customers.

Proposition 4. *Consider a system with uniform service for two indistinguishable RA customer types with linear delay costs $c_1 > c_2$ and CARA utility with r_1, r_2. Let $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*)$ be the first-best demand vector and suppose that $\lambda_i^* \in (0, \Lambda_i)$. By Proposition 2 the first-best tariff set is $P_i^*(W) = \alpha_i^* - \beta_i^* W$, with $\alpha_i^* = \underline{v}_i(\lambda_i^*)$ and $\beta_i^* = c_i$, for $i = 1, 2$. Let $p_i^* := \alpha_i^* - \beta_i^* \mathrm{E}[W(\lambda^*)]$, $i = 1, 2$.*

1. *If $\alpha_1^* \le \alpha_2^*$, the first-best tariff set is not IC for any risk-aversion levels.*

2. *If $p_1^* = p_2^*$, the first-best tariff set is IC for all risk-aversion levels.*

3. *If $\alpha_1^* > \alpha_2^*$ and $p_1^* \ne p_2^*$, the first-best tariff set is IC if and only if the type with the higher mean payment is sufficiently risk averse: If $p_i^* > p_j^*$, there is $\underline{r} \in (0, \infty)$ such that type i chooses tariff i if and only if $r_i \ge \underline{r}$; type j chooses tariff j regardless of her risk preference.*

Proposition 4 implies that, even in cases where the first-best tariff set is not IC, offering a menu of two linear lead-time-dependent tariffs may yield strictly higher revenue than the single linear tariff characterized in Proposition 3. The results of Proposition 4 have the following intuition.

In part 1, if $\alpha_1^* \le \alpha_2^*$, the tariff targeted to the impatient type 1 customers both charges the lower up-front fee and refunds the higher delay discount rate since $\beta_1^* = c_1 > \beta_2^* = c_2$. The patient type 2 customers therefore prefer the type 1 tariff regardless of their risk-aversion level, so the first-best tariff cannot be IC. By inspection of the FOC (9), part 1 may apply, for example, if congestion effects are minor; that is, the terms involving c_i are small. For illustration, if both types have uniform valuations on $[0, \bar{v}_i]$, then $\underline{v}_i(\lambda_i) = \bar{v}_i(1 - \lambda_i/\Lambda_i)$. If congestion effects are minor and there is enough capacity, then the FOC (9) imply

$$\underline{v}_i(\lambda_i^*) + \lambda_i^* \underline{v}_i'(\lambda_i^*) = \bar{v}_i(1 - 2\lambda_i^*/\Lambda_i) \approx 0, \quad i = 1, 2,$$

so that it is optimal to serve roughly the top half of each type, i.e., $\lambda_i^* \approx \Lambda_i/2$ and $\underline{v}_i(\lambda_i^*) \approx \bar{v}_i/2$. Whenever the impatient type 1 customers have the lower maximum valuation, i.e., $\bar{v}_1 < \bar{v}_2$, we have $\alpha_1^* < a_2^*$, and the first-best solution is not IC for any risk-aversion levels.

In part 2, in the exceptional case where $p_1^* = p_2^*$, the first-best tariff set is clearly IC for RN customers because they are indifferent between any two tariffs with the same mean payment. Uniform pricing with a single flat rate also attains the first-best revenue for RN customers, but flat-rate pricing is suboptimal for RA customers as established in this paper. Attaining the first-best revenue for RA customers requires the differentiated first-best tariff set; to see why it is IC for all risk-aversion levels, consider a type 1 customer's choice. Her CE of the type 1 tariff full price equals the up-front fee α_1^* regardless of her risk aversion, because this tariff eliminates type 1 customers' delay cost risk. By contrast, her CE of the type 2 tariff full price increases in her risk aversion. Type 1 customers are therefore indifferent between the tariffs if they are RN but otherwise prefer the type 1 tariff:

$$\alpha_1^* = \alpha_2^* + (c_1 - c_2)\,\mathrm{E}[W(\lambda^*)] \le \alpha_2^* + \frac{L(\lambda^*, r_1(c_1 - c_2))}{r_1},$$

where the equation follows since $p_1^* = p_2^*$ and the inequality is strict if and only if $r_1 > 0$. Similar reasoning explains why type 2 customers prefer their targeted tariff regardless of their risk attitude.

From (9), $p_1^* = p_2^*$ holds if and only if $\lambda_1^* \underline{v}_1'(\lambda_1^*) = \lambda_2^* \underline{v}_2'(\lambda_2^*)$. This means that at the first-best arrival rates, both types have the same ratio of marginal valuation to elasticity, where the elasticity function of the type i marginal value function is $\varepsilon_i(\lambda_i) = -\underline{v}_i(\lambda_i)/(\lambda_i \underline{v}_i'(\lambda_i))$. For example, consider exponential valuations with c.d.f. $F_i(v) = 1 - \exp(-k_i v)$ for $v \ge 0$, where $k_i > 0$. The marginal valuation function $\underline{v}_i(\lambda_i) = \ln(\Lambda_i/\lambda_i)/k_i$ and $\lambda_i \underline{v}_i'(\lambda_i) = -1/k_i$, so $p_1^* = p_2^*$ if and only if $k_1 = k_2$.

In part 3, suppose that $\alpha_1^* > \alpha_2^*$ and $p_1^* < p_2^*$. Then type 1 customers choose their targeted tariff because

$$\alpha_1^* < \alpha_2^* + (c_1 - c_2)\,\mathrm{E}[W(\lambda^*)] \le \alpha_2^* + \frac{L(\lambda^*, r_1(c_1 - c_2))}{r_1}.$$

If type 2 customers are RN, they clearly prefer the type 1 tariff because it yields the lower mean payment. The type 1 tariff charges the higher up-front fee but also refunds the larger discount. The more risk averse type 2 customers are, the less they value the larger discount. If they are sufficiently risk averse, they prefer the type 2 tariff because it eliminates all delay cost risk. Similar intuition applies if $\alpha_1^* > \alpha_2^*$ and $p_1^* > p_2^*$.

4. The Value of Optimal Lead-Time-Dependent Pricing

By Propositions 1 and 2, flat-rate pricing is optimal if and only if customers are RN. Flat-rate pricing is also practically appealing because of its simplicity and because it frees the provider of liability for delays. Providers that serve RA customers must weigh these practical benefits of flat-rate pricing against the revenue gains of optimal lead-time-dependent pricing.

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

This raises the third question posed in the introduction: What is the value of optimal lead-time-dependent pricing? In this section, we study this question for two settings. In §4.1 we compare the performance of a system with *fixed capacity* under the optimal lead-time-dependent tariff against its performance under optimal flat-rate pricing and under the simplest practical delay refund policy, which we call the *simple refund policy*. In §4.2 we consider the interplay between pricing and operations by comparing the *optimal capacity* and performance under optimal lead-time-dependent versus optimal flat-rate pricing. We focus on a system with uniform service (e.g., FIFO) that serves a single RA type with CARA utility and linear delay cost.

4.1. Performance vs. Flat-Rate and Simple Refund Policies: Fixed Capacity

By Proposition 2, for a linear delay cost $C(W) = cW$, the optimal lead-time-dependent tariff is given by $P^*(W) = \underline{v}(\lambda^*) - cW$, where λ^* is the revenue-maximizing demand rate for RN customers. For comparison with flat-rate pricing, consider linear tariffs of the form $P(W) = \alpha - \beta W$. For the optimal lead-time-dependent tariff, $\beta = c$. For flat-rate pricing, $\beta = 0$ and the provider chooses only α. The simple refund policy charges α for on-time delivery by a threshold lead time $\bar{w}$ and issues a full refund for late delivery, i.e., $P(W) = \alpha - \alpha \cdot I\{W > \bar{w}\}$, and the provider chooses α and $\bar{w}$.

We compare lead-time-dependent versus flat-rate pricing analytically, show that the revenues of these tariffs bound the revenue of the simple refund policy, and compare the three tariffs numerically.

4.1.1. Optimal Flat-Rate Pricing. By (4), the demand relationship for $P(W) = \alpha - \beta W$ is

$$\alpha = \underline{v}(\lambda) - \frac{L(\lambda, r(c-\beta))}{r}. \tag{13}$$

Let $\Pi(\lambda)$ be the revenue function under the optimal lead-time-dependent tariff. For this tariff $\beta = c$, so by (13) the expected payment as a function of λ is $\alpha - \beta\,\mathrm{E}[W(\lambda)] = \underline{v}(\lambda) - c\,\mathrm{E}[W(\lambda)]$. The provider solves

$$\max_{\lambda} \Pi(\lambda) = \lambda\big(\underline{v}(\lambda) - c\,\mathrm{E}[W(\lambda)]\big). \tag{14}$$

Let $\Pi^f(\lambda; r)$ be the revenue function under flat-rate pricing, where r expresses the dependence on risk aversion. In this case $\beta = 0$, so by (13) the flat rate as a function of λ is $\underline{v}(\lambda) - L(\lambda, rc)/r$. The provider solves

$$\max_{\lambda} \Pi^f(\lambda; r) = \lambda\left(\underline{v}(\lambda) - \frac{L(\lambda, rc)}{r}\right). \tag{15}$$

Let $\lambda^* := \arg\max_\lambda \Pi(\lambda)$ denote the optimal arrival rate under optimal pricing, $\Pi^* := \Pi(\lambda^*)$ the optimal revenue, and $P^*(W) := \alpha^* - cW$ the optimal price function, where $\alpha^* := \underline{v}(\lambda^*)$. Because the optimal price function eliminates customers' payoff risk, these quantities are independent of r. Let $\lambda^f(r) := \arg\max_\lambda \Pi^f(\lambda; r)$ denote the optimal arrival rate under flat-rate pricing and $\Pi^f(r) := \Pi^f(\lambda^f(r); r)$ the corresponding optimal revenue. The optimal flat rate is

$$\alpha^f(r) := \underline{v}(\lambda^f(r)) - \frac{L(\lambda^f(r), rc)}{r}. \tag{16}$$

For analytical convenience, we make the following mild technical assumptions. (We write g_x and g_{xy} for the first- and second-order partial derivatives of a bivariate function $g(x, y)$.)

Assumption A1. $\Pi'(0) > 0 > \lim_{\lambda \to \min(\mu, \Lambda)} \Pi'(\lambda)$. *This ensures an interior solution under optimal pricing.*

Assumption A2. *The functions* $\Pi(\lambda)$ *and* $\Pi^f(\lambda; r)$ *are strictly concave in* λ.

Assumption A3. $L_\lambda(\lambda, s)/s$ *increases in* s, *which ensures that* $\Pi^f_{\lambda r}(\lambda; r) < 0$.

The online supplement details sufficient conditions for Assumptions A2 and A3.

Proposition 5. *Suppose that the provider only charges a flat rate, and there is a single RA customer type with CARA utility and linear delay costs* $C(W) = cW$.

1. *For* $r > 0$, *the arrival rate and the revenue under the optimal flat rate* $\alpha^f(r)$ *are lower than under the optimal price function* $P^*(W) = \underline{v}(\lambda^*) - cW$: $\lambda^f(r) < \lambda^*$ *and* $\Pi^f(r) < \Pi^*$. *Moreover,* $\lim_{r\to 0} \lambda^f(r) = \lambda^*$, $\lim_{r\to 0} \Pi^f(r) = \Pi^*$, *and* $\lim_{r\to 0} \alpha^f(r) < \alpha^*$.

2. *The arrival rate* $\lambda^f(r)$ *and the revenue* $\Pi^f(r)$ *under the optimal flat rate are strictly positive and decreasing in* r *if* $r < \bar{r}$, *and they equal zero if* $r \geq \bar{r}$, *where*

$$0 < \bar{r} = \arg\left\{r \geq 0\colon \frac{L(0, rc)}{r} = \underline{v}(0)\right\} \quad \textit{and}$$
$$\bar{r} < \infty \quad \textit{if } \underline{v}(0) < \infty. \tag{17}$$

3. *The optimal flat rate* $\alpha^f(r)$ *need not be monotone in* r *and satisfies* $\lim_{r\to\bar{r}} \alpha^f(r) = 0$.

By Proposition 2, the optimal lead-time-dependent tariff eliminates customers' full price risk. In contrast, under flat-rate pricing, customers face some full price risk, so at every congestion level the flat rate is lower than the mean payment under the optimal lead-time-dependent tariff. Flat-rate pricing therefore yields a lower optimal utilization and revenue. Without the flexibility to offer a delay refund, the provider can lower full price variability only indirectly, by lowering delay cost variability, that is, decreasing utilization. At the extreme, for RA levels above the threshold $\bar{r}$, customers are not willing to pay a positive flat fee at any utilization, so it is unprofitable to operate the system under flat-rate pricing, even though the system is profitable under the optimal tariff.

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

4.1.2. Simple Refund Policy. Under the simplest lead-time-dependent tariff found in practice, customers receive a full refund if their actual lead time exceeds the quoted threshold. By Proposition 2, for RA customers this delay refund structure is optimal only if it mirrors their delay cost structure. If customer delay costs have a different structure, for example, linear as in this section, this policy is suboptimal because it insures only some of their delay cost risk. However, even in such cases this refund policy is practically appealing. For one, it is simple to implement because the firm only sets two controls: the price α and the lead-time threshold $\bar{w}$. Moreover, this policy limits providers' liability for delays while giving them the flexibility to insure delay cost risk at least partially.

Let $\Pi^s(\lambda, \bar{w}; r)$ be the revenue under the simple refund policy as a function of the arrival rate λ and the lead-time quote $\bar{w}$ for a given RA parameter r. Let $\alpha^s(\lambda, \bar{w}; r)$ denote the price as a function of λ and $\bar{w}$, which is determined from the demand relationship (3). The provider solves

$$\max_{\lambda, \bar{w}} \Pi^s(\lambda, \bar{w}; r) = \lambda \alpha^s(\lambda, \bar{w}; r) \Pr\{W(\lambda) \le \bar{w}\}, \quad (18)$$

where $\Pr\{W(\lambda) \le \bar{w}\}$ is the on-time probability. The expected payment equals that under the flat-rate policy as $\bar{w} \to \infty$, i.e., $\lim_{\bar{w}\to\infty} \alpha^s(\lambda, \bar{w}; r) \Pr\{W(\lambda) \le \bar{w}\} = \underline{v}(\lambda) - L(\lambda, rc)/r$. The revenues under the optimal lead-time-dependent tariff, the simple refund policy, and flat-rate pricing satisfy

$$\Pi(\lambda) \ge \Pi^s\big(\lambda, \bar{w}^*(\lambda; r); r\big) \ge \Pi^f(\lambda; r), \quad (19)$$

where $\bar{w}^*(\lambda; r)$ maximizes the expected payment $\alpha^s(\lambda, \bar{w}; r) \Pr\{W(\lambda) \le \bar{w}\}$ for fixed λ. The first inequality in (19) holds by Proposition 2; the second holds because the simple refund policy generalizes the flat-rate contract. The revenue ranking in (19) is intuitive: For given utilization and lead-time variability, the expected payment is higher the more the provider shares customers' delay cost risk.

4.1.3. Numerical Comparison. We illustrate Proposition 5 and the simple refund policy with a numerical example. Specifically, we compare the performance of the optimal lead-time-dependent tariff with that of optimal flat-rate pricing and the optimal simple refund policy.

Example 1. Consider an $M/M/1$ queue with capacity $\mu = 5$ and market size $\Lambda = 10$. The value distribution is uniform on $[0, 5]$, and the delay cost rate is $c = 1$. Recall from Proposition 2 that the optimal lead-time-dependent tariff is independent of the RA parameter r. The optimal arrival rate $\lambda^* = 3.3$ solves (14), where $\mathrm{E}[W(\lambda)] = 1/(\mu - \lambda)$ for the $M/M/1$ queue, the optimal tariff is $P^*(W) = \underline{v}^* - cW = 3.35 - W$, and the optimal profit $\Pi^* = 9.1$. Under flat-rate pricing, the optimal arrival rate solves (15), where $L(\lambda, rc) = \ln((\mu - \lambda)/(\mu - \lambda - rc))$ for the $M/M/1$ queue, and the RA threshold in (17) is $\bar{r} \approx 5$, where $\bar{r} < \mu/c$. Under the simple refund policy the optimal arrival rate and lead-time quote solve (18), where for the $M/M/1$ queue

$$\alpha^s(\lambda, \bar{w}; r) = \frac{1}{r} \ln\left(\frac{\exp(r\,\underline{v}(\lambda))/\tilde{W}(\lambda, rc) - \exp\{-\bar{w}(\mu - \lambda - rc)\}}{1 - \exp\{-\bar{w}(\mu - \lambda - rc)\}} \right) \quad (20)$$

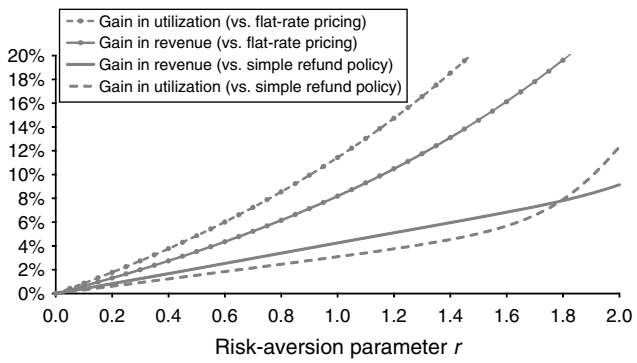
from (3),

$$\tilde{W}(\lambda, rc) = \frac{\mu - \lambda}{\mu - \lambda - rc}, \quad \text{and}$$

$$\Pr\{W(\lambda) \le \bar{w}\} = 1 - \exp(-\bar{w}(\mu - \lambda)).$$

By Proposition 5 the utilization and revenue under the optimal flat rate are lower than under the optimal lead-time-dependent tariff, and they decrease in the risk-aversion level. We observe these losses also under the optimal simple refund policy, but because it partially insures delay cost risk, they are not as large as under flat-rate pricing. Figure 1 shows for $r \in [0, 2]$ the percentage gains in revenue and utilization under the optimal lead-time-dependent tariff, relative to the optimal flat-rate and simple refund policies. Two observations stand out. First, these gains increase considerably in risk aversion, particularly relative to flat-rate pricing. (As we show in §4.2, even modest revenue gains when capacity is fixed can translate into significantly larger profit gains under capacity optimization.) Second, the simple refund policy performs well relative to optimal lead-time-dependent

Figure 1 **Percentage Gains in Revenue and Utilization Under the Optimal Lead-Time-Dependent Tariff, Relative to Optimal Flat-Rate Pricing and the Optimal Simple Refund Policy, as Functions of RA Parameter r**

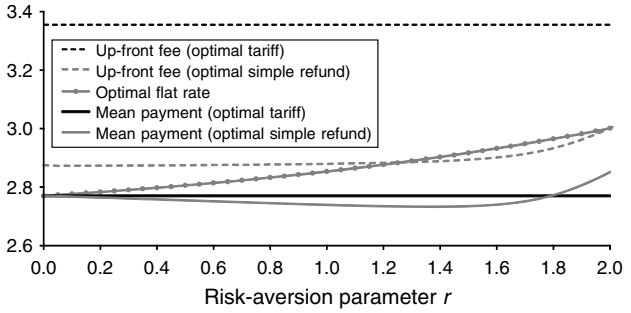


Note. $M/M/1$ queue with $\mu = 5$, $\Lambda = 10$, CARA utility, linear delay cost rate $c = 1$, and valuations $\sim U[0, 5]$.

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

Figure 2 Price Metrics Under the Optimal Lead-Time-Dependent Tariff, the Optimal Simple Refund Policy, and the Optimal Flat Rate, as Functions of RA Parameter r



Note. $M/M/1$ queue with $\mu = 5$, $\Lambda = 10$, CARA utility, linear delay cost rate $c = 1$, and valuations $\sim U[0, 5]$.

pricing and significantly better than flat-rate pricing. Given its practical benefits, the simple refund policy may therefore be the most attractive of the three tariffs.

The differences in utilization among the tariffs imply only minor lead-time performance differences. Specifically, the optimal lead-time quotes of the simple refund policy yield on-time probabilities of 94% or higher. Relative to these lead-time quotes, service levels are similar under the optimal flat rate (up to 2.8% higher) and the optimal lead-time-dependent tariff (up to 3.4% lower).

Figure 2 shows the price metrics for the three tariffs, for $r \in [0, 2]$. The two lead-time-dependent tariffs yield lower average and more variable payments, compared with flat-rate pricing. For $r < 1.2$, their up-front fees exceed the optimal flat rate. Increasing risk aversion has two countervailing effects on the optimal flat rate and the mean payment for the optimal simple refund policy. It yields a lower utilization, which increases these prices, but it also reduces the willingness to pay at any given utilization, which decreases these prices. Under flat-rate pricing the reduced-utilization effect dominates and the optimal flat rate increases in $r \in [0, 2]$, but by part 3 of Proposition 5, it eventually drops to zero as $r \to \bar{r}$. Under the simple refund policy, the mean payment initially decreases in r because the utilization loss is only significant at higher RA levels (see Figure 1).

4.2. Performance vs. Flat-Rate Pricing Under Capacity Optimization

The analysis has so far focused on pricing for a given capacity level. In this case, the provider reduces customers' payoff risk directly through the tariff structure, but the provider controls their delay cost risk from lead-time variability only indirectly by reducing demand and utilization. In this section, we discuss joint pricing and capacity decision. We compare the optimal capacity level and performance under the optimal lead-time-dependent tariff with these measures under the optimal flat rate.

For convenience, we consider a single-server queue and denote its capacity by μ. The results for the multiple-server case are similar. Let $\Pi^*(\mu)$ and $\Pi^{f*}(\mu)$ be, respectively, the maximum revenue as a function of capacity under the optimal lead-time-dependent tariff and the optimal flat rate. Recall that $\Pi^*(\mu)$ is independent of the RA parameter r (Proposition 2), whereas $\Pi^{f*}(\mu)$ depends on r (Proposition 5). Let $\mu^*(b) := \arg\max_{\mu \geq 0}(\Pi^*(\mu) - b\mu)$ and $\mu^{f*}(b) := \arg\max_{\mu \geq 0}(\Pi^{f*}(\mu) - b\mu)$, where $b\mu$ is the capacity cost per unit time and $b > 0$.

Proposition 6. *Consider a single-server system with linear capacity cost rate $b\mu$ and a single RA customer type with CARA utility and linear delay costs $C(W) = cW$. Let $\underline{v}(0) < \infty$.*

1. *Under the optimal lead-time-dependent tariff, the following holds*:

(a) *There is a threshold $\underline{\mu} > 0$ such that $\Pi^*(\mu) = 0$ for $\mu \leq \underline{\mu}$. Furthermore,*

$$\lim_{\mu \to \infty} \Pi^*(\mu) = \max_{\lambda \in [0, \Lambda]} \lambda \underline{v}(\lambda) \quad \textit{and}$$

$$\lim_{\mu \to \underline{\mu}} \Pi^{*\prime}(\mu) = 0 = \lim_{\mu \to \infty} \Pi^{*\prime}(\mu).$$

(b) *There is a threshold $\bar{b} \in (0, \infty)$ such that $\max_{\mu \geq 0}(\Pi^*(\mu) - b\mu) > 0$ if and only if $b < \bar{b}$. The optimal capacity decreases in b, $\mu^*(b) > \underline{\mu}$ for $b < \bar{b}$, and $\mu^*(b) = 0$ for $b > \bar{b}$.*

2. *Under the optimal flat rate, the following holds for any $r > 0$:*

(a) *There is a threshold $\underline{\mu}^f > \underline{\mu}$ such that $\Pi^{f*}(\mu) = 0$ for $\mu \leq \underline{\mu}^f$. Furthermore,*

$$\lim_{\mu \to \infty} \Pi^{f*}(\mu) = \lim_{\mu \to \infty} \Pi^*(\mu) \quad \textit{and}$$

$$\lim_{\mu \to \underline{\mu}^f} \Pi^{f*\prime}(\mu) = 0 = \lim_{\mu \to \infty} \Pi^{f*\prime}(\mu).$$

(b) *There is a threshold $\bar{b}^f \in (0, \bar{b})$ such that $\max_{\mu \geq 0}(\Pi^{f*}(\mu) - b\mu) > 0$ if and only if $b < \bar{b}^f$. The optimal capacity decreases in b, $\mu^{f*}(b) > \underline{\mu}^f$ for $b < \bar{b}^f$, and $\mu^{f*}(b) = 0$ for $b > \bar{b}^f$.*

Optimal lead-time-dependent pricing naturally yields a higher profit compared with optimal flat-rate pricing. By Proposition 6, the capacity cost at which the system just breaks even is higher ($\bar{b}^f < \bar{b}$). With either tariff the system exhibits scale economies and negative profits at small capacity levels. However, Proposition 6 suggests that these scale economies are weaker under optimal lead-time-dependent versus flat-rate pricing. Specifically, the system requires less capacity to generate positive revenue ($\underline{\mu} < \underline{\mu}^f$) and profit, and the revenue gain vanishes for ample

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

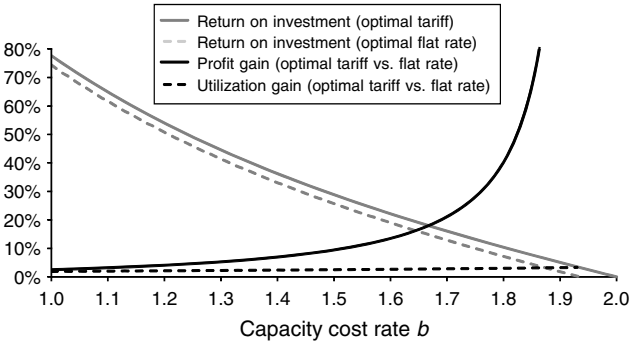
capacity. As a result, the marginal value of capacity under optimal lead-time-dependent versus flat-rate pricing is larger initially, at lower capacity levels, but lower eventually, at larger capacity levels (since $\lim_{\mu\to\infty}\Pi^{f*}(\mu) = \lim_{\mu\to\infty}\Pi^{*}(\mu)$). Intuitively, offering delay refunds and reducing delay cost variability are substitutable means to reduce customers' full price risk. Under flat-rate pricing, the provider can reduce this risk only by lowering the delay cost variability, that is, increasing capacity. Indeed, as discussed in Example 2, we find that optimal lead-time-dependent pricing yields a higher return on a lower optimal capacity investment compared with optimal flat-rate pricing.

The performance differences identified in Proposition 6 increase in risk aversion; that is, under optimal flat-rate pricing, the minimum capacity threshold $\underline{\mu}^f$ for positive revenue increases in r, and the cost threshold $\bar{b}^f$ for profitable operation decreases in r, as does the optimal profit. Proposition 6 generalizes in a natural way for any increasing and convex capacity cost function.

Example 2. We illustrate Proposition 6 numerically for an $M/M/1$ queue with $\Lambda = 10$, uniformly distributed valuations on $[0, 5]$, and $c = 1$ (as in Example 1). These parameters yield $\bar{b} = 2$ for the break-even capacity cost under optimal lead-time-dependent pricing (part 1(b) of Proposition 6). For $b \in (0, 2]$, we compute the jointly optimal pricing and capacity controls under the optimal lead-time-dependent tariff (which is independent of r by Proposition 2) and under optimal flat-rate pricing for risk-aversion levels $r = 0.2$, 1, 2. We find that flat-rate pricing yields a slightly higher arrival rate and revenue in some cases, but these differences are insignificant—the two tariffs yield approximately identical arrival rates, (expected) payments, and revenues, in contrast to the fixed capacity case (part 1 of Proposition 5). The key observation is that *optimal lead-time-dependent pricing results in a lower optimal capacity level compared with flat-rate pricing*, which implies a higher utilization and return on capacity investment (ROI) and a lower service level, that is, longer lead times. Specifically, as noted above, at the capacity $\mu^*(b)$, which is optimal under lead-time-dependent pricing, the marginal value of capacity under flat-rate pricing exceeds the marginal capacity cost. (Numerically, we find that $\Pi^{f*\prime}(\mu) > \Pi^{*\prime}(\mu)$ for all μ where $\Pi^{f*}(\mu)$ is concave.)

Figure 3 shows these profit and utilization gains for $r = 0.2$ and $b \in [1, 2]$, and the ROI for each tariff as a scale-free reference point of profitability. As b drops below 1, the ROI and optimal capacity levels grow excessively large and the difference between the tariffs vanishes. (For example, for $b = 0.5$, the ROI ≈ 200 for both tariffs and the profit gain of lead-time-dependent pricing $\approx 0.7\%$. As $b \to 0$, both tariffs operate with ample capacity and identical profits.) The key observation from Figure 3 is that the profit gain from optimal lead-time-dependent pricing can be quite significant for a small percentage gain in utilization. The larger the capacity cost, the tighter capacity and the higher this profit gain. For $r = 0.2$, the break-even threshold $\bar{b}^f = 1.94$ under optimal flat-rate pricing (part 2(b) of Proposition 6). Finally, these profit gains increase in risk aversion; for example, for capacity cost $b = 1$, the profit gains are 2.5%, 15.2%, and 39.6%, for $r = 0.2$, $r = 1$, and $r = 2$, respectively. (The threshold $\bar{b}^f$ equals 1.71 and 1.49 for $r = 1$ and $r = 2$, respectively.)

Figure 3 **ROI, and Percentage Gains in Profit and Utilization Under the Optimal Lead-Time-Dependent Tariff Relative to Optimal Flat-Rate Pricing, as Functions of the Capacity Cost Parameter b**



Note. $M/M/1$ queue with $\Lambda = 10$, CARA utility with $r = 0.2$, linear delay cost rate $c = 1$, and valuations $\sim U[0, 5]$.

5. Concluding Remarks

We show that if customers face lead-time variability and are risk averse with respect to delay cost and payment variability, tariffs that depend on realized lead times outperform the flat-rate pricing schemes that are standard throughout the lead-time pricing literature. Our model and results provide some theoretical support for the use of such lead-time-dependent tariffs in practice, and they suggest that their benefits can be significant, particularly under joint pricing and capacity optimization. We provide novel insights on how to structure such tariffs; refer to §1 for a summary. These results are general in that they hold for any system with lead-time variability.

Our findings also suggest that it is critical for providers to understand customer preferences with respect to delay cost and payment variability. As such, this paper points to the value of empirical research on customer risk preferences in queueing settings. Both the degree of risk aversion and the specific form of risk preferences are ultimately empirical questions.

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

Our results raise further questions involving pricing, operational, and information controls.

In terms of pricing, more work is needed on tariff design under important practical constraints.

For example, to limit complexity and provider liability there may be constraints on monetary transfers between providers and customers. The simple refund policy studied in §4.1 represents the simplest lead-time-dependent tariff with limited liability. For this pricing policy, Example 1 illustrates for a single tariff how transfer constraints reduce performance because they leave customers exposed to delay cost risk, although the simple refund policy performs well relative to the optimal lead-time-dependent tariff (with unlimited provider liability) and considerably better than flat-rate pricing. Similar analyses are of interest for different kinds of transfer constraints; for example, each payment must exceed an exogenous minimum amount or a percentage of the up-front fee. Another important issue is the impact of transfer constraints on a menu of tariffs. For example, FedEx faces this issue. Quite likely, the simple refund tariffs it offers do not exactly match customer delay costs, and FedEx does not know individual customers' preferences. Proposition 4 shows that the first-best menu of tariffs may be IC in the absence of transfer constraints. Under what types of transfer constraints does this result still hold? More generally, how do such constraints affect the first-best menu and the distortion and performance loss under the second-best menu?

A related issue is to consider constraints on customers' ex post utility. It is important to highlight that under the first-best tariffs (Propositions 2 and 4) all customers have nonnegative ex post utility, in contrast to the standard model with flat-rate pricing. Lead-time-dependent tariffs that insure delay costs only imperfectly (e.g., the simple refund policy, or uniform pricing for two types as in Proposition 3) leave customers exposed to some risk of negative ex post utility (although this risk is smaller compared with flat-rate pricing), which raises the question: How should tariffs be modified if customers can cancel their orders to ensure nonnegative ex post utility?

Limits on the rationality and enforcement abilities of the contracting parties may also constrain tariff design. On the one hand, the simple refund policy discussed in §4.1 exemplifies a common tariff that alleviates these implementation issues through simplicity, without sacrificing much performance relative to the optimal tariff. On the other hand, as discussed in §2.3.3, the increasing availability of detailed lead-time forecasts and the proliferation of third-party services make it possible to manage and enforce increasingly sophisticated contracts. Nevertheless, customers may not be able to accurately forecast their expected utility, for example, due to insufficient lead-time information or due to their bounded rationality. The design of tariffs under such constraints is an important emerging research issue. Huang et al. (2013) are the first to model bounded rationality in a queueing system; they assume RN customers, flat-rate pricing, and linear delay costs.

Another interesting research direction is to consider tariff design for other demand models, for example, by considering bivariate utility functions of lead times and payments that may capture different risk attitudes toward lead-time variability and payment variability.

Proposition 6 and Example 2 show that capacity and lead-time-dependent pricing can be viewed as substitutes. There are more opportunities for research on the interplay between operations and pricing. For example, suppose that a firm charges flat rates to risk-averse customers. Which strategy yields the larger improvement in profitability and under what conditions: switching from flat rates to performance-sensitive tariffs or offering differentiated flat rates and priority service?

Our analysis also raises questions on the interplay between pricing and the delay information available to customers. We assume that customers do not have real-time delay information. Giving customers such information reduces the coefficient of variation of their conditional lead-time distribution. For example, in the observable $M/M/1$ queue, the coefficient of variation of the waiting time when the queue length is n equals $1/\sqrt{n}$, so the significance of lead-time variability decreases in the queue length. This suggests that for RA customers with real-time delay information, the provider may benefit from dynamic pricing policies with a workload-dependent tariff structure, for example, by charging flat rates for longer queues and based on realized lead times for shorter queues.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0434.

Acknowledgments

The authors are grateful to Refael Hassin, Ignatius J. Horstmann, Costis Maglaras, Xuanming Su, the associate editor, and the referees for constructive comments that helped improve this paper significantly. This research was partly supported by the Natural Sciences and Engineering Research Council of Canada. Some of this research was conducted while Opher Baron was visiting the Faculty of Industrial Engineering and Management at the Technion–Israel Institute of Technology and while Yoav Kerner was a postdoctoral fellow at the Rotman School of Management.

References

Akan M, Ata B, Dana J (2013) Revenue management by sequential screening. Working paper, Northwestern University, Evanston, IL.

RIGHTSLINK

Downloaded from informs.org by [128.100.40.87] on 30 January 2014, at 08:25 . For personal use only, all rights reserved.

Allon G, Federgruen A (2007) Competition in service industries. *Oper. Res.* 55:37–55.

Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59:1382–1394.

Armony M, Maglaras C (2004) Contact center with a call-back option and real-time delay information. *Oper. Res.* 52:527–545.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonments. *Oper. Res.* 57:66–81.

Ata B, Olsen TL (2009) Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Oper. Res.* 57:753–768.

Baker K (1984) Sequencing rules and due-date assignments in a job shop. *Management Sci.* 30:1093–1104.

Bansal M, Maglaras C (2009a) Dynamic pricing when customers strategically time their purchase: Asymptotic optimality of a two-price policy. *J. Revenue Pricing Management* 8:42–66.

Bansal M, Maglaras C (2009b) Product design in a market with satisficing customers. Netessine S, Tang CS, eds. *Consumer-Driven Demand and Operations Models* (Springer, New York), 37–62.

Bohman R (2003) A new twist in guaranteed LTL services. *Logist. Management* 42(10):23.

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

Burnson R (2012) Ocean cargo carrier performance gaps outlined in new report. *Logist. Management* (July 18), http://www.logisticsmgmt.com/.

Çelik S, Maglaras C (2008) Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Sci.* 54:1132–1146.

Charnsirisakskul K, Griffin PM, Keskinocak P (2006) Pricing and scheduling decisions with leadtime flexibility. *Eur. J. Oper. Res.* 171:153–169.

Chen H, Frank M (2001) State dependent pricing with a queue. *IIE Trans* 33:847–860.

Courty P, Hao L (2000) Sequential screening. *Rev. Econom. Stud.* 67:697–717.

Debo L, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54:1497–1512.

Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. *Management Sci.* 36:1502–1517.

Dobson G, Pinker EJ (2006) The value of sharing lead time information. *IIE Trans.* 38:171–183.

Duenyas I (1995) Single facility due date setting with multiple customer classes. *Management Sci.* 41:608–619.

Duenyas I, Hopp WJ (1995) Quoting customer lead times. *Management Sci.* 41:43–57.

Feng J, Liu L, Liu X (2011) An optimal policy for joint dynamic price and lead-time quotation. *Oper. Res.* 59:1523–1527.

Friedlander MC (2001) Law and practice: Contractor marketing—Penalty clauses. Retrieved September 2012, http://www.schiffhardin.com/binary/law_and_practice.pdf.

Gallego G, Sahin O (2010) Revenue management with partially refundable fares. *Oper. Res.* 58:817–833.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53:962–970.

Ha A (1998) Incentive-compatible pricing for a service facility with joint production and congestion externality. *Management Sci.* 44:1623–1636.

Ha A (2001) Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Sci.* 47:915–930.

Harrison JM (2003) A broader view of Brownian networks. *Ann. Appl. Probab.* 13:1119–1150.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54:1185–1195.

Hassin R (1995) Decentralized regulation of a queue. *Management Sci.* 41:163–173.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer, Boston).

Ho T, Zheng Y (2004) Setting customer expectation in service delivery: An integrated marketing-operations perspective. *Management Sci.* 50:479–488.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92:1644–1655.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing Service Oper. Management* 15: 263–279.

Keskinocak P, Tayur S (2004) Due date management policies. *Handbook of Quantitative Supply Chain Analysis: Modeling in the eBusiness Era* (Kluwer Academic Publishers, Norwell, MA), 485–554.

Kittsteiner T, Moldovanu B (2005) Priority auctions and queue disciplines that depend on processing time. *Management Sci.* 51:236–248.

Kumar S, Randhawa R (2010) Exploiting market size in service systems. *Manufacturing Service Oper. Management* 12:511–526.

Landsberger M, Meilijson I (1999) A general model of insurance under adverse selection. *Econom. Theory* 14:331–352.

Liu Q, van Ryzin G (2008) Strategic capacity rationing to induce early purchases. *Management Sci.* 54:1115–1131.

Low D (1974) Optimal dynamic pricing policies for an $M/M/s$ queue. *Oper. Res.* 20:545–561.

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53:242–262.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Oper. Res.* 38:870–883.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37:15–24.

Plambeck E (2004) Optimal leadtime differentiation via diffusion approximations. *Oper. Res.* 52:213–228.

Plambeck E, Kumar S, Harrison JM (2001) Leadtime constraints in stochastic processing networks under heavy traffic conditions. *Queueing Systems* 39:23–54.

Rothschild M, Stiglitz JE (1976) Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quart. J. Econom.* 90:629–649.

Shang W, Liu L (2011) Promised delivery time and capacity games in time-based competition. *Management Sci.* 57:599–610.

So KC, Song JS (1998) Price, delivery time guarantees, and capacity selection. *Eur. J. Oper. Res.* 111:28–49.

Spearman ML, Zhang RQ (1999) Optimal lead time policies. *Management Sci.* 45:290–295.

Stiglitz JE (1977) Monopoly, non-linear pricing and imperfect information: The insurance market. *Rev. Econom. Stud.* 44: 407–430.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* 46:1249–1267.

Wein LM (1991) Due-date setting and priority sequencing in a multiclass $M/G/1$ queue. *Management Sci.* 37:834–850.

Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45:192–207.

RIGHTSLINK