Pricing Time-Sensitive Services Based on Realized Performance

Philipp Afèche\(^a\) • Opher Baron\(^a\) • Yoav Kerner\(^b\)

\(^a\)Rotman School of Management, University of Toronto
\(^b\)Ben-Gurion University of the Negev

Forthcoming in *Manufacturing & Service Operations Management*

Services like *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 in the standard model since Naor (1969). 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. Our main findings are: 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 vs. flat rate pricing.

1 Introduction

Motivation and research questions. 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).

While 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 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 revenues?
2. What are the properties of the optimal lead time-dependent pricing scheme?
3. What is the value of optimal lead time-dependent pricing?

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 makes 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, i.e., 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 her realized lead time. The standard model restricts attention to schemes that charge one or more flat rates, i.e., 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 the 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: 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 insure 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 sells uniform service (e.g., FIFO) to indistinguishable customer types, a menu of differentiated tariffs can have 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 quite 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 quite 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.
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.


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 2009, 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 which 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, e.g., based on the queue length (e.g., Low 1974, Chen and Frank 2001, Çelik and Maglaras 2008, Ata and Olsen 2009, and Feng et al. 2011). In these settings, different flat rates reflect performance fluctuations across service classes or across consecutive customers, but unlike in ours, each customer knows her payment exactly at the moment 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, Ha 2001, and 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, e.g., Courty and Li (2000), Gallego and Sahin (2010), and Akan et al. (2011). Liu and van Ryzin (2008), and Bansal and Maglaras (2009) 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.

**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 revenues, and the properties of the optimal lead time-dependent pricing schemes. In §4 we study the value of optimal lead time-dependent pricing for a single class and type, first for fixed capacity and then under joint price and capacity optimization. Concluding remarks are in §5. Proofs are in the Online Supplement.

## 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, i.e., 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 the parameter $\mu$. Potential customers have unit demand and arrive according to an exogenous stationary stochastic process with a finite rate $\Lambda$. The provider is risk neutral and makes static price and lead time decisions to maximize her long-run revenue rate. We say “optimal” for 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 quite 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 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 have the same delay cost and utility functions. The provider may offer one or more price-service classes. We say “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. We next introduce the basic single-type, single-class model and then outline how it extends to multiple types and/or classes.

**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 p.d.f. $f$. Let $F$ denote the c.d.f., $F = 1 - F$, and $F^{-1}$ be its inverse. If all customers with valuation higher than the marginal valuation $\underline{v}$ decide to buy, then their arrival rate is $\lambda (\underline{v}) := \Lambda F (\underline{v})$. We call $\lambda$ also the demand rate. Conversely, the marginal value function $v (\lambda) := F^{-1} (\lambda / \Lambda)$ maps arrival rates to marginal valuations. A customer with valuation $v$ who experiences lead time $w$ has net valuation $v - C(w)$, where the delay cost function $C : \mathbb{R}_+ \to \mathbb{R}$ 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 $P : \mathbb{R}_+ \to \mathbb{R}$ 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).

Due to 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, scheduling policy and stochastic properties of the arrival and service processes, the distribution of $W$ only depends on the arrival rate $\lambda$. We write $W(\lambda)$ when making this dependence explicit. For example, in a FIFO $M/M/1$ queue with service rate $\mu$ 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, i.e., the valuation distribution $F$, the delay cost function $C$, the utility function $U$, the rate $\Lambda$, and the statistical properties of the arrival and service processes. Based on this information and the relationship between $\lambda$ 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 $\lambda$. Customers cannot observe the queue length and evaluate their payoff distribution based on the announced tariff $P$ and distribution of $W$. A customer with valuation $v$ buys if and only if her expected utility $E[U(v - C(W(\lambda)) - P(W))]$ is nonnegative. Purchase decisions are irrevocable, i.e., we assume no reneging or retrials. We require that the announced distribution of $W$ matches the distribution of $W(\lambda)$, i.e., 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

$$\max_{\lambda, P} \lambda E[P(W(\lambda))]$$

$$s.t. \quad E[U(v(\lambda) - C(W(\lambda)) - P(W(\lambda)))] = 0.$$  \hspace{1cm} (1) \hspace{1cm} (2)

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

**CARA utility 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

$$1 - \exp(-r\bar{v}(\lambda)) E[\exp r(cW(\lambda) + P(W(\lambda)))] = 0.$$  \hspace{1cm} (3)
Consider the linear tariff \( P(W) = \alpha - \beta W \), with \( \alpha, \beta \geq 0 \) constants. Let \( \tilde{W}(\lambda, s) := E[\exp(sW(\lambda))] \) and \( L(\lambda, s) := \ln \tilde{W}(\lambda, s) \) denote, respectively, the moment-generating function (MGF) and the semi-invariant MGF of the r.v. \( W(\lambda) \) evaluated at \( s \). By (3) the equilibrium arrival rate \( \lambda \) satisfies

\[
\alpha = \nu(\lambda) - \frac{\ln (E[\exp(r(c - \beta)W(\lambda))])}{r} = \nu(\lambda) - \frac{L(\lambda, r(c - \beta))}{r}.
\]

That is, if the provider announces the tariff \( P(W) = \alpha - \beta W \) and the distribution of \( W(\lambda) \), then \( \lambda \) satisfies (4). Note that \( \alpha + L(\lambda, r(c - \beta)) / r \) is the certainty equivalent (CE) of the full price, i.e., 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 \( \mu \), \( L(\lambda, r(c - \beta)) / r = \ln ((\mu - \lambda) / (\mu - \lambda - r(c - \beta))) \).

**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 \( i \in \{1, 2, ..., N\} \). The functions \( \nu_i, C_i, \) and \( U_i \) specify the type-\( i \) attributes as explained above. An arrival is of type \( i \) with probability \( \Lambda_i / \Lambda \), where \( \Lambda_i \) is the arrival rate of potential type-\( i \) customers. Let \( \lambda_i \) denote the type-\( i \) arrival rate, \( \lambda = (\lambda_1, \lambda_2, ..., \lambda_N) \) and \( \lambda = \sum_{i=1}^{N} \lambda_i \). We index price-service classes by \( k \in \{1, 2, ..., K\} \), where \( W_k \) denotes the steady-state lead time of class \( k \) and \( P_k(W_k) \) its price function.

**Discussion.** The key modeling novelty is that we let demand be sensitive to both delay cost and payment variability, by modeling customers as risk averse with respect to their payoff.

**Risk averse customers.** 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, \( E[C(W)] + p \). A customer with valuation \( v \) makes a purchase if her expected payoff is nonnegative, i.e., \( v \geq E[C(W)] + p \). However, due to lead time variability customers face delay cost risk. A customer’s ex post payoff may be lower than its 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; e.g., in a FIFO \( M/M/1 \) queue 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, i.e., 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.

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 2009). In the standard model, customers with nonlinear delay costs are not risk neutral with respect to lead time uncertainty but still risk neutral “with respect to money”, i.e., the resulting cost of delay: they are indifferent among 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, i.e., the delay cost \( c > 0 \) is incurred only if the lead time exceeds the deadline \( \bar{w} > 0 \). That is, customers are satisficers with respect to their delay (Bansal and Maglaras 2009). In this case \( E[C(W)] = c \Pr\{W > \bar{w}\} \). 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 \( P(W) = p + c \cdot \Pr\{W > \bar{w}\} - c \cdot I\{W > \bar{w}\} \) which 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.

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 in order to evaluate their expected utility – which 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, (RA or RN) customers with a corresponding deadline delay cost structure only need to know the on-time probability. More sophisticated delay refund schedules, e.g., 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 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 revenues, and 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 risk neutral with respect to payoff variability. Then for any given number of price-service classes and any scheduling policy:

1. If customer 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 $P_k(W_k) = p_k$. 

9
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 do 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 where 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 multi-class $M/M/1$ 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 vs. business vs. government. Firms may also be able to distinguish among types if their preferences are correlated with product attributes; e.g., a firm that sells lower and higher value products may know that customers 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 for 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 \( \lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*) \) be a revenue-maximizing demand vector if all customer types were RN and distinguishable, i.e., \( \lambda^* \in \text{arg max}_\lambda \sum_{i=1}^{N} \lambda_i \cdot (\nu_i (\lambda_i) - E [C_i (W_i (\lambda))] ) \).

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:

\[
P_i^* (W_i) = \nu_i (\lambda_i^*) - C_i (W_i).
\]

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 are 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 $\lambda^*$, 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 $C(W) = c \cdot I \{W > \overline{w}\}$, 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.

### 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 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 $\alpha^*$ and $\beta^*$ be the optimal tariff parameters and $v_i^*$ the resulting marginal type-i valuation.

**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 $\alpha^*$ and $\beta^*$ be the optimal tariff parameters and $v_i^*$ 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, $v_2^* < \alpha^* < v_1^*$, and the delay discount rate exceeds the delay cost rate of the impatient type and is lower than that of the impatient type, $c_2 < \beta^* < c_1$. 

<table>
<thead>
<tr>
<th>Customer Type</th>
<th>Payment $P(W)$</th>
<th>Full Price $P(W) + C(W)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{RN}$</td>
<td>$\text{RA}$</td>
</tr>
<tr>
<td>Amount</td>
<td>$v^* - E[C(W^*)]$</td>
<td>$v^* - C(W^*)$</td>
</tr>
<tr>
<td>Mean</td>
<td>$v^* - E[C(W^*)]$</td>
<td>$v^* - E[C(W^*)]$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0</td>
<td>$\text{stdev}[C(W^*)]$</td>
</tr>
</tbody>
</table>

Table 1: Payments and full prices: RN vs. RA customers (one type, $\lambda^* = \lambda^*$, $v^* = v^*(\lambda^*)$, $W^* := W(\lambda^*)$.)
2. If \( r_1 > r_j = 0 \), it is optimal to eliminate the RA type’s delay cost risk: \( \alpha^* = v_j^* \) 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 \( \beta \) closer to \( c_j \). Setting \( \beta \) outside the interval \((c_2, c_1)\) is suboptimal as 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 up-front fee \( \alpha^* \) 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 \( \beta \), so it is optimal to eliminate the delay cost risk of the other, RA type.

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, i.e., 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 observable customer types with linear delay costs, the optimal tariff set satisfies \( P_i^* (W) = \alpha_i^* - \beta_i^* W \) for \( i = 1, 2 \), where \( \beta_i^* = c_i \) and \( \alpha_i^* = v_i (\lambda_i^*) \) equals the marginal type-\( i \) valuation at the optimal arrival rates with RN customers. Under this tariff, the first-best demand vector \( \lambda^* \) is the solution of

\[
\max_{\lambda} \sum_{i=1,2} \lambda_i (v_i (\lambda_i) - c_i E [W (\lambda)])
\]

s.t.

\[
0 \leq \lambda_i \leq \Lambda_i, \quad i = 1, 2,
\]

\[
\lambda_1 + \lambda_2 < \mu.
\]

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 FOC for the first-best demand vector are necessary and sufficient:

\[
v_1 (\lambda_1^*) - c_1 E [W (\lambda^*)] + \lambda_1^* v_1' (\lambda_1^*) = v_2 (\lambda_2^*) - c_2 E [W (\lambda^*)] + \lambda_2^* v_2' (\lambda_2^*) = \sum_{i=1,2} \lambda_i^* c_i \frac{dE [W (\lambda^*)]}{d\lambda}.
\]
Incentive-compatibility of the first-best tariff set. Suppose that the provider offers a menu of two linear tariffs $P_i(W) = \alpha_i - \beta_i W$ for $i = 1, 2$. Then, customers of each type purchase the service at the tariff targeted to their type, if and only if

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

i.e., the demand relationship (4) holds for each tariff (which ensures individual rationality), and

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

which ensure incentive-compatibility (IC). By (11), a type-$i$ customer prefers the type-$i$ tariff $(\alpha_i, \beta_i)$ only if the CE of her full price under this tariff (the LHS) does not exceed the CE of her full price for the type-$j$ tariff (the RHS). Substituting for $\alpha_1$ and $\alpha_2$ from (10) into (11) yields the IC constraints in a form that depends only on $\lambda$ and the delay discount rates:

$$\frac{L(\lambda, r_1(c_1 - \beta_1))}{r_1} - \frac{L(\lambda, r_2(c_2 - \beta_1))}{r_2} \leq \bar{v}_1(\lambda_1) - \bar{v}_2(\lambda_2) \leq \frac{L(\lambda, r_1(c_1 - \beta_2))}{r_1} - \frac{L(\lambda, r_2(c_2 - \beta_2))}{r_2}. \quad (12)$$

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

In the standard model with RA 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:

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

(Part 3 of Lemma 1 in the Online Appendix). This implies from (10) that the expected payment of the tariff chosen by type-$i$ must equal its expected marginal net value

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

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

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

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 $\lambda^* = (\lambda^*_1, \lambda^*_2)$ 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 $\alpha_i^* = \underline{v}(\lambda_i^*)$ and $\beta_i^* = c_i$, for $i = 1, 2$. Let $p_i^* := \alpha_i^* - \beta_i^*E[W(\lambda^*)], i = 1, 2$.

1. If $\alpha_1^* \leq \alpha_2^*$, 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 $p_1^* \neq p_2^*$, the first-best tariff set is IC if and only if the type with the higher mean payment is sufficiently risk averse: if $p_1^* > p_2^*$, there is $r \in (0, \infty)$ such that type $i$ chooses tariff $i$ if and only if $r_i \geq r$; 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.

Part 1. If $\alpha_1^* \leq \alpha_2^*$, 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 that the first-best tariff cannot be IC. By inspection of the FOC (9), Part 1 may apply, for example, if congestion effects are minor, i.e., the terms involving $c_i$ are small. For illustration, if both types have uniform valuations on $[0, \overline{\nu}_i]$, then $\underline{v}_i(\lambda_i) = \overline{\nu}_i(1 - \lambda_i/\Lambda_i)$. If congestion effects are minor and there is enough capacity, then the FOC (9) imply

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

so that it is optimal to serve roughly the top half of each type, i.e., $\lambda_i^* \approx \Lambda_i/2$ and $\underline{v}_i(\lambda_i^*) \approx \overline{\nu}_i/2$. Whenever the impatient type-1 customers have the lower maximum valuation, i.e., $\overline{\nu}_1 < \overline{\nu}_2$, we have $\alpha_1^* < \alpha_2^*$, and the first-best solution is not IC for any risk aversion levels.

Part 2. In the exceptional case where $p_1^* = p_2^*$, the first-best tariff set is clearly IC for RN customers since 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 $\alpha_1^*$ regardless of her risk aversion, since this tariff eliminates her delay cost risk.
By contrast, her CE of the type-2 tariff full price increases in her risk aversion. A type-1 customer is therefore indifferent between the tariffs if she is RN but otherwise prefers the type-1 tariff:

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

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 \( \lambda_1^* \varepsilon_1 (\lambda_1^*) = \lambda_2^* \varepsilon_2 (\lambda_2^*) \). This means that at the first-best arrival rates, both types have the same ratio of marginal valuation to elasticity, where the elasticity function for the type-\( i \) marginal value function is \( \varepsilon_i (\lambda_i) = -\varepsilon_i (\lambda_i) / (\lambda_i \varepsilon_i (\lambda_i)) \). For example, consider exponential valuations with c.d.f. \( F_i (v) = 1 - \exp (-k_i v) \) for \( v \geq 0 \), where \( k_i > 0 \). The marginal valuation function \( \varepsilon_i (\lambda_i) = \ln (\lambda_i / \lambda_i) / k_i \) and \( \lambda_i \varepsilon_i (\lambda_i) = -1 / k_i \), so \( p_1^* = p_2^* \) if and only if \( k_1 = k_2 \).

Part 3. Suppose that \( \alpha_1^* > \alpha_2^* \) and \( p_1^* < p_2^* \). Then type-1 chooses her targeted tariff since

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

If type-2 customers are RN, they clearly prefer the type-1 tariff since 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, the less they value the larger discount. If they are sufficiently risk averse, they prefer the type-2 tariff since 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 since 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. 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 vs. optimal flat rate pricing. We focus on the case of a single RA type with CARA utility and linear delay cost.
4.1 Performance versus 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 \( P^* (W) = \varphi (\lambda^*) - cW \), where \( \lambda^* \) 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 \( \alpha \). The simple refund policy charges \( \alpha \) for on-time delivery by a threshold lead time \( \overline{w} \) and issues a full refund for late delivery, i.e., \( P(W) = \alpha - \alpha \cdot I \{ W > \overline{w} \} \), and the provider chooses \( \alpha \) and \( \overline{w} \).

We compare lead time-dependent vs. 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.

**Optimal flat rate pricing.** By (4) the demand relationship for \( P(W) = \alpha - \beta W \) is

\[
\alpha = \varphi (\lambda) - \frac{L(\lambda, r(e - \beta))}{r}.
\]  

(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 \( \lambda \) is \( \alpha - \beta E[W(\lambda)] = \varphi(\lambda) - cE[W(\lambda)] \). The provider solves

\[
\max_{\lambda} \Pi(\lambda) = \lambda \left( \varphi(\lambda) - cE[W(\lambda)] \right).
\]  

(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 \( \lambda \) is \( \varphi(\lambda) - L(\lambda, rc)/r \). The provider solves

\[
\max_{\lambda} \Pi^f(\lambda; r) = \lambda \left( \varphi(\lambda) - \frac{L(\lambda, rc)}{r} \right).
\]  

(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 \( \alpha^* := \varphi(\lambda^*) \). Since 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, \( \Pi^f(r) = \Pi^f(\lambda^f(r); r) \) the corresponding optimal revenue, and the optimal flat rate is

\[
\alpha^f(r) := \varphi(\lambda^f(r)) - \frac{L(\lambda^f(r), rc)}{r}.
\]  

(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 \( g(x, y) \).

1. \( \Pi'(0) > 0 > \lim_{\lambda \to \min(\mu, A)} \Pi'(\lambda) \). This ensures an interior solution under optimal pricing.

2. The functions \( \Pi(\lambda) \) and \( \Pi^f(\lambda; r) \) are strictly concave in \( \lambda \).

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

The Online Appendix details sufficient conditions for A2-A3.
Proposition 5 Suppose that the provider only charges a flat rate, and there is a single risk-averse 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 $r < \tau$, and they equal zero if $r \geq \tau$, where

$$0 < \tau = \arg \left\{ r \geq 0 : \frac{L(0,rc)}{r} = \underline{v}(0) \right\} \text{ and } \tau < \infty \text{ if } \underline{v}(0) < \infty. \quad (17)$$

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

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, i.e., decreasing utilization. At the extreme, for RA levels above the threshold $\tau$, 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.

**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, e.g., linear as in this section, this policy is suboptimal as it insure 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 since the firm only sets two controls – the price $\alpha$ and the lead time threshold $\bar{\tau}$. 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 $\lambda$ 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 $\lambda$ and $\bar{w}$, which is determined from the demand relationship (3). The provider solves

$$\max_{\lambda,\bar{w}} \Pi^s(\lambda, \bar{w}; r) = \lambda \alpha^s(\lambda, \bar{w}; r) \Pr \{W(\lambda) \leq \bar{w}\}, \quad (18)$$
where \( \Pr \{ W(\lambda) \leq \bar{w} \} \) is the on-time probability. The expected payment equals that under the flat rate policy as \( \bar{w} \to \infty \), i.e., \( \lim_{\bar{w} \to \infty} \alpha^*(\lambda, \bar{w}; r) \Pr \{ W(\lambda) \leq \bar{w} \} = p(\lambda) - L(\lambda, rc) / r \). The revenues under the optimal lead time-dependent tariff, the simple refund policy and flat rate pricing satisfy:

\[
\Pi(\lambda) \geq \Pi^s(\lambda, \bar{w}^*(\lambda; r); r) \geq \Pi^f(\lambda; r),
\]

(19)

where \( \bar{w}^*(\lambda; r) \) maximizes the expected payment \( \alpha^*(\lambda, \bar{w}; r) \Pr \{ W(\lambda) \leq \bar{w} \} \). The first inequality in (19) holds by Proposition 2, the second since 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.

**Example 1.** We illustrate Proposition 5 and the simple refund policy with a numerical example for an \( M/M/1 \) queue. We compare the performance of the optimal lead time-dependent tariff with that of optimal flat rate pricing and the optimal simple refund policy. The capacity \( \mu = 5 \), the market size \( \Lambda = 10 \), the value distribution is uniform on \([0, 5]\), and the delay cost rate \( c = 1 \). 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 \( E[W(\lambda)] = 1 / (\mu - \lambda) \) for the \( M/M/1 \) queue, the optimal tariff is \( P^*(W) = \bar{w}^* - cW = 3.35 - W \) 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 \( \lambda \) and lead time quote \( \bar{w} \) solve (18), where for the \( M/M/1 \) queue

\[
\alpha^*(\lambda, \bar{w}; r) = \frac{1}{r} \ln \left( \frac{\exp(\bar{w}(\lambda))}{W(\lambda, rc)} - \exp \left\{ -\bar{w}(\mu - \lambda - rc) \right\} \frac{1 - \exp \left\{ -\bar{w}(\mu - \lambda - rc) \right\}}{1 - \exp \left\{ -\bar{w}(\mu - \lambda - rc) \right\}} \right)
\]

(20)

from (3), \( \bar{W}(\lambda, rc) = (\mu - \lambda) / (\mu - \lambda - rc) \), and \( \Pr \{ W(\lambda) \leq \bar{w} \} = 1 - \exp (-\bar{w}(\mu - \lambda)) \).

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 compared 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 quite well relative to optimal lead time-dependent 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 to 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 $r \to \infty$. Under the simple refund policy the mean payment initially decreases in $r$ since the utilization loss is only significant at higher RA levels (see Figure 1).

### 4.2 Performance versus 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 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 decisions, and we compare the optimal capacity level and performance under the optimal lead time-dependent tariff vs. the optimal flat rate.

For convenience we consider a single-server queue and denote its capacity by $\mu$. The results for the multiple-server case are similar. Let $\Pi^*(\mu)$ and $\Pi^{fs}(\mu)$ 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 $\Pi^{\mu^*} (\mu)$ 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 $b > 0$. For simplicity we assume that $\nu (0) < \infty$.

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

1. Under the optimal lead-time dependent tariff:
   
   (a) There is a threshold $\mu > 0$ such that $\Pi^* (\mu) = 0$ for $\mu \leq \underline{\mu}$. Furthermore, $\lim_{\mu \to \infty} \Pi^* (\mu) = \max_{\lambda \in [0, \Lambda]} \lambda \underline{\mu} (\lambda)$ and $\lim_{\mu \to \infty} \Pi^{f*} (\mu) = 0 = \lim_{\mu \to \infty} \Pi^{\mu^*} (\mu)$.
   
   (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, for any $r > 0$:
   
   (a) There is a threshold $\mu^f > \mu$ such that $\Pi^{f*} (\mu) = 0$ for $\mu \leq \mu^f$. Furthermore, $\lim_{\mu \to \infty} \Pi^{f*} (\mu) = \lim_{\mu \to \infty} \Pi^* (\mu)$ and $\lim_{\mu \to \infty} \Pi^{f*\mu} (\mu) = 0 = \lim_{\mu \to \infty} \Pi^{f*\mu} (\mu)$.
   
   (b) There is a threshold $\overline{b}^f (\bar{b})$ such that $\max_{\mu \geq 0} (\Pi^{f*} (\mu) - b\mu) > 0$ if and only if $b < \overline{b}^f (\bar{b})$. The optimal capacity decreases in $b$, $\mu^{f*} (b) > \underline{\mu}^f$ for $b < \overline{b}^f (\bar{b})$, and $\mu^{f*} (b) = 0$ for $b > \overline{b}^f (\bar{b})$.

Optimal lead time-dependent pricing naturally yields a higher profit compared to optimal flat rate pricing. By Proposition 6 the capacity cost at which the system just breaks even is higher ($\overline{b}^f (\bar{b}) < \overline{b}$). With either tariff the system exhibits negative profits at small capacity levels, and scale economies. However, Proposition 6 suggests that these scale economies are weaker under optimal lead time-dependent vs. flat rate pricing. Specifically, the system requires less capacity to generate...
positive revenue ($\mu < \mu^f$) and profit, and the revenue gain vanishes for ample capacity. As a result, the marginal value of capacity under optimal lead time-dependent vs. 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, i.e., increasing capacity. Indeed, as discussed in Example 2 below, we find that optimal lead time-dependent pricing yields a higher return on a lower optimal capacity investment compared to optimal flat rate pricing.

The performance differences identified in Proposition 6 increase in risk aversion; i.e., under optimal flat rate pricing the minimum capacity threshold $\mu^f$ for positive revenue increases in $r$, and the cost threshold $b^f$ 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 to flat rate pricing, which implies a higher utilization and return on capacity investment (ROI) and a lower service level, i.e., 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 $\Pi^{f^*'} (\mu) > \Pi^{*'} (\mu)$ for all $\mu$ 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. (E.g., for $b = 0.5$ the ROI $\approx 200\%$ 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: e.g., 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.)

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 quite 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.

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 how tariff constraints reduce performance because they leave customers exposed to delay cost risk, although the simple refund policy performs quite 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, e.g., 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’ references. Proposition 4 shows that the first-best solution may be IC in the absence of tariff 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 to 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, 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, e.g., 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. (2012) 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, e.g., by considering bivariate utility functions of lead times and payments that may capture different risk attitudes towards 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, e.g., by charging flat rates for longer queues and based on realized lead times for shorter queues.

References


Online Supplement: Proofs

Proof of Proposition 1. Suppose there are $N$ customer types. Without loss of generality, suppose that $K$ price-service classes are chosen by some customers. Write $\lambda_{ik}$ for the rate of type-$i$ customers who buy class $k$, where $\lambda_i = \sum_{k=1}^{K} \lambda_{ik}$ is the overall type-$i$ arrival rate.

Part 1. Suppose that for $i \in \{1, 2, ..., N\}$ and $k \in \{1, 2, ..., K\}$, the arrival rates $\{\lambda_{ik}\}$ and delays $\{W_k\}$ are feasible, and that $\{W_k\}$ and the price functions $\{P_k\}$ form a customer equilibrium, i.e., they satisfy the individual rationality (IR) constraints, and if types are indistinguishable by the provider, the incentive-compatibility (IC) constraints. Since the customer equilibrium conditions and the provider’s expected revenue rate depend on $\{P_k\}$ only through the expectations $\{E[P_k(W_k)]\}$, setting the flat rates $p_k = E[P_k(W_k)]$ for $k \in \{1, 2, ..., K\}$ establishes Part 1.

Part 2. If customer types have different service requirements, then the lead time in a given class may vary by type and there may not exist flat rates that ensure incentive-compatibility at the desired arrival rates. However, if types are distinguishable, then only the IR constraints need to hold. These constraints are easily satisfied since the provider can charge a different flat rate for each type.

Proof of Proposition 2. Fix a scheduling policy and a demand vector $\lambda$ such that the steady-state lead times of all classes have finite moments. Let $W_i(\lambda)$ denote the lead time of class $i$, which is targeted to type $i$. To prove the result we show that: for given $\lambda$ and $W_i(\lambda)$ the provider maximizes her expected type-$i$ revenue by charging $P_i(W_i) = \psi_i(\lambda_i) - C_i(W_i)$, and the resulting revenue rate is a function of $\lambda$ that is invariant to customers’ risk aversion levels.

Part 1. Because the provider can distinguish among types, he can limit each customer to the price-service class that is targeted to her type. It therefore suffices to prove the claim for a single type. For simplicity we write $W_i$, suppressing the dependence on $\lambda$. Given $\lambda_i$ and $W_i$, the provider maximizes the type-$i$ revenue by solving

$$\max_{P_i} \lambda_i E[P_i(W_i)]$$  \hspace{1cm} (21)

$$\text{s.t. } E[U(\psi_i(\lambda_i) - C_i(W_i) - P_i(W_i))] = 0.$$  \hspace{1cm} (22)

This problem is the multi-type analog of (1)-(2) for given $\lambda_i$ and $W_i$. We have

$$U(E[\psi_i(\lambda_i) - C_i(W_i) - P_i(W_i)]) \geq E[U(\psi_i(\lambda_i) - C_i(W_i) - P_i(W_i))] = 0.$$  \hspace{1cm} (23)

The inequality follows by Jensen’s inequality since $U$ is concave, and the equality follows from (22). Since $U(0) = 0$ and $U' > 0$, we have $U(x) \geq 0 \iff x \geq 0$, so that (23) implies the following upper bound on the expected payment:

$$E[P_i(W_i)] \leq \psi_i(\lambda_i) - E[C_i(W_i)]$$

with equality if and only if (23) holds with equality. If type $i$ is RN, then the inequality in (23) holds with equality, and every tariff $P_i$ is feasible if and only if $E[P_i(W_i)] = \psi_i(\lambda_i) - E[C_i(W_i)]$. If type $i$ is RA, i.e., $U$ is strictly concave, then the inequality in (23) holds with equality if and only if $\psi_i(\lambda_i) - C_i(W_i) - P_i(W_i) \equiv 0$, i.e., the marginal customer’s payoff is zero with probability one. It follows that $P_i(W_i) \equiv \psi_i(\lambda_i) - C_i(W_i)$ is the unique optimal tariff.
Since the optimal tariff has this form for arbitrary \( \lambda_i \) and \( W_i (\lambda) \), it also has this form at the optimal arrival rates and the corresponding lead time distributions.

**Part 2.** Given a scheduling policy, the tariffs \( P_i (W_i) \equiv v_i (\lambda_i) - C_i (W_i) \) maximize the revenue rate for fixed \( \lambda \) and ensure that (22) holds for all types. The resulting revenue rate function is

\[
\sum_{i=1}^{N} \lambda_i (v_i (\lambda_i) - E [C_i (W_i (\lambda))]),
\]

which is independent of customers’ risk aversion and is the same as for RN customers. [1]

Lemma 1 summarizes useful properties of \( \tilde{W} (\lambda, s) \) and \( L (\lambda, s) \) for Propositions 3-6.

**Lemma 1** Let \( \tilde{W} (\lambda, s) := E [\exp (s W (\lambda))] \) be the MGF of the r.v. \( W (\lambda) \) and \( L (\lambda, s) := \ln \tilde{W} (\lambda, s) \).

1. \( \tilde{W} (\lambda, s) \geq 0, \tilde{W} (\lambda, 0) = 1, \) and \( \tilde{W} (\lambda, s) \) is strictly increasing in \( s \) with \( \tilde{W}_s (\lambda, 0) = E [W (\lambda)] \), and strictly convex in \( s \).

2. \( L (\lambda, 0) = 0, L (\lambda, s) \) is strictly increasing in \( s \) with \( L_s (\lambda, 0) = E [W (\lambda)] \), and strictly convex in \( s \) with \( L_{ss} (\lambda, 0) = Var [W (\lambda)] \).

3. For any constant \( k \neq 0 \), \( L (\lambda, rk) / r \) is strictly increasing in \( r \geq 0 \) and \( \lim_{r \to 0} L (\lambda, rk) / r = kE [W (\lambda)] \).

4. If \( W (\lambda) \) is the sojourn time in a M/G/1 queue, then: (a) For \( s > 0 \) both \( \tilde{W} (\lambda, s) \) and \( L (\lambda, s) \) are strictly increasing and strictly convex in \( \lambda \), and for \( s < 0 \) both \( \tilde{W} (\lambda, s) \) and \( L (\lambda, s) \) are strictly decreasing and strictly concave in \( \lambda \). (b) The function \( L_{\lambda} (\lambda, s) \) is strictly convex in \( s \).

**Proof.** Parts 1-2. are standard. E.g., see Gallager (1996), Chapters 1 and 7 (problem 7.7).

**Part 3.** We first show that

\[
\frac{d}{dr} \frac{L (\lambda, rk)}{r} = \frac{rk L_s (\lambda, rk) - L (\lambda, rk)}{r^2} > 0 \text{ for } r \geq 0.
\]

(25)

For \( r > 0 \), (25) holds since \( L (\lambda, 0) = 0 \) and \( L (\lambda, s) \) is strictly convex in \( s \) by Part 2, which implies

\[
rk L_s (\lambda, rk) = \int_{0}^{rk} L_s (\lambda, x) \, dx > \int_{0}^{rk} L_s (\lambda, x) \, dx = L (\lambda, rk).
\]

For \( r = 0 \) we have

\[
\lim_{r \to 0} \frac{rk L_s (\lambda, rk) - L (\lambda, rk)}{r^2} = \lim_{r \to 0} \frac{rk^2 L_{ss} (\lambda, rk)}{2r} = \frac{k^2 VAR (W (\lambda))}{2} > 0.
\]

The first equality follows since \( L (\lambda, 0) = 0 \) (Part 2) and from l’Hôpital’s rule, and the second since \( L_{ss} (\lambda, 0) = Var [W (\lambda)] \) by Part 2. Finally, we have

\[
\lim_{r \to 0} \frac{L (\lambda, rk)}{r} = \lim_{r \to 0} \frac{d}{dr} L (\lambda, rk) = k \lim_{r \to 0} L_s (\lambda, rk) = kE [W (\lambda)].
\]

The first equality follows since \( L (\lambda, 0) = 0 \) (Part 2) and from l’Hôpital’s rule, and the last since \( L_s (\lambda, 0) = E [W (\lambda)] \) by Part 2.
Part 4. Let $X$ be the service time and $\tilde{X}(s)$ the MGF of its distribution. Let $X_e$ be the equilibrium residual service time and $\tilde{X}_e(s)$ the MGF of its distribution. Then $\tilde{X}_e(s) = \mu(\tilde{X}(s) - 1)/s$. From the Pollaczek-Khinchin formula, the MGF of the sojourn time is

$$\tilde{W}(\lambda, s) = \tilde{X}(s) \frac{1 - \rho}{1 - \rho X_e(s)} = \tilde{X}(s) \frac{\mu - \lambda}{\mu - \lambda X_e(s)},$$

and it is defined only for $s$ such that $1 > \rho \tilde{X}_e(s)$, or equivalently, $\mu - \lambda \tilde{X}_e(s) > 0$.

4(a) First consider $\tilde{W}(\lambda, s)$. We have

$$\tilde{W}_\lambda(\lambda, s) = \frac{\mu \tilde{X}(s)}{(\mu - \lambda \tilde{X}_e(s))} \left( \tilde{X}_e(s) - 1 \right),$$

$$\tilde{W}_{\lambda \lambda}(\lambda, s) = 2\frac{\mu \tilde{X}(s) \tilde{X}_e(s)}{(\mu - \lambda \tilde{X}_e(s))^3} \left( \tilde{X}_e(s) - 1 \right).$$

Both have the same sign as $\left( \tilde{X}_e(s) - 1 \right)$, which has the sign of $s$ because $X_e$ is nonnegative: for $x \geq 0$, $\text{sgn}(e^{sx} - 1) = \text{sgn}(s)$. This establishes 4(a) for $\tilde{W}(\lambda, s)$. For $L(\lambda, s)$ we get

$$L_\lambda(\lambda, s) = \frac{\tilde{W}_\lambda(\lambda, s)}{\tilde{W}(\lambda, s)} = \frac{\mu - \lambda \tilde{X}_e(s)}{\tilde{X}(s) (\mu - \lambda)} \left( \frac{\mu \tilde{X}(s)}{(\mu - \lambda \tilde{X}_e(s))} \right) \left( \tilde{X}_e(s) - 1 \right) = \frac{\mu - \lambda \tilde{X}_e(s)}{\mu - \lambda \tilde{X}_e(s)};$$

$$L_{\lambda \lambda}(\lambda, s) = \mu \frac{\mu - \lambda \tilde{X}_e(s) + \tilde{X}_e(s) (\mu - \lambda)}{(\mu - \lambda)^2 \left( \mu - \lambda \tilde{X}_e(s) \right)} \left( \tilde{X}_e(s) - 1 \right).$$

As above both have the sign of $s$.

4(b) We have

$$L_{\lambda \beta}(\lambda, s) = \frac{\mu}{\mu - \lambda} \frac{\tilde{X}_e'(s) \left( \mu - \lambda \tilde{X}_e(s) \right) \left( \tilde{X}_e(s) - 1 \right)}{(\mu - \lambda \tilde{X}_e(s))^2} = \frac{\mu \tilde{X}_e'(s)}{(\mu - \lambda \tilde{X}_e(s))^2}.$$

Noting that $\tilde{X}_e''(s) > 0$ by Part 1, it follows that

$$L_{\lambda \beta \beta}(\lambda, s) = \frac{\mu \tilde{X}_e''(s)}{(\mu - \lambda \tilde{X}_e(s))^2} + \frac{2\lambda \mu \left( \tilde{X}_e'(s) \right)^2}{(\mu - \lambda \tilde{X}_e(s))^3} > 0.$$

---

**Proof of Proposition 3.**

Let $\lambda = \lambda_1 + \lambda_2$. Since customers’ processing requirements are i.i.d., the distribution of $W$ depends on the rates $\lambda_i$ only through their sum $\lambda$. The demand system satisfies

$$\alpha \left\{ \begin{array}{ll} = \psi_i(\lambda_i) - \frac{L(\lambda, \tau_i, (c_i - \beta_i))}{\tau_i}, & \lambda_i > 0 \\ \geq \psi_i(\lambda_i) - \frac{L(\lambda, \tau_i, c_i - \beta_i)}{\tau_i}, & \lambda_i = 0 \end{array} \right.,$$

(26)
which is the 2-type version of (4). If it is optimal to serve both types, then the optimal linear price function \( \alpha^* - \beta^* W \) and the resulting demand vector \( \lambda^* \) must be a solution of the problem

\[
\max_{\alpha, \beta, \lambda} \Pi (\alpha, \beta, \lambda) = \sum_{i=1,2} \lambda_i \cdot (\alpha - \beta E [W(\lambda)])
\]

subject to

\[
\begin{align*}
0 & \leq \lambda_i \leq \Lambda_i, \quad i = 1, 2, \\
\lambda_1 + \lambda_2 & < \mu \\
h_i (\alpha, \beta, \lambda) & \triangleq \psi_i (\lambda_i) - \alpha - \frac{L(\lambda, r_i (c_i - \beta))}{r_i} = 0, \quad i = 1, 2.
\end{align*}
\]

We first show that the gradients of the constraint functions (29) are linearly independent. We have

\[
\nabla h_i (\alpha, \beta, \lambda) = \begin{bmatrix}
\frac{\partial h_i (\alpha, \beta, \lambda)}{\partial \alpha} \\
\frac{\partial h_i (\alpha, \beta, \lambda)}{\partial \beta} \\
\frac{\partial h_i (\alpha, \beta, \lambda)}{\partial \lambda_1} \\
\frac{\partial h_i (\alpha, \beta, \lambda)}{\partial \lambda_2}
\end{bmatrix}
= \begin{bmatrix}
-1 \\
L_\lambda (\lambda, r_i (c_i - \beta)) \\
\frac{\psi_i' (\lambda_i) 1_{i=1}}{r_1} - \frac{L_\lambda (\lambda, r_i (c_i - \beta))}{r_1} \\
\frac{\psi_i' (\lambda_i) 1_{i=2}}{r_2} - \frac{L_\lambda (\lambda, r_i (c_i - \beta))}{r_2}
\end{bmatrix}
\]

Let \( s_i = r_i (c_i - \beta) \). \( \nabla h_1 (\alpha, \beta, \lambda) \) and \( \nabla h_2 (\alpha, \beta, \lambda) \) are linearly dependent if and only if

\[
1 = \frac{L_s (\lambda, s_1)}{L_s (\lambda, s_2)} = \frac{\psi_1' (\lambda_1) - \frac{L_\lambda (\lambda, s_1)}{r_1}}{-\frac{L_\lambda (\lambda, s_2)}{r_2}} = \frac{\frac{L_\lambda (\lambda, s_1)}{r_1}}{-\frac{L_\lambda (\lambda, s_2)}{r_2}}.
\]

The first equation holds if and only if \( s_1 = s_2 \), since \( L_s (\lambda, s) \) is strictly increasing in \( s \) by Part 2 of Lemma 1. Suppose that \( s_1 = s_2 \), and let \( k = L_\lambda (\lambda, s_1) = L_\lambda (\lambda, s_2) \). If \( k = 0 \) (which holds iff \( s_1 = s_2 = 0 \)), then the last ratio equals zero, which cannot hold. If \( k \neq 0 \) then the second-to-last and the last ratio equal one if and only if, respectively,

\[
k \left( \frac{1}{r_1} - \frac{1}{r_2} \right) = \psi_1' (\lambda_1) \quad \text{and} \quad k \left( \frac{1}{r_1} - \frac{1}{r_2} \right) = -\psi_2' (\lambda_2).
\]

Since \( \psi_i' (\lambda_i) < 0 \), these equations cannot both hold, and we conclude that \( \nabla h_1 (\alpha, \beta, \lambda) \) and \( \nabla h_2 (\alpha, \beta, \lambda) \) are linearly independent.

Since \( h_1 (\alpha, \beta, \lambda) = h_2 (\alpha, \beta, \lambda) = 0 \) for any solution that serves both types, and since \( \nabla h_1 (\alpha, \beta, \lambda) \) and \( \nabla h_2 (\alpha, \beta, \lambda) \) are linearly independent, the solution \( (\alpha^*, \beta^*, \lambda^*) \) must satisfy the Karush-Kuhn-Tucker (KKT) conditions. Suppose that \( (\alpha^*, \beta^*, \lambda^*) \) is a solution that serves some, but not all, customers of each type, so \( \lambda_i^* \in (0, \Lambda_i) \). It cannot be optimal to operate at capacity, so \( \lambda_i^* + \lambda_j^* < \mu \).

Let \( \gamma_i \) be the Lagrange multiplier of the constraint \( h_i (\alpha, \beta, \lambda) = 0 \). Suppressing the superscript \( * \) and the arguments of \( \Pi \) and \( h_i \), the KKT conditions are:

\[
\lambda = \frac{\partial \Pi}{\partial \alpha} = \sum_{k=1,2} \gamma_k \frac{\partial h_k}{\partial \alpha} = - (\gamma_1 + \gamma_2), \quad (30)
\]

\[
-\lambda E [W(\lambda)] = \frac{\partial \Pi}{\partial \beta} = \sum_{k=1,2} \gamma_k \frac{\partial h_k}{\partial \beta} = \gamma_1 L_s (\lambda, s_1) + \gamma_2 L_s (\lambda, s_2), \quad (31)
\]

\[
\frac{\partial \Pi}{\partial \lambda_i} = \sum_{k=1,2} \gamma_k \frac{\partial h_k}{\partial \lambda_i} = \sum_{k=1,2} \gamma_k \left( \psi_i' (\lambda_i) 1_{i=k} - \frac{L_\lambda (\lambda, s_k)}{r_k} \right), \quad i = 1, 2, \quad (32)
\]

\[
h_i = \frac{\psi_i (\lambda_i) - \alpha - \frac{L(\lambda, r_i (c_i - \beta))}{r_i}}{r_i} = 0, \quad i = 1, 2, \quad (33)
\]
where
\[
\frac{\partial \Pi}{\partial \lambda_1} = \frac{\partial \Pi}{\partial \lambda_2} = \alpha - \beta EW(\lambda) - \beta \lambda \frac{dE[W(\lambda)]}{d\lambda}.
\]

It follows from (32) that
\[
\sum_{k=1,2} \gamma_k \frac{\partial h_k}{\partial \lambda_1} = \sum_{k=1,2} \gamma_k \frac{\partial h_k}{\partial \lambda_2},
\]
which implies \(\gamma_1 \frac{v_1'}{(\lambda_1)} = \gamma_2 \frac{v_2'}{(\lambda_2)}\).

Together with (30), this implies that \(\gamma_i < 0\) for \(i = 1, 2\). Specifically, \(\gamma_2 = \frac{\lambda}{v_2'(\lambda_2) / v_1'(\lambda_1) + 1} < 0\) and \(\gamma_1 = -\frac{\lambda}{1 + v_1'(\lambda_1) / v_2'(\lambda_2)} < 0\).

Next, multiply (30) by \(E[W(\lambda)]\) and add it to (31) to obtain
\[
\gamma_1 (L_s(\lambda, s_1) - E[W(\lambda)]) + \gamma_2 (L_s(\lambda, s_2) - E[W(\lambda)]) = 0.
\]

Noting that \(\gamma_i < 0\) for \(i = 1, 2\), this holds if and only if one of the following conditions is satisfied:
\[
L_s(\lambda, r_1 (c_1 - \beta)) > E[W(\lambda)] > L_s(\lambda, r_2 (c_2 - \beta))
\]
or
\[
L_s(\lambda, r_1 (c_1 - \beta)) < E[W(\lambda)] < L_s(\lambda, r_2 (c_2 - \beta))
\]
or
\[
L_s(\lambda, r_1 (c_1 - \beta)) = E[W(\lambda)] = L_s(\lambda, r_2 (c_2 - \beta)).
\]

**Part 1.** We show that for \(r_1, r_2 > 0\), (34) holds if and only if \(c_2 < \beta < c_1\). If \(\beta \geq c_1 > c_2\), we have by Part 2 of Lemma 1:
\[
L_s(\lambda, r_1 (c_1 - \beta)) \leq E[W(\lambda)] \quad \text{and} \quad L_s(\lambda, r_2 (c_2 - \beta)) < E[W(\lambda)].
\]

Since \(\gamma_i < 0\), it follows that the LHS of (34) is strictly positive; so that \(\beta < c_1\). A similar argument shows that \(\beta > c_2\). With \(c_2 < \beta < c_1\), we have \(L_s(\lambda, r_1 (c_1 - \beta)) > 0 > L_s(\lambda, r_2 (c_2 - \beta))\), so that \(v_2 < \alpha < v_1\) follows from (29).

**Part 2.** If \(r_j = 0 < r_i\), then \(L_s(\lambda, r_j (c_j - \beta)) = E[W(\lambda)]\), and (34) reduces to
\[
\gamma_j (L_s(\lambda, r_i (c_i - \beta)) - E[W(\lambda)]) = 0.
\]

Since \(\gamma_j < 0\), it holds if and only if \(\beta = c_i\). The type-i demand constraint in (29) implies \(\alpha = v_1\).

**Proof of Proposition 4.** The analysis in Section 3.3 implies that the first-best tariff set is IC if and only if \(\lambda^*\) satisfies (12) for \(\beta_1 = c_1\) and \(\beta_2 = c_2\). Noting that \(L(\lambda, r_i (c_i - \beta_i)) / r_i = 0\) if \(\beta_i = c_i\), these conditions are equivalent to
\[
- \frac{L(\lambda^*, r_2 (c_2 - c_1))}{r_2} \leq \frac{v_1 (\lambda_1^*) - v_2 (\lambda_2^*)}{\lambda_1^* - \lambda_2^*},
\]
where type-2 (type-1) chooses her targeted tariff if and only if the first (second) inequality holds. Furthermore, since \(L(\lambda, s) < 0\) for \(s < 0\) by Part 2 of Lemma 1, and since \(L(\lambda, r (c_1 - c_2)) / r \leq 0\...
is strictly increasing in \( r \geq 0 \) with \( \lim_{r \to 0} L(\lambda, r (c_1 - c_2)) / r = (c_1 - c_2) E[W(\lambda)] \) by Part 3 of Lemma 1, it follows that for any \( r_1 \) and \( r_2 \),

\[
0 < -\frac{L(\lambda^*, r_2 (c_2 - c_1))}{r_2} \leq (c_1 - c_2) E[W(\lambda^*)] \leq \frac{L(\lambda^*, r_1 (c_1 - c_2))}{r_1},
\]

where the second inequality is strict if and only if \( r_2 > 0 \), and the last if and only if \( r_1 > 0 \).

We establish the claims by relating them to (35) via (36).

**Part 1.** Because \( \alpha_1^* = v_1(\lambda_1^*) \), if \( \alpha_1^* \leq \alpha_2^* \), then \( v_1(\lambda_1^*) - v_2(\lambda_2^*) \leq 0 \). Then the first inequality in (36) implies that the first inequality in (35) cannot hold, so that IC is violated for any \( r_i \).

**Part 2.** If \( p_1^* = \alpha_1^* - c_1 E[W(\lambda^*)] = p_2^* = \alpha_2^* - c_2 E[W(\lambda^*)] \), then because \( c_1 > c_2 \), we have \( \alpha_1^* > \alpha_2^* \) and \( v_1(\lambda_1^*) - v_2(\lambda_2^*) = (c_1 - c_2) E[W(\lambda^*)] \). From (36) we have that (35) holds for all \( r_i \).

**Part 3.** If \( \alpha_1^* > \alpha_2^* \), then \( v_1(\lambda_1^*) - v_2(\lambda_2^*) > 0 \), and we have two cases with \( p_1^* \neq p_2^* \).

(i) Suppose that \( p_1^* > p_2^* \). Then \( v_1(\lambda_1^*) - v_2(\lambda_2^*) > (c_1 - c_2) E[W(\lambda^*)] \). From (36) we have

\[
0 < -\frac{L(\lambda^*, r_2 (c_2 - c_1))}{r_2} \leq (c_1 - c_2) E[W(\lambda^*)] < v_1(\lambda_1^*) - v_2(\lambda_2^*),
\]

so the first inequality in (35) holds and type-2 chooses her targeted tariff for all \( r_2 \). If type-1 is RN, she also prefers the type-2 tariff: we have \( \lim_{r_1 \to 0} L(\lambda^*, r_1 (c_1 - c_2)) / r_1 = (c_1 - c_2) E[W(\lambda^*)] < v_1(\lambda_1^*) - v_2(\lambda_2^*) \), where the equality follows from (36) and the inequality from (37), so the second inequality in (35) is violated. Because \( L(\lambda^*, r_1 (c_1 - c_2)) / r_1 \) strictly increases in \( r_1 \) by Part 3 of Lemma 1, and \( \lim_{r_1 \to \infty} L(\lambda, r_1 (c_1 - c_2)) / r_1 = \infty \) since \( c_1 > c_2 \), there is an unique \( r \in (0, \infty) \) such that

\[
v_1(\lambda_1^*) - v_2(\lambda_2^*) = \frac{L(\lambda^*, r_1 (c_1 - c_2))}{r_1}
\]

and the second inequality in (35) holds if and only if \( r_1 \geq r \).

(ii) Suppose that \( p_1^* < p_2^* \). Then \( v_1(\lambda_1^*) - v_2(\lambda_2^*) < (c_1 - c_2) E[W(\lambda^*)] \). From (36) we have

\[
0 < v_1(\lambda_1^*) - v_2(\lambda_2^*) < (c_1 - c_2) E[W(\lambda^*)] \leq \frac{L(\lambda^*, r_1 (c_1 - c_2))}{r_1},
\]

where \( \alpha_1^* > \alpha_2^* \) implies the first inequality. It follows that the second inequality in (35) holds and type-1 chooses her targeted tariff for all \( r_1 \). If type-2 is RN, she also prefers the type-1 tariff: we have \( \lim_{r_2 \to 0} -L(\lambda^*, r_2 (c_2 - c_1)) / r_2 = (c_1 - c_2) E[W(\lambda^*)] > v_1(\lambda_1^*) - v_2(\lambda_2^*) \), where the equality follows from (36) and the inequality from (38), so that the first inequality in (35) is violated. Since \( -L(\lambda^*, r_2 (c_2 - c_1)) / r_2 \) strictly decreases in \( r_2 \) by Part 3 of Lemma 1, and \( \lim_{r_2 \to \infty} L(\lambda, r_2 (c_2 - c_1)) / r_2 = 0 \) because \( c_2 < c_1 \), there is an unique \( r \in (0, \infty) \) such that

\[
v_1(\lambda_1^*) - v_2(\lambda_2^*) = -\frac{L(\lambda^*, r_2 (c_2 - c_1))}{r_2}
\]

and the first inequality in (35) holds if and only if \( r_2 \geq r \).

**Sufficient conditions for Assumptions A2 and A3.**

**A2.** Two mild conditions are sufficient for \( \Pi(\lambda) \) and \( \Pi^f(\lambda; r) \) to be strictly concave in \( \lambda \).

(i) The gross revenue function \( \lambda_E(\lambda) \) is strictly concave. Many common distributions satisfy this assumption, e.g., the uniform, normal, logistic, Laplace and power function distributions, and the gamma and Weibull distributions with shape parameter \( \geq 1 \).
(ii) The function \( \lambda L(\lambda, s) \) is convex in \( \lambda \) for \( s \geq 0 \), i.e., \( 2L_\lambda(\lambda, s) + \lambda L_{\lambda\lambda}(\lambda, s) \geq 0 \). This assumption also implies that \( \lambda E[W(\lambda)] \) is convex in \( \lambda \) since \( L(\lambda, r)/r \) converges uniformly to \( E[W(\lambda)] \) as \( r \to 0 \). This assumption seems reasonable for many queueing models. A sufficient condition is that \( W(\lambda) \) is stochastically increasing in \( \lambda \) and that its c.d.f. is concave in \( \lambda \); for \( G/G/1 \) queues both properties follow from a similar argument as in Weber (1983).

A3. \( L_\lambda(\lambda, s)/s \) increases in \( s \), i.e., \( sL_{\lambda s}(\lambda, s) - L_\lambda(\lambda, s) \geq 0 \). By Lemma 1, this holds for the \( M/G/1 \) queue: \( L_\lambda(\lambda, 0) = 0 \) by Part 2 of Lemma 1 and \( L_\lambda(\lambda, s) \) is strictly convex in \( s \) by Part 4(b).

Proof of Proposition 5. Part 1. That \( \Pi^f(\lambda) < \Pi^* \) for \( r > 0 \) follows from the uniqueness of the optimal price function by Part 1 of Proposition 1. We next show that \( \lambda^f(\lambda) < \lambda^* \) for \( r > 0 \). We first establish that \( \Pi'(\lambda) > \Pi^f_\lambda(\lambda; r) \). From (14) and (15)

\[
\Pi'(\lambda) = \varphi(\lambda) + \lambda \varphi'(\lambda) - c\left( E[W(\lambda)] + \lambda \frac{dE[W(\lambda)]}{d\lambda} \right),
\]

(39)

\[
\Pi^f_\lambda(\lambda; r) = \varphi(\lambda) + \lambda \varphi'(\lambda) - \frac{L(\lambda, rc) + \lambda L_\lambda(\lambda, rc)}{r}.
\]

(40)

The difference satisfies

\[
\Pi'(\lambda) - \Pi^f_\lambda(\lambda; r) = \left( \frac{L(\lambda, rc)}{r} - cE[W(\lambda)] \right) + \lambda \left( \frac{L_\lambda(\lambda, rc)}{r} - \frac{dE[W(\lambda)]}{d\lambda} \right) > 0,
\]

(41)

because the first bracket is positive by Part 3 of Lemma 1, and the second is non-negative as we show next. A3 implies that \( L_\lambda(\lambda, rc)/r \) increases in \( r \). Furthermore,

\[
\lim_{r \to 0} \frac{L_\lambda(\lambda, rc)}{r} = \frac{d}{d\lambda} \lim_{r \to 0} \frac{L(\lambda, rc)}{r} = \frac{dE[W(\lambda)]}{d\lambda},
\]

(42)

where the first equality holds since \( L_\lambda(\lambda, 0) = 0 \) (Part 2 of Lemma 1) and from l’Hôpital’s rule, and the second by Part 3 of Lemma 1. Using (41) we have \( \Pi^f_\lambda(\lambda^*; r) < \Pi'(\lambda^*) = 0 \) where the equality follows by A1. Finally, because \( \Pi(\lambda) \) is strictly concave by A2, the optimal \( \Pi^f(\lambda; r) \) is obtained for \( \lambda < \lambda^* \).

Since Part 3 of Lemma 1 and (42) imply that \( \lim_{r \to 0} \Pi^f_\lambda(\lambda; r) = \Pi'(\lambda) \) for all \( \lambda \), \( \lim_{r \to 0} \lambda^f(\lambda) = \lambda^* \) and \( \lim_{r \to 0} \Pi^f(r) = \Pi^* \). That \( \lim_{r \to 0} \alpha^f(r) < \alpha^* \) follows from (16) because

\[
\lim_{r \to 0} \alpha^f(r) = \lim_{r \to 0} \left( \varphi(\lambda^f(r)) - \frac{L(\lambda^f(r), rc)}{r} \right) = \varphi(\lambda^*) - cE[W(\lambda^*)] < \alpha^* = \varphi(\lambda^*).
\]

Part 2. First note that since \( \lambda^f(\lambda) < \lambda^* \), if \( \lambda^f(\lambda) \geq \lambda^* \) it must satisfy \( \Pi^f_\lambda(\lambda^f(\lambda); r) = 0 \).

Next, since \( \Pi^f(\lambda; r) \) is strictly concave in \( \lambda \) for fixed \( r \) (by A2) and since \( \Pi^f_\lambda(\lambda; r) < 0 \) as shown in (43), it follows that if \( \lambda^f(\lambda) > 0 \) then \( \lambda^f(\lambda) \) and \( \Pi^f(r) \) are strictly decreasing at \( r \), and that if \( \lambda^f(\lambda^0) = 0 \) for some \( \lambda^0 \) then \( \lambda^f(\lambda) = \Pi^f(r) = 0 \) for all \( r \geq \lambda^0 \). From (40) we have

\[
\Pi^f_\lambda(\lambda; r) = -\frac{c(\lambda L_\lambda(\lambda, rc) - L(\lambda, rc) + \lambda L_\lambda(\lambda, rc))}{r^2} < 0,
\]

(43)

where the first bracket in the numerator is positive since \( L(\lambda, rc)/r \) is strictly increasing in \( r \geq 0 \) by Part 3 of Lemma 1, and the second is non-negative by A3.
It remains to establish the threshold \( \tau \) in (17). Since \( \Pi^f(\lambda;r) \) is strictly concave in \( \lambda \) for fixed \( r \), the arrival rate \( \lambda^f(r) > 0 \) if and only if \( \Pi^f(0;r) > 0 \). Note that \( \lim_{r \to 0} \Pi^f_{\lambda}(0;r) = \Pi^f(0) > 0 \) (the equality follows from Part 1, the inequality from (11)), and that from (40)

\[
\lim_{r \to \infty} \Pi^f_{\lambda}(0;r) = \nu(0) - \lim_{r \to \infty} \frac{L(0,rc)}{r} = \nu(0) - c \lim_{r \to \infty} L_s(0,rc) = -\infty.
\]

Since \( \Pi^f_{\lambda}(0;r) < 0 \), there is an unique risk aversion parameter \( \tau \in (0,\infty) \) such that \( \Pi^f_{\lambda}(0,\tau) = 0 \), or equivalently, \( \frac{L(0,\tau c)}{\tau} = \nu(0) \), and \( \Pi^f_{\lambda}(0;r) > 0 \) if and only if \( r < \tau \).

**Part 3.** See Figure 2 for Example 1 which shows that \( \alpha^f(r) > 0 \) and increases in \( r \in [0,2] \). That \( \alpha^f(r) \) is nonmonotone in \( r \) follows from (16) because

\[
\lim_{r \to \tau} \alpha^f(r) = \lim_{r \to \tau} \left( \frac{\nu(\lambda^f(r)) - \frac{L(\lambda^f(r),rc)}{r}}{\nu(0) - \frac{L(0,\tau c)}{\tau}} \right) = 0.
\]

The second equality holds since \( \lim \inf_{r \to \tau} \lambda^f(r) = 0 \) by Part 2, and the third by (17).

**Proof of Proposition 6.** Let \( \Pi^f(\lambda,\mu;r) \) be the revenue function under flat rate pricing. With RN customers, flat rate and optimal lead time-dependent pricing yield the same revenue, so \( \Pi^f(\lambda,\mu;0) \) is the revenue function under optimal lead time-dependent pricing. In the proof we first derive properties of \( \Pi^f(\lambda,\mu;r) \) for \( r > 0 \) and then take limits as \( r \to 0 \).

Let the r.v. \( W(\lambda,\mu) \) denote the steady-state lead time, \( \tilde{W}(\lambda,\mu,s) := E[\exp (sW(\lambda,\mu))] \) its MGF, and \( L(\lambda,\mu,s) := \ln \tilde{W}(\lambda,\mu,s) \) its semi-invariant MGF. We have from (14) and (15) that

\[
\Pi^f(\lambda,\mu;0) = \lambda \left( \nu(\lambda) - cE[W(\lambda,\mu)] \right), \quad (44)
\]

\[
\Pi^f(\lambda,\mu;r) = \lambda \left( \nu(\lambda) - \frac{1}{r}L(\lambda,\mu,rc) \right), \quad (45)
\]

\[
\Pi^f_{\lambda}(\lambda,\mu;r) = \nu(\lambda) + \lambda \gamma^f(\lambda) - \frac{L(\lambda,\mu,rc) + \lambda \nu(\lambda,\mu,rc)}{r}, \quad (46)
\]

\[
\Pi^f_{\mu}(\lambda,\mu;r) = -\lambda \frac{L(\lambda,\mu,rc)}{r}. \quad (47)
\]

Let \( \gamma^f(\mu;r) := \text{arg max}_{\lambda} \Pi^f(\lambda,\mu;r) \), so \( \Pi^*_{\mu}(\mu;\mu;r) = \Pi^f(\gamma^f(\mu;r),\mu;r) \) and \( \Pi^*_{\mu}(\mu) = \Pi^f(\gamma^f(\mu;0),\mu;0) \).

**Parts 1(a)/2(a).** We have \( \lim_{\mu \to \infty} \Pi^*_{\mu}(\mu) = \lim_{\mu \to \infty} \Pi^*_{\mu}(\mu) = \text{max}_{\lambda \in [0,\Lambda]} \lambda \nu(\lambda) \) since \( \lim_{\mu \to \infty} L(\lambda,\mu,rc) = 0 \) and so \( \lim_{\mu \to \infty} \Pi^f(\lambda,\mu;r) = \lambda \nu(\lambda) \) for all \( \lambda, r \). We have that

\[
L\lambda(\lambda,\mu,rc) \geq 0 \geq L\mu(\lambda,\mu,rc) \text{ for } \lambda < \mu, \text{ rc } \geq 0,
\]

because \( W(\lambda,\mu) \) is stochastically increasing in \( \lambda \) (decreasing in \( \mu \)), and therefore the expectation of every increasing function of \( W(\lambda,\mu) \), such as the MGF with a positive argument \( (rc \geq 0) \), is increasing in \( \lambda \) (decreasing in \( \mu \)).

It follows that the bracketed term in (45) decreases in \( \lambda \) and increases in \( \mu \), so \( \Pi^f(\mu) = 0 \) for

\[
\mu \leq \underline{\mu} := \text{arg} \{ \mu \geq 0 : \nu(0) - \frac{1}{r}L(0,\mu,rc) = 0 \}. \quad (49)
\]

(Note that \( L(0,\mu,rc) \) is the semi-invariant MGF of the service time, evaluated at \( rc \).) Similarly, we have from (44) that \( \Pi^*_{\mu}(\mu) = 0 \) for

\[
\mu \leq \underline{\mu} := \text{arg} \{ \mu \geq 0 : \nu(0) - cE[W(0,\mu)] = 0 \} = c/\nu(0), \quad (50)
\]
where the last equality holds because \(E[W(0, \mu)] = 1/\mu\). Note that \(\mu < \mu^-\) because \(cE[W(0, \mu)] < \frac{1}{r}L(0, \mu, rc)\) for \(r > 0\) by Part 3 of Lemma 1.

For \(\mu > \mu^-\) we have \(\lambda^{f*}(\mu; r) > 0\), and furthermore

\[
\frac{d\Pi^f(\lambda^{f*}(\mu; r), \mu; r)}{d\mu} = \Pi^f(\lambda^{f*}(\mu; r), \mu; r) = -\lambda^{f*}(\mu; r) \frac{L(\lambda^{f*}(\mu; r), \mu, rc)}{r} \geq 0. \tag{51}
\]

The first equality in (51) holds since \(\Pi^f(\lambda^{*}(\mu; r), \mu; r) = 0\) by \(A1\), and \(|d\lambda^{f*}(\mu; r)/d\mu| < \infty\) since \(\Pi^f(\lambda^{*}(\mu; r)) < 0\) by \(A2\), the second equality holds by (47), and the inequality follows from (48).

Furthermore, we have

\[
\lim_{\mu \to \mu^+} \Pi^{f*}(\mu) = 0 = \lim_{\mu \to \infty} \Pi^{f*}(\mu). \tag{52}
\]

The first equality holds by (51) since \(\lambda^{f*}(\mu^f; r) = 0\), and \(L(0, \mu^f, rc) = v(0) < \infty\) implies that \(|L(\mu^f, \mu^f, rc)| < \infty\). The second equality holds since \(\Pi^f(\mu)\) is increasing with \(\lim_{\mu \to \infty} \Pi^f(\mu) = \max_{\lambda \in [0, \Lambda]} \lambda^f(\lambda).

A similar argument (as \(r \to 0\)) shows \(\Pi^{*'}(\mu) \geq 0\) for \(\mu > \mu\) and \(\Pi^{*'}(\mu) = 0 = \lim_{\mu \to \infty} \Pi^{*'}(\mu).

**Parts 1(b)/2(b)** We omit the straightforward proof that \(0 < \mu^- < \mu^f < \infty\). That \(\mu^* (b) > \mu\) for \(b < \mu^-\) \(\mu^f (b) > \mu^f\) for \(b < \mu^f\) is immediate from 1(a) (from 2(a)). That \(\mu^* (b)\) and \(\mu^f (b)\) decrease in \(b\) follows since the profit functions \((\Pi^*(\mu) - b\mu)\) and \((\Pi^{f*}(\mu) - b\mu)\) are submodular in \((b, \mu)\).