M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 3, Summer 2013, pp. 423–443
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)
http://dx.doi.org/10.1287/msom.2013.0449

Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay

Philipp Afèche
Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada,
afeche@rotman.utoronto.ca

How should a firm design a price/lead-time menu and scheduling policy to maximize revenues from heterogeneous time-sensitive customers with private information about their preferences? We consider this question for a queueing system with two customer types and provide the following results. First, we develop a novel problem formulation and solution method that combines the achievable region approach with mechanism design. This approach extends to menu design problems for other systems. Second, the work conserving $c\mu$ priority rule, known to be delay cost minimizing, incentive-compatible, and socially optimal, need not be revenue maximizing. A *strategic delay* policy may be optimal: It prioritizes impatient customers, but artificially inflates the lead times of patient customers. This suggests a broader guideline: Revenue-maximizing firms that lack customer-level demand information should also consider customer incentives, not only operational constraints, in their scheduling policies. Third, we identify general necessary and sufficient conditions for optimal strategic delay: a *price*, a *lead-time*, and a *segment-size* condition. We translate these into demand and capacity parameter conditions for cases with homogeneous and heterogeneous valuations for each type. In some cases strategic delay is optimal if capacity is relatively abundant, in others if it is relatively scarce.

Key words: congestion; delay; incentives; lead times; mechanism design; pricing; priorities; quality of service; queueing systems; revenue management; scheduling; service differentiation
History: Received: January 3, 2011; accepted: January 18, 2013.

1. Introduction

1.1. Motivation, Example, and Research Question

How should a capacity-constrained firm design a price/lead-time menu and scheduling policy to maximize its revenues from heterogeneous time-sensitive customers with private information about their preferences? This question is relevant for service and manufacturing firms whose customers' willingness to pay for a product or service depends on the lead time between order placement and delivery. Recognizing that some customers value speedy service more than others, Federal Express and United Parcel Service offer a menu of differentiated price/lead-time options, for example, same-day or two-day service. Such lead-time-based price and service differentiation can also be a valuable revenue management tool for manufacturing firms, particularly those with a make- or assemble-to-order process. For example, Beta LAYOUT (beta-layout.com), a printed circuit board supplier with headquarters in Germany, offers a price/lead-time menu to its over 28,000 customers. The problem of designing the revenue-maximizing menu and the corresponding scheduling policy is significantly complicated if the provider *does not know individual customers' preferences*, but only has aggregate information about their attributes, for example, based on market research. In such settings, each customer chooses her preference among the menu options, and the provider must account for these service class choices, which give rise to *incentive-compatibility* constraints. (We write "IC" for "incentive-compatibility" and "incentive-compatible.")

We consider this problem in the context of a queueing model of a monopoly firm. IC pricing and scheduling in queueing systems is well understood under social optimization, but not so under revenue/profit optimization. Social optimization is a key objective for a government service, for an internal service center, or from a regulatory perspective, but a commercial firm serving external customers is primarily concerned with its own profitability. As this paper shows, revenue-maximizing and IC price/lead-time menus yield optimal scheduling policies with novel features. Consider the following example, which also illustrates our model. We study the problem for two customer segments or types, indexed by $i \in \{1, 2\}$, and model the operation as an $M/M/1$ system. Let μ be the capacity in jobs per unit time. For maximum simplicity, this example assumes *ample* capacity: $\mu = \infty$. This means that every work conserving scheduling policy yields zero lead times. (The paper specifies the optimal menu as a function of $\mu \leq \infty$.) The types differ in their arrival rates, valuation distributions, and

Table 1 If Types Are Not Observable, Then the Revenue-Maximizing Policy Uses *Strategic Delay*

Policy	FB		SB-wc		SB	
Provider observes types	Yes		No		No	
Scheduling policy	Optimal: Absolute priority to 1, work conserving		Not optimal: Absolute priority to 1, work conserving		Optimal: Absolute priority to 1, *strategic delay*	
Customer type	1	2	1	2	1	2
Price	1.5	0.5	0.64	0.64	1.36	0.48
Lead time	0	0	0	0	0	0.44
Demand rate	2.5	5.0	3.93	3.57	2.74	4.32
Revenue rate	3.75	2.5	2.53	2.30	3.72	2.08
Total revenue rate	6.25		4.82 (−22.8% vs. FB)		5.79 (−7.3% vs. FB)	

Notes. Valuations: type 1 $\sim U[0, 3]$, type 2 $\sim U[0, 1]$. Delay cost rates: $c_1 = 2$, $c_2 = 0.2$. Arrival rates: $\Lambda_1 = 5$, $\Lambda_2 = 10$.

delay cost rates. Type i arrive at a rate Λ_i per unit time; we call Λ_i the segment size. Let $\Lambda_1 = 5$ and $\Lambda_2 = 10$. A customer's valuation represents her willingness to pay for zero lead time. Type 1 valuations are uniformly distributed on the interval $[0, 3]$, and those of type 2 are uniformly distributed on $[0, 1]$. Customers are time sensitive: The *net* value of a type i with valuation v and lead time w is $v - c_i w$, where $c_i > 0$ is the type i delay cost rate per unit of lead time. Type 1 customers are more impatient: let $c_1 = 2 > c_2 = 0.2$. Customers do not observe the queue. Based on the prices and lead times, they choose which service class to purchase, if any, to maximize their net value minus price from service. Taking this choice behavior into account, the provider chooses a static price/lead-time menu and a scheduling policy to maximize her revenue rate. Table 1 summarizes the revenue-maximizing price/lead-time menu and demand and revenue rates for three policies.

The *first-best* policy (FB) is the one that maximizes the revenue if the provider *can* observe types. It charges different prices for zero lead times. For $\mu = \infty$, every work conserving scheduling policy is first-best. The one shown in Table 1 is the standard $c\mu$ priority rule, since it is the unique first-best policy for $\mu < \infty$. It assigns static priorities in increasing order of jobs' $c_i\mu_i$ index, so type 1 get priority since $c_1 > c_2$. The $c\mu$ policy minimizes the system's delay cost rate. The $c\mu$ policy is also *socially optimal and IC* (Mendelson and Whang 1990). This is clear in this example: with ample capacity, it is socially optimal and IC to serve everyone with a price and lead time of zero.

However, under revenue maximization, the $c\mu$ policy is *not* in general part of the *second-best* policy, that is, the optimal one if the provider *cannot* observe types. In the example, the menu under FB is not IC since type 1 would not pay more for the same lead time. If a provider is *restricted* to work conserving policies, IC requires charging a single price; the second-best policy with this restriction (SB-wc) yields a revenue loss of over 20% versus FB.

However, the provider *can* charge the impatient type 1 a premium *if* she artificially inflates the lead time targeted to the patient type 2: doing so deters type 1 from the slower, cheaper service preferred by type 2. This is the unrestricted second-best policy (SB) in this example. We call this artificial delay policy *strategic delay*, because its rationale is to manipulate customers' strategic service class choices, and its operational impact is that scheduling is no longer work conserving. Strategic delay *increases* the delay cost, which sets it apart from standard instances of optimal server idleness (see Kanet and Sridharan 2000). The value of strategic delay can be significant. The revenue loss of SB versus FB drops to roughly 7%, and the gain versus SB-wc exceeds 20%. However, strategic delay is not always optimal. This paper focuses on the question: *Under what demand and capacity conditions is strategic delay optimal*?

1.2. Summary of Main Results and Contributions

This paper develops a new analytical approach and identifies novel solution properties for the problem of designing revenue-maximizing and IC price/lead-time menus and scheduling policies for queueing systems.

1. *Problem formulation and solution method*. We use mechanism design and adapt the achievable region approach to formulate the scheduling control problem as a nonlinear program in arrival rates and lead times. Two sets of constraints ensure IC and operationally achievable lead times. We solve this problem in two steps. First, we show how the lead-time constraints partition the arrival rate space and identify the optimal policy for each set in the partition. Second, we optimize the resulting piecewise revenue function over arrival rates. This approach for designing price/lead-time menus is novel and can

be applied to systems with different operational or demand attributes.

2. *Implications of optimal strategic delay for the design of scheduling policies*. A key insight of this paper is that strategic delay can be optimal under a range of plausible conditions. Since strategic delay is neither work conserving nor delay cost minimizing, this result implies a more general guideline: In designing IC and revenue-maximizing price/lead-time menus, providers should not restrict attention to standard scheduling policies that minimize delay costs or related measures—in our model, equivalent to the work conserving $c\mu$ rule. Such policies are operationally appealing and maximize revenues in the absence of IC constraints, but they ignore customer incentives. The optimality of strategic delay also raises implementation issues. We discuss three approaches: idling the server before, reducing its speed during, and delaying the delivery after processing.

3. *Demand and capacity conditions for optimal strategic delay*. We identify three general necessary and sufficient conditions for optimal strategic delay: a *price*, a *lead-time*, and a *segment-size* condition. We apply these to specific valuation distributions and obtain explicit demand and capacity parameter conditions. The types' maximum valuation-to-delay cost ratios, segment sizes, and the demand elasticities corresponding to their valuations play key roles in these conditions. (i) In the special case of homogeneous valuations for each type, strategic delay is optimal if and only if the patient type has the higher valuation-to-delay cost ratio, the lower valuation, its segment is not too large relative to the impatient segment, and the capacity exceeds a threshold. (ii) Valuation heterogeneity for each type yields different results, as we show for uniformly distributed valuations. Strategic delay can be optimal for every ranking of the types' maximum valuation-to-delay cost ratios. Moreover, strategic delay is not necessarily a "large capacity phenomenon": If impatient customers have the higher maximum valuation-to-delay cost ratio, which is quite plausible, then under mild conditions strategic delay may be optimal only with relatively scarce but not with ample capacity, or only at low and high but not at intermediate capacity. (iii) For general valuation distributions, if impatient (patient) customers have the higher maximum valuation-to-delay cost ratio, then strategic delay is (not) optimal at the lowest capacity levels for which it is second-best to serve both types.

1.3. Literature and Positioning

This paper builds on queueing control and economic mechanism design tools. Traditionally, analysis, design, and control problems for queueing systems assume that the system manager is fully informed about and controls all job flows. See Stidham (2002) for a survey. Minimizing the delay cost or related measures is a prevalent optimality criterion in these settings. We adapt the achievable region approach to multiclass stochastic scheduling problems, pioneered by Coffman and Mitrani (1980). In contrast to the standard approach, we extend the achievable region to policies that are not work conserving, and we restrict it by IC constraints.

Mechanism design tools have been applied to many resource allocation problems under private information. Myerson (1981) provides a seminal analysis of optimal auction design. Among studies of screening or adverse selection problems, papers on the design of price/quality menus (e.g., Mussa and Rosen 1978, Rochet and Chonè 1998) are closest to ours. Although some form of quality degradation similar to strategic delay is known to be optimal in these "standard" screening problems, two important features distinguish our setup from the standard one. First, the capacity constraint, queueing, and delay costs imply externalities among service classes. Second, the provider controls these externalities through the price/lead-time menu and her scheduling policy.

Like in this paper, Su and Zenios (2006) use mechanism design and the achievable region approach to solve a problem with a capacity constraint and externalities, namely, allocating a fixed total supply rate of quality-differentiated kidneys to decoupled queues of risk-differentiated, privately informed transplant candidate types with fixed arrival rates. Unlike in this paper, their analysis involves no pricing, focuses on social welfare criteria, uses fluid approximations that ignore queueing constraints, and characterizes the achievable region of expected kidney quality vectors.

This paper fits in a research stream on pricing and operational decisions for queueing systems with self-interested customers. Naor (1969) started this stream. See Hassin and Haviv (2003) for an excellent survey. We consider static price/lead-time menus, unlike papers on dynamic price and/or lead-time quotation (e.g., Plambeck 2004, Çelik and Maglaras 2008, Ata and Olsen 2013).

Three characteristics jointly distinguish our problem from most others on static price/lead-time menus: revenue maximization, schedule optimization, and customers who choose their class.

In problems of *socially optimal* and *IC* pricing and scheduling, the solution is the same as that without IC constraints, and the optimal scheduling policy is work conserving (Mendelson and Whang 1990, Van Mieghem 2000, Hsu et al. 2009). Most studies of *revenue/profit maximization* restrict the scheduling policy, customers' service class choices, or both (Lederer and Li 1997, Rao and Petersen 1998, Boyaci and Ray 2003, Maglaras and Zeevi 2003, Afèche and

Mendelson 2004, Maglaras and Zeevi 2005, Allon and Federgruen 2009, Jayaswal et al. 2011, Zhao et al. 2012).

This paper is part of a stream of studies that design static revenue-maximizing and IC price/lead-time menus and corresponding scheduling policies. The problem formulation and solution method presented in this paper, and the concept of strategic delay and its potential value, were first identified and illustrated in early drafts of this work leading up to Afèche (2004). He derives for the two-type model considered here partial sufficient demand conditions under which strategic delay may be optimal. He also shows that it may be optimal to alter priority assignments if types have different service requirements. His analysis assumes fixed unit capacity and uniform valuations. This paper identifies necessary and sufficient demand and capacity conditions for optimal strategic delay under general valuation distributions, and it relates strategic delay to the first- and second-best solutions.

Cui et al. (2012) extend the analysis of Afèche (2004) for a special case of his model, by allowing the provider to choose a static admission probability for each class; such rationing may yield different lead times for customers with the same delay cost. Katta and Sethuraman (2005) and Afèche and Pavlin (2011) show for a multitype model that it may optimal to pool multiple types into a common service class. The latter show that strategic delay may also be optimal and characterize the optimal menu as a function of demand parameters and the capacity. Yahalom et al. (2006) perform an approximate analysis of the optimal menu and scheduling policy for fixed arrival rates, under convex increasing delay costs. Maglaras et al. (2013) derive structural insights on the IC and revenue-maximizing menu based on an asymptotic analysis for multiserver systems.

There are interesting connections between strategic delay and policies of discretionary task completion and service inducement in queueing systems. The benefits of these policies derive from increasing *service* times of jobs. In contrast, strategic delay requires increasing the *entire lead time*, regardless of its effect on service times. In discretionary task completion, the value of a job increases in its service time (Hopp et al. 2007), whereas strategic delay does not affect job values. Debo et al. (2008) study service inducement, whereby an expert provides unnecessary services that add no value to the customer but that allow the expert to increase revenues by charging based on her service time. Strategic delay and service inducement have in common that they increase revenues by adding unnecessary delays, but there are important differences between these concepts. Strategic delay arises only if customers are heterogeneous and can choose from a menu of classes. In Debo et al. (2008), customers are identical and the provider offers a single first-in, first-out (FIFO) class; customers only choose whether to buy or not. The benefit of service inducement derives from its state-dependent nature: Customers observe whether the provider is idle or busy, and she only induces service for those who arrive when she is idle. These modeling differences imply differences in the results. Service inducement increases revenues on delayed customers, whereas strategic delay reduces revenues on delayed (low priority) customers, but increases revenues on high priority customers. Furthermore, whenever it is optimal, service inducement reduces welfare. In contrast, optimal strategic delay may yield a Pareto improvement versus the second-best work conserving policy (Example 2 in §6.4).

1.4. Plan of the Paper

In §2 we describe the model and formulate the problem. In §3 we specify admissible scheduling policies and strategic delay. In §4 we transform the problem by adapting the achievable region approach, present the solution method, and summarize the main notation. In §5 we present the first-best solution. The main results are in §6: We solve the second-best problem, develop the necessary and sufficient conditions for optimal strategic delay, translate these into demand parameter and capacity conditions, and explain within our framework why the $c\mu$ policy is socially optimal and IC. In §7 we offer concluding remarks. Proofs are in the online supplement (available at http://dx.doi.org/10.1287/msom.2013.0449).

2. Model and Problem Formulation

2.1. Model Primitives

We model a capacity-constrained firm that faces a population of small price- and delay-sensitive potential customers as an $M/M/1$ system. "Delay" or "lead time" interchangeably refer to the entire time interval between the placement and delivery of an order. The marginal cost of service is zero. Customers differ ex ante in two attributes, their valuation and delay cost rate, but have independent and identically distributed (i.i.d.) nominal service times as explained below. They are grouped based on their delay cost rates into two types or segments, indexed by $i \in \{1, 2\}$. Type i arrivals are Poisson with fixed rate or segment size Λ_i; let $\mathbf{\Lambda} \triangleq (\Lambda_1, \Lambda_2)$. (We write all vectors in boldface.) The arrival rate of any customer is infinitesimal relative to Λ_i. Each arrival is for one unit of service. We specify purchase decisions in §2.2. The arrival processes and the distributions of customer attributes are mutually independent.

Valuations. A customer's valuation v represents her willingness to pay for instant delivery. Type i

valuations are i.i.d. draws from a continuous distribution with cumulative distribution function F_i and continuous probability density function f_i, where $f_i(v) > 0$ for $v \in [\underline{v}_i, \bar{v}_i]$ and $0 \le \underline{v}_i < \bar{v}_i < \infty$. Let $\bar{F}_i = 1 - F_i$, and let $\bar{F}_i^{-1}$ be its inverse.

Delay Costs. All type i customers have the same constant delay cost rate $c_i > 0$ per unit time in the system. We assume without loss of generality (w.l.o.g.) that $c_1 > c_2$ and refer to type 1 as *impatient* customers and to type 2 as *patient* customers. For service with lead time w, a type i customer with valuation v is willing to pay $v - c_i \cdot w$; we call this amount her *net value*.

Nominal Service Times. A job's *nominal* service time is defined as its total service time if the server works at its maximum rate while dedicated to that job. As discussed in §3, if the server slows down for certain jobs, then their *effective* service times exceed their nominal values. Nominal service times are i.i.d. draws from an exponential distribution with mean $1/\mu$. Let μ denote the nominal capacity in jobs per unit time. The results specify how the optimal menu depends on μ. We assume $\mu^{-1} < \mu_0^{-1} \triangleq \min(\bar{v}_1/c_1, \bar{v}_2/c_2)$, which rules out the trivial case where it is unprofitable to serve type i regardless of λ_j for $j \ne i$. It may still be *optimal* not to serve type i.

Information Structure. The provider knows the following *aggregate* demand statistics, for example, based on market research and/or past purchase data: the arrival process statistics including $\boldsymbol{\Lambda}$, the value distributions F_i, delay cost rates c_i, and the nominal service time distribution and its mean $1/\mu$. However, the delay cost rates and valuations of *individual* customers are the private information of each customer, based on the notion that the provider either cannot track individuals' purchase histories or that such information is inadequate to estimate their future purchase preferences. A customer only knows her nominal mean service time when deciding on her purchase; nominal service time realizations become known only once jobs are completed. Customers do not see the queue.

2.2. Mechanism Design Problem Formulation

The provider designs a price/lead-time menu and a scheduling policy to maximize her revenue rate subject to customers' choices. We formalize this problem as a mechanism design game.

Decisions and Timing. The provider first designs and announces a static *menu* of up to two service classes, indexed by $k \in \{1, 2\}$, and a scheduling policy r, which specifies how to process customers in each class. We outline the set of admissible scheduling policies below and specify it in detail in §3. "Class" refers to the attributes of a menu option; "type" refers to those of a customer. Class k has two attributes, a per-job price p_k and an expected lead time W_k, both scalar constants. We often refer to W_k simply as the class k lead time or delay. Let $\mathbf{p} \triangleq (p_1, p_2)$ and $\mathbf{W} \triangleq (W_1, W_2) \in \mathbb{R}_+^2$.

Customer arrival times are exogenous. However, customers are self-interested and strategic in their purchase decisions. They are risk neutral with respect to lead-time uncertainty and seek to maximize their expected utility. Upon her arrival a customer chooses based on the menu $(\mathbf{p}, \mathbf{W})$ which service class to purchase, if any, as specified below. We assume no reneging and no retrials. Let $\lambda_k(\mathbf{p}, \mathbf{W})$ be the class k arrival or demand rate as a function of the menu, and $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2)$. The provider serves all class k requests at the price p_k and schedules them based on the policy r.

Purchase Decisions and Demand Rates. Given a menu $(\mathbf{p}, \mathbf{W})$, the expected class k utility of a type i customer with valuation v is $v - c_i W_k - p_k$; her expected class k *full price* is $p_k + c_i W_k$. Customers who do not buy get zero utility. We restrict attention w.l.o.g. (see §2.3) to IC menus that target class i to type i customers, such that they (weakly) prefer class i or no service over service in class $j \ne i$. Given an IC menu, a type i with valuation v buys class i if $v - c_i \cdot W_i - p_i \ge 0$ and otherwise does not buy at all, which captures the individual rationality (IR) constraints and implies $\lambda_i = \Lambda_i \cdot \bar{F}_i(p_i + c_i W_i)$, $i = 1, 2$. We call class i *open* if $\lambda_i > 0$ and *closed* if $\lambda_i = 0$. To close class i, we set $p_i = \bar{v}_i - c_i W_i$; with this convention, the full prices satisfy $p_i + c_i W_i \le \bar{v}_i$ for $i = 1, 2$. To ensure that type i customers do not buy class $j \ne i$, the menu must satisfy the IC constraints $c_i \cdot W_i + p_i \le c_i \cdot W_j + p_j$ if $\lambda_j > 0$, $i \ne j$; the condition $\lambda_j > 0$ indicates that the constraint is only in force if class j is open.

Posted Lead Times, Realized Lead Times, and Admissible Scheduling Policies. Customers base their decisions on the *posted expected* delays $\mathbf{W}$. They do not possess information about queue lengths, scheduling policy, arrival rates, capacity, etc., to reliably forecast their mean delays. Let $w_i^r(\boldsymbol{\lambda}, \mu)$ denote the realized class i mean steady-state delay given a scheduling policy r, as a function of the demand vector $\boldsymbol{\lambda}$ and the capacity μ, and $\mathbf{w}^r(\boldsymbol{\lambda}, \mu) \triangleq (w_1^r(\boldsymbol{\lambda}, \mu), w_2^r(\boldsymbol{\lambda}, \mu))$. Although the realized lead times of individual customers deviate from the posted averages, we require that $\mathbf{W}$ agree with the system's realized *average* steady-state delays, based on the notion that reputation effects and third-party auditors instill in the provider the commitment to perform in line with her announcements. That is, we impose the *consistency* constraints $\mathbf{W} = \mathbf{w}^r(\boldsymbol{\lambda}, \mu)$, where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{p}, \mathbf{W})$ are the demand rates induced by the purchase decisions given the menu $(\mathbf{p}, \mathbf{W})$.

Let $\mathcal{A}$ be the set of admissible scheduling policies. Price/lead-time optimization studies typically assume a given policy such as FIFO, a priority discipline, or processor sharing, and restrict attention

to work conserving policies. This paper imposes neither restriction. To complete the problem formulation without delving into scheduling details, we defer the detailed definition and discussion of $\mathcal{A}$ until §3 (see Definition 2). Here we simply state two necessary and sufficient conditions on $\mathcal{A}$ for the provider's optimization problem to be well defined: If $\lambda_1 + \lambda_2 < \mu$, then, (i) for every admissible policy $r \in \mathcal{A}$, the mean steady-state delays $\mathbf{w}^r(\boldsymbol{\lambda}, \mu)$ are well defined and finite, and (ii) the minimum of the system-wide delay cost rate $\lambda_1 c_1 w_1^r(\boldsymbol{\lambda}, \mu) + \lambda_2 c_2 w_2^r(\boldsymbol{\lambda}, \mu)$ over $r \in \mathcal{A}$ exists.

Provider Problem. In summary, the provider solves

$$\max_{\mathbf{p}\in\mathbb{R}^2,\, \mathbf{W}\in\mathbb{R}^2_+,\, r\in\mathcal{A}} \quad \sum_{i=1}^{2} p_i \cdot \lambda_i \tag{1}$$

$$\text{s.t.} \quad \lambda_i = \Lambda_i \cdot \bar{F}_i(p_i + c_i \cdot W_i), \quad i = 1, 2, \tag{2}$$

$$c_i \cdot W_i + p_i \le c_i \cdot W_j + p_j \quad \text{if } \lambda_j > 0, \quad i \ne j, \tag{3}$$

$$\lambda_1 + \lambda_2 < \mu, \tag{4}$$

$$\mathbf{W} = \mathbf{w}^r(\boldsymbol{\lambda}, \mu), \tag{5}$$

where (2) is for IR, (3) is for IC, (4) is for stability, and (5) is for consistency. We call (1)–(5) the *second-best* problem. The *first-best* problem is the second-best problem without the IC constraints (3).

2.3. Discussion

(i) *Two delay costs*. We restrict attention to $N = 2$ delay costs, $c_i \in \{c_1, c_2\}$. This yields the *simplest* setup for our results. It approximates settings where jobs can be clustered into two segments based on their delay costs, for example, regular versus urgent, such that the delay cost variance within each segment is small relative to the variance between segments. Mendelson and Whang (1990) show for an arbitrary number $N \ge 2$ of delay costs that the work conserving $c\mu$ priority policy is *socially optimal* and IC. In §6.5 we revisit this result within our framework, which clarifies why the problem of IC revenue optimization for $N > 2$ (Katta and Sethuraman 2005, Afèche and Pavlin 2011) is more challenging than the problem of IC social optimization.

(ii) *Restriction to a single service class for customers with the same delay cost rate.* We restrict attention to menus that target a single class to all type i customers even though they differ in their valuations. This is w.l.o.g. under our assumption that the provider accepts all purchase requests: It can be shown that among all such static menus, the provider cannot increase revenues with a menu that targets two or more distinct service classes to type i customers based on their valuations.

(iii) *Restriction on prices.* Since service times are i.i.d., it is w.l.o.g. that prices are independent of service times. Charging based on realized service times may be optimal if $\mu_1 \ne \mu_2$ (Afèche 2004).

(iv) *Equivalence of mechanism specified in* §2.2 *to a direct revelation mechanism.* Based on the revelation principle (e.g., Myerson 1981), mechanism design problems restrict attention w.l.o.g. to IC direct revelation mechanisms in which each customer directly reveals her type. Although the mechanism specified in §2.2 is strictly speaking an "indirect mechanism," it is de facto equivalent to a direct revelation mechanism and more naturally describes how services are typically sold.

3. Admissible Scheduling Policies and Strategic Delay

This section defines the set of admissible scheduling policies $\mathcal{A}$ with a focus on extending the standard space of work conserving policies to include simple policies that are *not* work conserving.

3.1. Work Conserving Policies

As a reference point for $\mathcal{A}$, consider the following definition.

Definition 1. Consider a scheduling policy r that serves jobs in each class FIFO and does not serve multiple jobs simultaneously. It is *work conserving* if it satisfies the following conditions.

1. Nonidling: It does not idle the server when there is unfinished work in the system.
2. Service time invariance: It does not affect the total service time of any job.
3. Immediate delivery: It delivers each job to the customer as soon as it is completed.
4. Nonanticipative and regenerative: It assigns the server to jobs on the basis of the time elapsed and the history of the process since the last epoch at which the system became empty.

The restrictions to FIFO in each class and the exclusion of processor sharing are w.l.o.g. and made to simplify the exposition. Conditions 1 and 2 are standard. Condition 2 is equivalent to requiring that the total service time of every job equal its *nominal* service time, as defined in §2.1. Condition 2 allows, for example, preemptive-resume priority policies with zero switchover times, but rules out policies that slow down the server. Condition 3 is implicitly assumed as part of any notion of work conservation, but usually not explicitly stated; we do so here to highlight the possible differences between the features of policies that are work conserving and those that are not in our setting. Condition 4 is similar to the one in Federgruen and Groenevelt (1988) and appears to be

the most general, easily describable restriction under which the existence of long-run averages of waiting times is verifiable. Condition 4 excludes policies that depend on prior knowledge of customers' *actual* (remaining) nominal service times (we assume that this information is unavailable), or where decisions in one busy period depend on information concerning prior busy periods. Condition 4 does allow decisions to be based on the *expected* (remaining) or the accumulated nominal service times.

3.2. Strategic Delay and Admissible Policies

Conditions 1–3 in Definition 1 capture the essence of work conserving policies: they do not affect the evolution of the total work-in-system inventory. As a result, various performance vectors obey invariance principles known as conservation laws. Here, given $(\boldsymbol{\lambda}, \mu)$ and $\lambda_1 + \lambda_2 < \mu$, the mean lead times under a work conserving policy r satisfy

$$\frac{\lambda_1}{\mu} w_1^r(\boldsymbol{\lambda}, \mu) + \frac{\lambda_2}{\mu} w_2^r(\boldsymbol{\lambda}, \mu) = \frac{1}{\mu} \frac{\lambda_1 + \lambda_2}{\mu - \lambda_1 - \lambda_2}. \quad (6)$$

That is, the average work-in-system inventory equals a constant that is independent of r. Restricting attention to work conserving policies is intuitively appealing since they clear work from the system as quickly as "operationally feasible." However, as shown in §1, in our setting it may be optimal to intentionally inflate the delay of one class in a controlled fashion without altering the delay of the other class, which is clearly *not* work conserving. We call this artificial inflation *strategic delay* and consider three approaches for its implementation; each violates one condition of Definition 1:

1. Idling the server *before* commencing service on a waiting job, which violates Condition 1.

2. Reducing the server speed *during* processing, which violates Condition 2. Define a job's *effective* service time to be its total service time given the server's actual processing rate while working on that job. If the server slows down while processing a job, then its effective service time exceeds its nominal service time. We allow no other deviations from Condition 2; for example, customers do not renege, and any service preemptions are preemptive-resume and with zero switchover times. Therefore, *nominal* service times must be independent of the scheduling policy; *effective* service times may increase but only due to a reduced server speed.

3. Delaying the delivery of a job *after* its processing is completed, which violates Condition 3.

Under a policy r that uses any combination of these delay tactics, the left-hand side of (6) exceeds the constant on its right-hand side. Although one can conceive of arbitrarily sophisticated controls for these delay tactics, the focus of this paper is to identify the concept of strategic delay and conditions for its optimality, not to analyze its most sophisticated implementations. With this in mind, we restrict attention to the following class of policies. We discuss implementation criteria in §7.

Definition 2. For $\lambda_1 + \lambda_2 < \mu$, a scheduling policy r is *admissible* ($r \in \mathcal{A}$) if it serves jobs in each class FIFO, does not serve multiple jobs simultaneously, and satisfies the following conditions.

1. Server idleness: The provider chooses a random variable $I_i \geq 0$ with $E(I_i)^2 < \infty$ for $i = 1, 2$. It assigns to the nth class i job a cumulative server idle time I_i^n such that $\{I_i^n\}_{n=1}^{\infty}$ are i.i.d. as I_i and independent of all other system processes. If the set of jobs requiring processing is nonempty at time $t \in [0, \infty)$, then the server is assigned to exactly one job in this set. Whenever the server is assigned to the nth class i job, the server processes that job if and only if it has already idled for a total amount of time I_i^n while being assigned to that job.

2. Server speed: The provider chooses a constant server speed $k_i \in (0, 1]$ for $i = 1, 2$. It processes class i jobs at a fraction k_i of its maximum processing rate, so the effective class i capacity is $\mu \cdot k_i \leq \mu$. The scheduling policy does not affect the nominal service time of any job.

3. Delivery delays: The provider chooses a random variable $D_i \geq 0$ with $E(D_i) < \infty$ for $i = 1, 2$. It assigns to the nth class i job a delivery delay D_i^n such that $\{D_i^n\}_{n=1}^{\infty}$ are i.i.d. as D_i and independent of all other system processes. After its processing is completed, the nth class i job is put into a delay node for an amount of time D_i^n before delivery.

4. Nonanticipative and regenerative: It assigns the server to jobs on the basis of the time elapsed and the history of the process since the last epoch at which the system became empty.

5. Stability: The effective server utilization satisfies $\sum_{i=1}^{2} (\lambda_i E I_i + \lambda_i / (k_i \mu)) < 1$.

In Condition 1, notice the distinction between the server being *assigned* to and *processing* a job. While processing a job the server is also assigned to it, but while assigned to a job the server may be idling; key is that it is unavailable to *other* jobs. The condition that all of a job's cumulative server idle time be injected prior to processing is for simplicity and operational efficiency in that it minimizes processing interruptions. Call the sum of the cumulative server idle time plus the effective service time of a job its *server assignment time*. For given I_i and k_i, the server assignment times of class i jobs are i.i.d. and independent of the sequencing policy with mean $EI_i + (k_i \mu)^{-1}$, where $(k_i \mu)^{-1}$ is the effective class i mean service time. If $EI_i = 0$ and $k_i = 1$, the class i mean server assignment time equals the nominal mean service

time. Definition 2 allows preemption, that is, interruption of server assignment times. Preemption is instant and costless during the server idle time portion of a job's server assignment time. Whether preemption is feasible during processing depends on the operational context. For simplicity we allow preemption during processing.

4. Problem Transformation and Solution Method

In this section we first transform the formulation (1)–(5) from a problem in $(\mathbf{p}, \mathbf{W}, r)$ to an equivalent one in $(\boldsymbol{\lambda}, \mathbf{W})$. We then outline the solution method. Table 2 summarizes the main notation.

Define the marginal value functions

$$v_i(\lambda_i) \triangleq \bar{F}_i^{-1}\left(\frac{\lambda_i}{\Lambda_i}\right), \quad \lambda_i \in [0, \Lambda_i], \quad i = 1, 2, \tag{7}$$

where $v_i(\lambda_i)$ is the valuation of the marginal type i customer corresponding to λ_i. The properties of F_i imply $0 \le \underline{v}_i = v_i(\Lambda_i) < v_i(0) = \bar{v}_i < \infty$ and $-\infty < v_i'(\lambda_i) < 0$ on $[0, \Lambda_i]$. Let $\varepsilon_i(\lambda_i) \triangleq -v_i(\lambda_i)/\lambda_i v_i'(\lambda_i)$ denote the elasticity function corresponding to $v_i(\lambda_i)$. We further assume the following.

Assumption A1. *The minimum type i valuation $v_i(\Lambda_i) = 0$, $i = 1, 2$. This ensures that it is not optimal to serve all type i customers and rules out menus that satisfy $v_i(\Lambda_i) > c_i W_i + p_i$.*

Assumption A2. *Let $R_i(\lambda_i) \triangleq \lambda_i \cdot v_i(\lambda_i)$ be the type i gross revenue function; we assume that $R_i'' < 0$.*

Table 2 Summary of Main Notation

$\bar{v}_i, c_i$	Type i maximum value, delay cost rate
Λ_i, λ_i	Type i segment size, arrival rate
v_i, ε_i, R_i	Type i marginal value, elasticity, gross revenue functions
μ, μ_0	Capacity, minimum capacity
M	Set of feasible arrival rates
$\{M_0, M_1, M_2\}$	Partition of M for IC under $c\mu$ policy
$\mathbf{W}^f, \mathbf{W}^s$	Lead-time functions under first-, second-best policies
$\mathbf{w}^{c\mu}, \mathbf{p}^{c\mu}$	Lead-time, price functions under $c\mu$ policy
$\mathbf{w}^{sd}, \mathbf{p}^{sd}$	Lead-time, price functions under strategic delay policy
$\bar{W}$	Indifference threshold function for IC, where $\bar{W} = w_2^{sd}$
Π^f, Π^s, Π^{sd}	Revenue functions under first-, second-best, strategic delay policies
$\Pi^f_{\lambda_i}, \Pi^f_{\lambda_i\lambda_j}$	First-, second-order partial derivatives of Π^f (similarly for Π^s, Π^{sd})
$\boldsymbol{\lambda}^f, \boldsymbol{\lambda}^s, \boldsymbol{\lambda}^{sd}$	Optimal arrival rates under first-, second-best, strategic delay policies
λ_1°	Optimal type 1 arrival rate for $\lambda_2 = 0$
$\boldsymbol{\lambda}^{sd}(\lambda_1)$	Optimal arrival rates under strategic delay for fixed λ_1
$\lambda_2^{sd}(\lambda_1)$	Optimal type 2 arrival rate under strategic delay for fixed λ_1
μ^f, μ^s	Infimum of capacity levels where first-, second-best policy serves both types
μ^{sd}	Infimum of capacity levels where strategic delay is optimal

The demand functions (2) with $p_i \le \bar{v}_i - c_i W_i$ and A1 imply the inverse demand functions

$$p_i(\lambda_i, W_i) \triangleq v_i(\lambda_i) - c_i W_i, \quad \lambda_i \in [0, \Lambda_i], \ i = 1, 2. \tag{8}$$

The marginal type i customer has zero expected utility, that is, the price equals the marginal net value.

4.1. Operationally Achievable Lead Times

The problem (1)–(5) depends on an admissible scheduling policy $r \in \mathcal{A}$ only through its steady-state mean lead times $\mathbf{w}^r(\boldsymbol{\lambda}, \mu)$. We transform the control problem of choosing a policy $r \in \mathcal{A}$ into the simpler optimization problem of choosing a vector $\mathbf{W}$ in the corresponding *achievable region* $\mathrm{OA}(\boldsymbol{\lambda}, \mu) \triangleq \{\mathbf{w}^r(\boldsymbol{\lambda}, \mu)\colon r \in \mathcal{A}\}$ given $(\boldsymbol{\lambda}, \mu)$. Lemma 1 is immediate from the achievable region for work conserving policies (Coffman and Mitrani 1980).

Lemma 1. *Fix $\boldsymbol{\lambda}$ and μ such that $\lambda_1 + \lambda_2 < \mu$. Call a lead-time vector $\mathbf{W}$ operationally achievable for $(\boldsymbol{\lambda}, \mu)$ if $\mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda}, \mu)$. A vector $\mathbf{W}$ is operationally achievable for $(\boldsymbol{\lambda}, \mu)$ if and only if*

$$W_i \ge \frac{1}{\mu - \lambda_i}, \quad i = 1, 2, \tag{9}$$

$$\frac{\lambda_1}{\mu} W_1 + \frac{\lambda_2}{\mu} W_2 \ge \frac{1}{\mu} \frac{\lambda_1 + \lambda_2}{\mu - \lambda_1 - \lambda_2}. \tag{10}$$

The work conserving policies that give absolute preemptive priority to class 1 or 2 correspond to the extreme points of $\mathrm{OA}(\boldsymbol{\lambda}, \mu)$ and yield the lower bounds (9). For the standard achievable region, which is defined for work conserving policies, (10) holds with equality and corresponds to the conservation law (6).

4.2. IC Lead Times

Let $M(\mu) \triangleq \{\boldsymbol{\lambda}\colon \mathbf{0} \le \boldsymbol{\lambda} \le \boldsymbol{\Lambda}, \lambda_1 + \lambda_2 < \mu\}$ denote the set of feasible arrival rates. The IC constraints (3) and the inverse demand functions (8) yield IC constraints that only depend on $(\boldsymbol{\lambda}, \mathbf{W})$.

Definition 3. Fix μ and $\boldsymbol{\lambda} \in M(\mu)$.

1. Define the *indifference threshold*

$$\bar{W}(\boldsymbol{\lambda}) \triangleq \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2}, \quad \text{where } \bar{W}(\mathbf{0}) = \frac{\bar{v}_1 - \bar{v}_2}{c_1 - c_2}. \tag{11}$$

2. A lead-time vector $\mathbf{W}$, or a scheduling policy r with $\mathbf{w}^r(\boldsymbol{\lambda}, \mu) = \mathbf{W}$, is *IC for* $(\boldsymbol{\lambda}, \mu)$ if and only if $\mathbf{W} \in \mathrm{IC}(\boldsymbol{\lambda})$, where

$$\mathrm{IC}(\boldsymbol{\lambda}) \triangleq \left\{\mathbf{W}\colon \lambda_1(\bar{W}(\boldsymbol{\lambda}) - W_1) \ge 0 \ge \lambda_2(\bar{W}(\boldsymbol{\lambda}) - W_2)\right\}. \tag{12}$$

The indifference threshold $\bar{W}(\boldsymbol{\lambda})$ defined in (11) plays a key role in customers' service class choices: Both types have the same marginal net value for the lead time $\bar{W}(\boldsymbol{\lambda})$. We discuss the lead-time constraints (12) that define the IC region $\mathrm{IC}(\boldsymbol{\lambda})$ in §6.1.

4.3. Transformed Problem

We obtain the following equivalent problem from (1)–(5):

$$\max_{\boldsymbol{\lambda},\mathbf{W}} \Pi(\boldsymbol{\lambda},\mathbf{W}) \triangleq \sum_{i=1}^{2} \lambda_i \cdot p_i(\lambda_i, W_i)$$
$$= \sum_{i=1}^{2} \lambda_i \cdot (v_i(\lambda_i) - c_i \cdot W_i) \quad (13)$$
$$\text{s.t.} \quad \boldsymbol{\lambda} \in M(\mu), \quad (14)$$
$$\mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda},\mu) \cap \mathrm{IC}(\boldsymbol{\lambda}). \quad (15)$$

4.4. Solution Method

For fixed μ we solve the second-best problem (13)–(15) in two steps.

Step 1. Fix $\boldsymbol{\lambda} \in M(\mu)$. Determine the second-best lead-time vector

$$\mathbf{W}^s(\boldsymbol{\lambda},\mu) \triangleq \arg\min_{\mathbf{W}} \{\lambda_1 c_1 W_1 + \lambda_2 c_2 W_2 \text{ s.t. } \mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda},\mu) \cap \mathrm{IC}(\boldsymbol{\lambda})\}. \quad (16)$$

It maximizes the revenue and minimizes the delay cost rate over all operationally achievable and IC lead times for $(\boldsymbol{\lambda},\mu)$. It is unique if $\mathrm{OA}(\boldsymbol{\lambda},\mu) \cap \mathrm{IC}(\boldsymbol{\lambda}) \neq \varnothing$. Call a policy $r \in \mathcal{A}$ "second-best for $(\boldsymbol{\lambda},\mu)$" if $\mathbf{w}^r(\boldsymbol{\lambda},\mu) = \mathbf{W}^s(\boldsymbol{\lambda},\mu)$. Step 1 yields the second-best revenue function $\Pi^s(\boldsymbol{\lambda},\mu) \triangleq \Pi(\boldsymbol{\lambda},\mathbf{W}^s(\boldsymbol{\lambda},\mu))$.

Step 2. Determine the jointly second-best arrival rates $\boldsymbol{\lambda}^s(\mu) \triangleq \arg\max_{\boldsymbol{\lambda}\in M(\mu)} \Pi^s(\boldsymbol{\lambda},\mu)$ and lead times $\mathbf{W}^s(\boldsymbol{\lambda}^s(\mu),\mu)$. Any scheduling policy $r \in \mathcal{A}$ that achieves these lead times is second-best for $(\boldsymbol{\lambda}^s(\mu),\mu)$; constructing such a policy is the synthesis problem. The inverse demand functions (8) determine the unique optimal prices from $\boldsymbol{\lambda}^s(\mu)$ and $\mathbf{W}^s(\boldsymbol{\lambda}^s(\mu),\mu)$.

In §6 we execute these two steps to solve the second-best problem and characterize its solution depending on the demand parameters and capacity. In §5 we first use the same two-step approach to solve the benchmark first-best problem, that is, (13)–(15) without the IC constraints $\mathbf{W} \in \mathrm{IC}(\boldsymbol{\lambda})$:

1. Determine for fixed μ and $\boldsymbol{\lambda} \in M(\mu)$ the first-best lead-time vector

$$\mathbf{W}^f(\boldsymbol{\lambda},\mu) \triangleq \arg\min_{\mathbf{W}} \{\lambda_1 c_1 W_1 + \lambda_2 c_2 W_2 \text{ s.t. } \mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda},\mu)\}. \quad (17)$$

It minimizes the delay cost rate and yields the first-best revenue function $\Pi^f(\boldsymbol{\lambda},\mu) \triangleq \Pi(\boldsymbol{\lambda},\mathbf{W}^f(\boldsymbol{\lambda},\mu))$.

2. Determine the first-best arrival rates $\boldsymbol{\lambda}^f(\mu) \triangleq \arg\max_{\boldsymbol{\lambda}\in M(\mu)} \Pi^f(\boldsymbol{\lambda},\mu)$.

4.5. Relationship to Standard Achievable Region Approach

Our solution method is based on the achievable region approach to multiclass stochastic scheduling problems, which was pioneered by Coffman and Mitrani (1980); see also Federgruen and Groenevelt (1988), and Stidham (2002). However, the second-best problem calls for important modifications to the standard approach: (i) Our achievable region $\mathrm{OA}(\boldsymbol{\lambda},\mu)$ allows for policies that are *not* work conserving, unlike in standard problems. (ii) For fixed $\boldsymbol{\lambda}$, the optimization in $\mathbf{W}$ (in Step 1) is over the *intersection* $\mathrm{OA}(\boldsymbol{\lambda},\mu) \cap \mathrm{IC}(\boldsymbol{\lambda})$. The standard approach does not consider IC constraints. (iii) We optimize the revenue over $\boldsymbol{\lambda} \in M(\mu)$ (in Step 2). Standard problems consider a fixed $\boldsymbol{\lambda}$.

5. Benchmark: The First-Best Problem

For fixed μ and $\boldsymbol{\lambda} \in M(\mu)$, the first-best lead times (17) and the corresponding scheduling policy are well known and follow from Lemma 1: In $M/M/1$ systems with linear delay costs, the work conserving preemptive $c\mu$ priority rule minimizes the system's delay cost rate. Here, type 1 get absolute priority since $c_1 > c_2$ and $\mu_1 = \mu_2$. We say "$c\mu$ policy" to refer both to its priority ranking and to its work conserving property, and we denote related quantities by superscript $c\mu$.

Lemma 2. *Fix μ and $\boldsymbol{\lambda} \in M(\mu)$. The work conserving preemptive $c\mu$ priority policy and its mean steady-state lead times $\mathbf{w}^{c\mu}(\boldsymbol{\lambda},\mu)$ are first-best for $(\boldsymbol{\lambda},\mu)$:*

$$\mathbf{W}^f(\boldsymbol{\lambda},\mu) = \mathbf{w}^{c\mu}(\boldsymbol{\lambda},\mu) = \begin{bmatrix} w_1^{c\mu}(\lambda_1,\mu) \\ w_2^{c\mu}(\boldsymbol{\lambda},\mu) \end{bmatrix} \triangleq \begin{bmatrix} \dfrac{1}{\mu-\lambda_1} \\ \dfrac{\mu}{(\mu-\lambda_1)(\mu-\lambda_1-\lambda_2)} \end{bmatrix}. \quad (18)$$

Lemma 2 highlights a simple but important fact: *without IC constraints, revenue maximization and delay cost minimization are equivalent* for any $\boldsymbol{\lambda}$. The first-best prices under the $c\mu$ policy are

$$p_1^{c\mu}(\lambda_1,\mu) \triangleq v_1(\lambda_1) - c_1 w_1^{c\mu}(\lambda_1,\mu) \quad \text{and} \quad p_2^{c\mu}(\boldsymbol{\lambda},\mu) \triangleq v_2(\lambda_2) - c_2 w_2^{c\mu}(\boldsymbol{\lambda},\mu). \quad (19)$$

Proposition 1. *Fix $\mu > \mu_0$. The first-best revenue function $\Pi^f(\boldsymbol{\lambda},\mu) = \lambda_1 p_1^{c\mu}(\lambda_1,\mu) + \lambda_2 p_2^{c\mu}(\boldsymbol{\lambda},\mu)$ is strictly concave and submodular in $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}^f(\mu) = \arg\max_{\boldsymbol{\lambda}\in M(\mu)} \Pi^f(\boldsymbol{\lambda},\mu)$ is unique.*

It is optimal to serve type i if it satisfies one of two conditions:

(i) *It has the weakly higher $\bar{v}_i/c_i$ ratio, where*

$$\frac{\Pi^f_{\lambda_1}(\boldsymbol{\lambda},\mu)}{c_1} - \frac{\Pi^f_{\lambda_2}(\boldsymbol{\lambda},\mu)}{c_2} \begin{cases} > \dfrac{\bar{v}_1}{c_1} - \dfrac{\bar{v}_2}{c_2}, & \lambda_1 = 0 < \lambda_2; \\ < \dfrac{\bar{v}_1}{c_1} - \dfrac{\bar{v}_2}{c_2}, & \lambda_1 > 0 = \lambda_2. \end{cases} \quad (20)$$

(ii) *It has the higher net value at zero utilization, where*

$$\Pi^f_{\lambda_1}(\mathbf{0},\mu)-\Pi^f_{\lambda_2}(\mathbf{0},\mu)=\left[\bar{v}_1-\frac{c_1}{\mu}\right]-\left[\bar{v}_2-\frac{c_2}{\mu}\right]=(c_1-c_2)\left[\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}-\frac{1}{\mu}\right]. \quad (21)$$

Furthermore, $\mu^f \triangleq \inf\{\mu \geq \mu_0\colon\ \boldsymbol{\lambda}^f(\mu) > \mathbf{0}\} < \infty$ *and* $\boldsymbol{\lambda}^f(\mu) > \mathbf{0} \Leftrightarrow \mu > \mu^f$.

1. *If* $\bar{v}_1/c_1 = \bar{v}_2/c_2$, *then*

$$\frac{1}{\mu^f}=\frac{1}{\mu_0}=\frac{\bar{v}_1}{c_1}=\frac{\bar{v}_2}{c_2}=\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2},$$

$\boldsymbol{\lambda}^f(\mu^f)=\mathbf{0}$, *and* $\boldsymbol{\lambda}^f(\mu)>\mathbf{0}$ *for* $\mu>\mu_0$.

2. *If* $\bar{v}_1/c_1 > \bar{v}_2/c_2$, *then*

$$\frac{1}{\mu^f}<\frac{1}{\mu_0}=\frac{\bar{v}_2}{c_2}<\frac{\bar{v}_1}{c_1}<\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}$$

and $\lambda^f_1(\mu)>0=\lambda^f_2(\mu)$ *for* $\mu \leq \mu^f$.

3. *If* $\bar{v}_2/c_2 > \bar{v}_1/c_1$, *then*

$$\frac{\bar{v}_2}{c_2}>\frac{1}{\mu_0}=\frac{\bar{v}_1}{c_1}>\frac{1}{\mu^f}>\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}$$

and $\lambda^f_2(\mu)>0=\lambda^f_1(\mu)$ *for* $\mu \leq \mu^f$.

That $\Pi^f(\boldsymbol{\lambda},\mu)$ is submodular in $\boldsymbol{\lambda}$ follows because the low priority lead time $w_2^{c\mu}(\boldsymbol{\lambda},\mu)$ increases in the high priority arrival rate. The sufficient conditions (i) and (ii) for serving a type are intuitive. (i) By (20) the type with the weakly higher $\bar{v}_i/c_i$ ratio has the higher ratio of marginal revenue to delay cost if only the other type is served. (ii) The marginal revenue of the type that has the higher net value (i.e., is more profitable) at zero utilization further increases, relative to the other type's marginal revenue, if only the other type is served. By (21), at zero utilization the net value of the impatient type 1 (patient type 2) is higher if the mean service time is below (above) the IC indifference threshold, where $\bar{W}(\mathbf{0}) = (\bar{v}_1-\bar{v}_2)/(c_1-c_2)$ by (11). Conditions (i) and (ii) and the ranking of the $\bar{v}_i/c_i$ ratios imply parts 1–3 of Proposition 1.

The first-best solution also provides some intuition for the second-best solution for $\mu \in (\mu_0,\mu^f]$. In particular, at *zero* utilization the difference in the types' marginal revenues equals the difference in their net values, as shown in (21). Since both classes offer the same lead time at zero utilization, both types would prefer the (lower price) class targeted to the lower net value type, *if* it were opened. However, for $\mu \in (\mu_0,\mu^f]$ opening this class is *not* first-best by Proposition 1. Therefore, at zero utilization, the type with the higher net value buys her targeted class, whereas the type with the lower net value has no incentive to purchase the class targeted to the higher net value type. As shown in §6.4, for $\mu \in (\mu_0,\mu^f]$, these preferences at zero utilization prevail at the *optimal* utilization, so the first-best solution is second-best.

6. The Second-Best Problem

In this section we solve the second-best problem (13)–(15) with the method outlined in §4.4. In §6.1 we execute Step 1, that is, we determine the second-best lead times for fixed $\boldsymbol{\lambda}$. We show that *strategic delay* is optimal for $\boldsymbol{\lambda}$ in one of three regions that partition the $\boldsymbol{\lambda}$-space. In §6.2 we execute Step 2, that is, we characterize the second-best $(\boldsymbol{\lambda},\mathbf{W})$. We develop necessary and sufficient conditions for optimal strategic delay at the second-best arrival rates. We then apply these conditions to identify demand and capacity parameters that yield optimal strategic delay, in §6.3 for the special case of homogeneous valuations for each type, and in §6.4 for our model with heterogeneous valuations for each type. In §6.5 we explain within our framework why the $c\mu$ policy is socially optimal and IC.

6.1. Second-Best Lead-Time Vector for Fixed $\boldsymbol{\lambda}$

In Step 1 we determine for fixed $\boldsymbol{\lambda} \in M(\mu)$ the second-best lead times by solving (16):

$$\mathbf{W}^s(\boldsymbol{\lambda},\mu)=\arg\min_{\mathbf{W}}\{\lambda_1 c_1 W_1+\lambda_2 c_2 W_2 \text{ s.t. } \mathbf{W}\in \mathrm{OA}(\boldsymbol{\lambda},\mu)\cap \mathrm{IC}(\boldsymbol{\lambda})\}.$$

Consider the region $\mathrm{IC}(\boldsymbol{\lambda})$ defined in (12). IC requires $W_1 \leq \bar{W}(\boldsymbol{\lambda})$ if class 1 is open and $W_2 \geq \bar{W}(\boldsymbol{\lambda})$ if class 2 is open, where $\bar{W}(\boldsymbol{\lambda}) = (v_1(\lambda_1)-v_2(\lambda_2))/(c_1-c_2)$ is the IC indifference threshold in (11). These constraints reflect the fact that an impatient customer's net value decreases more sharply in the lead time, compared to that of a patient customer, and both types have the same marginal net value for the lead time $\bar{W}(\boldsymbol{\lambda})$. Hence, if $W_2 \geq \bar{W}(\boldsymbol{\lambda})$, then the impatient type has a (weakly) lower marginal net value for class 2 than the patient type, and therefore has no incentive to buy class 2: $v_1(\lambda_1)-c_1W_2 \leq v_2(\lambda_2)-c_2W_2 = p_2$. The equality holds since by (8) the marginal customer of each type has zero expected utility in its targeted class. Conversely, if $W_2 < \bar{W}(\boldsymbol{\lambda})$, impatient customers prefer class 2 to their own class. The intuition for the IC constraint $W_1 \leq \bar{W}(\boldsymbol{\lambda})$ is similar.

Now consider the $c\mu$ policy. By Lemma 2 it is first-best for all μ and $\boldsymbol{\lambda} \in M(\mu)$, so it is second-best for $(\boldsymbol{\lambda},\mu)$ if and only if its lead-time vector is IC, that is, $\mathbf{w}^{c\mu}(\boldsymbol{\lambda},\mu) \in \mathrm{IC}(\boldsymbol{\lambda})$. Partition $M(\mu)$ as follows:

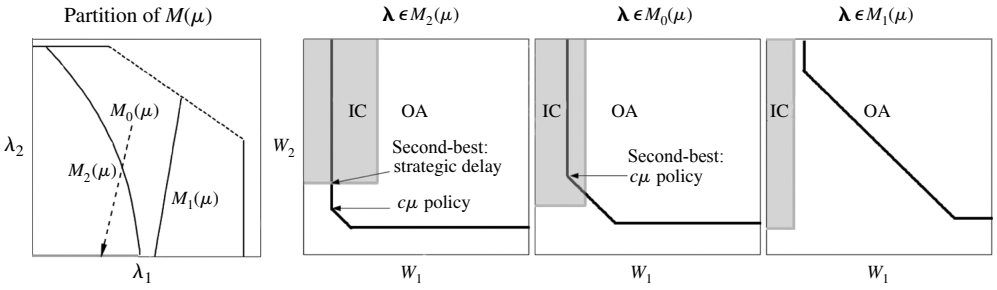
$$M_0(\mu) \triangleq \{\boldsymbol{\lambda}\in M(\mu)\colon\ \lambda_1(\bar{W}(\boldsymbol{\lambda})-w_1^{c\mu}(\lambda_1,\mu)) \geq 0 \geq \lambda_2(\bar{W}(\boldsymbol{\lambda})-w_2^{c\mu}(\boldsymbol{\lambda},\mu))\}, \quad (22)$$

$$M_1(\mu) \triangleq \{\boldsymbol{\lambda}\in M(\mu)\colon\ \lambda_1(\bar{W}(\boldsymbol{\lambda})-w_1^{c\mu}(\lambda_1,\mu)) < 0\}, \quad (23)$$

$$M_2(\mu) \triangleq \{\boldsymbol{\lambda}\in M(\mu)\colon\ \lambda_2(\bar{W}(\boldsymbol{\lambda})-w_2^{c\mu}(\boldsymbol{\lambda},\mu)) > 0\}. \quad (24)$$

The $c\mu$ policy is second-best for $(\boldsymbol{\lambda},\mu)$ if and only if $\boldsymbol{\lambda} \in M_0(\mu)$, where the subscript 0 indicates that neither

Figure 1 Proposition 2: Partition of $M(\mu)$ and Second-Best Lead Times for Representative $\boldsymbol{\lambda}$ in Each Set of the Partition



Notes. The grey line on the boundary $\lambda_2 = 0$ is part of $M_0(\mu)$. For fixed μ and $\lambda \in M(\mu)$, Lemma 1 defines the OA region in (9) and (10), and Definition 3 defines the IC region in (11) and (12).

IC constraint is violated: If the high priority class is open ($\lambda_1 > 0$), then $w_1^{c\mu}(\lambda_1, \mu) \le \bar{W}(\boldsymbol{\lambda})$, and if the low priority class is open, then $w_2^{c\mu}(\boldsymbol{\lambda}, \mu) \ge \bar{W}(\boldsymbol{\lambda})$, so each type prefers her targeted class. For $\boldsymbol{\lambda} \in M_i$, $i = 1, 2$, the lead time of priority class i under the $c\mu$ policy violates the IC constraint. For $\boldsymbol{\lambda} \in M_1(\mu)$, the high priority lead time $w_1^{c\mu}(\lambda_1, \mu)$ exceeds $\bar{W}(\boldsymbol{\lambda})$, so patient customers prefer class 1 to class 2 if $\mathbf{W} = \mathbf{w}^{c\mu}(\boldsymbol{\lambda}, \mu)$. Deterring them from class 1 requires a *shorter* lead time, but this is operationally impossible since $W_1 \ge w_1^{c\mu}(\lambda_1, \mu)$ for all achievable $\mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda}, \mu)$. To induce demand rates $\boldsymbol{\lambda}' \in M_0(\mu)$, the provider can either close class 1 or reduce its appeal to patient customers, by raising its price so that $\lambda_1' < \lambda_1$ and/or by reducing the class 2 price so that $\lambda_2' > \lambda_2$. For $\boldsymbol{\lambda} \in M_2(\mu)$, the low priority lead time $w_2^{c\mu}(\boldsymbol{\lambda}, \mu)$ is below $\bar{W}(\boldsymbol{\lambda})$, so impatient customers prefer class 2 to class 1 if $\mathbf{W} = \mathbf{w}^{c\mu}(\boldsymbol{\lambda}, \mu)$. Since the $c\mu$ policy yields the maximum class 2 lead time among work conserving policies, to deter impatient customers from class 2, the provider must *artificially increase* W_2 above its operationally achievable level, $w_2^{c\mu}(\boldsymbol{\lambda}, \mu)$, by inserting *strategic delay*. This discussion implies Proposition 2, which is illustrated in Figure 1.

Proposition 2. *Fix μ and $\boldsymbol{\lambda} \in M(\mu)$.*

1. *If $\boldsymbol{\lambda} \in M_0(\mu)$, the work conserving preemptive $c\mu$ policy is second-best for* $(\boldsymbol{\lambda}, \mu)$: $\mathbf{W}^s(\boldsymbol{\lambda}, \mu) = \mathbf{w}^{c\mu}(\boldsymbol{\lambda}, \mu)$.

2. *If $\boldsymbol{\lambda} \in M_1(\mu)$, no operationally achievable lead times are* IC *for* $(\boldsymbol{\lambda}, \mu)$: $\mathrm{OA}(\boldsymbol{\lambda}, \mu) \cap \mathrm{IC}(\boldsymbol{\lambda}) = \varnothing$.

3. *If $\boldsymbol{\lambda} \in M_2(\mu)$, no work conserving policy is* IC *for* $(\boldsymbol{\lambda}, \mu)$. *The second-best lead times are*

$$\mathbf{W}^s(\boldsymbol{\lambda}, \mu) = \mathbf{w}^{sd}(\boldsymbol{\lambda}, \mu) = \begin{bmatrix} w_1^{sd}(\lambda_1, \mu) \\ w_2^{sd}(\boldsymbol{\lambda}) \end{bmatrix} \triangleq \begin{bmatrix} w_1^{c\mu}(\lambda_1, \mu) \\ \bar{W}(\boldsymbol{\lambda}) \end{bmatrix}, \tag{25}$$

where $w_2^{sd}(\boldsymbol{\lambda}) > w_2^{c\mu}(\boldsymbol{\lambda}, \mu)$, and "sd" denotes a strategic delay policy with lead times given by (25). *It sequences jobs in the $c\mu$ order, giving preemptive priority to impatient over patient customers, but uses server idleness, server speed, and/or delivery delays to artificially inflate the mean low priority lead time to $\bar{W}(\boldsymbol{\lambda})$. The low priority lead times exceed operationally achievable levels by an average strategic delay of*

$$w_2^{sd}(\boldsymbol{\lambda}) - w_2^{c\mu}(\boldsymbol{\lambda}, \mu) = \frac{p_1^{c\mu}(\lambda_1, \mu) + c_1 w_1^{c\mu}(\lambda_1, \mu) - [p_2^{c\mu}(\boldsymbol{\lambda}, \mu) + c_1 w_2^{c\mu}(\boldsymbol{\lambda}, \mu)]}{c_1 - c_2} > 0. \tag{26}$$

Remark 1. *Strategic delay increases the delay cost* to establish IC, relative to the minimum under the first-best $c\mu$ policy (Lemma 2), but it allows the provider to target arrival rates $\boldsymbol{\lambda} \in M_2(\mu)$, which can be optimal as shown in §§6.2–6.4.

Remark 2. *How much strategic delay is optimal for fixed $\boldsymbol{\lambda}$?* Increasing the class 2 lead time by ΔW_2 while dropping its price by $\Delta p_2 = -c_2 \cdot \Delta W_2$ keeps $p_2 + c_2 W_2$ constant. This leaves the class 1 full price for impatient customers constant but increases their class 2 full price by $(c_1 - c_2) \cdot \Delta W_2$. By (26), the optimal strategic delay for fixed $\boldsymbol{\lambda} \in M_2(\mu)$ therefore equals the impatient type's full price premium for class 1 relative to class 2 under the $c\mu$ policy, divided by the delay cost difference $c_1 - c_2$. With this delay added to the class 2 lead time, impatient customers are indifferent between the two classes.

6.2. Strategic Delay Optimality: Necessary and Sufficient Conditions

We turn to Step 2 outlined in §4.4: find the second-best arrival rates $\boldsymbol{\lambda}^s(\mu) \in \arg\max_{\boldsymbol{\lambda} \in M(\mu)} \Pi^s(\boldsymbol{\lambda}, \mu)$ and lead times $\mathbf{W}^s(\boldsymbol{\lambda}^s(\mu), \mu)$. By Proposition 2 strategic delay is optimal for *fixed* $\boldsymbol{\lambda} \in M_2(\mu)$. We develop necessary and sufficient conditions for strategic delay to be optimal at $\boldsymbol{\lambda}^s(\mu)$.

Let $\Pi^{sd}(\boldsymbol{\lambda}, \mu) \triangleq \Pi(\boldsymbol{\lambda}, \mathbf{w}^{sd}(\boldsymbol{\lambda}, \mu))$ be the revenue function under strategic delay. By Proposition 2,

$$\Pi^s(\boldsymbol{\lambda}, \mu) = \begin{cases} \Pi^f(\boldsymbol{\lambda}, \mu) = \lambda_1 p_1^{c\mu}(\lambda_1, \mu) + \lambda_2 p_2^{c\mu}(\boldsymbol{\lambda}, \mu), & \boldsymbol{\lambda} \in M_0(\mu); \\ 0, & \boldsymbol{\lambda} \in M_1(\mu); \\ \Pi^{sd}(\boldsymbol{\lambda}, \mu) = \lambda_1 p_1^{c\mu}(\lambda_1, \mu) + \lambda_2 p_2^{sd}(\boldsymbol{\lambda}), & \boldsymbol{\lambda} \in M_2(\mu), \end{cases} \tag{27}$$

where Π^f is the first-best revenue function (Proposition 1), $p_1^{c\mu}(\lambda_1, \mu)$ and $p_2^{c\mu}(\boldsymbol{\lambda}, \mu)$ are given by (19), and by (25) the class 2 price under strategic delay is

$$p_2^{sd}(\boldsymbol{\lambda}) \triangleq v_2(\lambda_2) - c_2 w_2^{sd}(\boldsymbol{\lambda}) = v_2(\lambda_2) - c_2 \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2}. \quad (28)$$

This price *increases* in λ_1: Serving more impatient customers requires dropping the class 1 full price, which reduces their marginal value $v_1(\lambda_1)$ and the lead time $w_2^{sd}(\boldsymbol{\lambda})$ required to deter them from class 2. As a result, $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ is *supermodular* in $\boldsymbol{\lambda}$.

Definition 4. We say that *strategic delay is optimal* if $\arg\max_{\boldsymbol{\lambda} \in M(\mu)} \Pi^s(\boldsymbol{\lambda}, \mu) \subset M_2(\mu)$.

Proposition 3. *Fix $\mu > \mu_0$, and suppose that $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ is strictly concave in $\boldsymbol{\lambda}$.*

1. *Strategic delay is optimal and $\boldsymbol{\lambda}^s(\mu)$ is the unique second-best demand vector, if and only if*

$$\boldsymbol{\lambda}^s(\mu) = \arg\max_{\boldsymbol{\lambda} \in M(\mu)} \Pi^{sd}(\boldsymbol{\lambda}, \mu) \in M_2(\mu),$$

which holds if and only if $\boldsymbol{\lambda} = \lambda^s(\mu)$ satisfies the following:

$$\Pi_{\lambda_1}^{sd}(\boldsymbol{\lambda}, \mu) = p_1^{c\mu}(\lambda_1, \mu) + \lambda_1 \frac{\partial p_1^{c\mu}(\lambda_1, \mu)}{\partial \lambda_1} + \lambda_2 \frac{\partial p_2^{sd}(\boldsymbol{\lambda})}{\partial \lambda_1} = 0, \quad (29)$$

$$\Pi_{\lambda_2}^{sd}(\boldsymbol{\lambda}, \mu) = p_2^{sd}(\boldsymbol{\lambda}) + \lambda_2 \frac{\partial p_2^{sd}(\boldsymbol{\lambda})}{\partial \lambda_2} = 0, \quad (30)$$

$$\boldsymbol{\lambda} > \mathbf{0}, \quad (31)$$

$$w_2^{sd}(\boldsymbol{\lambda}) - w_2^{c\mu}(\boldsymbol{\lambda}, \mu) > 0, \quad \text{and } \mu > \lambda_1 + \lambda_2. \quad (32)$$

2. *If strategic delay is optimal, then the first-best arrival rates $\boldsymbol{\lambda}^f(\mu) \in M_2(\mu)$.*

The converse of part 2 is not true; for example, if $\boldsymbol{\lambda}^f(\mu) \in M_2(\mu)$, it may be second-best to close class 2 (Proposition 7.1(a)).

Next, to translate (29)–(32) into more specific conditions, we characterize the maximum revenue $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ as a function of the impatient customer arrival rate λ_1. By (27), $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ depends on the arrival rate of patient customers only through their revenue $\lambda_2 p_2^{sd}(\boldsymbol{\lambda})$, which is independent of capacity. We make the following mild assumptions.

Assumption A3. *The function $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ is strictly concave in $\boldsymbol{\lambda}$ for fixed $\mu \le \infty$, and $v_2'/R_2'' < c_1/c_2$.*

Lemma 3. *For fixed λ_1, the optimal type 2 arrival rate under strategic delay, that is,*

$$\lambda_2^{sd}(\lambda_1) \triangleq \arg\max_{\lambda_2 \in [0, \Lambda_2]} \lambda_2 p_2^{sd}(\boldsymbol{\lambda}),$$

is unique. Let $\boldsymbol{\lambda}^{sd}(\lambda_1) \triangleq (\lambda_1, \lambda_2^{sd}(\lambda_1))$.

1. *If $\bar{v}_2/c_2 \le v_1(\lambda_1)/c_1$, then opening class 2 is not profitable with strategic delay: $\lambda_2^{sd}(\lambda_1) = 0$.*

If $\bar{v}_2/c_2 > v_1(\lambda_1)/c_1$, then $\lambda_2^{sd}(\lambda_1) \in (0, \Lambda_2)$, $p_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1)) > 0$, and

$$\Pi_{\lambda_2}^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1), \mu) = 0 \Leftrightarrow \frac{R_2'(\lambda_2^{sd}(\lambda_1))}{c_2} = \frac{v_2(\lambda_2^{sd}(\lambda_1))}{c_2}\left(1 - \frac{1}{\varepsilon_2(\lambda_2^{sd}(\lambda_1))}\right) = \frac{v_1(\lambda_1)}{c_1}. \quad (33)$$

2. *The rate $\lambda_2^{sd}(\lambda_1)$ is nondecreasing, and $\lambda_2^{sd\prime}(\lambda_1) > 0$ for $\lambda_1 > \underline{\lambda}_1 \triangleq \min\{\lambda_1 \ge 0\colon v_1(\lambda_1)/c_1 \le \bar{v}_2/c_2\}$.*

3. *The lead time $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1))$ and the strategic delay $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1)) - w_2^{c\mu}(\boldsymbol{\lambda}^{sd}(\lambda_1), \mu)$ strictly decrease in λ_1.*

4. *For fixed λ_1, the segment share $\lambda_2^{sd}(\lambda_1)/\Lambda_2$ and the lead time $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1))$ are independent of $\Lambda_2 > 0$, and the strategic delay $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1)) - w_2^{c\mu}(\boldsymbol{\lambda}^{sd}(\lambda_1), \mu)$ is nonincreasing in Λ_2.*

In Lemma 3.1, the condition for opening class 2 follows by (28). Under strategic delay,

$$p_2^{sd}(\boldsymbol{\lambda}) > 0 \Leftrightarrow \frac{v_2(\lambda_2)}{c_2} > \frac{v_1(\lambda_1)}{c_1} > \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} = w_2^{sd}(\boldsymbol{\lambda}), \quad (34)$$

that is, the class 2 price is positive if and only if the patient type has the higher marginal value-to-delay cost ratio. This ensures that increasing the class 2 lead time to $w_2^{sd}(\boldsymbol{\lambda})$, which makes the impatient type indifferent between the classes, leaves the patient type with positive marginal net value.

The optimality condition (33) implies that the more elastic the patient type's marginal value function, the lower the price and the higher the lead time of their class under strategic delay.

In Lemma 3.2, that $\lambda_2^{sd}(\lambda_1)$ is increasing follows since $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ is supermodular in $\boldsymbol{\lambda}$.

In Lemma 3.3, the condition $v_2'/R_2'' < c_1/c_2$ in A3 ensures that the impatient type's marginal value $v_1(\lambda_1)$ drops more sharply in λ_1 than the patient type's marginal value $v_2(\lambda_2^{sd}(\lambda_1))$. As a result, the optimal class 2 lead time and strategic delay decrease in λ_1. A sufficient condition for $v_2'/R_2'' < c_1/c_2$ is that the elasticity $\varepsilon_2(\lambda_2)$ be nondecreasing. For linear $v_i(\lambda_i)$, discussed in §6.4, $v_2'/R_2'' = 1/2$.

In Lemma 3.4, that $\lambda_2^{sd}(\lambda_1)/\Lambda_2$ and $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1))$ are constant in Λ_2 follows because by (7) and (28), the class 2 lead time and price depend on λ_i and Λ_i only through the fraction of type i served, λ_i/Λ_i. Therefore, $\lambda_2^{sd}(\lambda_1)$ is proportional to Λ_2, and the strategic delay is nonincreasing in Λ_2.

By (19), (28), (29), and Lemma 3, the total revenue derivative with respect to λ_1 satisfies

$$\frac{d\Pi^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1),\mu)}{d\lambda_1} = \Pi^{sd}_{\lambda_1}(\boldsymbol{\lambda}^{sd}(\lambda_1),\mu) = \left[R_1'(\lambda_1) - \frac{c_1\mu}{(\mu-\lambda_1)^2}\right] + \left[\lambda_2^{sd}(\lambda_1)\cdot c_2\frac{-v_1'(\lambda_1)}{c_1-c_2}\right]. \quad (35)$$

It strictly decreases in λ_1 since $\Pi^{sd}(\boldsymbol{\lambda},\mu)$ is strictly concave in $\boldsymbol{\lambda}$. The term in the first bracket of (35) is the marginal revenue from impatient customers, which decreases in λ_1 since $R_1'' < 0$ by A2. The term in the second bracket is the marginal change in the maximum revenue from patient customers with respect to λ_1, which is *positive* if $\lambda_2^{sd}(\lambda_1) > 0$ since $\partial p_2^{sd}(\boldsymbol{\lambda})/\partial\lambda_1 = -c_2 v_1'(\lambda_1)/(c_1-c_2) > 0$.

Let λ_1^{sd} be the impatient type arrival rate that solves $\Pi^{sd}_{\lambda_1}(\boldsymbol{\lambda}^{sd}(\lambda_1),\mu) = 0$. By Proposition 3.1, strategic delay is optimal if and only if (31) and (32) hold for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{sd}(\lambda_1^{sd})$, that is, both types are served and strategic delay is required to deter impatient customers from class 2. We translate these into more specific conditions, based on the following key system properties under strategic delay.

First, by (35), serving impatient customers is profitable, and λ_1^{sd} is at least as large as the optimal arrival rate of impatient customers in the absence of patient customers; let λ_1° be this rate. Then by (35), $\lambda_1^{sd} \geq \lambda_1^\circ > 0$ and

$$\frac{R_1'(\lambda_1^\circ)}{c_1} = \frac{v_1(\lambda_1^\circ)}{c_1}\left(1 - \frac{1}{\varepsilon_1(\lambda_1^\circ)}\right) = \frac{\mu}{(\mu-\lambda_1^\circ)^2}. \quad (36)$$

Second, serving patient customers is optimal if and only if it is profitable to open class 2 at λ_1°, i.e., $\bar{v}_2/c_2 > v_1(\lambda_1^\circ)/c_1$ (Lemma 3.1). This holds since $\Pi^{sd}(\boldsymbol{\lambda},\mu)$ is strictly concave and supermodular in $\boldsymbol{\lambda}$.

Third, if serving patient customers is optimal, the demand vector $\boldsymbol{\lambda}^{sd}(\lambda_1^{sd})$ increases in their segment size Λ_2: By (35), increasing $\lambda_1 > \lambda_1^\circ$ lowers the revenue from impatient customers but increases the revenue from patient ones, and this gain is proportional to $\lambda_2^{sd}(\lambda_1)$, which increases in Λ_2 (Lemma 3.4). This implies (Lemma 3.3 and 3.4) that the *optimal* strategic delay $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^{sd})) - w_2^{c\mu}(\boldsymbol{\lambda}^{sd}(\lambda_1^{sd}),\mu)$ decreases in Λ_2. Therefore, strategic delay can be optimal if and only if it is optimal for a small patient segment, i.e., $\Lambda_2 \approx 0$. In this case, $\boldsymbol{\lambda}^{sd}(\lambda_1^{sd}) \approx (\lambda_1^\circ, 0)$ by (35) and (36), so that

$$w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^{sd})) \approx w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^\circ)) \quad \text{and}$$
$$w_2^{c\mu}(\boldsymbol{\lambda}^{sd}(\lambda_1^{sd}),\mu) = \frac{\mu}{(\mu-\lambda_1^{sd})(\mu-\lambda_1^{sd}-\lambda_2^{sd}(\lambda_1^{sd}))} \approx \frac{\mu}{(\mu-\lambda_1^\circ)^2}.$$

This discussion implies three necessary and sufficient conditions for optimal strategic delay, which Proposition 4 formalizes as the price, the lead-time, and the segment-size condition.

Proposition 4. *Fix $\mu > \mu_0$. Then*

$$\lambda_1^\circ \triangleq \arg\max_{\lambda_1}\{\lambda_1 p_1^{c\mu}(\lambda_1,\mu) \text{ s.t. } \lambda_1 \in [0,\Lambda_1],\ \lambda_1 < \mu\} > 0.$$

Strategic delay is optimal if and only if the following conditions hold.

1. *Price condition. At λ_1° it is profitable to open class* 2 *with strategic delay*:

$$\frac{\bar{v}_2}{c_2} > \frac{v_1(\lambda_1^\circ)}{c_1}. \quad (37)$$

2. *Lead-time condition. IC requires strategic delay if the patient segment is small* ($\Lambda_2 \approx 0$):

$$w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^\circ)) = \frac{v_1(\lambda_1^\circ) - v_2(\lambda_2^{sd}(\lambda_1^\circ))}{c_1-c_2} > \frac{\mu}{(\mu-\lambda_1^\circ)^2}, \quad (38)$$

where (38) *is independent of Λ_2. If* (37) *holds, then* (38) *is equivalent to the condition $\varepsilon_2(\lambda_2^{sd}(\lambda_1^\circ)) > \varepsilon_1(\lambda_1^\circ)c_2/(c_1-c_2)+1$.*

3. *Segment-size condition. The patient segment is smaller than a threshold $\bar{\Lambda}_2$: $0 < \Lambda_2 < \bar{\Lambda}_2 < \infty$.*

The price and lead-time conditions are independent of Λ_2. Either one *must* hold.[1] By the equivalent condition for (38), the patient type's marginal value function must be sufficiently elastic: By (33) the more elastic this function, the lower the class 2 price, and the higher its lead time.

Proposition 4 provides a specific test for strategic delay optimality. In §§6.3 and 6.4 we apply this test to identify explicit demand and capacity parameter conditions for optimal strategic delay.

6.3. Strategic Delay Optimality: Homogeneous Valuations for Each Type

We start with the simplest valuation model in which all type i customers have the same valuation: $v_i(\lambda_i) = \bar{v}_i$ for $\lambda_i \in [0,\Lambda_i]$. In this case, the class 2 lead time under strategic delay, which equals the threshold $\bar{W}(\boldsymbol{\lambda})$ defined in (11), is independent of $\boldsymbol{\lambda}$, i.e., $w_2^{sd}(\boldsymbol{\lambda}) = (\bar{v}_1-\bar{v}_2)/(c_1-c_2)$. The class 2 price is $p_2^{sd}(\boldsymbol{\lambda}) = \bar{v}_2 - c_2((\bar{v}_1-\bar{v}_2)/(c_1-c_2))$ by (28), so the revenue is additively separable in λ_1 and λ_2 by (27):

$$\Pi^{sd}(\boldsymbol{\lambda},\mu) = \lambda_1 p_1^{c\mu}(\lambda_1,\mu) + \lambda_2 p_2^{sd}(\boldsymbol{\lambda}) = \lambda_1\left[\bar{v}_1 - c_1\frac{1}{\mu-\lambda_1}\right] + \lambda_2\left[\bar{v}_2 - c_2\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}\right]. \quad (39)$$

[1] If (37) does not hold, then $\lambda_2^{sd}(\lambda_1^\circ) = 0$ and (38) holds:

$$\frac{v_1(\lambda_1^\circ)-\bar{v}_2}{c_1-c_2} \geq \frac{v_1(\lambda_1^\circ)}{c_1} > \frac{R_1'(\lambda_1^\circ)}{c_1} = \frac{\mu}{(\mu-\lambda_1^\circ)^2}.$$

Under strategic delay the optimal impatient type arrival rate λ_1^{sd} is independent of λ_2, and $\lambda_1^{sd} = \lambda_1^\circ$, where

$$\lambda_1^\circ = \Lambda_1 \text{ if } \frac{\bar{v}_1}{c_1} > \frac{\mu}{(\mu-\Lambda_1)^2}, \quad \text{and}$$
$$\lambda_1^\circ = \arg\left\{\lambda_1 \geq 0;\ \frac{\bar{v}_1}{c_1} = \frac{\mu}{(\mu-\lambda_1)^2}\right\} \quad \text{otherwise.} \tag{40}$$

Since $R_1'(\lambda_1) = \bar{v}_1$ in this model, (40) is the counterpart of (36).

In this model, A1 (i.e., $v_i(\Lambda_i) = 0$) does not hold, but the necessary and sufficient conditions of Proposition 4 apply. In particular, the price condition (37) and the lead-time condition (38) do not hinge on A1, and we specialize the segment-size condition to account for the fact that $v_i(\Lambda_i) = \bar{v}_i$.

1. The *price condition* (37) requires that the patient type has the higher $\bar{v}_i/c_i$ ratio: $\bar{v}_2/c_2 > \bar{v}_1/c_1$. This implies $\bar{v}_2/c_2 > (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2)$, so by (39) it is optimal to serve all patient customers: $\lambda_2^{sd} = \Lambda_2$.

2. If $\bar{v}_2/c_2 > \bar{v}_1/c_1$, then $\bar{v}_1/c_1 > (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2)$, so that by (40) the *lead-time condition* (38) requires

$$\frac{\bar{v}_1 - \bar{v}_2}{c_1 - c_2} > \frac{\mu}{(\mu-\Lambda_1)^2}. \tag{41}$$

That is, (38) can only hold if it is optimal to serve all impatient customers: $\lambda_1^{sd} = \Lambda_1$. (The price and lead-time conditions cannot both hold if $\bar{v}_1/c_1 = \mu/(\mu-\lambda_1^\circ)^2$.) It is necessary for (41) that patient customers have lower valuations, i.e., $\bar{v}_1 > \bar{v}_2$, and capacity is sufficiently large: $\mu > (c_1 - c_2)/(\bar{v}_1 - \bar{v}_2) > \mu_0$.

3. The *segment-size condition* specializes to the requirement

$$\Lambda_2 < \bar{\Lambda}_2 = \min\left\{\mu - \Lambda_1 - \frac{\mu}{\mu-\Lambda_1}\left(\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}\right)^{-1}, \Lambda_1\left(\frac{c_1}{c_2}-1\right)\right\}, \tag{42}$$

where $\bar{\Lambda}_2 > 0$ if (41) holds. The first argument on the right-hand side of (42) ensures $w_2^{c\mu}(\boldsymbol{\Lambda}, \mu) < (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2)$, i.e., $\boldsymbol{\Lambda} \in M_2(\mu)$ by (24). That is, when all customers are served, impatient customers prefer class 2 under the $c\mu$ policy (Proposition 2.3). The second argument on the right-hand side of (42) ensures that strategic delay maximizes the second-best revenue in this case. Specifically, if $W_2 < (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2)$, then one option to make impatient customers indifferent between the classes is to reduce the class 1 price so that $p_1 + c_1 W_1 = p_2 + c_1 W_2$. In this situation,

$$p_2 = \bar{v}_2 - c_2 W_2 \quad \text{and}$$
$$p_1 = p_2 + c_1(W_2 - W_1) = \bar{v}_2 + (c_1 - c_2)W_2 - c_1 W_1 < \bar{v}_1 - c_1 W_1.$$

The inequality implies that impatient customers have positive expected utility.[2] Instead of reducing the class 1 price, using strategic delay to inflate the class 2 lead time by ΔW_2 is more profitable if and only if $\Lambda_2 < \Lambda_1(c_1/c_2 - 1)$: the revenue gain on impatient customers is $\Lambda_1(c_1 - c_2)\Delta W_2$; the loss on patient ones is $\Lambda_2 c_2 \Delta W_2$.

To summarize, strategic delay is optimal if and only if the *patient type has the higher $\bar{v}_i/c_i$ ratio* and the lower valuation; its segment is not too large relative to that of the impatient type, that is,

$$1 < \frac{\bar{v}_1}{\bar{v}_2} < \frac{c_1}{c_2} \quad \text{and} \quad \frac{\Lambda_2}{\Lambda_1} < \frac{c_1}{c_2} - 1; \tag{43}$$

and the *capacity μ is sufficiently large*, that is, it exceeds a threshold which (42) yields in closed form.

6.4. Strategic Delay Optimality: Heterogeneous Valuations for Each Type

As we show in this section, valuation heterogeneity at each delay cost level yields different results, compared to the case of homogeneous valuations. For one, strategic delay can be optimal for any ranking of the $\bar{v}_i/c_i$ ratios. Moreover, strategic delay is not necessarily a "large capacity phenomenon": If the impatient type has the higher $\bar{v}_i/c_i$ ratio, which is quite plausible, then under mild conditions strategic delay may be optimal *only at relatively scarce, but not at ample capacity* (Proposition 7).

The results in this section hold for any value distributions that satisfy the assumptions in §2.1 and A1–A3, except for Propositions 5.2, 6.2, and 7.2, which assume linear v_i functions that satisfy the following.

Assumption A4. *Let $F_i(v) = v/\bar{v}_i$, $v \in [0, \bar{v}_i]$, so $v_i(\lambda_i) = \bar{v}_i(1 - \lambda_i/\Lambda_i)$, $\lambda_i \in [0, \Lambda_i]$, and $v_i'/R_i'' = 1/2$. We assume $\Lambda_1/\Lambda_2 > \bar{v}_1/\bar{v}_2/(4c_1/c_2(c_1/c_2 - 1))$, which ensures that $\Pi^{sd}(\boldsymbol{\lambda}, \mu)$ is strictly concave in $\boldsymbol{\lambda}$ for $\mu \leq \infty$, so* A3 *holds.*

6.4.1. Ample Capacity. As a building block for the finite capacity results, consider the limiting case of ample capacity, i.e., $\mu = \infty$. In this case, work conserving policies yield zero lead times, so that price discrimination *requires* strategic delay. Lemma 3 and Proposition 4 imply Corollary 1.

Corollary 1. *Fix $\mu = \infty$. Then $\lambda_1^\circ = \arg\{\lambda_1 \in [0, \Lambda_1]: R_1'(\lambda_1) = 0\} > 0$. Strategic delay is optimal if and only if the following conditions hold.*

1. *Price condition. At λ_1° it is profitable to open class* 2 *with strategic delay*:

$$\frac{\bar{v}_2}{c_2} > \frac{v_1(\lambda_1^\circ)}{c_1}, \quad \textit{where } \varepsilon_1(\lambda_1^\circ) = 1. \tag{44}$$

2. *Lead-time condition. IC requires strategic delay if the patient segment is small* ($\Lambda_2 \approx 0$):

$$w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^\circ)) = \frac{v_1(\lambda_1^\circ) - v_2(\lambda_2^{sd}(\lambda_1^\circ))}{c_1 - c_2} > 0, \tag{45}$$

[2] In the more plausible main model of the paper, it is not optimal to serve all customers of either type, which rules out a solution where the marginal customer of one type gets positive expected utility.

where (45) *is independent of* Λ_2. *If* (44) *holds, then* (45) *is equivalent to the condition* $\varepsilon_2(\lambda_2^{sd}(\lambda_1^\circ)) > c_1/(c_1-c_2)$.

3. *Segment-size condition. The patient segment size* $\Lambda_2 < \bar{\Lambda}_2$. *If* (44) *and* (45) *hold, then*

$$\bar{\Lambda}_2 = \frac{R_1'(x_1)}{v_1'(x_1)}\frac{c_1-c_2}{c_2}\frac{1}{\lambda_2^{sd}(x_1)/\Lambda_2} > 0, \tag{46}$$

where

$$x_1 \triangleq \arg\{\lambda_1 \in [0,\lambda_1]\colon w_2^{sd}(\boldsymbol{\lambda}^{sd}(x_1)) = 0\} \in (\lambda_1^\circ, \Lambda_1),$$

and $\lambda_2^{sd}(x_1)/\Lambda_2$ *is independent of* Λ_2.

For linear $v_i(\lambda_i)$, Corollary 1 yields the following conditions for optimal strategic delay:

$$\frac{c_1}{c_1-c_2/2} < \frac{\bar{v}_1}{\bar{v}_2} < 2\frac{c_1}{c_2} \quad \text{and}$$
$$\Lambda_2 < \bar{\Lambda}_2 = 2\Lambda_1\left(\frac{c_1}{c_2}\left(1-\frac{\bar{v}_2}{\bar{v}_1}\right)-\frac{1}{2}\right), \tag{47}$$

where $\bar{v}_1/\bar{v}_2 < 2(c_1/c_2)$ is the price condition. These conditions are similar to those in (43), but the price condition in (47) allows a higher $\bar{v}_i/c_i$ ratio for the impatient type. Serving both types is profitable only if the price condition holds. If the lead-time condition is violated, i.e., $\bar{v}_1/\bar{v}_2 \le c_1/(c_1-c_2/2)$, there is insufficient value differentiation to warrant price discrimination. This applies to the special case $\bar{v}_1 = \bar{v}_2$, where the single price $p = \bar{v}_i/2$ is first- and second-best. If the segment-size condition is violated, i.e., $\Lambda_2 \ge \bar{\Lambda}_2$, strategic delay yields a larger loss in the patient segment compared to the gain in the impatient segment, so charging one price is optimal.

6.4.2. Impact of Capacity on the Second-Best Solution: Preliminaries. Define the thresholds $\mu^s \triangleq \inf\{\mu \ge \mu_0\colon \boldsymbol{\lambda}^s(\mu) > \mathbf{0}\}$ and $\mu^{sd} \triangleq \inf\{\mu \ge \mu_0\colon \boldsymbol{\lambda}^s(\mu) \in M_2(\mu)\}$, where $\inf\varnothing = \infty$ and $\mu^s \le \mu^{sd}$ since strategic delay is optimal only if it is second-best to serve both types (Proposition 3.1). Recall that $\mu^f = \inf\{\mu \ge \mu_0\colon \boldsymbol{\lambda}^f(\mu) > \mathbf{0}\} < \infty$ by Proposition 1. For each ranking of the $\bar{v}_i/c_i$ ratios, we identify demand conditions that yield optimal strategic delay for *some* capacity (i.e., $\mu^{sd} < \infty$), characterize μ^{sd} and the capacity interval(s) in $(\mu^{sd}, \infty]$ with optimal strategic delay, and relate μ^{sd} to μ^f and μ^s. The discussion builds on Propositions 1 and 4, Corollary 1, and on Lemma 4.

Lemma 4. *Suppose that*

$$\frac{1}{\mu} < \min\left(\frac{1}{\mu^f}, \frac{\bar{v}_1-\bar{v}_2}{c_1-c_2}\right).$$

1. *It is second-best to serve the impatient type.*

2. *It is second-best to serve the patient type if and only if the price condition* (37) *holds*: $\bar{v}_2/c_2 > v_1(\lambda_1^\circ)/c_1$.

For the intuition to Lemma 4, recall that for $\mu > \mu^f$, the first-best $c\mu$ policy serves both types (Proposition 1), and for $\mu > \mu_0$ the impatient type is profitable under the strategic delay policy (Proposition 4). Part 1 of Lemma 4 holds because, for $1/\mu < (\bar{v}_1-\bar{v}_2)/(c_1-c_2)$, the class 1 lead time under the $c\mu$ policy is IC at low arrival rates (i.e., $\boldsymbol{\lambda} \notin M_1(\mu)$ for $\lambda_1 \approx 0$ by (11) and (23)). Part 2 holds because, if it is not IC under the $c\mu$ policy to open class 2, then it is second-best to open class 2 if and only if its price with strategic delay is positive at λ_1°.

6.4.3. The Types Have the Same $\bar{v}_i/c_i$ Ratio. This special case develops intuition for the other cases.

Proposition 5. *Fix* $\mu > \mu_0$ *and suppose that the types have the same* $\bar{v}_i/c_i$ *ratio.*

1. *Then*

$$\frac{\bar{v}_1-\bar{v}_2}{c_1-c_2} = \frac{\bar{v}_1}{c_1} = \frac{\bar{v}_2}{c_2} = \frac{1}{\mu_0} = \frac{1}{\mu^f} = \frac{1}{\mu^s} \ge \frac{1}{\mu^{sd}}.$$

For every profitable system, i.e., for $\mu > \mu_0$, *it is first- and second-best to serve both types, and strategic delay may be optimal.*

2. *For linear* $v_i(\cdot)$ *there exist capacity levels at which strategic delay is optimal, if and only if it is optimal at ample capacity* ($\mu = \infty$), *i.e.,* $\Lambda_2/\Lambda_1 < 2(c_1/c_2 - 1.5)$. *In this case*: (a) *Strategic delay is optimal for all* $\mu > \mu^{sd}$. (b) *If* $\Lambda_2/\Lambda_1 \le 2(c_1/c_2-1.5)/(1+\bar{v}_1/c_1\Lambda_1(c_1/c_2-1))$, *then* $\mu^{sd} = \mu_0$. *Otherwise,* $\mu^{sd} > \mu_0$.

Consider the conditions of Proposition 4. For fixed μ, strategic delay is optimal for some Λ_2 if and only if the price condition (37) and the lead-time condition (38) hold, which are, respectively,

$$\frac{\bar{v}_2}{c_2} > \frac{v_1(\lambda_1^\circ(\mu))}{c_1} \quad \text{and} \tag{48}$$

$$w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^\circ(\mu))) = \frac{v_1(\lambda_1^\circ(\mu)) - v_2(\lambda_2^{sd}(\lambda_1^\circ(\mu)))}{c_1-c_2} > \frac{R_1'(\lambda_1^\circ(\mu))}{c_1} = \frac{\mu}{(\mu-\lambda_1^\circ(\mu))^2}. \tag{49}$$

Serving impatient customers is not profitable at the minimum capacity ($\lambda_1^\circ(\mu_0) = 0$).

The price condition (48) holds for all $\mu > \mu_0$: Since the types have the same $\bar{v}_i/c_i$ ratio, serving impatient customers reduces their marginal valuation-to-delay cost ratio $v_1(\lambda_1^\circ(\mu))/c_1$ below $\bar{v}_2/c_2$. It follows from Lemma 4 that it is second-best to serve both types for every $\mu > \mu_0$.

The lead-time condition (49) holds for all $\mu > \mu_0$ if $c_1/c_2 > 1.5$, and for no μ otherwise. Fixing $\bar{v}_1/c_1 = \bar{v}_2/c_2$ and $\lambda_1^\circ(\mu) > 0$, the class 2 lead time $w_2^{sd}(\boldsymbol{\lambda}^{sd}(\lambda_1^\circ(\mu)))$ increases in $\bar{v}_1$ and c_1: Intuitively, the higher the impatient type's net value, the more it prefers the class targeted to the patient type.

For $c_1/c_2 > 1.5$, consider the interaction between the capacity μ and the patient segment size Λ_2. Increasing μ yields higher optimal arrival rates, but also speeds up service. The former effect reduces the strategic delay, the latter increases it, and the net effect depends on Λ_2. For fixed μ, the optimal strategic delay decreases in Λ_2 (Proposition 4). If Λ_2 violates the segment-size condition for ample capacity, i.e., $\Lambda_2 \geq \Lambda_1 2(c_1/c_2 - 1.5)$, strategic delay is not optimal at any capacity. If Λ_2 is below the threshold in part 2(b), strategic delay is optimal for every system ($\mu^{sd} = \mu_0$). For Λ_2 in between these thresholds, strategic delay is optimal only for high enough capacity ($\mu^{sd} > \mu_0$).

6.4.4. The Patient Type Has the Higher $\bar{v}_i/c_i$ Ratio. In this case strategic delay can be optimal only for large enough capacity, consistent with the result for the homogeneous valuation model in §6.3.

Proposition 6. *Fix $\mu > \mu_0$ and suppose that the patient type has the higher $\bar{v}_i/c_i$ ratio.*

1. *Then*

$$\frac{\bar{v}_2}{c_2} > \frac{\bar{v}_1}{c_1} = \frac{1}{\mu_0} > \frac{1}{\mu^f} > \frac{1}{\mu^s} \geq \frac{\bar{v}_1 - \bar{v}_2}{c_1 - c_2}.$$

(a) *For $\mu \leq \mu^f$, the first-best policy is second-best and serves only patient customers. For $\mu \in (\mu^f, \mu^s)$ it is first-best to serve both types but second-best to serve only patient customers; they prefer class 1 at the first-best solution. For $1/\mu < (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2)$ it is second-best to serve both types.*

(b) *If strategic delay is optimal for some capacity, then $\mu^s < \mu^{sd} < \infty$, and there is a $\mu' \in (\mu^s, \mu^{sd})$ such that for μ' the first-best policy is second-best and serves both types.*

2. *For linear $v_i(\cdot)$ there exist capacity levels at which strategic delay is optimal, if and only if it is optimal at ample capacity ($\mu = \infty$), i.e., $c_1/(c_1 - c_2/2) < \bar{v}_1/\bar{v}_2 < 2c_1/c_2$ and $\Lambda_2 < 2\Lambda_1((c_1/c_2)(1 - \bar{v}_2/\bar{v}_1) - \frac{1}{2})$. In this case strategic delay is optimal for all $\mu > \mu^{sd}$, and*

$$\mu^{sd} > \mu^*(\underline{x}) \triangleq \underline{x} + \frac{1 + \sqrt{1 + 4\underline{x}(\bar{v}_1/c_1)(1 - 2\underline{x}/\Lambda_1)}}{2(\bar{v}_1/c_1)(1 - 2\underline{x}/\Lambda_1)} > \frac{c_1 - c_2}{\bar{v}_1 - \bar{v}_2} > \mu_0, \quad \text{where}$$
$$\underline{x} \triangleq \frac{\Lambda_1}{2} \frac{(c_1/c_2)(\bar{v}_2/\bar{v}_1) - 1}{c_1/c_2 - 1.5}. \tag{50}$$

At low arrival rates, both types' $\bar{v}_i/c_i$ ratios exceed the IC indifference threshold, because $\bar{v}_2/c_2 > \bar{v}_1/c_1 > (\bar{v}_1 - \bar{v}_2)/(c_1 - c_2) = \bar{W}(\mathbf{0})$. At low capacity, the lead times therefore exceed the indifference threshold, so the patient type is more profitable and prefers the high priority class targeted to the impatient type. Proposition 6.1(a) follows from these properties, Proposition 1, and Lemma 4. If strategic delay is optimal, then these lead-time preferences are reversed at higher capacity: The impatient type prefers the low priority class at the first-best solution, which implies Proposition 6.1(b), that is, the first-best solution must be second-best at some intermediate capacity level.

Part 2 is similar to the case of homogeneous valuations in §6.3. Since the patient type has the higher $\bar{v}_i/c_i$ ratio, the price condition holds for all capacity levels $\mu > \mu_0$, but the minimum capacity at which the lead-time condition can hold exceeds μ_0, as shown in (50). The threshold μ^{sd} attains $\mu^*(\underline{x})$ for a negligibly small patient segment (i.e., $\Lambda_2 \to 0$), but μ^{sd} increases in Λ_2 since a larger patient segment implies higher arrival rates and delays.

Example 1. We illustrate Proposition 6 for linear v_i functions that satisfy the conditions in part 2. Figure 2 shows key metrics as functions of the capacity $\mu \geq (c_1 - c_2)/(\bar{v}_1 - \bar{v}_2) = 0.9$, for three policies: first-best (FB), second-best with strategic delay allowed (SB), and second-best with restriction to the work conserving $c\mu$ policy (SB-wc). The threshold $\mu^{sd} = 6.3$. For $\mu \in [0.9, 5.4]$, the three policies agree, which illustrates part 1(b). For $\mu > 5.4$ the first-best is not second-best. For $\mu \in (5.4, 6.3]$, strategic delay is suboptimal, so SB and SB-wc agree. For $\mu > 6.3$, strategic delay is optimal, and SB deviates from SB-wc: To ensure IC, SB inflates the class 2 lead time through strategic delay, whereas SB-wc drops the price of class 1 and raises that of class 2. As a result, SB achieves almost as much price discrimination as FB, and much more than SB-wc. For example, at $\mu = 14$ the class 1 price *premium* versus the class 2 price is 193% under FB, 189% under SB, but only 16% under SB-wc. The value of strategic delay can be dramatic: the revenue of SB exceeds that of SB-wc by approximately 10% at $\mu = 10$ and 23% at $\mu = 20$. (Over all parameters with $\bar{v}_1/c_1 < \bar{v}_2/c_2$ and $\mu = \infty$, the *maximum* revenue gain of strategic delay versus work conserving scheduling approaches 100%.)

6.4.5. The Impatient Type Has the Higher $\bar{v}_i/c_i$ Ratio. In contrast to the other cases, here it may be that strategic delay is optimal only at relatively scarce but not at ample capacity, or only at low and high but not at intermediate capacity.

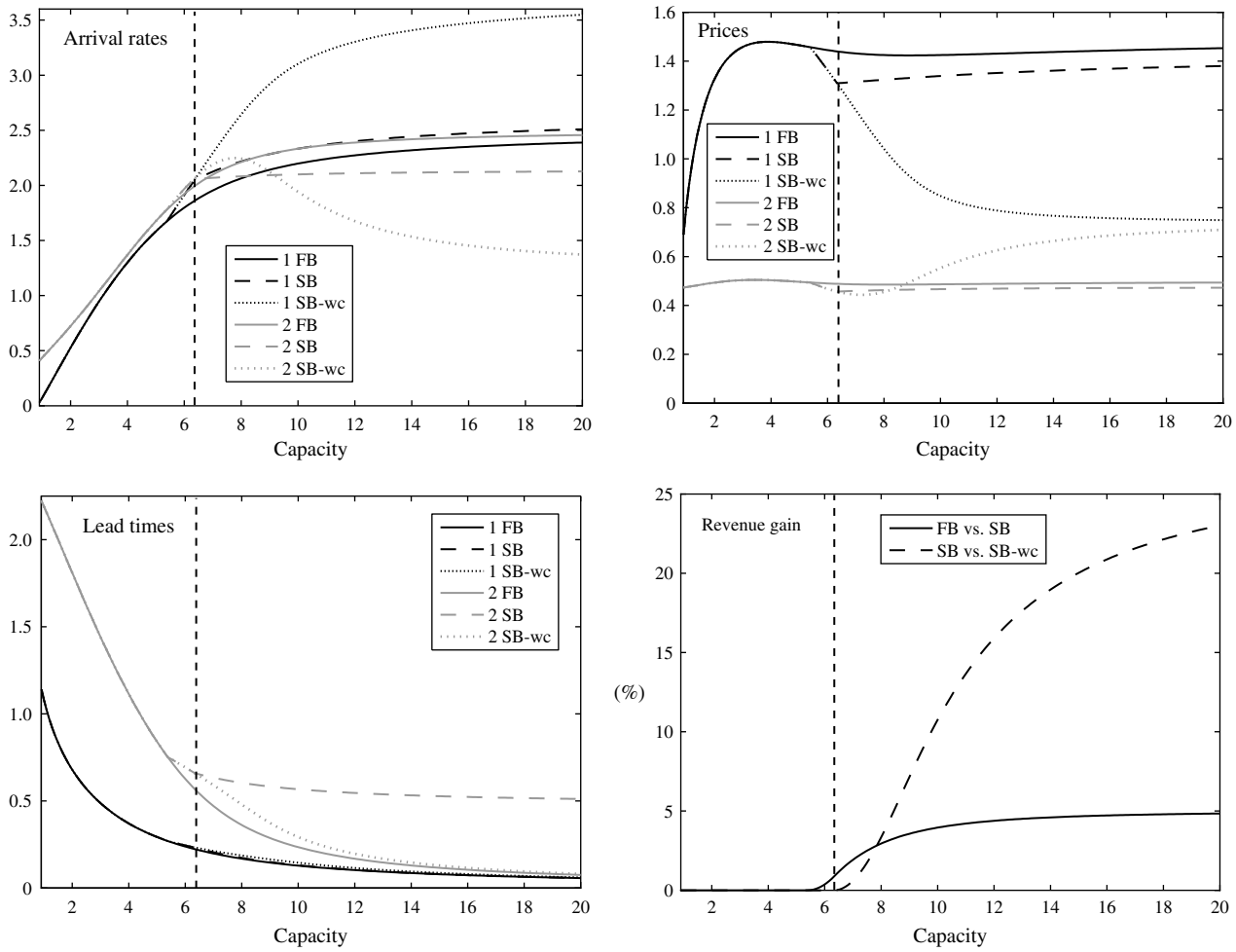
Proposition 7. *Fix $\mu > \mu_0$, and suppose that the impatient type has the higher $\bar{v}_i/c_i$ ratio.*

1. *Then*

$$\frac{\bar{v}_1 - \bar{v}_2}{c_1 - c_2} > \frac{\bar{v}_1}{c_1} > \frac{\bar{v}_2}{c_2} = \frac{1}{\mu_0} > \frac{1}{\mu^f} > \frac{1}{\mu^s}.$$

(a) *For $\mu \leq \mu^f$, the first-best policy is second-best and serves only impatient customers. For $\mu \in (\mu^f, \mu^s)$, it is first-best to serve both types, but second-best to serve only impatient customers; they prefer class 2 at the first-best solution. For $\mu > \mu^s$, it is second-best to serve both types.*

Figure 2 Illustration of Proposition 6: Strategic Delay Is Optimal Only if Capacity Is Sufficiently Large



Notes. $\bar{v}_1 = 3$, $\bar{v}_2 = 1$, $c_1 = 2$, $c_2 = 0.2$, and $\Lambda_1 = \Lambda_2 = 5$. Capacity thresholds: $\mu^f < \mu^s < (c_1 - c_2)/(\bar{v}_1 - \bar{v}_2) = 0.9 < \mu^{sd} = 6.3$.

(b) *There exist capacity levels at which strategic delay is optimal if and only if* $\varepsilon_1(\underline{\lambda}_1) > 1$, *i.e.*, $R_1'(\underline{\lambda}_1) > 0$, *where* $\underline{\lambda}_1 = \arg\{\lambda_1 \geq 0\colon v_1(\lambda_1)/c_1 = \bar{v}_2/c_2\} \in (0, \Lambda_1)$. *In this case,*

$$\mu^{sd} = \mu^s = \underline{\lambda}_1 + \frac{1 + \sqrt{1 + 4\underline{\lambda}_1 \cdot R_1'(\underline{\lambda}_1)/c_1}}{2R_1'(\underline{\lambda}_1)/c_1}. \tag{51}$$

If $R_1'(\underline{\lambda}_1) \leq 0$, *it is second-best to serve only impatient customers for all* μ.

2. *For linear* $v_i(\cdot)$, $R_1'(\underline{\lambda}_1) > 0 \Leftrightarrow \bar{v}_1/\bar{v}_2 < 2(c_1/c_2)$. *In this case,* $\mu^{sd} = \mu^s$ *is given by* (51) *with* $\underline{\lambda}_1 = \Lambda_1(1 - (c_1/c_2)\bar{v}_2/\bar{v}_1)$ *and* $R_1'(\underline{\lambda}_1)/c_1 = 2(\bar{v}_2/c_2) - \bar{v}_1/c_1$.

Strategic delay need not be optimal for all $\mu > \mu^{sd}$:

(a) *If* $\bar{v}_1/\bar{v}_2 \leq c_1/(c_1 - c_2/2)$ *or* $\Lambda_2 > 2\Lambda_1((c_1/c_2)(1 - \bar{v}_2/\bar{v}_1) - 1/2)$, *then strategic delay is optimal for* $\mu \in (\mu^{sd}, \bar{\mu})$, *but not for* $\mu \geq \bar{\bar{\mu}}$, *where* $\bar{\mu}$ *and* $\bar{\bar{\mu}}$ *are thresholds that satisfy* $\mu^{sd} < \bar{\mu} \leq \bar{\bar{\mu}} < \infty$.

(b) *If* $c_1/(c_1 - c_2/2) < \bar{v}_1/\bar{v}_2 < 2(c_1/c_2)$, *there is a threshold* $\underline{\Lambda}_2 \in (0, 2\Lambda_1((c_1/c_2)(1 - \bar{v}_2/\bar{v}_1) - 1/2)]$ *such that the following holds.*

If $\Lambda_2 \in (0, \underline{\Lambda}_2)$, *then strategic delay is optimal for all* $\mu > \mu^{sd}$.

If $\underline{\Lambda}_2 < \Lambda_2 < 2\Lambda_1((c_1/c_2)(1 - \bar{v}_2/\bar{v}_1) - 1/2)$, *then strategic delay is optimal if and only if* $\mu \in (\mu^{sd}, \bar{\mu}) \cup (\underline{\underline{\mu}}, \infty]$, *where* $\bar{\mu}$ *and* $\underline{\underline{\mu}}$ *are thresholds that satisfy* $\mu^{sd} < \bar{\mu} < \underline{\underline{\mu}} < \infty$.

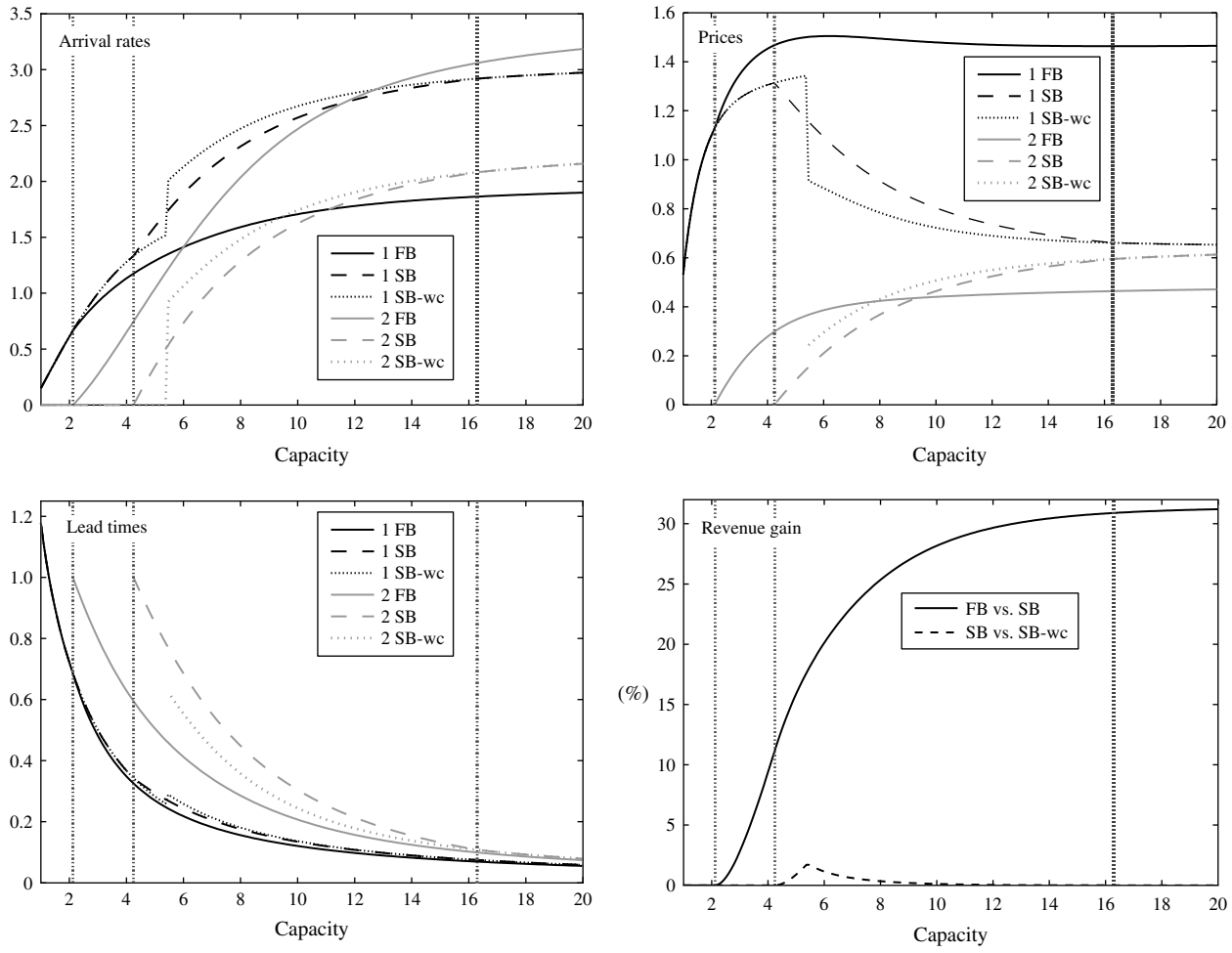
Since $(\bar{v}_1 - \bar{v}_2)/(c_1 - c_2) > \bar{v}_1/c_1 > \bar{v}_2/c_2$, the lead-time condition holds, but the price condition is violated for lower capacity levels $\mu > \mu_0$. Specifically, at lower capacity, the lead times are below the IC indifference threshold, so the impatient type prefers class 2 targeted to the patient type; that is, the lead-time condition holds and strategic delay is needed if class 2 is opened. Strategic delay reduces the class 2 price relative to the $c\mu$ policy, so the second-best capacity threshold for serving both types exceeds the first-best threshold, i.e., $\mu^s > \mu^f$. By Lemma 4, opening class 2 is second-best if and only if the price condition is satisfied. By Proposition 7.1(b), the price condition holds at *some* capacity if and only if $R_1'(\underline{\lambda}_1) > 0$, i.e., if it is optimal to serve more than $\underline{\lambda}_1$ of the impatient type at ample capacity, where $\underline{\lambda}_1$ is the minimal arrival rate that yields a nonnegative class 2 price under strategic delay. In this case, strategic delay is optimal for

a set of larger capacity levels around the threshold μ^{sd} in (51). This result shows that optimal strategic delay arises naturally if the impatient type has the higher $\bar{v}_i/c_i$ ratio: The condition $R_1'(\underline{\lambda}_1) > 0$ is easily satisfied, loosely speaking, whenever the types' $\bar{v}_i/c_i$ ratios are not too far apart. This condition is also easy to evaluate.

For capacity $\mu > \mu^{sd}$, the price condition holds, so that optimal strategic delay depends only on the lead-time and segment-size conditions (Proposition 4). In contrast to cases where $\bar{v}_1/c_1 \le \bar{v}_2/c_2$, here strategic delay need not be optimal for all $\mu > \mu^{sd}$: Around μ^{sd}, optimal strategic delay arises because the class 2 price and arrival rate are inherently low, so that the lead-time and segment-size conditions hold in general. At ample capacity, optimal strategic delay also hinges on the elasticities and segment sizes. Proposition 7.2 for linear $v_i(\cdot)$ follows by part 1, and by (47) for ample capacity. In part 2(a), the lead time and/or the segment-size conditions in (47) are violated, so strategic delay is not optimal at larger capacity. In part 2(b) both conditions hold, so strategic delay is optimal for smaller capacity close to μ^{sd} and for sufficiently large capacity; however, for intermediate capacity levels, strategic delay is optimal only if the segment size Λ_2 is below the threshold $\underline{\Lambda}_2$.

Example 2. Figure 3 illustrates Proposition 7.2(a). It shows the same metrics and policies as in Example 1. The thresholds are $\mu_0 = 1$, $\mu^f = 2.1$, $\mu^{sd} = \mu^s = 4.2$, and $\bar{\mu} = \bar{\bar{\mu}} = 16.3$. For $\mu \in (4.2, 5.5)$, serving patient customers is second-best only if strategic delay is allowed (SB). If it is precluded (SB-wc), then opening class 2 reduces profits because it requires a large class 1 price drop to ensure IC. In this capacity range, optimal strategic delay yields a *Pareto improvement* versus SB-wc: The impatient customer arrival rate is higher under SB, which implies a lower full price by (19) and a higher customer surplus. For $\mu \ge 5.5$, class 2 is profitable under SB-wc: the drop in class 1 revenues is offset by additional class 2 revenues. Strategic delay

Figure 3 Illustration of Proposition 7.2(a): Strategic Delay Is Optimal if Capacity Is Relatively Scarce, But Not if It Is Ample



Notes. $\bar{v}_1 = 3$, $\bar{v}_2 = 1$, $c_1 = 2$, $c_2 = 1$, and $\Lambda_1 = 4$, $\Lambda_2 = 7$. Capacity thresholds: $\mu^{fb} = 2.1$, $\mu^{sd} = \mu^S = 4.2$, and $\bar{\mu} = \bar{\bar{\mu}} = 16.3$.

is optimal for $\mu \in (4.2, 16.3)$ and yields more price discrimination than SB-wc, as in Example 1. Strategic delay is suboptimal for $\mu \geq 16.3$: Since the patient segment is relatively large and time sensitive, it is more profitable to reduce the class 2 lead time and raise its price at larger capacity. Strategic delay yields less value than in Example 1: The SB revenue exceeds that under SB-wc by at most 2%, at $\mu = 5.5$. (Over all $\bar{v}_1/c_1 > \bar{v}_2/c_2$ and $\mu = \infty$, the *maximum* revenue gain of strategic delay approaches 33%.)

6.5. IC Social Optimization vs. IC Revenue Maximization

Mendelson and Whang (1990) characterize the *socially optimal* and IC price/lead-time menu, allowing $N > 2$ delay cost rates. Their main result, that the work conserving $c\mu$ priority policy is socially optimal and IC, has an intuitive geometric interpretation in our analytical framework. Let $\mathrm{NV}(\lambda, \mathbf{W})$ denote the net value rate. The social optimization problem without IC constraints is

$$\max_{\boldsymbol{\lambda} \in M(\mu), \mathbf{W} \in \mathrm{OA}(\boldsymbol{\lambda}, \mu)} \mathrm{NV}(\lambda, \mathbf{W}) = \sum_{i=1}^{2} \left[\int_0^{\lambda_i} v_i(x)\, dx - c_i \lambda_i W_i \right].$$

Since the $c\mu$ policy is first-best (Lemma 2), let

$$\mathrm{NV}^{c\mu}(\boldsymbol{\lambda}, \mu) \triangleq \sum_{i=1}^{2} \left[\int_0^{\lambda_i} v_i(x)\, dx - c_i \lambda_i w_i^{c\mu}(\boldsymbol{\lambda}, \mu) \right]$$

be the first-best net value function. That the $c\mu$ policy is socially optimal and IC holds because the *socially optimal arrival rate vector is in* $M_0(\mu)$: $\arg\max_{\boldsymbol{\lambda} \in M(\mu)} \mathrm{NV}^{c\mu}(\boldsymbol{\lambda}, \mu) \in M_0(\mu)$. The following stronger property also implies this result. At every $\boldsymbol{\lambda}$ where type i customers have an incentive to buy class $j \neq i$ under the $c\mu$ policy, they also have the higher marginal net value, so $\boldsymbol{\lambda}$ cannot be socially optimal: $\boldsymbol{\lambda} \in M_1(\mu) \Rightarrow \mathrm{NV}^{c\mu}_{\lambda_1}(\boldsymbol{\lambda}, \mu) < \mathrm{NV}^{c\mu}_{\lambda_2}(\boldsymbol{\lambda}, \mu)$ and $\boldsymbol{\lambda} \in M_2(\mu) \Rightarrow \mathrm{NV}^{c\mu}_{\lambda_1}(\boldsymbol{\lambda}, \mu) > \mathrm{NV}^{c\mu}_{\lambda_2}(\boldsymbol{\lambda}, \mu)$; that is, the types' marginal *net value* contributions are aligned with their incentives.

In contrast, under revenue optimization, the types' marginal *revenue* contributions are not necessarily aligned with their incentives. Therefore, the revenue-maximizing first-best arrival rates need *not* be in $M_0(\mu)$; for example, if strategic delay is optimal, then $\boldsymbol{\lambda}^f(\mu) \in M_2(\mu)$ by Proposition 3.2. This is also why the problem of IC revenue maximization for $N > 2$ is significantly more challenging than the problem of IC social optimization.

7. Concluding Remarks

We present a novel problem formulation and solution method for designing revenue-maximizing and IC price/lead-time menus in queueing systems. This framework, based on Afèche (2004), combines mechanism design and the achievable region approach. It can be applied to systems with different operational or demand attributes (e.g., Katta and Sethuraman 2005, Yahalom et al. 2006, Afèche and Pavlin 2011, Cui et al. 2012, Maglaras et al. 2013). We show that a strategic delay policy is optimal for a broad range of demand and capacity conditions; see §1.2 for a summary of these results.

Strategic delay runs counter to conventional work conserving and delay-cost-minimizing scheduling policies. A general implication is that firms that use lead-time-based price differentiation should also consider customer incentives, not only operational constraints, in their scheduling policies.

The optimality of strategic delay also raises implementation issues. The following criteria may be helpful in choosing among the delay tactics described in §3.2, idling the server before, reducing its speed during, and delaying the delivery after processing.

(i) *Preemption flexibility*. The provider will want the ability to use strategic delay without slowing down the higher priority class(es). Slowing down the server meets this criterion only if low priority jobs can be preempted, for example, in standard tax preparation services with minimal customer interaction between order placement and delivery. Otherwise, it is preferable to use server idleness before and/or delivery delays after processing. The latter give the most control over both low and high priority lead times.

(ii) *Processing rate flexibility*. Similarly, if it is costly or infeasible to vary the server speed, for example, in laboratory tests that must meet stringent technical requirements, then the choice is between server idleness and delivery delays.

(iii) *Customer interaction and information*. The right delay tactic also depends on what customers know about their delays. In manufacturing operations without customer interaction between order and delivery, all three approaches are available. Similarly, in services such as package delivery, providers can implement generalized notions of pre- and postprocessing delays at hubs close to source and destination. However, in services where customers interact with the server during processing, delivery delays after processing may be infeasible. In such cases, the choice depends on how much customers know about the system state and service requirements. Idling the server will upset customers if they can see the server, for example, in car rental branches. Slowing down the server may upset customers, unless they are unfamiliar with service requirements, for example, in technical support services. Therefore, in call centers with standard transactions, it may be best to implement strategic delay through server idleness before processing.

(iv) *Lead-time variability*. Reducing the lead-time variance for a given target mean may be desirable if customers are averse to delay cost risk, for example, when sourcing critical electronic components. This may be easiest to achieve by choosing delivery delays after processing to minimize the difference between realized and quoted lead times.

Our results raise further questions.

(i) A potentially fruitful avenue is to study some of the implementation issues discussed above. For example, what is the optimal policy to implement strategic delay via server idleness only? How should strategic delay be implemented if customers or the firm incur different costs for delays before, during, and after processing?

(ii) An interesting challenge is the multitype version of our problem. It is analytically tractable under certain restrictions on the valuation–delay cost distribution (Katta and Sethuraman 2005, Afèche and Pavlin 2011). The problem for an unrestricted valuation–delay cost distribution remains open. The difficulty arises because the number of IC constraints is quadratic in the number of types, and with two-dimensional types, local IC between neighboring types does not ensure global IC. This challenge does not arise under IC social optimization, as explained in §6.5, and IC constraints are absent in standard applications of the achievable region approach.

(iii) We assume the provider cannot distinguish patient from impatient types. If types correspond to identifiable segments such as residential versus business customers, the provider faces the first-best problem, and strategic delay is not optimal. The case with more than two types, only some of which can be distinguished, is practically relevant and potentially of theoretical interest in that IC constraints only apply to a subset of classes.

(iv) We assume i.i.d. service times. Afèche (2004) shows for types with heterogeneous service requirements that optimal strategic delay can also arise if impatient customers have the lower mean service time, but that other delay tactics may be optimal if patient customers have the higher $c\mu$ index; namely, it may be optimal to alter priorities relative to the $c\mu$ policy, in some cases prioritizing them in the *reverse* $c\mu$ order. These results assume that service requirements are fixed. Another interesting problem arises if quality depends on service times, which is characteristic of discretionary services (Hopp et al. 2007): How should firms design price/lead-time/quality menus?

(v) Last but not least, although it is known that strategic delay may be optimal in a duopoly (Afèche 2008), IC and revenue-maximizing price/lead-time design under competition is only partially understood.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0449.

Acknowledgments

The author is grateful to the associate editor and the referees for numerous constructive comments that helped improve this paper significantly.

References

Afèche P (2004) Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delay tactics. Working paper, University of Toronto, Toronto.

Afèche P (2008) Revenue management and delay tactics under competition and customer choice. Presentation, Manufacturing and Service Operations Management Conference, June 5, University of Maryland, College Park, MD.

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

Afèche P, Pavlin M (2011) Optimal price-lead time menus for queues with customer choice: Priorities, pooling and strategic delay. Working paper, University of Toronto, Toronto.

Allon G, Federgruen A (2009) Competition in service industries with segmented markets. *Management Sci.* 55(4):619–634.

Ata TL, Olsen M (2013) Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. *Queueing Systems* 73(1):35–78.

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

Çelik S, Maglaras C (2008) Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Sci.* 54(6):1132–1146.

Coffman EG Jr, Mitrani I (1980) A characterization of waiting time performance realizable by single-server queues. *Oper. Res.* 28(3):810–821.

Cui T, Chen Y, Shen ZM (2012) Revenue-maximizing pricing, scheduling, and probabilistic admission control for queues under information asymmetry. Working paper, University of California, Berkeley, Berkeley.

Debo L, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54(8):1497–1512.

Federgruen A, Groenevelt H (1988) Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* 36(5):733–741.

Hassin R, Haviv M (2003) *To Queue or Not to Queue* (Kluwer, Boston).

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Hsu VN, Xu SH, Jukic B (2009) Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing Service Oper. Management* 11(3):375–396.

Jayaswal S, Jewkes E, Ray S (2011) Product differentiation and operations strategy in a capacitated environment. *Eur. J. Oper. Res.* 210(3):716–728.

Kanet JJ, Sridharan V (2000) Scheduling with inserted idle time: Problem taxonomy and literature review. *Oper. Res.* 49(1):99–110.

Katta A, Sethuraman J (2005) Pricing strategies and service differentiation in queues—A profit maximization perspective. Working paper, Columbia University, New York.

Lederer PJ, Li L (1997) Pricing, production, scheduling and delivery-time competition. *Oper. Res.* 45(3):407–420.

Maglaras C, Zeevi A (2003) Pricing and performance analysis for a system with differentiated services and customer choice. Srikant R, Voulgaris G, eds. *Proc. 41st Annual Allerton Conf. on Communication, Control, and Comput., Allerton, IL.*

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53(2):242–262.

Maglaras C, Yao J, Zeevi A (2013) Revenue maximization in queues via service differentiation. Working paper, Columbia University, New York.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Oper. Res.* 38(5):870–883.

Mussa M, Rosen S (1978) Monopoly and product quality. *J. Econom. Theory* 18(2):301–317.

Myerson RB (1981) Optimal auction design. *Math. Oper. Res.* 6(1):58–73.

Naor P (1969) On the regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Plambeck E (2004) Optimal leadtime differentiation via diffusion approximations. *Oper. Res.* 52(2):213–228.

Rao S, Petersen ER (1998) Optimal pricing of priority services. *Oper. Res.* 46(1):46–56.

Rochet J, Chonè P (1998) Ironing, sweeping, and multidimensional screening. *Econometrica* 66(4):783–826.

Stidham S Jr (2002) Analysis, design and control of queueing systems. *Oper. Res.* 50(1):197–216.

Su X, Zenios S (2006) Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Sci.* 52(11):1647–1660.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* 46(9):1249–1267.

Yahalom T, Harrison JM, Kumar S (2006) Designing and pricing incentive compatible grades of service in queueing systems. Working Paper, Stanford University, Stanford, CA.

Zhao X, Stecke KE, Prasad A (2012) Lead time and price quotation mode selection: Uniform or differentiated? *Production Oper. Management* 21(1):177–193.