Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Idleness and other Delaying Tactics

Philipp Afèche
p-afeche@kellogg.northwestern.edu
Kellogg School of Management · Northwestern University
Evanston, IL 60208

January 2004

Abstract

How should a capacity-constrained firm design an incentive-compatible price-scheduling mechanism to maximize revenues from a heterogeneous pool of time-sensitive customers with private information on their willingness to pay, time-sensitivity and processing requirement? We consider this question in the context of a queueing system that serves two customer types. We provide the following insights. First, the familiar $c\mu$ priority rule, known to minimize the system-wide expected delay cost and to be incentive-compatible under social optimization, need not be optimal in this setting. This specific fact suggests a more general guideline: in designing incentive-compatible and revenue-maximizing scheduling policies, delay cost-minimization, which plays a prominent role in controlling and pricing queueing systems, should not be the dominant criterion ex ante. Second, we identify optimal scheduling policies with novel features. One such policy prioritizes the more time-sensitive customers but voluntarily delays the completed orders of low-priority customers. This insertion of strategic idleness deters time-sensitive customers from purchasing the low-priority class. In other situations, it is optimal to appropriately randomize priority assignments, in one extreme case serving customers in the reverse $c\mu$ order, which maximizes the system delay cost among all work conserving policies. Compared to the $c\mu$ rule, these optimal policies increase, decrease or reverse the delay differentiation between customer types. We show how the optimal level of delay differentiation systematically emerges from a trade-off between operational constraints and customer incentives. Third, our stepwise solution approach can be adapted for designing revenue-maximizing and incentive-compatible mechanisms in systems with different customer attributes or operational properties.

Key words: Congestion, Delay, Incentives, Mechanism Design, Pricing, Priorities, Quality, Queuing Systems, Revenue Management, Scheduling, Service Differentiation.

1 Introduction

How should a capacity-constrained firm design a price-scheduling mechanism to maximize its revenues from a heterogeneous pool of time-sensitive customers with private information about their willingness to pay, delay cost and processing requirement? Many service and manufacturing firms face this question. For example, transportation service providers such as Federal Express and UPS offer customers a menu of price-delivery options (same-day, overnight, two-day, etc.), recognizing that some customers value speedy delivery more than others. Such delay-based price differentiation can also be a valuable revenue management tool for a manufacturer, particularly if it produces or assembles products to order. A make-to-order process typically allows a firm to increase product variety and lower inventories, but the inherent lead time between order placement and fulfillment may cause it to lose the business of impatient customers. By offering the option to pay more for faster delivery while charging less for longer lead times, it may be possible to accommodate customers with different preferences by segmenting the market and to allocate the processing resource(s) to orders based on their time-sensitivity and profitability. The problem of determining the revenue-maximizing price and scheduling policy is significantly complicated if the provider only has aggregate information about customer attributes, e.g., based on market research, but cannot tell apart individual customers. In this common scenario, all customers can choose among all options on the provider's price-delay menu and do so in line with their own self-interest. This behavior gives rise to *incentive-compatibility* constraints, which the provider must take into account when designing her price-scheduling mechanism.

We analyze this problem in the context of a queueing model. Incentive-compatible pricing and scheduling in queueing systems is well understood under the *social optimization* objective of maximizing the sum of customer plus provider benefits. The socially optimal mechanism is of interest for an operation that serves customers *within* an organization, and to provide guidelines for economic policy. However, a firm that deals with external customers is mainly concerned with its own revenues. Interestingly, the design of *revenue-maximizing and incentive-compatible* price-scheduling mechanisms for queueing systems has only received limited attention so far. As we show in this paper, the revenue-maximizing scheduling policy has novel features and may significantly differ from that under social optimization.

We consider a firm that serves two segments or types of customers, each characterized by three attributes: a value or willingness to pay for one unit of the product or service (drawn from a general continuous distribution), a linear delay cost rate c_i, and a mean service time $1/\mu_i$. Without loss of generality, we assume that type 1 have a larger ratio of delay cost rate to mean service time than type 2 customers ($c_1\mu_1 > c_2\mu_2$). Customers arrive according to independent Poisson processes and have exponentially distributed service times. The firm offers two service classes. It chooses a pair of prices that may depend on actual service times and a scheduling policy that specifies how customers are processed and which determines the expected delay of each class as a function of the arrival rates. Customers are self-interested and behave strategically. They do not observe the queue and decide, based on the prices and expected delays of the two service classes, which one to purchase, if any, by maximizing their value minus total cost (price plus expected delay cost) of service.

This article aims to contribute two sets of insights, which we discuss in turn: it (i) derives optimal scheduling policies with novel features, and (ii) proposes a solution approach that should be of value in designing mechanisms for systems with different properties.

First, we show that maximizing revenues from customers with private information gives rise to optimal scheduling policies that do *not* minimize the system's expected delay cost rate. One such policy involves the insertion of *optimal strategic idleness,* and the other assigns priorities in a

non-standard way. Delay cost minimization plays a prominent role in the literature on scheduling and pricing multi-class queueing systems, both ex ante as an optimization criterion and ex post as a property of the optimal policy under other criteria. The familiar $c\mu$ *priority rule* minimizes the delay cost rate in systems with Poisson arrivals and linear delay cost rates (cf. Cox 1961, Kakalik 1969). It assigns static priority levels to customers in increasing order of their index $c_i \cdot \mu_i$. In our model this implies giving type 1 absolute priority over type 2 customers. (Other common scheduling policies such as the shortest remaining processing time discipline, which minimizes the average system inventory in single-server systems, are also consistent with delay cost minimization.) The delay cost-minimizing property of the $c\mu$ rule directly implies that it is both revenue-maximizing and socially optimal, *if* the provider can tell customer types apart. If types are indistinguishable to the provider, the $c\mu$ rule is also incentive-compatible under social optimization, as shown by Mendelson and Whang (1990) for $N \geq 2$ customer types. That is, at the socially optimal arrival rates, the $c\mu$ rule yields a level of *delay differentiation* between the different service classes that induces customers to choose priority classes in a manner consistent with the $c\mu$ rule. These results are robust under convex delay costs: in heavy traffic, a dynamic version of the $c\mu$ rule, the generalized $c\mu$ (or $Gc\mu$) rule, is asymptotically delay cost-minimizing (Van Mieghem 1995) and incentive-compatible under social optimization (Van Mieghem 2000).

However, we show that the $c\mu$ rule is incentive-compatible only for certain arrival rates, since customer incentives depend on system congestion. More importantly, the $c\mu$ rule need not be incentive-compatible at the revenue-maximizing arrival rates. The features of the optimal policies depend as follows on customer attributes (see Figure 1).

Customer Attributes	Type 1 vs. Type 2 customers $(c_1\mu_1 > c_2\mu_2)$		
Time Sensitivity	Higher $(c_1 > c_2)$	Higher $(c_1 > c_2)$	Lower $(c_1 < c_2)$
Mean Service Time	Higher $(\mu_1 < \mu_2)$	Equal or Lower $(\mu_1 \geqslant \mu_2)$	Lower $(\mu_1 > \mu_2)$
Socially Optimal & Incentive-compatible Scheduling Policy	**$c\mu$ rule**	**$c\mu$ rule**	**$c\mu$ rule**
Revenue-maximizing & Incentive-compatible Scheduling Policy	**$c\mu$ rule** always optimal	**Strategic Idleness** may be optimal: More delay differentiation, compared to $c\mu$ rule: • equal type 1 delay • larger type 2 delay	**Randomized Priorities or Reversed $c\mu$ Order** may be optimal: Less or reversed delay differentiation, compared to $c\mu$ rule: • larger type 1 delay • smaller type 2 delay

Figure 1: Revenue-maximizing and incentive-compatible scheduling policy may differ from the socially optimal and delay cost minimizing $c\mu$ rule, which gives absolute priority to type 1 customers.

If type 1 are both more time-sensitive and have a higher mean service time than type 2 customers ($c_1 > c_2$ and $\mu_1 < \mu_2$), the $c\mu$ rule is always optimal. Otherwise, the provider may be better

off by increasing or reducing the degree of delay differentiation, compared to the $c\mu$ rule. If type 1 are more time-sensitive and have an equal or smaller mean processing requirement than type 2 customers ($c_1 > c_2$ and $\mu_1 \geq \mu_2$), it may be revenue-maximizing to *increase the delay differentiation* among the two types, relative to the $c\mu$ rule: give type 1 customers absolute priority but intentionally delay type 2 customers' completed orders by an optimal amount of time. We refer to this idleness of completed jobs as *strategic idleness* since it aims to manipulate the incentives of strategic customers. It has the effect of degrading the quality of the low priority service, making it relatively less appealing to the more impatient type 1 customers and allowing the provider to charge them a higher price for high-priority service. Inserting strategic idleness is optimal under *relatively low* system congestion, and we give explicit conditions for when this course of action is advised under linear demand. It is worth noting that strategic idleness is not optimal because it reduces, but despite the fact that it *increases* the system's expected delay cost rate. In this sense, strategic idleness fundamentally differs from the server idleness that is a property of delay cost-minimizing policies in certain other settings. In those cases, it is optimal to keep a server idle at times when the momentary loss of system utilization is offset by the expected option value of better future job-to-server allocations. For example, in systems with nonpreemptive priorities and arrivals that are not Poisson, it may be optimal to idle the server when a long job is awaiting service but the arrival of a shorter job is impending (cf. Wolff 1989, p. 445). Rubinovitch (1985) shows that it may be optimal in multi-server systems to idle a slower server in anticipation that a faster server is about to become available.

If type 1 are less time-sensitive and have a smaller mean service time than type 2 customers ($c_1 < c_2$ and $\mu_1 > \mu_2$), the optimal policy may require *reducing or reversing the delay differentiation,* compared to the $c\mu$ rule, by lowering the expected delay of type 2 customers (who have low priority under the $c\mu$ rule) while raising that of type 1 customers. Altering the delays in this way reduces the class 1 and raises the class 2 price, making class 1 service more appealing to type 1 customers. The optimal delays can be attained by a policy that appropriately randomizes priority assignments. In certain extreme cases, it may be optimal to prioritize customers in the reverse $c\mu$ order, giving type 2 priority over type 1 customers, which *maximizes the delay cost rate among all work conserving policies.*

These findings support the following general guidelines for designing revenue-maximizing and incentive-compatible price-scheduling mechanisms. (*i*) Delay cost minimization (in our model equivalent to using the $c\mu$ rule) should not ex ante be a dominant criterion for choosing a scheduling policy since the optimal policy does not generally have this property. (*ii*) The optimal level of delay differentiation between the service classes systematically depends on customer attributes and on the system structure and congestion level.

A second set of insights derives from our solution approach for determining the optimal scheduling policy: it can be adapted for designing revenue-maximizing and incentive-compatible mechanisms in systems with different operational or customer attributes. The steps are: (*i*) use the incentive-compatibility constraints to obtain bounds that the expected delays must satisfy independent of the price and scheduling policy, (*ii*) combine these bounds with the linear constraints which define the (operationally) achievable set of expected delays and transform the scheduling control problem into a constrained optimization problem, (*iii*) partition the space of arrival rates into regions by applying the bounds derived from the incentive constraints to the delays under the $c\mu$ rule, (*iv*) determine the optimal policy in each region, and (*v*) identify the region that contains the revenue-maximizing arrival rates.

This paper bridges streams of research on queueing system control and pricing, and on mechanism design. A vast literature considers the analysis, design and control of queueing systems in settings where, unlike in this paper, the system manager is omniscient and omnipotent, able to

fully observe all information and determine all job flows. See Conway et al. (1967) for a classical or Stidham (2002) for a recent survey. As noted, minimizing the delay cost or related measures such as average waiting time or system inventory is a prevalent optimality criterion in these settings. We use the achievable-region approach, pioneered by Coffman and Mitrani (1980) and extended by Federgruen and Groenevelt (1986), Shanthikumar and Yao (1992) and others, to characterize the feasible set of expected delays.

Our analysis also draws on standard approaches from the theory of mechanism design, which studies optimal resource allocation problems under private information. Myerson (1981) provides a seminal analysis of optimal auction design. Jehiel et al. (1996) consider a problem with externalities. It is worth noting that unlike in their setting, the structure of the externality is endogenous in our model since it depends on the scheduling policy.

Numerous papers study pricing and scheduling for queueing systems with strategic agents. See Hassin and Haviv (2003) for an excellent survey. Naor (1969) is widely credited with the first published analysis in this area. He shows for a FIFO $M/M/1$ queue that individual customers' decisions are not socially optimal and that this problem can be remedied with a static admission price. There is no question of incentive-compatibility in his nor in other papers that consider customers with *identical* time-sensitivities and mean service times (e.g., Yechiali 1971, Balachandran 1972, Lippman and Stidham 1977 and Dewan and Mendelson 1990). Among papers that do consider *heterogeneous* customers, Kleinrock (1967) first studied priority pricing by ignoring customer incentives, whereas Marchand (1974), Ghanem (1975), Dolan (1978), Mendelson and Whang 1990 and Van Mieghem (2000) focus on and jointly provide a thorough understanding of the incentive-compatible and socially optimal mechanism. Ha (2002) derives incentive-compatible and socially optimal prices in systems where each customer chooses a service rate. Among recent studies on revenue-maximizing price and scheduling policies in the absence of customer choice, Maglaras and Zeevi (2003) consider a system without queueing that offers a Guaranteed Service class and one with Best-effort that shares capacity among customers, and Caldentey and Wein (2003) study the problem for a make-to-stock queue that serves a long term contract and spot market demand. By contrast, only partial insights appear to be available on *incentive-compatible* revenue-maximizing mechanisms. Considering two customer types with equal mean service times but different linear delay cost rates, Plambeck (2004) uses diffusion approximations to study the joint problem of dynamic lead time quotation, static pricing and capacity sizing for an $M/M/1$ queue in heavy traffic. A key assumption to justify the use of heavy traffic theory is that the patient customers tolerate long lead times. She proposes a policy that is approximately incentive-compatible at utilizations below 100% and is asymptotically optimal in the limit as the patient customers become very delay-tolerant and the utilization is near 100%. Rao and Petersen (1998) and Lederer and Li (1997) focus on incentive-compatible pricing to maximize revenues but *assume* (as opposed to derive) certain expected delay functions. The former consider a generic congestion model without specific queueing structure: the expected delay for each of a fixed number of priority classes is given by an exogenous function of flow rates. Lederer and Li (1997) study price-delay equilibria under perfect competition and assume that firms use the $c\mu$ rule for scheduling, which they justify by making reference to the delay cost-minimizing property discussed above.

Papers that consider competing firms and assume FIFO scheduling include Cachon and Harker (2002), Kalai et al. (1992) and Li and Lee (1994). Gupta et al. (1996) study a general equilibrium model of congestion in a network setting. Shumsky and Pinker (2003) consider incentive issues that arise between the gatekeeper of a service who may refer jobs to several specialists with private information.

The plan of this paper is as follows. Section 2 presents the model and problem formulation. Section 3 studies the case of homogenous service times, which is the simplest setup that gives rise

to optimal strategic idleness. Section 4 studies the general model with heterogeneous service times, where strategic idleness or priority service in the reverse $c\mu$ order may be optimal. Section 5 offers concluding remarks. Most proofs are in the appendix.

2 Model and Problem Formulation

We model a capacity-constrained firm as an $M/M/1$ queueing system that serves two types or market segments of delay-sensitive customers, indexed by $i = 1, 2$, who have private information about their preferences. For simplicity we assume a zero marginal cost of service. type i customers arrive according to an exogenous Poisson process with finite rate or market size Λ_i per unit time. Each segment comprises a pool of "atomistic" customers whose individual demands are infinitesimal relative to the arrival rate. Each customer has three attributes, a value, a delay cost rate and a service time, which we discuss next.

Customer attributes. Customers have a positive value, or willingness to pay in the absence of delay, for one unit of the product or service. type i customers' values are i.i.d. draws from a continuous distribution F_i with p.d.f. f_i, assumed strictly positive and continuous on the interval $V_i := [\underline{v}_i, \overline{v}_i]$, where $\underline{v}_i \geq 0$. Let $\overline{F}_i = 1 - F_i$. If all type i customers with value $\geq v$ request service, then their actual arrival (or demand) rate is $\lambda_i = \Lambda_i \overline{F}_i(v)$. Conversely, the marginal value of a type i customer corresponding to arrival rate λ_i equals $\overline{F}_i^{-1}(\frac{\lambda_i}{\Lambda})$, where $\overline{F}_i^{-1}$ is the inverse of $\overline{F}_i$. Define the downward-sloping marginal value (or inverse gross demand) functions as

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

where $v_i(\lambda_i)$ is a one-to-one mapping between the demand rate λ_i and the corresponding marginal value (cf. Lippman and Stidham 1977, Mendelson 1985). Observe that $v_i(0) = \overline{v}_i > v_i(\Lambda_i) = \underline{v}_i$, $v_i(\lambda_i) > 0$ and $v_i'(\lambda_i) < 0$ for $\lambda_i < \Lambda_i$. Let $\lambda = (\lambda_1, \lambda_2)$.

A type i customer incurs a constant delay cost rate $c_i > 0$ per unit time in the system, including her service time. The service times of type i customers are i.i.d. draws from an exponential distribution with mean μ_i^{-1}. We assume without loss of generality that $c_1\mu_1 > c_2\mu_2$. Until Section 4, we assume that type 1 are more time-sensitive than type 2 customers ($c_1 > c_2$) and that all customers have a unit mean service time ($\mu_1 = \mu_2 = 1$).

We make the following additional assumptions: *A*1. The arrival processes, service time and value distributions are mutually independent. *A*2. To avoid the case where customers are not profitable even in the absence of congestion, assume $v_i(0) > \frac{c_i}{\mu_i}$ for $i = 1, 2$. This still allows the possibility of only one type being served at the optimal solution. *A*3. For simplicity, we assume that it is not optimal to serve all customers of either type. This holds for all value distributions F_i if $\Lambda_i \geq \mu_i$. Otherwise, it is sufficient to require that $v_i(\Lambda_i) \leq \frac{c_i}{\mu_i}$.

Information structure. The arrival processes, value distributions, delay cost parameters and service time distributions are common knowledge. A customer's attributes are her private information and are observed neither by the provider nor by other customers. Specifically, a customer knows her actual value and delay cost rate but only her expected service time when making her purchase decision. Her actual service time becomes known - to her and to the provider - only once her order is completed. Thus, the private information of a type i customer with value v at the time of her purchase decision can be summarized by the pair (i, v), where i denotes a delay cost rate c_i and a mean service time μ_i^{-1}. Two type i customers with value v who experience different service times are identical ex ante, i.e., when making their purchase decision. The set of ex ante distinct customers is therefore $T := \{(i, v) : i \in \{1, 2\}, v \in V_i\}$. Only the provider observes the system state; customers lack this information when making their decisions.

Admissible mechanisms. The provider's problem is to design a mechanism that maximizes her expected revenue per unit time. The timing of decisions is as follows. The provider first chooses and announces a mechanism, which is defined by the number of service classes and by decision rules on how to price, admit and schedule customers in each class. (The terms "class" or "service class" refer to an option on the provider's service menu, whereas "type" refers to customer attributes.) Upon arriving to the facility, customers decide which service class to purchase, if any, as described below, and are accordingly admitted, charged and scheduled as prescribed by the announced mechanism. Customers cannot renege and those who do not purchase do not affect the subsequent evolution of the arrival process.

We restrict our attention to the following set of *admissible mechanisms.* (*i*) We focus on static pricing policies. Prices may depend on the service class and on customers' actual service times. (*ii*) We consider static and deterministic admission policies, whereby the provider admits every customer who requests service at the announced price and scheduling policies. Most queueing models where customers do not observe the system state implicitly assume static pricing and admission policies. (*iii*) The scheduling policy or rule is a control that specifies how admitted customers are processed at any time and determines the expected steady state delay of each service class as a function of the arrival rates to all classes. FIFO, LIFO and priority disciplines are commonly-assumed policies. In this paper we do not a priori assume any particular scheduling policy. We define $\mathcal{A}$ to be the set of admissible scheduling policies, which comprises all stationary policies for which the expected steady state delays of all service classes are well defined. We discuss the properties of admissible scheduling policies in more detail in Section 3.2. (*iv*) As we show below, there is no loss of generality in only offering up to two service classes. However, for the sake of transparency, we start with the most general case whereby the provider may target one class to each ex ante distinct customer, i.e., to each pair $(i, v) \in T$. We then show how customers' incentives and private information limit the provider's choices.

Mechanism design formulation. We introduce the notation to describe an admissible mechanism. Following the mechanism design literature, it is convenient to restrict attention to *direct revelation mechanisms.* Under such a mechanism, customers announce, possibly dishonestly, a type $(i, v) \in T$ and the provider bases admission, pricing and scheduling decisions on these announcements. As we discuss below, there is no loss of generality in considering only such mechanisms. An admissible direct revelation mechanism is described by the following functions. The indicator function $a_i : V_i \to \{0, 1\}$ summarizes the admission of type i customers, where $a_i(v) = 1$ if a type i customer with value v is admitted and $a_i(v) = 0$ otherwise. Let $a = (a_1, a_2)$. The corresponding type i arrival rate satisfies

$$\lambda_i = \Lambda_i \int_{v \in V_i} a_i(v) f_i(v) \, dv \tag{2}$$

and the set of feasible arrival rates is given by

$$M := \{\lambda : 0 \leq \lambda_i < \Lambda_i, \ \lambda_1 + \lambda_2 < 1\}, \tag{3}$$

where $\lambda_i < \Lambda_i$ follows from assumption *A*3 and $\lambda_1 + \lambda_2 < 1$ is necessary for system stability.

The function $p_i : V_i \to \mathbb{R}$ denotes the prices paid by type i customers, where we define $p_i(v) := 0$ if $a_i(v) = 0$. Let $p = (p_1, p_2)$. If customers have the same service time distribution, as is the case until Section 4, there is no loss of generality in restricting attention to prices that are independent of service time. The reader may want to think of $p_i(v)$ as the expected payment of a type i customer with value v at the time of her purchase decision.

Let $r \in \mathcal{A}$ denote an admissible scheduling policy. The function $W_i(\cdot|a, r) : V_i \to \mathbb{R}$ specifies the expected steady state delays (wait in queue plus service time) of type i customers under admission

rule a and scheduling policy r. Let $W(\cdot|a,r) = (W_1(\cdot|a,r), W_2(\cdot|a,r))$ and define $W_i(v|a,r) := 0$ if $a_i(v) = 0$. Under an admissible scheduling policy the expected delays $W_i(v|a,r)$, for $(i,v) \in T$, are well defined given any admission rule a that yields arrival rates $\lambda \in M$. Since each customer is infinitesimal, these expected delays are not affected by the actions of an individual customer. The scheduling policy r may be dynamic, but $W(\cdot|a,r)$ is a static function.

Under a mechanism (a,p,r), the service class of a customer who declares type (i,v) is characterized by three attributes: an admission indicator $a_i(v)$, a price (or expected payment) $p_i(v)$ and an expected steady state delay $W_i(v|a,r)$. Customers are self-interested and choose their type announcement strategically, seeking to maximize their expected utility. Since they have private information about their type, the targeted service classes must be compatible with customer incentives. A type i customer with value v who pays P for service and experiences a delay of t time units has a *net value* (net of delay cost) of $v - c_i \cdot t$, and her utility is $v - c_i \cdot t - P$. Since customers do not observe the system state, they forecast their delay assuming that the system is in steady state. If customers are admitted based on a, charged according to p and scheduled following r, a type i customer with value v who truthfully reports her type has expected utility

$$u_i(v|a,p,r) := (v - c_i \cdot W_i(v|a,r) - p_i(v)) \cdot a_i(v) \tag{4}$$

when admitted, charged and scheduled according to her targeted service class.

We call an admissible direct revelation mechanism (a,p,r) *feasible* if $\lambda \in M$ (where λ is given by (2)) and no customer has an incentive to misrepresent her private information, which holds if the expected utilities (4) satisfy the *individual rationality* (IR) constraints

$$u_i(v|a,p,r) \geq 0, \qquad (i,v) \in T, \tag{5}$$

and the *incentive-compatibility* (IC) constraints

$$u_i(v|a,p,r) \geq (v - c_i W_j(x|a,r) - p_j(x)) \cdot a_j(x) \qquad (i,v) \neq (j,x) \in T. \tag{6}$$

Thus, each customer maximizes her expected utility by choosing her designated service class, where the RHS of (6) is the expected utility of a type i customer with value v who chooses the service class tailored to a type j customer with value x. Under a feasible mechanism (a,p,r), it is a Nash equilibrium for customers to truthfully report their type.

It is worth making the following observations about this mechanism specification. First, the service classes chosen by distinct customers can but need not have different prices and expected delays. Our specification accommodates any degree of price and delay differentiation, ranging from uniform pricing and service for all classes (which can be implemented for example by charging all customers the same price and scheduling them FIFO) to setting for each service class a unique price-expected delay pair, and the optimal degree of price and delay differentiation is endogenously determined. Second, there is an immediate equivalence between a feasible direct revelation mechanism (a,p,r), in which the provider assigns service classes to customers based on their direct (and truthful) type announcements, and a corresponding "indirect" mechanism that offers the same number of service classes with the same attributes, but where customers select a service class and thereby signal their type indirectly. The same allocation of customers to service classes, with $a_i(v)$, $p_i(v)$ and $W_i(v|a,r)$ being the service class attributes of a customer with type $(i,v) \in T$, forms a Nash equilibrium for either mechanism. Third, there is no loss of generality in restricting attention to direct revelation mechanisms in the set of admissible mechanisms. In general, the provider could design other kinds of mechanisms where customers do not directly report their type or choose a service class. For example, they may be admitted, charged and prioritized based on bids that they submit before joining and seeing the system state (cf. Hassin 1995, Afeche and

Mendelson 2004). Any such mechanism is admissible if customers communicate their "signal" to the provider prior to joining and seeing the system state, and the expected admission, price and delay are well defined static functions of customer signals. Due to the *revelation principle* (cf. Myerson 1981), for any such feasible mechanism, there is a feasible direct revelation mechanism which gives to the provider and to all customers the same expected payoffs. Therefore, we may consider without loss of generality only admissible direct revelation mechanisms. The expected revenue rate from a mechanism (a, p, r) is

$$\Pi(a, p, r) := \sum_{i=1}^{2} \Lambda_i \int_{v \in V_i} a_i(v)\, p_i(v)\, f_i(v)\, dv. \tag{7}$$

The provider's problem is to choose functions $a_i : V_i \to \{0, 1\}$, $p_i : V_i \to \mathbb{R}$, $i = 1, 2$, and a scheduling rule $r \in \mathcal{A}$ so as to maximize (7) subject to (2-3) and (5-6). We call an admissible mechanism optimal if it is revenue-maximizing and feasible.

Discussion. One way to visualize the model is to consider a setting where a firm serves two distinct customer segments, each consisting of a large pool of small residential or business customers who arrive at random. The firm has aggregate information about the distributions of customer attributes, e.g., based on market research, but cannot tell individual customers apart and thus considers their values, delay cost parameters and service times as random samples from these distributions. The assumption that all type i customers have the same delay cost rate c_i adequately approximates settings where the differences in time-sensitivity are significant across segments (e.g., regular vs. premium customers, or leisure vs. business travellers), but less significant within each segment. The case with two distinct delay cost rates is also the *simplest* setup that yields our results. Our model does not fit a setting with a small number of large customers, since each large customer may significantly affect the system's delay distribution. Similar assumptions are implicit in most queueing models. Mendelson and Whang (1990) study social optimization for this model with $N \geq 2$ types.

3 Homogeneous Service Times

We start with the case where type 1 are more time-sensitive than type 2 customers ($c_1 > c_2$) and both types have equal mean service times. The plan is as follows. First, we use the IR and IC constraints to provide a simplified characterization of feasible admissible mechanisms that includes price-independent bounds on the expected delays, to characterize admissible mechanisms that yield the same revenue and to show that restricting attention to mechanisms that offer up to two distinct service classes involves no loss of generality. Second, we define the class of admissible scheduling policies, characterize the set of achievable expected delays under such policies and show how the scheduling control problem can be transformed into the problem of choosing expected delay vectors in the achievable set. Third, we use the delay bounds implied by the IC constraints to partition the demand rate set M into three regions, and we characterize the conditionally optimal mechanism for λ in each region. We define the notion of *strategic idleness* and show that it is optimal for λ in one of these regions. Finally, we characterize the jointly optimal arrival rates, prices and scheduling policy and identify for the linear demand case parameter combinations for which strategic idleness is optimal at the revenue-maximizing arrival rates.

3.1 Incentive-Compatible and Revenue Equivalent Mechanisms

The following Lemma gives a simple characterization of feasible admissible mechanisms.

Lemma 1 *Take an admissible mechanism (a, p, r) such that $\lambda \in M$. Such a mechanism is feasible, i.e., (5) and (6) hold, if and only if:*

1. *A type i customer with value x_i is admitted ($a_i(x_i) = 1$) if and only if $x_i \geq v_i(\lambda_i) > \underline{v}_i$.*

2. *All admitted type i customers have the same expected total cost of service:*

$$v_i(\lambda_i) = c_i W_i(x_i|a, r) + p_i(x_i) \text{ for } x_i \geq v_i(\lambda_i). \tag{8}$$

The marginal customer has zero expected utility. The utilities of type i customers are

$$u_i(x_i|a, p, r) = x_i - v_i(\lambda_i) \text{ for } x_i \geq v_i(\lambda_i). \tag{9}$$

3. *Call an admissible scheduling rule r "incentive-compatible at λ" or "IC at λ" if there exist prices such that (6) holds. A rule r is IC at λ if and only if the delays satisfy:*

$$\lambda_1 > 0 \Rightarrow \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \geq W_1(x_1|a, r) \text{ for } x_1 \geq v_1(\lambda_1) \tag{10}$$

$$\lambda_2 > 0 \Rightarrow \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \leq W_2(x_2|a, r) \text{ for } x_2 \geq v_2(\lambda_2). \tag{11}$$

By Lemma 1 there is a one-to-one mapping from admission rules a to arrival rates λ, and the incentive-compatible prices $p_i(v)$ are uniquely determined for $(i, v) \in T$, given arrival rates λ and expected delay $W_i(v|a, r)$. Since the provider cannot discriminate among type i customers with different values, the expected total cost (price plus expected delay cost) of all admitted type i customers must be the same in equilibrium and equal the value of the marginal customer who has zero expected utility. The conditions (10-11) are necessary and sufficient for *any* admissible rule r to be incentive-compatible at given arrival rates λ and give intuitive, price-independent bounds on the expected steady state delays of all admitted customers. As specified by (10), the delays of service classes targeted at type 1 customers must be "relatively small" (and the corresponding prices relatively high), i.e., lower than the ratio of marginal value to delay cost differences, to prevent the more patient type 2 customers from switching to one of the service classes targeted at type 1 customers. Similarly, the expected delays of type 2 customers must exceed the ratio of marginal value to delay cost differences, to prevent the more time-sensitive type 1 customers from purchasing a class targeted at a type 2 customer.

We have so far allowed for a unique service class for each distinct customer. Lemma 2 implies that there is no loss of generality in only considering up to two distinct classes.

Lemma 2 *For any feasible admissible mechanism (a, p, r) such that $\lambda \in M$, there is a feasible admissible mechanism (a, p', r') with the same expected revenue, where:*

1. *All admitted type i customers ($i = 1, 2$) have the same expected delay, given by*

$$W_i\left(v|a, r'\right) = W_i\left(\lambda, r'\right) := \frac{\Lambda_i}{\lambda_i} \int_{v=V_i} a_i(v) W_i(v|a, r) f_i(v) \, dv \text{ for } v \geq v_i(\lambda_i). \tag{12}$$

2. *All admitted type i customers ($i = 1, 2$) pay the same price, given by*

$$p_i(v) = p_i\left(\lambda, r'\right) := v_i(\lambda_i) - c_i W_i\left(\lambda, r'\right) \text{ for } v \geq v_i(\lambda_i). \tag{13}$$

Since there is no need for more than two distinct classes we simplify the notation as follows. We denote an admissible mechanism by (λ, r), where (2) and Lemma 1 define a one-to-one mapping from admission rules a to arrival rates λ. Write $W_i(\lambda, r)$ for the expected class i steady state delay given arrival rates $\lambda \in M$ and scheduling policy $r \in \mathcal{A}$, and let $W(\lambda, r) = (W_1(\lambda, r), W_2(\lambda, r))$. Let p_i be the class i price and $p = (p_1, p_2)$. For given (λ, r), prices are uniquely determined by the inverse demand relationships (13). The expected utility of a type i customer with value v who chooses class i service is $u_i(v|\lambda, r) = v - c_i \cdot W_i(\lambda, r) - p_i = v - v_i(\lambda_i)$. The provider's problem is

$$\max_{\lambda \in M,\ r \in \mathcal{A}} \Pi(\lambda, r) := \sum_{i=1}^{2} \lambda_i \cdot p_i(\lambda, r) = \sum_{i=1}^{2} \lambda_i \cdot (v_i(\lambda_i) - c_i \cdot W_i(\lambda, r)), \tag{14}$$

subject to (10-11). This formulation accommodates the case of equal attributes (price and expected delay) for both service classes, in which case (10-11) require that $W_i(\lambda, r) = (v_1(\lambda_1) - v_2(\lambda_2)) / (c_1 - c_2)$ for $i = 1, 2$.

3.2 Admissible Scheduling Policies and Achievable Performance

As we show below, the standard scheduling policies often assumed or shown to be optimal, such as work conserving FIFO or static absolute priority policies, need not be optimal in our setting. It is evident from (14) that the expected revenue rate depends on an admissible scheduling policy $r \in \mathcal{A}$ only through the corresponding expected delay function $W(\lambda, r)$. In this Section we define the class of admissible scheduling policies $\mathcal{A}$, define and characterize the set $D(\lambda)$ of expected delays that are achievable by admissible policies for $\lambda \in M$, and specify a family of policies that attain all vectors in $D(\lambda)$. We use this correspondence between scheduling policies and attainable expected delays to transform the control problem (14) into the simpler problem of choosing expected delays in the achievable set.

Definition 1. Let $\overline{\mathcal{A}}$ be the class of *admissible work conserving scheduling policies*. It consists of all policies that (a) are stationary, (b) do not idle the server when there are jobs waiting to be served, (c) do not affect arrival processes or service requirements, and (d) are nonanticipative, i.e., only make use of past history and the current state of the system (but cannot be based on actual remaining service times). We impose no further restrictions.

Condition (a) ensures that all customers forecast the same expected steady state delays, regardless of their arrival times, while $(b) - (d)$ guarantee that the expected steady state delays $W(\lambda, r)$ are well defined for $\lambda \in M$ and imply the following standard conservation law. For a proof, see Gelenbe and Mitrani (1980), Section 6.2.

Lemma 3 *Fix $\lambda \in M$. Under an admissible work conserving scheduling rule $r \in \overline{\mathcal{A}}$, the expected steady state delay of all admitted customers satisfies:*

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot W_1(\lambda, r) + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot W_2(\lambda, r) = \frac{1}{1 - \lambda_1 - \lambda_2}. \tag{15}$$

Property (15) concerns the average delay over *all* admitted customers. The expected steady state delays of individual service classes are also bounded by the expected delay vectors under *absolute* (or strict) priority disciplines, which give static preemptive priority to all customers of one class over all others and schedule customers of a given class FIFO (preemptive-resume and -repeat rules perform the same since service times are exponential). Naturally, with two service classes there are two such priority orders.

Definition 2. The $c\mu$ rule, denoted by $r = c\mu$, gives absolute preemptive priority to customers in increasing order of their product of delay cost rate by service rate, $c_i \cdot \mu_i$, i.e., a customer's priority

level increases in her time sensitivity and decreases in her mean service time. With $c_1 > c_2$, equal mean service times for all customers, and class i service targeted at type i customers, the $c\mu$ rule gives absolute priority to class 1 over class 2 customers. Similarly, the *reverse* $c\mu$ *rule*, denoted by $r = Rc\mu$, gives absolute preemptive priority to class 2 over class 1 customers.

Lemma 4 *Fix $\lambda \in M$. Under an admissible work conserving scheduling rule $r \in \overline{\mathcal{A}}$, the expected steady state delays of both service classes are bounded as follows:*

$$\frac{1}{1-\lambda_1} = W_1(\lambda, c\mu) \leq W_1(\lambda, r) \leq W_1(\lambda, Rc\mu) = \frac{1}{(1-\lambda_2)(1-\lambda_1-\lambda_2)} \tag{16}$$

$$\frac{1}{1-\lambda_2} = W_2(\lambda, Rc\mu) \leq W_2(\lambda, r) \leq W_2(\lambda, c\mu) = \frac{1}{(1-\lambda_1)(1-\lambda_1-\lambda_2)}. \tag{17}$$

The linear constraints of Lemmas 3 and 4 completely characterize the set of achievable expected delays under admissible work conserving scheduling policies. Shanthikumar and Yao (1992) call these constraints strong conservation laws (see also Coffman and Mitrani 1980 and Federgruen and Groenevelt 1986 for characterizations of the achievable performance space for multi-class queueing systems.) We augment the set $\overline{\mathcal{A}}$ as follows.

Definition 3. Let $\mathcal{A}$ be the class of *admissible scheduling policies*. A policy r is admissible if and only if there is a work conserving policy $r' \in \overline{\mathcal{A}}$ such that r differs from r' only in that it delays completed class i jobs on average by $d_i \geq 0$ time units. For an admissible work conserving rule $r \in \overline{\mathcal{A}}$ with expected delays $W(\lambda, r)$, infinitely many scheduling rules r' exist with $W(\lambda, r') = W(\lambda, r) + d$, where $d = (d_1, d_2) \geq 0$. We say that a policy $r \in \mathcal{A}$ inserts *job idleness* if $d_1 > 0$ and/or $d_2 > 0$ since completed jobs sit idle prior to leaving the system. Lemmas 3-4 imply the following bounds for the achievable expected delays:

Lemma 5 *Fix $\lambda \in M$. Under an admissible scheduling policy $r \in \mathcal{A}$, the expected steady state delay of all admitted customers satisfies:*

$$\frac{\lambda_1}{\lambda_1+\lambda_2} \cdot W_1(\lambda, r) + \frac{\lambda_2}{\lambda_1+\lambda_2} \cdot W_2(\lambda, r) \geq \frac{1}{1-\lambda_1-\lambda_2}. \tag{18}$$

The expected steady state delays of individual service classes are bounded below as follows:

$$W_1(\lambda, c\mu) = \frac{1}{1-\lambda_1} \leq W_1(\lambda, r) \tag{19}$$

$$W_2(\lambda, Rc\mu) = \frac{1}{1-\lambda_2} \leq W_2(\lambda, r). \tag{20}$$

For $\lambda \in M$, let $D(\lambda)$ denote the set of expected delay vectors $W = (W_1, W_2)$ that satisfy (18)-(20). Observe that $D(\lambda)$ is a simple polyhedron whose extreme points are attained by the expected delay vectors under the $c\mu$ and $Rc\mu$ priority rules, respectively. Any $W \in D(\lambda)$ can be attained by a member of the following family of admissible scheduling policies.

Definition 4. Let $r = (\alpha, d)$ denote the following *randomized static preemptive priority policy with inserted job idleness*. There is a high- and a low-priority queue, where customers in the former are given absolute preemptive priority over those in the latter, and customers within a queue are served in FIFO order. Under an (α, d) policy, class 1 (class 2) customers are placed in the high-priority queue with probability α $(1-\alpha)$, and in the low-priority queue with probability $1-\alpha$ (α), where $\alpha \in (0, 1)$. Upon completing service, class i customers are on average delayed by a finite $d_i \geq 0$

extra time units. Notice that an (α, d) policy is admissible. It type i customers choose class i service, the expected delays are

$$W_1(\lambda,(\alpha,d)) = \frac{\alpha}{1-\alpha\lambda_1-(1-\alpha)\lambda_2} + \frac{1-\alpha}{(1-\alpha\lambda_1-(1-\alpha)\lambda_2)(1-\lambda_1-\lambda_2)} + d_1 \quad (21)$$

$$W_2(\lambda,(\alpha,d)) = \frac{1-\alpha}{1-\alpha\lambda_1-(1-\alpha)\lambda_2} + \frac{\alpha}{(1-\alpha\lambda_1-(1-\alpha)\lambda_2)(1-\lambda_1-\lambda_2)} + d_2. \quad (22)$$

The $c\mu$ rule ($Rc\mu$ rule) corresponds to $\alpha = 1$ ($\alpha = 0$) and $d = 0$. Lemma 6 is easily verified.

Lemma 6 *For $\lambda \in M$ and $W \in D(\lambda)$, there is an (α, d)-policy such that $W(\lambda,(\alpha,d)) = W$.*

That is, for given $\lambda \in M$ the expected delays under (α, d) policies span the space $D(\lambda)$ of expected delays that are attainable by admissible policies $r \in \mathcal{A}$. The control problem (14) can therefore be transformed into the following optimization problem:

$$\max_{\lambda\in M,\ W\in D(\lambda)} \Pi(\lambda, W) := \sum_{i=1}^{2} \lambda_i \cdot (v_i(\lambda_i) - c_i \cdot W_i), \quad (23)$$

subject to (10-11). In summary, the expected delay vector W at feasible arrival rates $\lambda \in M$ is bounded by two sets of constraints: (10-11), implied by customer incentives, and (18)-(20), derived from the system properties.

3.3 Conditionally Optimal Scheduling Policy at Fixed λ

We now characterize for fixed $\lambda \in M$ the optimal expected delays and corresponding scheduling rules. We take the $c\mu$ priority rule as a starting point to partition M into three regions as discussed below. We show that the solutions have the same structure in each region and differ across regions. The $c\mu$ rule plays a prominent role in scheduling multi-class queueing systems with Poisson arrivals and *linear* delay cost since it has the following properties.

Average delay cost minimization and work conservation. Cox and Smith (1961) showed for a multi-class $M/G/1$ system with nonpreemptive priorities that the $c\mu$ rule minimizes the average delay cost over all nonpreemptive static policies, not allowing for server idleness. Kakalik (1969) showed that this policy is dynamically optimal as well, even if inserting idleness is permitted. For an $M/M/1$ system, the preemptive $c\mu$ rule minimizes the average delay cost over all policies. We state this classic result for equal mean service times (it is immediate from Lemmas 3-4) as:

Lemma 7 *Fix $\lambda \in M$. For $c_1 > c_2 \geq 0$ and $\mu_1 = \mu_2 = 1$, the $c\mu$ rule minimizes the expected aggregate delay cost per unit time over all admissible scheduling policies:*

$$\lambda_1 \cdot c_1 \cdot W_1(\lambda, c\mu) + \lambda_2 \cdot c_2 \cdot W_2(\lambda, c\mu) \leq \lambda_1 \cdot c_1 \cdot W_1(\lambda, r) + \lambda_2 \cdot c_2 \cdot W_2(\lambda, r) \quad \text{for } \forall\ r \in \mathcal{A}. \quad (24)$$

The $c\mu$ rule allows neither server idleness (Definition 1) nor job idleness (Definition 3). In light of (23), Lemma 7 implies that the $c\mu$ rule maximizes the revenue-rate for any $\lambda \in M$. However, it need not be the optimal policy as we show below, since it is not IC at all $\lambda \in M$.

Incentive-compatibility and social optimality. Mendelson and Whang (1990) derive the incentive-compatible mechanism that is *socially optimal*, i.e., that maximizes the expected aggregate net value rate, for an $M/M/1$ system with $N \geq 2$ customer types who have private information about their preferences (our customer model corresponds to theirs with $N = 2$). They show that the $c\mu$ rule is *socially optimal* and incentive-compatible. They do not consider the revenue-maximization problem.

Van Mieghem generalizes these results for convex delay costs and general arrival and service processes. In heavy traffic, the generalized $c\mu$ (or $Gc\mu$) rule, a dynamic version of the static $c\mu$ rule, is asymptotically delay cost-minimizing (Van Mieghem 1995) and socially optimal and incentive-compatible (Van Mieghem 2000). Lederer and Li (1997) characterize incentive-compatible price-delay equilibria in markets with multiple customer types and perfect competition, *assuming* that firms schedule based on the $c\mu$ rule.

Applying the bounds (10-11) implied by the IC constraints to the $c\mu$ rule partitions the set of arrival rate M into three regions as follows.

Region U_1 contains all λ where $\lambda_2 > 0$ and the expected low priority (class 2) delay under the $c\mu$ rule is smaller than the ratio of the marginal customers' value to delay cost differences:

$$U_1 := \left\{ \lambda \in M : \lambda_2 \cdot \left(W_2(\lambda, c\mu) - \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \right) < 0 \right\}. \tag{25}$$

Region U_0 comprises all λ where this ratio is in between the high- and low-priority delays:

$$U_0 = \left\{ \lambda \in M : \lambda_1 \left(W_1(\lambda, c\mu) - \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \right) \leq 0 \leq \lambda_2 \left(W_2(\lambda, c\mu) - \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \right) \right\}. \tag{26}$$

Region U_2 comprises all λ where $\lambda_1 > 0$ and the expected high-priority (class 1) delay under the $c\mu$ rule exceeds the ratio of marginal value to delay cost differences:

$$U_2 := \left\{ \lambda \in M : \lambda_1 \cdot \left(W_1(\lambda, c\mu) - \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \right) > 0 \right\}. \tag{27}$$

Let $W^*(\lambda)$ be the optimal expected delay vector at $\lambda \in M$, where $W^*(\lambda) = \arg \max_{W \in D(\lambda)} \Pi(\lambda, W)$ subject to (10-11). Let $r^*(\lambda)$ be an optimal scheduling rule, i.e., $W(\lambda, r^*(\lambda)) = W^*(\lambda)$.

Proposition 1 *For fixed $\lambda \in M$, the optimal delays and scheduling rules are as follows:*

1. *At $\lambda \in U_1$, neither the $c\mu$ rule nor any other admissible work conserving policy $r \in \overline{\mathcal{A}}$ is incentive-compatible. The optimal expected delays satisfy*

$$W_1^*(\lambda) = W_1(\lambda, c\mu I) = \frac{1}{1 - \lambda_1} = W_1(\lambda, c\mu) \tag{28}$$

$$W_2^*(\lambda) = W_2(\lambda, c\mu I) = \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} > W_2(\lambda, c\mu) = \frac{1}{(1 - \lambda_1)(1 - \lambda_1 - \lambda_2)}, \tag{29}$$

   where $r^*(\lambda) = c\mu I$ denotes the $c\mu I$ rule, or $c\mu$ rule with optimal strategic idleness, which is defined as follows: it gives type 1 preemptive priority over type 2 customers, but it artificially delays low-priority (type 2) customers by idling their completed jobs such that their mean delay equals the ratio of marginal value to delay cost differences.

2. *At $\lambda \in U_0$, the $c\mu$ rule is optimal ($r^*(\lambda) = c\mu$) and the optimal expected delays are*

$$W_1^*(\lambda) = W_1(\lambda, c\mu) = \frac{1}{1 - \lambda_1} \tag{30}$$

$$W_2^*(\lambda) = W_2(\lambda, c\mu) = \frac{1}{(1 - \lambda_1)(1 - \lambda_1 - \lambda_2)}. \tag{31}$$

3. *At $\lambda \in U_2$, no admissible scheduling policy is incentive-compatible.*

By Proposition 1 the delay cost minimizing $c\mu$ rule need not be revenue maximizing. The provider may be better off increasing the delay cost through insertion of job idleness. The optimal scheduling policy differs across regions since the incentive constraints (10-11) and the operational constraints (18)-(20) depend on λ. We discuss each region in turn.

Region U_1: Optimal Strategic Idleness. At arrival rates $\lambda \in U_1$, no work conserving policy $r \in \overline{\mathcal{A}}$ can be IC since the maximum expected delay of type 2 customers under such policies, which is attained under the $c\mu$ rule, is smaller than the ratio of marginal customers' value to delay cost differences. To see why, recall that for a scheduling policy to be IC, each price must equal the respective marginal customer's expected net value

$$p_i(\lambda, r) = v_i(\lambda_i) - c_i W_i(\lambda, r), \quad i = 1, 2, \tag{32}$$

and type 1 customers must not have an incentive to use class 2 service:

$$v_1(\lambda_1) = p_1(\lambda, r) + c_1 W_1(\lambda, r) \leq p_2(\lambda, r) + c_1 W_2(\lambda, r). \tag{33}$$

Combining (32) and (33) gives the following equivalent condition:

$$v_1(\lambda_1) - c_1 W_2(\lambda, r) \leq v_2(\lambda_2) - c_2 W_2(\lambda, r) \Leftrightarrow v_1(\lambda_1) - v_2(\lambda_2) \leq (c_1 - c_2) W_2(\lambda, r). \tag{34}$$

That is, the marginal type 1 customer's expected net value must be lower than that of the marginal type 2 customer if *both* use class 2 service. In region U_1, system congestion is *relatively low* in the sense that the expected class 2 delay under all work conserving policies $r \in \overline{\mathcal{A}}$ is too small to satisfy (34). As a result, type 1 customers have a *higher* expected net value in *either* service class than type 2 customers, although they are more impatient ($c_1 > c_2$). (Type 2 customers have no incentive to purchase class 1 service since $v_2(\lambda_2) - c_2 W_1(\lambda, r) \leq v_1(\lambda_1) - c_1 W_1(\lambda, r)$ at $\lambda_1 \in U_1$ for all work conserving policies $r \in \overline{\mathcal{A}}$.)

The shape of region U_1 (and whether it is empty or not) depends on the properties of the marginal value functions $v_i(\lambda_i)$, $i = 1, 2$, and on the underlying value distributions. If $v_1(0) - c_1 > v_2(0) - c_2$, then U_1 is nonempty and includes the origin ($\lambda = 0$), i.e., (34) is violated in the absence of congestion (when the expected delays under any work conserving policy equal the unit mean service time.) This condition is sufficient but not necessary.

By (34), the maximum class 1 price that type 1 customers are willing to pay equals the class 2 price plus a type 1 customer's expected delay cost difference between the two classes:

$$\overline{p}_1(\lambda, W) := p_2(\lambda, r) + c_1(W_2 - W_1) = v_2(\lambda_2) + (c_1 - c_2) W_2 - c_1 W_1. \tag{35}$$

As shown above, under the $c\mu$ rule (and any other work conserving policy) the marginal type 1 customer's expected net value exceeds this upper bound:

$$v_1(\lambda_1) - c_1 W_1(\lambda, c\mu) > \overline{p}_1(\lambda, W(\lambda, c\mu)), \tag{36}$$

implying that no price-expected delay combination can be a Nash equilibrium. To restore incentive-compatibility, the provider must make class 1 service *relatively* more attractive to type 1 customers, compared to class 2 service. Increasing the class 2 delay (while keeping the class 1 delay constant) raises the price upper bound $\overline{p}_1(\lambda, W)$, making class 2 service relatively less attractive as an outside option. Increasing the expected class 2 delay by ΔW_2 time units raises $\overline{p}_1(\lambda, W)$ by $\Delta W_2 \cdot (c_1 - c_2)$, whereas the class 2 price must be lowered by $\Delta W_2 \cdot c_2$ to counterbalance the increase in type 2 customers' delay cost.

The expected delay of class 2 customers can be increased through voluntary insertion of job idleness, i.e., by delaying their completed jobs by a positive amount of time. We call this job

idleness *strategic idleness* since it degrades the quality of the low priority service with the aim of manipulating the incentives of strategic customers. At the point where the class 2 delay equals $W_2(\lambda, c\mu I) = \frac{v_1(\lambda_1)-v_2(\lambda_2)}{c_1-c_2}$, type 1 customers' total expected cost from class 2 service equals that under class 1 service, establishing incentive-compatibility. Further delaying class 2 customers would only harm revenues. The optimal amount of strategic job idleness for class 2 customers is

$$d_2^*(\lambda) = \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} - \frac{1}{(1-\lambda_1)(1-\lambda_1-\lambda_2)} \quad \text{for } \lambda \in U_1. \tag{37}$$

Infinitely many scheduling rules yield the optimal expected delays (28-29): the distribution of extra idle time may differ across scheduling rules that attain these mean delays.

In conclusion, inserting strategic idleness reduces the value rate *generated* by the system (the "size of the pie") by increasing the delay cost rate, compared to the delay cost minimizing $c\mu$ rule, but allows the provider to *capture* a portion of this net value as revenue (its "share of the pie") by setting prices that leave the marginal customers of each type with zero expected utility. Note that *strategic idleness is optimal not because it reduces but despite the fact that it increases the average delay cost rate* and thereby reduces system efficiency. However, it alters the delay differentiation between the service classes in a way that restores incentive-compatibility. In our setting with preemptive priorities, idleness insertion can never be optimal under the criterion of delay cost minimization, even if arrivals are not Poisson.

At $\lambda \in U_1$, strategic idleness leaves customers with the same utility as the $c\mu$ rule: type 1 have the same delay and price while type 2's higher delay cost is offset by a lower price.

Region U_0: $c\mu$ rule is incentive-compatible. At arrival rates $\lambda \in U_0$, the $c\mu$ rule is incentive-compatible. (Notice that region U_0 is always nonempty.) By definition (26), if $\lambda_1 > 0$ ($\lambda_2 > 0$) the expected high priority class 1 (low priority class 2) delay is smaller (larger) than the ratio of marginal customers' value to delay cost differences, or:

$$\lambda_1 > 0 \Rightarrow v_1(\lambda_1) - c_1 W_1(\lambda, c\mu) = p_1(\lambda, c\mu) \geq v_2(\lambda_2) - c_2 W_1(\lambda, c\mu) \tag{38}$$

$$\lambda_2 > 0 \Rightarrow v_2(\lambda_2) - c_2 W_2(\lambda, c\mu) = p_2(\lambda, c\mu) \geq v_1(\lambda_1) - c_1 W_2(\lambda, c\mu). \tag{39}$$

That is, the high-priority delay is "low enough" so that the more time-sensitive marginal type 1 customer has a higher expected net value from class 1 service than the marginal type 2 customer. Similarly, the low-priority delay is "high enough" for type 2 customers to be better off with class 2 service than type 1 customers.

Region U_2: no admissible policy is incentive-compatible. At arrival rates $\lambda \in U_2$, the minimum expected delay of type 1 customers, which is attained under the $c\mu$ rule, is larger than the ratio of marginal customers' value to delay cost differences. (Region U_2 is always nonempty: the smallest class 1 expected delay is unbounded as $\lambda_1 + \lambda_2 \to 1$.) As a result, no admissible scheduling policy $r \in \mathcal{A}$ can be IC. System congestion is *relatively high* in region U_2: for any $r \in \mathcal{A}$, the more patient marginal type 2 customer ($c_2 < c_1$) has a *higher* expected net value than the marginal type 1 customer if both use class 1 service:

$$v_2(\lambda_1) - c_2 W_1(\lambda, r) > v_1(\lambda_1) - c_1 W_1(\lambda, r) \Leftrightarrow v_1(\lambda_1) - v_2(\lambda_2) < (c_1 - c_2) W_1(\lambda, r). \tag{40}$$

To highlight the contrast with region U_1, consider the maximum class 2 price that type 2 customers are willing to pay. It must not exceed the class 1 price plus a type 2 customer's expected delay cost difference between the two classes:

$$\overline{p}_2(\lambda, W) := p_1(\lambda, r) + c_2(W_1 - W_2) = v_1(\lambda_1) + (c_2 - c_1) W_1 - c_2 W_2. \tag{41}$$

Under the $c\mu$ rule the marginal type 2 customer's expected net value exceeds this bound:

$$v_2(\lambda_2) - c_2 W_2(\lambda, c\mu) > \overline{p}_2(\lambda, W(\lambda, c\mu)). \tag{42}$$

Unlike in region U_1, here the price bound $\overline{p}_2(\lambda, W)$ *decreases* in the class 1 delay W_1 (since $c_1 > c_2$): the higher the expected delay and the lower the price of class 1, the more attractive it is for the patient type 2 customers. Increasing W_1 can therefore not restore incentive-compatibility. Making class 1 less attractive to type 2 customers by reducing W_1 (and raising its price) is physically impossible since W_1 already is at its lower bound under the $c\mu$ rule.

3.4 When is Strategic Idleness Optimal?

We have so far focused on the conditionally optimal mechanism, given an arrival rate vector $\lambda \in M$. We now characterize the jointly optimal arrival rates, prices and scheduling policy, focusing on the question: when is strategic idleness optimal?

Let $\Pi(\lambda) := \Pi(\lambda, W^*(\lambda))$ be the optimal expected revenue rate at λ. By Proposition 1,

$$\Pi(\lambda) := \begin{cases} \Pi(\lambda, c\mu I) & \lambda \in U_1 \\ \Pi(\lambda, c\mu) & \lambda \in U_0 \\ 0 & \lambda \in U_2 \end{cases}, \tag{43}$$

where $\Pi(\lambda, r)$ is shorthand for $\Pi(\lambda, W(\lambda, r))$ and the revenue functions satisfy

$$\Pi(\lambda, c\mu I) = \lambda_1 \left(v_1(\lambda_1) - \frac{c_1}{1-\lambda_1} \right) + \lambda_2 \left(v_2(\lambda_2) - c_2 \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \right) \tag{44}$$

$$\Pi(\lambda, c\mu) = \lambda_1 \left(v_1(\lambda_1) - \frac{c_1}{1-\lambda_1} \right) + \lambda_2 \left(v_2(\lambda_2) - \frac{c_2}{(1-\lambda_1)(1-\lambda_1-\lambda_2)} \right). \tag{45}$$

Let λ^* be the revenue-maximizing arrival rates. Strategic idleness is optimal if and only if λ^* lies in region U_1. We offer two insights on when this occurs. For concreteness, we assume uniform value distributions

$$F_i(v) = \frac{v}{A_i}, \quad v \in [0, A_i], \ i = 1, 2, \tag{46}$$

which imply linear marginal value functions

$$v_i(\lambda_i) := A_i - B_i \cdot \lambda_i, \quad \lambda_i \in \left[0, \frac{A_i}{B_i} \right], \quad \Lambda_i = A_i / B_i.$$

Proposition 2 gives necessary and sufficient conditions for optimality of strategic idleness. Proposition 3 identifies parameter combinations that satisfy the sufficient conditions of Proposition 2. (Similar results can be obtained for other value distributions.)

Proposition 2 *Let $\lambda^{c\mu I}$ be a maximum of $\Pi(\lambda, c\mu I)$ on M. Strategic idleness is strictly optimal if and only if $\lambda^{c\mu I}$ is unique, in the interior of M and in U_1:*

$$\lambda_2^{c\mu I} > 0 \text{ and } \frac{A_1 - B_1 \cdot \lambda_1^{c\mu I} - \left(A_2 - B_2 \cdot \lambda_2^{c\mu I} \right)}{c_1 - c_2} > \frac{1}{\left(1 - \lambda_1^{c\mu I} \right) \left(1 - \lambda_1^{c\mu I} - \lambda_2^{c\mu I} \right)}. \tag{47}$$

If so, then $\lambda^{c\mu I}$ is the unique revenue-maximizing arrival rate vector:

$$\lambda^* = \arg\max_{\lambda \in M} \Pi(\lambda) = \lambda^{c\mu I} = \arg\max_{\lambda \in M} \Pi(\lambda, c\mu I). \tag{48}$$

Proposition 2 implies that one can determine whether strategic idleness is optimal by only considering the problem of maximizing $\Pi(\lambda, c\mu I)$ over U_1; there is no need to compare its solution to the maximum revenue under the $c\mu$ rule. Next, we characterize the pairs of marginal value curves and delay cost parameters that satisfy the conditions of Proposition 2. Based on the previous discussion, strategic idleness is optimal if λ^* lies in region U_1, which holds if the provider finds it optimal to (i) serve both types, but (ii) not so many customers that the expected low-priority delay under the $c\mu$ rule exceeds the ratio of the marginal customers' value to delay cost differences. Proposition 3 makes these conditions precise. The partial derivatives of $\Pi(\lambda, c\mu I)$ are

$$\Pi_{\lambda_1}(\lambda, c\mu I) = p_1(\lambda, c\mu I) + \sum_{i=1}^{2} \lambda_i \frac{\partial p_i(\lambda, c\mu I)}{\partial \lambda_1} = A_1 - 2B_1\lambda_1 - \frac{c_1}{(1-\lambda_1)^2} + \frac{\lambda_2 c_2 B_1}{c_1 - c_2} \tag{49}$$

$$\Pi_{\lambda_2}(\lambda, c\mu I) = p_2(\lambda, c\mu I) + \lambda_2 \frac{\partial p_2(\lambda, c\mu I)}{\partial \lambda_2} = \frac{(A_2 - B_2\lambda_2)c_1 - (A_1 - B_1\lambda_1)c_2}{c_1 - c_2} - \frac{\lambda_2 c_1 B_2}{c_1 - c_2}. \tag{50}$$

The function $\Pi(\lambda, c\mu I)$ need not be jointly concave. Proposition 3 identifies parameter combinations for which the FOC $\nabla\Pi(\lambda, c\mu I) = 0$ have a unique solution $\lambda^{c\mu I}$ that lies in region U_1 and is a strict maximum of $\Pi(\lambda, c\mu I)$, which implies that the sufficient conditions of Proposition 2 for optimality of strategic idleness hold and $\lambda^{c\mu I} = \lambda^*$.

Proposition 3 *Fix $A_1 > c_1 > 0$. Strategic Idleness is optimal at the revenue-maximizing arrival rate vector λ^* for the following parameter combinations that satisfy $A_1 - c_1 > A_2 - c_2$.*

1. *For (A_2, c_2) in the triangle T_1, strategic idleness is optimal for $B_1 > B_1^*$ and $B_2 > B_{21} > 0$, where*

$$T_1 = \left\{(A_2, c_2) : 0 < A_2 \le \frac{A_1}{2}; \max\left(0, A_2 - \frac{A_1 - c_1}{2}\right) < c_2 < \frac{A_2 c_1}{A_1}\right\}, \tag{51}$$

the threshold B_{21} depends on all parameters, and

$$B_1^* := \frac{A_1 c_2 - A_2 c_1}{(2c_2 - c_1)\left(1 - \sqrt{\frac{2c_2 - c_1}{2A_2 - A_1}}\right)} > 0. \tag{52}$$

2. *For (A_2, c_2) in the triangle T_2, strategic idleness is optimal for $B_1 > B_1^o$ and $B_2 > B_{22} > 0$, where*

$$T_2 = \left\{(A_2, c_2) : 0 < A_2 \le \frac{A_1 + c_1}{2}; \max\left(\frac{A_2 c_1}{A_1}, A_2 - \frac{A_1 - c_1}{2}\right) \le c_2 < \frac{2A_2 c_1}{A_1 + c_1}\right\}, \tag{53}$$

the threshold B_{22} depends on all parameters, and

$$B_1^o := \frac{A_1 c_2 - A_2 c_1}{c_2\left(1 - \sqrt{\frac{c_1 c_2}{2A_2 c_1 - A_1 c_2}}\right)} \ge 0. \tag{54}$$

3. *For (A_2, c_2) in the triangle T_3, strategic idleness is optimal for $B_1 \in (B_1^o, B_1^*)$ and $B_2 > B_{23} > 0$, where*

$$T_3 = \left\{(A_2, c_2) : \frac{A_1}{2} < A_2 < A_1; \frac{A_2 c_1}{A_1} < c_2 < \min\left(A_2 - \frac{A_1 - c_1}{2}, c_1\right)\right\} \tag{55}$$

and the threshold B_{23} depends on all parameters.

The triplets (A_i, B_i, c_i) for each of the two types represent a point in six-dimensional parameter space, but all such points can be normalized by setting $A_1 = 1$, since the maximizer $\lambda^{c\mu I}$ of $\Pi(\lambda, c\mu I)$ and the region U_1 are invariant to scaling of all parameters. Proposition 3 identifies subsets of this five-dimensional parameter space that yield optimal strategic idleness. We illustrate and discuss the result using Figure 2, which shows the triangles T_1, T_2 and T_3 for $A_1 = 1$ and $c_1 = 0.2$. Each point (A_2, c_2) represents two of the three attributes of type 2 customers, the maximum of the value distribution and the delay cost rate. For any point (A_2, c_2) within one of these triangles, there are parameters B_1 and B_2 (the slopes of the marginal value curves) such that strategic idleness is strictly optimal. Proposition 3 specifies the respective interval for B_1 analytically. The relevant interval for B_2 can be easily obtained numerically.

The area formed by the triangles T_1, T_2 and T_3 implies that for relatively small A_2, i.e., if the type 2 segment has relatively low value, strategic idleness can be optimal for relatively small c_2, i.e., type 2 customers are "not too impatient". Conversely, if the type 2 segment is relatively valuable, strategic idleness can be optimal if c_2 is large enough, i.e., type 2 customers are "sufficiently impatient." The following intuition explains this result. As noted above, for strategic idleness to be profitable, it must be optimal to serve both types, and the system congestion at the optimal arrival rates must be relatively low, so that the low-priority delay under the $c\mu$ rule is relatively small. In this case, increasing the class 2 delay establishes incentive-compatibility by making the class 2 service less attractive to type 1 customers. If A_2 is relatively small (here, for $A_2 < 0.6$), it is optimal to serve type 2 customers only if their delay cost c_2 is lower than a certain threshold. In Figure 2, this holds for (A_2, c_2) below the line L_2 (the upper edge of the triangle T_2.) For $A_2 \geq 0.6$, it is optimal to serve type 2 customers for all $c_2 < c_1$, as assumed by the model.

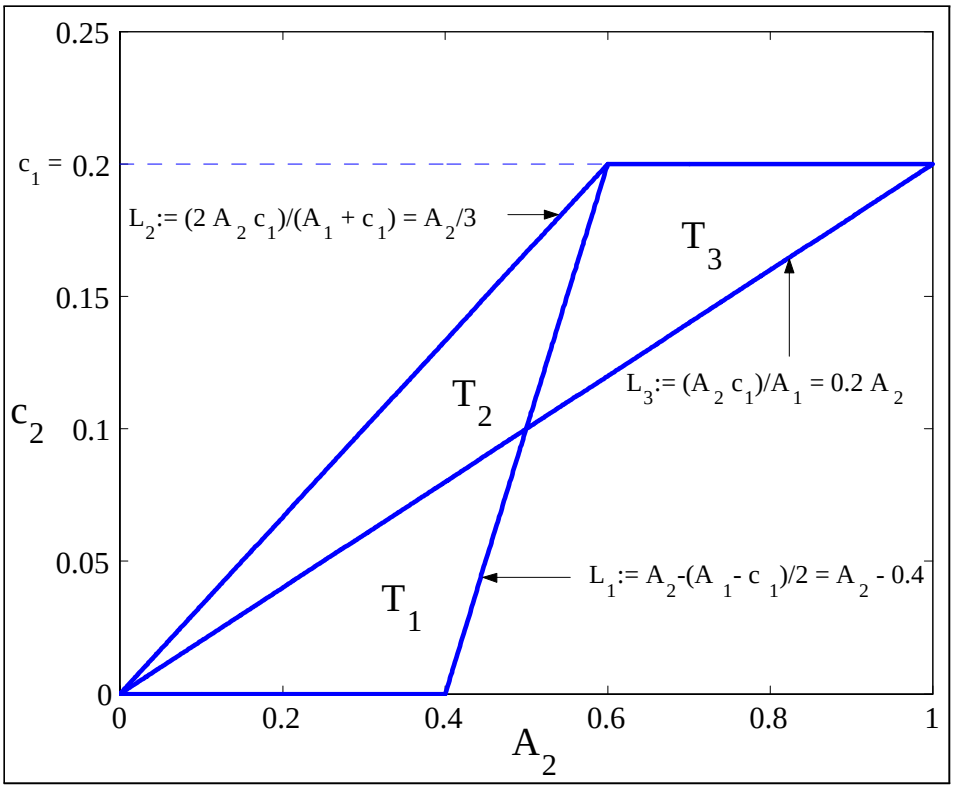

Figure 2: Illustration of Proposition 3 for $A_1 = 1$ and $c_1 = 0.2$. For (A_2, c_2) in triangles T_1, T_2 or T_3, there exist parameters B_1 and B_2 such that strategic idleness is strictly optimal.

When the type 2 segment is relatively valuable (here, for $A_2 > 0.4$), then unless the delay cost c_2 is above a certain threshold, it is optimal to serve so many of these customers that idleness is not profitable. In Figure 2, this requires (A_2, c_2) to be above the lines L_1 or L_3 (the lower edges of T_1 and T_3, respectively.)

3.5 Example: Optimal Strategic Idleness vs. $c\mu$ Rule

We illustrate our results and the value of optimal strategic idleness with the following numerical example. The parameters of type 1 customers' marginal value function are $A_1 = 100$ and $B_1 = 180$. Let $c_1 = 20$ be their delay cost per unit time. For type 2 customers, set $A_2 = 20$, $B_2 = 25$ and $c_2 = 5$. Type 1 customers have an average value of 50 and are on average five times as valuable and four times as impatient as type 2 customers. Strategic idleness is strictly optimal in this case. (By normalizing these parameters such that $A_1 = 1$, we obtain $(A_2, c_2) = (0.2, 0.05)$, which is in triangle T_2 of Figure 2. The corresponding minimum value of B_1 for optimality of strategic idleness, given by (54), is $B_1^o = 0.47$.) We compare the performance of the optimal mechanism, which uses strategic idleness, with the revenue-maximizing and incentive-compatible prices and arrival rates under the $c\mu$ rule.

Mechanism	$c\mu$ **rule**		**Strategic Idleness** ($c\mu I$)	
Metric	Type 1	Type 2	Type 1	Type 2
Arrival Rate	0.26	0.18	0.20	0.08
E[Delay]	1.36	2.46	1.26	3.02
E[Delay Cost]	27.20	12.31	25.12	15.12
Price	25.16	3.11	38.18	2.83
Price + E[Delay Cost]	52.36	15.42	63.31	17.95
E[Revenue Rate]	6.66	0.57	7.78	0.23
E[Total Revenue Rate]	7.23		8.02 (+10.8%)	

Table 1: Example with $(A_1, B_1, c_1) = (100, 180, 20)$ and $(A_2, B_2, c_2) = (20, 25, 5)$. Optimal mechanism, with strategic idleness, vs. conditionally optimal mechanism under the $c\mu$ rule.

Let λ^* denote the optimal arrival rates and $\lambda^{c\mu}$ be the revenue-maximizing rates when scheduling customers using the $c\mu$ rule. They satisfy $\lambda^* = \arg\max_{\lambda \in M} \Pi(\lambda) = \arg\max_{\lambda \in U_1} \Pi(\lambda, c\mu I)$ and $\lambda^{c\mu} = \arg\max_{\lambda \in U_0 \cup U_2} \Pi(\lambda) = \arg\max_{\lambda \in U_0} \Pi(\lambda, c\mu)$. Table 1 compares these mechanisms. The following observations are of interest. First, optimal strategic idleness yields a revenue gain of over 10%, compared to the delay cost minimizing $c\mu$ rule. The revenue gain on type 1 customers exceeds the loss on type 2 customers, who make up less than 10% of revenues under the $c\mu$ rule and about 3% of revenues under optimal strategic idleness. Second, the arrival rates of both customer types drop under strategic idleness, compared to the levels under the $c\mu$ rule. This effect is more pronounced for type 2 customers. Under the $c\mu$ rule, their arrival rate is 0.18, which amounts to about 41% of throughput. As a result of optimal strategic idleness, the type 2 arrival rate drops by over 50% to 0.08 or 28% of throughput. The total utilization drops from 44% to 28%. Third, both service classes have a higher expected total cost (price + expected delay cost) under strategic idleness. type 1 customers pay a dramatically higher price than under the $c\mu$ rule - it goes up by 13.12 or 51.8%. Their expected wait drops by only 0.1 time units, which translates to a delay cost reduction of 2.08 per customer. On balance, the expected total cost of class 1 service increases by 21%, from 52.36 to 63.31. The price for type 2 customers drops a bit under strategic idleness, compared to the $c\mu$ rule, from 3.11 to 2.83. Their expected wait increases, from 2.46 to 3.02 time units, even

though the throughput is lower for both service classes, compared to the $c\mu$ rule. The expected total cost of class 2 service increases by 16%, from 15.42 to 17.95.

It is worth highlighting the intuition for the fact that the expected class 2 delay *increases* under strategic idleness, compared to its level under the $c\mu$ rule, although the throughput is lower for both high- and low-priority customers. Under optimal strategic idleness, the class 2 delay is governed by incentives, not by system utilization. It equals the ratio of the marginal customers' value to delay cost differences and *decreases* in λ_1:

$$W_2(\lambda, c\mu I) = \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} = \frac{(A_1 - B_1\lambda_1) - (A_2 - B_2\lambda_2)}{c_1 - c_2}. \tag{56}$$

The *larger* λ_1, the larger the delay and the lower the price for high-priority (class 1) service, on balance making this class more attractive since its expected total cost equals the marginal type 1 value:

$$p_1(\lambda, c\mu I) + c_1 \cdot W_1(\lambda, c\mu) = v_1(\lambda_1) = A_1 - B_1\lambda_1. \tag{57}$$

Conversely, class 1 service becomes more expensive when λ_1 drops, requiring a higher expected class 2 delay to ensure incentive-compatibility.

3.6 Revenue Management vs. Social Optimization

It is instructive to compare our result on optimal strategic idleness under revenue management to that of Mendelson and Whang (1990) for social optimization. They show (Theorems 1 and 2) that the $c\mu$ rule is incentive-compatible under the social optimization objective of maximizing the system's expected net value rate, given by

$$NV(\lambda, r) := \sum_{i=1}^{2} \int_0^{\lambda_i} v_i(s)\, ds - \lambda_i \cdot c_i \cdot W_i(\lambda, r). \tag{58}$$

That is, there are arrival rates $\lambda^{**} \in M$ such that $NV(\lambda^{**}, c\mu) \geq NV(\lambda, r)$ for all $\lambda \in M$ and admissible policies $r \in \mathcal{A}$, and such that (10-11) and (18)-(20) are satisfied.

Their result has an intuitive geometric interpretation when viewed in our framework with regions U_0, U_1 and U_2: the socially optimal arrival rate λ^{**} *must be* in region U_0, since the $c\mu$ rule is not incentive-compatible in regions U_1 and U_2. It turns out that at each vector λ in region U_1, the marginal type 1 customer's contribution to the system net value exceeds that of the marginal type 2 customer, and vice versa for each λ in region U_2:

$$U_1 \quad : \quad W_2(\lambda, c\mu) < \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \Rightarrow NV_{\lambda_1}(\lambda, c\mu) > NV_{\lambda_2}(\lambda, c\mu) \tag{59}$$

$$U_2 \quad : \quad W_1(\lambda, c\mu) > \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2} \Rightarrow NV_{\lambda_1}(\lambda, c\mu) < NV_{\lambda_2}(\lambda, c\mu). \tag{60}$$

By contrast, the revenue function under the $c\mu$ rule is

$$\Pi(\lambda, c\mu) = \sum_{i=1}^{2} \lambda_i \cdot v_i(\lambda_i) - \lambda_i \cdot c_i \cdot W_i(\lambda_i, c\mu). \tag{61}$$

In cases where strategic idleness is optimal, the maximum of $\Pi(\lambda, c\mu)$ *cannot* be in U_0.

To establish (59-60), use the conservation law (15) to write the net value function as

$$NV(\lambda, c\mu) = \sum_{i=1}^{2} \int_0^{\lambda_i} v_i(s)\, ds - (c_1 - c_2) \cdot N_1(\lambda, c\mu) - c_2 \cdot N(\lambda, c\mu), \tag{62}$$

where $N_1(\lambda, c\mu) = \lambda_1 W_1(\lambda, c\mu)$ is the average inventory of class 1 customers under the $c\mu$ rule, and $N(\lambda, c\mu) = N_1(\lambda, c\mu) + N_2(\lambda, c\mu)$ is the total customer inventory in the system. A marginal increase in the class i arrival rate has a positive value effect and a negative delay cost effect. Increasing either arrival rate equally changes the total system inventory $N(\lambda, c\mu)$, but since type 1 customers have preemptive priority, the inventory of high-priority customers $N_1(\lambda, c\mu)$ increases only in λ_1. Therefore, the delay cost effect of a type 1 exceeds that of a type 2 customer and the difference only depends on the marginal change in $N_1(\lambda, c\mu)$. It is bounded by the class 1 and class 2 delays:

$$W_1(\lambda, c\mu) \leq \frac{\partial N_1(\lambda, c\mu)}{\partial \lambda_1} = \frac{1}{(1-\lambda_1)^2} \leq W_2(\lambda, c\mu), \tag{63}$$

and increases in the type 1 rate. Therefore the partial derivatives of $NV(\lambda, c\mu)$ satisfy

$$NV_{\lambda_1}(\lambda, c\mu) - NV_{\lambda_2}(\lambda, c\mu) = v_1(\lambda_1) - v_2(\lambda_2) - \frac{c_1 - c_2}{(1-\lambda_1)^2}, \tag{64}$$

which implies (59-60). Intuitively, congestion is low for $\lambda \in U_1$, and the marginal type 1 customer generates more *net* value than the marginal type 2 customer since its higher value offsets its larger delay cost impact. Conversely, at $\lambda \in U_2$ the system is relatively congested and a less time-sensitive type 2 customer is more valuable at the margin under the $c\mu$ rule.

4 Heterogeneous Service Times

We have so far assumed identical mean service times for all customers. In this Section we generalize our analysis to allow for different service time distributions across customer types. Let the (exponentially distributed) random variable $\widetilde{s}_i$ denote the service time of type i customers and let μ_i^{-1} be its mean. Recall: $c_1\mu_1 > c_2\mu_2$ (so type 1 get priority over type 2 customers under the $c\mu$ rule and vice versa under the $Rc\mu$ rule), neither customers nor the provider know actual service times ex ante (they become known only upon order completion), and the provider cannot distinguish among type 1 and -2 customers. Notice that any service time realization $t \in [0, \infty)$ could be a draw from either service time distribution.

In addition to showing that strategic idleness may still be optimal in this setting, we obtain a new insight: it may be optimal under certain conditions to *increase the type 1 delay and to reduce the type 2 delay, compared to the $c\mu$ rule.* This *reduces* (or reverses) the delay differentiation between the two types, compared to the $c\mu$ rule. By contrast, strategic idleness *increases* this delay differentiation.

Our plan is similar to that of Section 3. First, we adapt the class of admissible mechanisms to this setting. Second, we characterize incentive-compatible mechanisms. Third, we identify three distinct cases that yield structurally different results for the conditionally optimal scheduling policies at fixed λ. Finally, we use an example to illustrate the case in which it is optimal to serve customers in the reverse $c\mu$ order, prioritizing type 2 over type 1 customers, which *maximizes* the delay cost rate among all work conserving policies.

Admissible Mechanisms. It follows from an argument similar to the one in Lemma 2 that there is no loss of generality in only considering mechanisms with up to two distinct service classes. We use the same definition of admissible mechanisms (cf. Sections 2, 3.2), except that we limit our attention to nonpreemptive scheduling. Let $\mathcal{A}^{NP}$ denote this class of admissible policies, which includes all nonpreemptive scheduling policies $r \in \mathcal{A}$. Write $W_{k,i}(\lambda, r)$ for the expected delay of a type i customer who purchases class k service under policy r and arrival rates λ, and let $W(\lambda, r) := (W_{i,i}(\lambda, r))_{i=1,2}$. The restriction to nonpreemptive policies simplifies the analysis

(without sacrificing insight): it implies that the difference between the expected delays of both customer types in the same service class equals the difference in their mean service times:

$$W_{j,i}(\lambda, r) - W_{j,j}(\lambda, r) = \frac{1}{\mu_i} - \frac{1}{\mu_j} \text{ for } i \neq j,\ r \in \mathcal{A}^{NP}. \tag{65}$$

With different mean service times, the set of feasible arrival rates is defined as

$$M := \{\lambda : 0 \leq \lambda_i < \Lambda_i,\ \lambda_1/\mu_1 + \lambda_2/\mu_2 < 1\}. \tag{66}$$

Let $\rho_i := \lambda_i/\mu_i$ be the utilization due to type i customers. We state the following Lemmas, which are natural extensions of their counterparts in Section 3, without proof. Lemma 8 is the analogue of Lemma 5 and characterizes the set of achievable expected delay vectors:

Lemma 8 *Fix $\lambda \in M$. Under an admissible nonpreemptive scheduling policy $r \in \mathcal{A}^{NP}$, the expected steady state delay of all admitted customers satisfies:*

$$\rho_1 \cdot W_{1,1}(\lambda, r) + \rho_2 \cdot W_{2,2}(\lambda, r) \geq \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_1 - \rho_2}. \tag{67}$$

The expected steady state delays of individual service classes are bounded below as follows:

$$W_{1,1}(\lambda, c\mu) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_1} + \frac{1}{\mu_1} \leq W_{1,1}(\lambda, r) \tag{68}$$

$$W_{2,2}(\lambda, Rc\mu) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_2} + \frac{1}{\mu_2} \leq W_{2,2}(\lambda, r). \tag{69}$$

In this section, $r = c\mu$ and $r = Rc\mu$ denote the nonpreemptive versions of the $c\mu$ rule and the reverse $c\mu$ rule, respectively. For $\lambda \in M$, denote by $D^{NP}(\lambda)$ the set of expected delay vectors $W = (W_1, W_2)$ that satisfy (67)-(69). Lemma 9 is the analogue of Lemma 6 and states that the family of *nonpreemptive* (α, d) scheduling policies with randomized static priority assignment and inserted job idleness (cf. Definition 4), spans the set $D^{NP}(\lambda)$.

Lemma 9 *For $\lambda \in M$ and $W \in D^{NP}(\lambda)$, there is a nonpreemptive (α, d)-policy with $\alpha \in [0, 1]$ and $d \geq 0$ such that if type i uses class i service, then:*

$$W_1 = W_{1,1}(\lambda, (\alpha, d)) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \alpha\rho_1 - (1-\alpha)\rho_2} \left(\alpha + \frac{1-\alpha}{1 - \rho_1 - \rho_2}\right) + \frac{1}{\mu_1} + d_1 \tag{70}$$

$$W_2 = W_{2,2}(\lambda, (\alpha, d)) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \alpha\rho_1 - (1-\alpha)\rho_2} \left(1 - \alpha + \frac{\alpha}{1 - \rho_1 - \rho_2}\right) + \frac{1}{\mu_2} + d_2. \tag{71}$$

Since different types expect different service times, the provider may gain from charging based on both service class and realized service times. Let $P_k : \mathbb{R}_+ \to \mathbb{R}$ be the class k price as a function of service time, let $P = (P_1, P_2)$, and denote by $p_{k,i} = E[P_k(\widetilde{s}_i)]$ the expected payment of a type i customer for class k service. Define $\mathcal{P}$ to be the special class of price functions that are (weakly) increasing in service times. That is, $P \in \mathcal{P}$ if and only if $P_i(t_1) \leq P_i(t_2) \Leftrightarrow t_1 \leq t_2$ for $i = 1, 2$ and $0 \leq t_1 \leq t_2 < \infty$. Of course, $\mathcal{P}$ includes the special case of service time independent prices. We refer to members of $\mathcal{P}$ as $CDIT$ (class-dependent and increasing in time) price functions. Which scheduling policies are IC for given λ depends on whether we restrict attention to $\mathcal{P}$.

4.1 Incentive-Compatible Mechanisms

An admissible mechanism (λ, P, r) is incentive-compatible (IC) at $\lambda \in M$ if and only if type i customers' expected payment for class i service satisfies the inverse demand relationship

$$p_{i,i} = p_i(\lambda, r) := v_i(\lambda_i) - c_i \cdot W_{i,i}(\lambda, r), \quad i = 1, 2, \tag{72}$$

and if no type k customer has an incentive to purchase service in class $i \neq k$:

$$p_{i,k} \geq v_k(\lambda_k) - c_k \cdot W_{k,i}(\lambda, r) \text{ if } \lambda_i > 0. \tag{73}$$

Lemma 10 *An admissible mechanism (λ, P, r) with $\lambda \in M$ satisfies (72-73) as follows:*

1. *Absent restrictions on the price functions $P : \mathbb{R}^2_+ \to \mathbb{R}^2$, all $r \in \mathcal{A}^{NP}$ are IC at λ.*
2. *For CDIT price functions, i.e., for $P \in \mathcal{P}$:*

   (a) *For higher type 1 mean service time ($\mu_1^{-1} > \mu_2^{-1}$), policy $r \in \mathcal{A}^{NP}$ is IC at λ iff*

   $$(c_1 - c_2) W_{1,1}(\lambda, r) \leq v_1(\lambda_1) - v_2(\lambda_2) + c_2\left(\mu_2^{-1} - \mu_1^{-1}\right) \text{ for } \lambda_1 > 0. \tag{74}$$

   In this case, IC holds for every price function $P \in \mathcal{P}$ that satisfies (72) and

   $$p_{1,2} \in [v_2(\lambda_2) - c_2 W_{1,2}(\lambda, r), p_{1,1}] \text{ if } \lambda_1 > 0, \tag{75}$$
   $$p_{2,1} \geq \max(v_1(\lambda_1) - c_1 W_{2,1}(\lambda, r), p_{2,2}) \text{ if } \lambda_2 > 0. \tag{76}$$

   (b) *For higher type 2 mean service time ($\mu_1^{-1} \leq \mu_2^{-1}$), policy $r \in \mathcal{A}^{NP}$ is IC at λ iff*

   $$(c_1 - c_2) W_{2,2}(\lambda, r) \geq v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right) \text{ for } \lambda_2 > 0. \tag{77}$$

   In this case, IC holds for every price function $P \in \mathcal{P}$ that satisfies (72) and

   $$p_{1,2} \geq \max(v_2(\lambda_2) - c_2 W_{1,2}(\lambda, r), p_{1,1}) \text{ if } \lambda_1 > 0, \tag{78}$$
   $$p_{2,1} \in [v_1(\lambda_1) - c_1 W_{2,1}(\lambda, r), p_{2,2}] \text{ if } \lambda_2 > 0. \tag{79}$$

PROOF: 1. There always exist functions $P : \mathbb{R}^2_+ \to \mathbb{R}^2$ that satisfy (72)-(73).
2. However, it may be that (72)-(73) only hold for price functions that are not $CDIT$, i.e., for $P \notin \mathcal{P}$. Consider case (a): It follows from $\mu_1^{-1} > \mu_2^{-1}$ that $\widetilde{s}_1$ is stochastically larger than $\widetilde{s}_2$. This implies that $E[f(\widetilde{s}_1)] \geq E[f(\widetilde{s}_2)]$ for all increasing functions $f : \mathbb{R} \to \mathbb{R}$ for which the expectations exist (cf. Shaked and Shanthikumar 1994). Hence, if $p_{i,1} = E[P_i(\widetilde{s}_1)] < E[P_i(\widetilde{s}_2)] = p_{i,2}$, then $P \notin \mathcal{P}$. For $P \in \mathcal{P}$, a type 2 customer can be prevented from purchasing class 1 service if and only if conditions (72)-(73) hold for $(i, k) = (1, 2)$ and

$$v_1(\lambda_1) - c_1 \cdot W_{1,1}(\lambda, r) = p_{1,1} \geq p_{1,2} \geq v_2(\lambda_2) - c_2 \cdot W_{1,2}(\lambda, r) \text{ if } \lambda_1 > 0, \tag{80}$$

i.e., the expected net value of the marginal type 1 customer exceeds that of the marginal type 2 customer if both use class 1 service. Noting that $W_{1,2}(\lambda, r) = W_{1,1}(\lambda, r) + \mu_2^{-1} - \mu_1^{-1}$ and rearranging (80) yields (74) and (75). By contrast, for every scheduling policy, there is a $CDIT$ price function that prevents type 1 customers from using class 2 service: simply choose an increasing function $P_2(t)$ that satisfies (77). The proof of (b) is similar.□

For simplicity, we henceforth limit our attention to $CDIT$ price functions. It is also quite natural to charge customers that require more service a higher price. In this case, the IC conditions

imply price-independent bounds on the expected delays that must be satisfied under any admissible scheduling policy $r \in \mathcal{A}^{NP}$. However, unlike in the case of homogeneous service times, here there is only a *single* bound in each case, (74) for $\mu_1^{-1} > \mu_2^{-1}$ and (77) for $\mu_1^{-1} \leq \mu_2^{-1}$: the provider's ability to charge based on actual service times eliminates the other bound.

Using the inverse demand relationships (72), the control problem formulation is:

$$\max_{\lambda \in M,\ r \in \mathcal{A}^{NP}} \Pi(\lambda, r) = \sum_{i=1}^{2} \lambda_i \cdot p_i(\lambda, r) = \sum_{i=1}^{2} \lambda_i \cdot (v_i(\lambda_i) - c_i \cdot W_{i,i}(\lambda, r)), \tag{81}$$

subject to the respective expected delay bound implied by the IC constraints: (74) if $\mu_1^{-1} > \mu_2^{-1}$ or (77) if $\mu_1^{-1} \leq \mu_2^{-1}$. Observe that if the expected delays satisfy (74) (or (77), respectively), there are infinitely many $CDIT$ price functions that satisfy (75-76) (or (78-79), resp.). Following Section 3, we transform (81) into the optimization problem:

$$\max_{\lambda \in M,\ W \in D^{NP}(\lambda)} \Pi(\lambda, W) = \sum_{i=1}^{2} \lambda_i \cdot (v_i(\lambda_i) - c_i \cdot W_i), \tag{82}$$

subject to (74) if $\mu_1^{-1} > \mu_2^{-1}$ or (77) if $\mu_1^{-1} \leq \mu_2^{-1}$.

4.2 Conditionally Optimal Delay Differentiation at Fixed λ

We now identify the expected delays and scheduling rules that are conditionally optimal for fixed $\lambda \in M$. Lemma 11 is the analogue of Lemma 7 for nonpreemptive scheduling policies.

Lemma 11 *Fix $\lambda \in M$. For $c_1\mu_1 > c_2\mu_2 \geq 0$, the nonpreemptive $c\mu$ rule minimizes the expected aggregate delay cost rate over all admissible nonpreemptive scheduling policies:*

$$\rho_1 \cdot c_1 \cdot W_{1,1}(\lambda, c\mu) + \rho_2 \cdot c_2 \cdot W_{2,2}(\lambda, c\mu) \leq \rho_1 \cdot c_1 \cdot W_1(\lambda, r) + \rho_2 \cdot c_2 \cdot W_2(\lambda, r),\ r \in \mathcal{A}^{NP}. \tag{83}$$

By Lemma 11, the $c\mu$ rule maximizes the revenue rate (82) for any $\lambda \in M$, but it need not be optimal since it is not IC for λ in certain subsets of M. We show that Lemma 10 gives rise to different solutions in each of the three possible cases with $c_1\mu_1 > c_2\mu_2$. As in Section 3, we partition in each case the arrival rate set M into certain regions by applying the bounds of Lemma 10 to the expected delays under the $c\mu$ rule. Using Lemmas 8 and 9, we identify the optimal expected delays and scheduling policies for each region. Proposition 4 summarizes the optimal expected delays $W^*(\lambda)$ and corresponding scheduling policies $r^*(\lambda)$ (where $W(\lambda, r^*(\lambda)) = W^*(\lambda)$).

Proposition 4 *Optimal expected delays and scheduling under $CDIT$ pricing at $\lambda \in M$.*

1. *Type 1 more impatient and require more service than type 2: $\frac{c_1}{c_2} > \frac{\mu_2}{\mu_1} > 1$. Define*

$$U_{c\mu} = \left\{ \lambda \in M : \lambda_1 = 0 \text{ or } W_{1,1}(\lambda, c\mu) \leq \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_2\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2} \right\}$$
$$U_{No} = M \backslash U_{c\mu}.$$

   (a) *If $\lambda \in U_{c\mu}$, then the $c\mu$ rule is optimal and $W(\lambda, c\mu) = W^*(\lambda)$.*

   (b) *If $\lambda \in U_{No}$, then no admissible nonpreemptive policy $r \in \mathcal{A}^{NP}$ is IC.*

2. *Type* 1 *more impatient but require equal/less service than type* 2: $\frac{c_1}{c_2} > 1 \geq \frac{\mu_2}{\mu_1}$. *Define*

$$U_{c\mu} = \left\{ \lambda \in M : \lambda_2 = 0 \text{ or } W_{2,2}(\lambda, c\mu) \geq \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2} \right\}$$
$$U_{c\mu I} = M \backslash U_{c\mu}.$$

(a) *If* $\lambda \in U_{c\mu}$, *then the* $c\mu$ *rule is optimal and* $W(\lambda, c\mu) = W^*(\lambda)$.

(b) *If* $\lambda \in U_{c\mu I}$, *then no work conserving nonpreemptive policy is IC. A* $c\mu I$ *policy, which strategically idles type* 2 *jobs, attains the optimal expected delays*

$$W_1^*(\lambda) = W_{1,1}(\lambda, c\mu I) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1-\rho_1} + \frac{1}{\mu_1} = W_{1,1}(\lambda, c\mu) \quad (84)$$

$$W_2^*(\lambda) = W_{2,2}(\lambda, c\mu I) = \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2}. \quad (85)$$

3. *Type* 1 *less impatient and require less service than type* 2: $1 > \frac{c_1}{c_2} > \frac{\mu_2}{\mu_1}$. *Define*

$$U_{c\mu} = \left\{ \lambda \in M : \lambda_2 = 0 \text{ or } W_{2,2}(\lambda, c\mu) \leq \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2} \right\}$$
$$U_{c\mu - Rc\mu} = \left\{ \lambda \in M : \lambda_2 > 0, W_{2,2}(\lambda, Rc\mu) \leq \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2} < W_{2,2}(\lambda, c\mu) \right\}$$
$$U_{No} = M \backslash \left\{ U_{c\mu} \cup U_{c\mu - Rc\mu} \right\}.$$

(a) *If* $\lambda \in U_{c\mu}$, *then the* $c\mu$ *rule is optimal and* $W(\lambda, c\mu) = W^*(\lambda)$.

(b) *If* $\lambda \in U_{c\mu - Rc\mu}$ *then the optimal expected delays satisfy*

$$W_1^*(\lambda) = W_{1,1}(\lambda, (\alpha^*(\lambda), d^*)) = \frac{1}{\rho_1}\left(\frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1-\rho_1-\rho_2} - \rho_2 \cdot W_2^*(\lambda) \right) \quad (86)$$

$$W_2^*(\lambda) = W_{2,2}(\lambda, (\alpha^*(\lambda), d^*)) = \frac{v_1(\lambda_1) - v_2(\lambda_2) + c_1\left(\mu_2^{-1} - \mu_1^{-1}\right)}{c_1 - c_2}, \quad (87)$$

and are attained by a work conserving (α, d) *policy without job idleness* ($d^* = 0$) *and that gives static nonpreemptive priority to type 1 (type 2) customers with probability* $\alpha^*(\lambda)$ ($1 - \alpha^*(\lambda)$), *where* $\alpha^*(\lambda) \in [0, 1)$ *is such that* $W_{i,i}(\lambda, (\alpha^*(\lambda), 0)) = W_i^*(\lambda)$ *for* $i = 1, 2$. *The optimal expected delay of type 1 (type 2) customers is higher (lower) than under under the* $c\mu$ *rule:* $W_{1,1}(\lambda, c\mu) < W_1^*(\lambda)$ *and* $W_2^*(\lambda) < W_{2,2}(\lambda, c\mu)$. *If* $\alpha^*(\lambda) = 0$, *then the nonpreemptive reverse* $c\mu$ *rule is optimal.*

(c) *If* $\lambda \in U_{No}$, *then no admissible nonpreemptive policy* $r \in \mathcal{A}^{NP}$ *is IC.*

Proposition 4 gives a framework for the revenue-maximizing and incentive-compatible level of delay differentiation between the two service classes. The intuition for the three cases is as follows. If type 1 are more impatient and have a higher mean service time than type 2 customers (case 1), they can be prevented from purchasing class 2 service under any scheduling policy by using a $CDIT$ price function P_2 with a prohibitive penalty for long processing times. By contrast, such a price penalty is ineffective in preventing type 2 customers from purchasing class 1 service. Since these customers are relatively patient, class 1 service is unattractive to them if and only if its expected

delay is "sufficiently low" (i.e., below a threshold equal to the ratio of marginal value to delay cost differences, as with equal mean service times, plus a term that adjusts for the different mean service times) and its price correspondingly high. Among all admissible nonpreemptive policies $r \in \mathcal{A}^{NP}$, the $c\mu$ rule minimizes the expected class 1 delay (Lemma 8) and the average delay cost rate (Lemma 11): it is therefore optimal if it is IC, and no policy is IC otherwise.

If type 1 customers have an equal or smaller mean service time than type 2 customers (cases 2 and 3), then $CDIT$ pricing is effective in preventing type 2 customers from purchasing class 1 service under any scheduling policy. However, keeping type 1 customers from buying class 2 service does require a scheduling policy that yields the "right" mean delay level for class 2 service. Specifically, if type 1 are more impatient than type 2 customers (case 2), the expected class 2 delay must be "sufficiently high", i.e., it must exceed the critical ratio shown in the definition of $U_{c\mu}$. If it is too low under the $c\mu$ rule (at $\lambda \in U_{c\mu I}$), inserting strategic idleness to delay class 2 customers is optimal, as discussed in Section 3. Conversely, if type 1 are less time-sensitive and require on average less service than type 2 customers (case 3), their expected delay cost from class 2 service is smaller than that of type 2 customers. It follows that class 2 service is attractive to type 1 customers unless the expected class 2 delay is below a threshold (the critical ratio that defines region $U_{c\mu}$). The optimal expected class 2 delay $W_2^*(\lambda)$ therefore depends on two opposite factors: (i) incentive-compatibility requires that W_2 not exceed the critical ratio, but (ii) the $c\mu$ rule maximizes the revenue rate (82) (and minimizes the delay cost rate) by maximizing W_2 among all work conserving policies. This gives rise to the three regions of case 3. The $c\mu$ rule is optimal if it yields a low enough W_2 (for $\lambda \in U_{c\mu}$). Otherwise, the scheduling policy must be altered to *reduce* the delay and raise the price of class 2, which increases the class 1 delay by the conservation law and lowers its price, making class 2 unattractive to type 1 customers. An IC scheduling policy exists at $\lambda \in M$ if and only if the expected class 2 delay under the $Rc\mu$ rule (which prioritizes class 2 customers) is sufficiently low in the sense discussed above, since the $Rc\mu$ rule minimizes the expected class 2 delay among admissible nonpreemptive policies $r \in \mathcal{A}^{NP}$. This holds for $\lambda \in U_{c\mu-Rc\mu}$, where the critical ratio lies between the expected class 2 delays under the $Rc\mu$ and the $c\mu$ rules. In this case, the optimal expected class 2 delay equals the critical ratio and is attained by a workconserving policy with randomized priority assignment as stated in Proposition 4. At the extreme, it may be optimal to prioritize customers in the reverse $c\mu$ order (see Section 4.3). For λ outside $U_{c\mu}$ and $U_{c\mu-Rc\mu}$, the class 2 delay is too high under all admissible policies.

In summary, maximizing revenues from customers with private information about their attributes requires carefully choosing the delay differentiation between the service classes. The optimal level of delay differentiation systematically depends on customer attributes and the system structure and congestion. While the $c\mu$ rule is optimal when the provider can tell customer types apart, when she cannot, the revenue-maximizing and incentive-compatible scheduling policy may increase or decrease the delay differentiation, compared to the $c\mu$ rule.

4.3 Example: Optimality of Reverse $c\mu$ Rule

Proposition 4 defines the maximum expected revenue rate $\Pi(\lambda) := \Pi(\lambda, W^*(\lambda))$ as a function of λ. Which scheduling policy is optimal at the revenue-maximizing arrival rates λ^* depends on the specific distributions of customer attributes. This Section presents an example where case 3 of Proposition 4 applies at λ^*. It illustrates the extreme case in which it is optimal to serve customers in the reverse $c\mu$ order. Consider linear marginal value curves $v_i(\lambda_i) = A_i - \lambda_i B_i$, where $(A_1, B_1) = (28.2, 20)$ and $(A_2, B_2) = (21.7, 0.1)$. The type 1 value distribution has a significantly higher maximum and variance than that of type 2 customers. Let $c_1 = \mu_1^{-1} = 1$, $c_2 = 1.5$ and $\mu_2^{-1} = 1.6$: type 2 are 50% more time-sensitive and have a 60% higher mean service time than type

1 customers (and $c_1\mu_1 > c_2\mu_2$). The revenue function is

$$\Pi(\lambda) = \sum_{i=1}^{2} \lambda_i \left(v_i(\lambda_i) - \lambda_1 c_1 W_i^*(\lambda) \right), \text{ for } \lambda \in U_{c\mu} \cup U_{c\mu-Rc\mu}, \tag{88}$$

and $\Pi(\lambda) = 0$ for $\lambda \in U_{No}$, where $W^*(\lambda)$ is defined by case 3 of Proposition 4. Figure 3 shows the three regions $U_{c\mu}$, $U_{c\mu-Rc\mu}$ and U_{No} that partition M. Observe that the $c\mu$ rule is only IC for relatively high type 1 and low type 2 arrival rates (see the boundary between $U_{c\mu}$ and $U_{c\mu-Rc\mu}$ in Fig. 3). At low arrival rates, the marginal type 1 value is much higher than the marginal type 2 value. Hence, type 1 customers have an incentive to purchase class 2 service under all scheduling policies (see region U_{No}). For sufficiently high λ_1 and low λ_2 the marginal values of both types are at comparable levels, and the expected class 2 delay under the $c\mu$ rule is low enough to prevent type 1 customers from buying class 2 service. The optimal arrival rate, conditional on scheduling based on the $c\mu$ rule, is $\lambda^*_{c\mu} = (0.507, 0.097)$, where type 1 customers are just indifferent between the two service classes. However, since the marginal type 2 value hardly declines in λ_2, it is profitable to serve more of them. To do so and maintain IC, it is necessary to speed up class 2 (and slow down class 1) service while increasing λ_2, relative to the delay levels under the $c\mu$ rule. The optimal arrival rate vector $\lambda^* = (0.422, 0.166)$ is on the boundary between $U_{c\mu-Rc\mu}$ and U_{No}, where the $Rc\mu$ rule is optimal: type 2 get priority over type 1 customers. Compared to the $c\mu$ rule, it is optimal to serve more type 2 and fewer type 1 customers and to reverse the delay differentiation by delaying type 1 more than type 2 customers. As a result, the provider increases her type 2 and loses some type 1 revenue, increasing her total revenue.

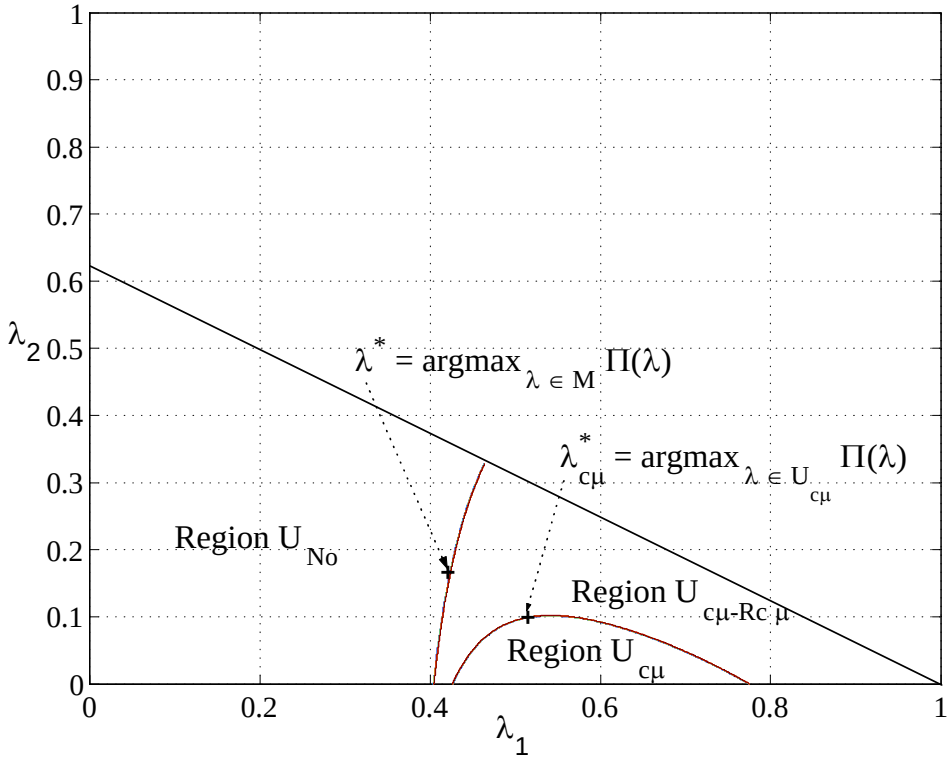

Figure 3: The reverse $c\mu$ rule is optimal for linear marginal value curves with $\left(A_1, B_1, c_1, \mu_1^{-1}\right) = (28.2, 20, 1, 1)$ and $\left(A_2, B_2, c_2, \mu_2^{-1}\right) = (21.7, 0.1, 1.5, 1.6)$. The $c\mu$ rule yields a higher type 1 and lower type 2 arrival rate.

Mechanism	$c\mu$ **rule**		**Reverse** $c\mu$ **rule**	
Metric	Type 1	Type 2	Type 1	Type 2
Arrival Rate	0.507	0.097	0.422	0.166
E[Delay]	2.531	6.127	4.650	2.743
E[Delay Cost]	2.531	9.191	4.650	4.114
Price	15.502	12.506	15.024	17.576
Price + E[Delay Cost]	18.033	21.697	19.674	21.690
E[Revenue Rate]	7.863	1.206	6.390	2.873
E[Total Revenue Rate]	9.071		9.263 (+2.1%)	

Table 2: Example with $\left(A_1, B_1, c_1, \mu_1^{-1}\right) = (28.2, 20, 1, 1)$ and $\left(A_2, B_2, c_2, \mu_2^{-1}\right) = (21.7, 0.1, 1.5, 1.6)$. Optimal mechanism, wich schedules in the reverse $c\mu$ order, and conditionally optimal mechanism under the $c\mu$ rule.

5 Concluding Remarks

In this paper we have studied the design of revenue-maximizing and incentive-compatible price scheduling mechanisms for a capacity-constrained firm that serves heterogeneous time-sensitive customers with private information on their willingness to pay, time-sensitivity and service requirement. We draw the following conclusions. First, the familiar $c\mu$ priority rule, known to minimize the system-wide expected delay cost and to be incentive-compatible under social optimization, need not be optimal in this setting. This suggests a more general guideline: in designing incentive-compatible and revenue-maximizing scheduling policies, delay cost-minimization, which plays a prominent role in controlling and pricing queueing systems, should not be the dominant criterion ex ante. Second, this principle is illustrated by optimal scheduling policies with novel features that depend on customer attributes. When one market segment is more time-sensitive but has an equal or smaller service time than the other, it may be optimal to insert strategic idleness, and we have shown for linear demand when this course of action is advised. However, in cases where the customer type that is prioritized according to the $c\mu$ rule is both less time-sensitive and requires less processing, it may be optimal to appropriately randomize priority assignments, or even to serve customers in the reverse $c\mu$ order, which maximizes the system delay cost among all work conserving policies. Compared to the $c\mu$ rule, these optimal policies increase, decrease or reverse the delay differentiation between customers types. The optimal level of delay differentiation systematically emerges from a trade-off between operational constraints and customer incentives. Third, our stepwise solution approach can be adapted for designing revenue-maximizing and incentive-compatible mechanisms in systems with different customer attributes or operational properties.

6 Appendix: Proofs

6.1 Proof of Lemma 1

The IC constraint (6) for $i = j \in \{1, 2\}$ is equivalent to

$$u_i\left(v|a,p,r\right) \geq u_i\left(x|a,p,r\right) + \left(v - x\right) a_i\left(x\right) \text{ for } v \neq x \in V_i. \tag{89}$$

In equilibrium, the expected utility of a type i customer with value v who truthfully reports her type must exceed her expected utility from declaring value $x \neq v$, which equals the expected utility of such a type i customer, plus the difference between the expected net values of these customers if *both* declare a value x. Thus, the expected utilities of type i customers are increasing in their

values. Switching the roles of v and x yields

$$(v-x)\,a_i(v) \geq u_i(v|a,p,r) - u_i(x|a,p,r) \geq (v-x)\,a_i(x) \text{ for } v \neq x \in V_i,\ i=1,2, \tag{90}$$

which implies that $a_i(v) \geq a_i(x)$ if $v > x$. It follows from (1) and $\lambda_i < \Lambda_i$ that there is a unique marginal value $v_i(\lambda_i) > \underline{v}_i$ as stated by 1. Using (90) with $v = v_i(\lambda_i)$ and $x < v_i(\lambda_i)$ implies that the marginal customer has zero expected utility: $u_i(v_i(\lambda_i)|a,p,r) = v_i(\lambda_i) - p_i(v_i(\lambda_i)) - c_i W_i(v_i(\lambda_i)|a,r) = 0$. Applying (90) with $v = v_i(\lambda_i)$ and $x > v_i(\lambda_i)$ implies (8) and (9). The IC constraint (6) also requires that no type i customer have an incentive to declare (j, x_j) with $j \neq i$ and $x_j \in V_j$ if $\lambda_2 > 0$. This holds for type 1 if

$$v_1(\lambda_1) \leq c_1 W_2(x_2|a,r) + p_2(x_2), \quad \text{for } x_2 \geq v_2(\lambda_2), x_2 \in V_2, \tag{91}$$

where the RHS is the cost of a type 1 customer who declares type $(2, x_2)$. Noting that

$$p_2(x_2) = v_2(\lambda_2) - c_2 W_2(x_2|a,r) \text{ for } x_2 \geq v_2(\lambda_2) \tag{92}$$

and combining with (91) yields the lower bound (11) on the expected delays of type 2 customers. The same argument for type 2 customers yields (10).□

6.2 Proof of Lemma 2

By Lemma 1 and (7), the expected revenue rate given mechanism (a,p,r) is

$$\Pi(a,p,r) = \sum_{i=1}^{2} \lambda_i v_i(\lambda_i) - \sum_{i=1}^{2} \Lambda_i c_i \int_{v \in V_i} a_i(v)\, W_i(v|a,r)\, f_i(v)\, dv. \tag{93}$$

Two admissible mechanisms (a,p,r) and (a,p',r') that yield the same average mean delay for type i customers are revenue equivalent, i.e., they generate the same revenue rate. If r is an admissible policy and r' randomly assigns admitted type i customers to service class (i,v) with probability density $\Lambda_i f_i(v)/\lambda_i$ and then serves them as specified by r, then r' is also admissible and yields expected delays $W_i(\lambda, r')$ as weighted averages specified by (12). Since the (10-11) hold for policy r, they also hold for r': giving equal expected delays to all type i customers minimizes (maximizes) the maximum (minimum) over type 1 (type 2) customers' expected delays. The prices in (13) follow from (8) and (12).□

6.3 Proof of Lemma 4

Under the preemptive $c\mu$ rule, class 1 customers are neither slowed down by class 2 customers nor by server idleness, and on every sample path their delay must be smaller than or equal to that under any other rule r, yielding the lower bound in (16). The upper bound in (17) follows from (15) and (16), where the formulae for $W_1(\lambda, c\mu)$ and $W_2(\lambda, c\mu)$ are standard. Applying the same argument to the $Rc\mu$ rule establishes the upper bound in (16) and lower bound in (17).□

6.4 Proof of Proposition 1

For fixed $\lambda \in M$, problem (23) is equivalent to delay cost minimization:

$$\min_{W} \sum_{i=1}^{2} \lambda_i \cdot c_i \cdot W_i, \tag{94}$$

subject to the incentive constraints (10-11) and the operational constraints (18)-(20).

1. By definition (25), the $c\mu$ rule is not incentive-compatible at $\lambda \in U_1$ since the low priority (class 2) expected delay violates (11) of Lemma 1: it is smaller than the lower bound imposed by the IC constraints. Together with (17) of Lemma 4, this implies that no admissible work conserving scheduling rule $r \in \overline{\mathcal{A}}$ is incentive-compatible at $\lambda \in U_1$:

$$W_2(\lambda, r) \leq W_2(\lambda, c\mu) < \frac{v_1(\lambda_1) - v_2(\lambda_2)}{c_1 - c_2}. \tag{95}$$

To restore incentive-compatibility, i.e., satisfy the lower bound in (11), the class 2 expected delay must be increased up to the point where it equals the ratio in (29). Lemma 4 implies that this requires job idleness. By (19), the expected class 1 delay under the $c\mu$ rule attains the lower bound among all $r \in \mathcal{A}$, so the expected delays (28-29) are the solution of (94).

2. By definition (26), the $c\mu$ rule is incentive-compatible at $\lambda \in U_0$ since the expected delays of both service classes satisfy (10-11) of Lemma 1. Since the $c\mu$ rule minimizes the delay cost rate by Lemma 7 and the expected delays satisfy (18)-(20), it is the optimal policy.

3. By definition (27) , the $c\mu$ rule is not incentive-compatible at $\lambda \in U_2$ since the high priority (class 1) expected delay violates (10) of Lemma 1: it is larger than the upper bound imposed by the IC constraints. Together with (19) of Lemma 5, this implies that no admissible scheduling rule $r \in \mathcal{A}$ is incentive-compatible at $\lambda \in U_2$.□

6.5 Proof of Proposition 2

1. Suppose that $\lambda^{c\mu I}$ is unique, in the interior of M and in U_1. Since $\Pi_{\lambda_1}(0, c\mu I) > 0$, $\Pi\left(\lambda^{c\mu I}, c\mu I\right) > 0$. Take $\lambda \neq \lambda^{c\mu I}$ in M. By Proposition 1: if $\lambda \in U_1$, then $\Pi^*(\lambda) = \Pi(\lambda, c\mu I)$. If $\lambda \in U_0$, then $\Pi^*(\lambda) = \Pi(\lambda, c\mu) \leq \Pi(\lambda, c\mu I)$ by definition of U_0. If $\lambda \in U_2$, then $\Pi^*(\lambda) = 0$. Hence

$$\Pi^*\left(\lambda^{c\mu I}\right) = \Pi\left(\lambda^{c\mu I}, c\mu I\right) > \Pi^*(\lambda) \text{ for all } \lambda \neq \lambda^{c\mu I} \text{ in } M. \tag{96}$$

2. If strategic idleness is strictly optimal, then U_1 is non-empty by Proposition 1 and

$$\max_{\lambda \in U_1} \Pi^*(\lambda) = \max_{\lambda \in U_1} \Pi(\lambda, c\mu I) > \max_{\lambda \in U_0 \cup U_2} \Pi^*(\lambda). \tag{97}$$

Let $\lambda^{c\mu I}$ be a maximum of $\Pi(\lambda, c\mu I)$ on U_1. (a) We show that it is in the interior of M:

$$\left\{\lambda : 0 < \lambda_1 < \min\left(\frac{A_1}{B_1}, 1 - \lambda_2\right), 0 < \lambda_2 < \min\left(1, \frac{A_2}{B_2}\right)\right\}. \tag{98}$$

We have $\lambda_2^{c\mu I} > 0$ by definition of U_1, and $\lambda_1^{c\mu I} > 0$ since

$$\Pi_{\lambda_1}(\lambda, c\mu I)|_{\lambda_1 = 0} = A_1 - c_1 + \lambda_2 \cdot \frac{c_2 B_1}{c_1 - c_2} > 0 \text{ for all } \lambda_2. \tag{99}$$

For $\lambda_2 > 0$, we have $\lambda \in U_1$ if and only if $G(\lambda) < 0$, where

$$G(\lambda) := \frac{1}{(1 - \lambda_1)(1 - \lambda_1 - \lambda_2)} - \frac{A_1 - B_1\lambda_1 - (A_2 - B_2\lambda_2)}{c_1 - c_2}. \tag{100}$$

Fix $\lambda_2 \in (0, 1)$. Since $G_{\lambda_1}(\lambda) > 0$ and

$$\lim_{\lambda_1 \to \min\left(\frac{A_1}{B_1}, 1 - \lambda_2\right)} G(\lambda) > 0, \tag{101}$$

there is no $\lambda \in U_1$ with $\lambda_1 \geq \min\left(\frac{A_1}{B_1}, 1-\lambda_2\right)$. If $\frac{A_2}{B_2} \geq 1$, then $\lambda \in M$ for all $\lambda \in U_1$: since $\lim_{\lambda_2 \to 1-\lambda_1} G(\lambda) > 0$, the boundary of U_1 is strictly below the line $\lambda_2 = 1$. If $\frac{A_2}{B_2} < 1$, then $\lambda_2^{c\mu I} < \frac{A_2}{B_2}$ since

$$\Pi_{\lambda_2}(\lambda, c\mu I)|_{\lambda_2 = \frac{A_2}{B_2}} = \frac{-A_2 c_1 - c_2(A_1 - \lambda_1 B_1)}{c_1 - c_2} < 0. \tag{102}$$

(b) The point $\lambda^{c\mu I}$ is the unique maximum of $\Pi(\lambda, c\mu I)$ on the positive orthant. By (a), $\nabla\Pi\left(\lambda^{c\mu I}, c\mu I\right) = 0$. The FOC $\nabla\Pi(\lambda, c\mu I) = 0$ has at most two solutions, only one of which can be a maximum. For $B_i > 0$, define

$$h_1(\lambda_1) = 2\lambda_1\left(\frac{c_1}{c_2} - 1\right) - \left(A_1 - \frac{c_1}{(1-\lambda_1)^2}\right)\frac{c_1 - c_2}{c_2 B_1} \tag{103}$$

$$h_2(\lambda_1) = \frac{A_2 c_1 - A_1 c_2 + \lambda_1 c_2 B_1}{2 c_1 B_2}, \tag{104}$$

where $\Pi_{\lambda_i}(\lambda, c\mu I)|_{\lambda_2 = h_i(\lambda_1)} \equiv 0$. No interior λ with $\lambda_2 > h_1(\lambda_1)$ or $\lambda_2 < h_2(\lambda_1)$ is a maximum since $\Pi_{\lambda_1\lambda_2}(\lambda, c\mu I) > 0$ and $\Pi_{\lambda_2\lambda_2}(\lambda, c\mu I) < 0$:

$$\Pi_{\lambda_1}(\lambda, c\mu I) > 0 \Leftrightarrow \lambda_2 > h_1(\lambda_1) \tag{105}$$

$$\Pi_{\lambda_2}(\lambda, c\mu I) > 0 \Leftrightarrow \lambda_2 < h_2(\lambda_1). \tag{106}$$

An interior λ is a maximum if and only if

$$\lambda_2 = h_1(\lambda_1) = h_2(\lambda_1) \tag{107}$$

and the function $\Pi(\lambda, c\mu I)$ is concave. Since $\Pi_{\lambda_2\lambda_2}(\lambda, c\mu I) < 0$, $\Pi_{\lambda_1\lambda_2}(\lambda, c\mu I) > 0$, and

$$h_1'(\lambda_1) = -\frac{\Pi_{\lambda_1\lambda_1}(\lambda, c\mu I)}{\Pi_{\lambda_1\lambda_2}(\lambda, c\mu I)}, \quad h_2'(\lambda_1) = -\frac{\Pi_{\lambda_1\lambda_2}(\lambda, c\mu I)}{\Pi_{\lambda_2\lambda_2}(\lambda, c\mu I)}, \tag{108}$$

the Hessian of $\Pi(\lambda, c\mu I)$ is negative semi-definite if and only if

$$\Pi_{\lambda_1\lambda_1}(\lambda, c\mu I)\,\Pi_{\lambda_2\lambda_2}(\lambda, c\mu I) - \left(\Pi_{\lambda_2\lambda_1}(\lambda, c\mu I)\right)^2 = \left(h_2'(\lambda_1) - h_1'(\lambda_1)\right)\Pi_{\lambda_2\lambda_2}(\lambda, c\mu I)\,\Pi_{\lambda_2\lambda_1}(\lambda, c\mu I) \geq 0. \tag{109}$$

This requires $h_1'(\lambda) \geq h_2'(\lambda)$. Both h_1 and h_2 are strictly increasing. Since h_1 is strictly convex and h_2 is linear, there is at most one point that satisfies (107) and (109). By hypothesis, it is $\lambda^{c\mu I}$, which establishes the claim. If $h_1(\lambda^o) = h_2(\lambda^o)$ at $\lambda^o \neq \lambda^{c\mu I}$, then $h_1'(\lambda^o) < h_2'(\lambda^o)$ and λ^o is a saddle point.

(c) The point $\lambda^{c\mu I}$ is the unique maximum of $\Pi(\lambda, c\mu I)$ on M. If there is a maximum λ' with $\lambda_2' = 0$, then $\Pi_{\lambda_1}\left(\lambda', c\mu I\right) = 0 > \Pi_{\lambda_2}\left(\lambda', c\mu I\right)$. To show that $\Pi\left(\lambda^{c\mu I}, c\mu I\right) > \Pi\left(\lambda', c\mu I\right)$, it suffices to show that $\lambda' \in U_0$. For $\overline{\lambda}_1$ which satisfies

$$G(\lambda)|_{\lambda_1 = \overline{\lambda}_1, \lambda_2 = 0} = \frac{1}{\left(1 - \overline{\lambda}_1\right)^2} - \frac{A_1 - A_2 - B_1\overline{\lambda}_1}{c_1 - c_2} = 0, \tag{110}$$

we have $\Pi_{\lambda_1}(\lambda, c\mu I) - \Pi_{\lambda_2}(\lambda, c\mu I)|_{\lambda_1 = \overline{\lambda}_1, \lambda_2 = 0} = -\overline{\lambda}_1 B_1 < 0$. Since $\Pi_{\lambda_1\lambda_1}(\lambda, c\mu I) - \Pi_{\lambda_2\lambda_1}(\lambda, c\mu I) < 0$, we must have $\lambda_1' < \overline{\lambda}_1$ and so $\lambda' \in U_0$. Since $\lambda_2' = 0$ and strategic idleness is strictly optimal, we have $\Pi\left(\lambda^{c\mu I}, c\mu I\right) = \Pi^*\left(\lambda^{c\mu I}\right) > \Pi^*\left(\lambda'\right) = \Pi\left(\lambda', c\mu I\right)$.□

6.6 Proof of Proposition 3

Fix $B_1 > 0$. Let $\lambda_{11}(B_1)$ and $\lambda_{12}(B_1)$ be the rates that satisfy:

$$\Pi_{\lambda_1}(\lambda, c\mu I)|_{\lambda_1=\lambda_{11}(B_1),\lambda_2=0} = 0 \tag{111}$$

$$\Pi_{\lambda_2}(\lambda, c\mu I)|_{\lambda_1=\lambda_{12}(B_1),\lambda_2=0} = 0. \tag{112}$$

From (49)-(50), $\lambda_{11}(B_1)$ and $\lambda_{12}(B_1)$ are unique for $B_1 > 0$. Notice that

$$h_1(\lambda_{11}(B_1)) = h_2(\lambda_{12}(B_1)) = 0, \tag{113}$$

where h_1 and h_2 are defined by (103-104). If $\lambda_{12}(B_1) < \lambda_{11}(B_1)$, then $\Pi(\lambda, c\mu I)$ has no maximum on the boundary $\lambda_2 = 0$ since $\Pi_{\lambda_1}(\lambda, c\mu I)|_{\lambda_2=0} > 0$ for $\lambda_1 \leq \lambda_{12}(B_1)$ and $\Pi_{\lambda_2}(\lambda, c\mu I)|_{\lambda_2=0} > 0$ for $\lambda_1 > \lambda_{12}(B_1)$. Since $h_1' > 0$, $h_1'' > 0$, $h_2' > 0$ and h_2 is linear, it follows that the FOC $\nabla\Pi(\lambda, c\mu I) = 0$ have at most one solution in the positive orthant, and that such a solution is the unique maximum of $\Pi(\lambda, c\mu I)$. Hence, the sufficient conditions of Proposition 2 for optimality of strategic idleness hold if and only if

$$\nabla\Pi\left(\lambda^{c\mu I}, c\mu I\right) = 0 \Longleftrightarrow \lambda_2^{c\mu I} = h_1\left(\lambda_1^{c\mu I}\right) = h_2\left(\lambda_1^{c\mu I}\right), \tag{114}$$

$$G\left(\lambda^{c\mu I}\right) < 0, \tag{115}$$

for some $\lambda^{c\mu I} > 0$, where G is defined by (100). Let $\lambda_{1G}(B_1)$ be the rate that satisfies

$$G(\lambda)|_{\lambda_1=\lambda_{1G}(B_1),\lambda_2=0} = 0. \tag{116}$$

For $A_1 - c_1 > A_2 - c_2$, it is unique and satisfies $0 < \lambda_{1G}(B_1) < \min\left(1, \frac{A_1}{B_1}\right)$. The proof involves two steps. Step 1: Show that if B_1 satisfies

$$\lambda_{12}(B_1) < \lambda_{11}(B_1) < \lambda_{1G}(B_1), \tag{117}$$

then (114)-(115) are satisfied for $B_2 \geq B_2^o$, where B_2^o is a threshold which depends on all parameters. Step 2: Identify parameter combinations for which (117) holds.

1. Suppose that (117) holds for a given B_1. Observe that the rates $\lambda_{11}(B_1)$, $\lambda_{12}(B_1)$ and $\lambda_{1G}(B_1)$ do not depend on B_2. For all B_1, we have $0 < \lambda_{11}(B_1) < \min\left(1, \frac{A_1}{B_1}\right)$.

For $B_2 = 0$, the points on the interior boundary of U_1, defined by $G(\lambda) = 0$, satisfy

$$\lambda_2 = g(\lambda_1) := 1 - \lambda_1 - \frac{(c_1 - c_2)}{(1-\lambda_1)(A_1 - A_2 - B_1\lambda_1)} \text{ for } \lambda_1 \in [0, \lambda_{1G}(B_1)], \tag{118}$$

where $g(\lambda_1)$ is a strictly decreasing function and $g(\lambda_{1G}(B_1)) = 0$. Notice that $g(0) > 0$, $h_1(0) < 0$ and $h_1(\lambda_1) > 0$ for $\lambda_1 > \lambda_{11}(B_1)$. If $\lambda_{11}(B_1) < \lambda_{1G}(B_1)$, then since $h_1(\lambda_1)$ is strictly increasing, $h_1(\lambda)$ and $g(\lambda)$ intersect exactly once at some point $\lambda^o(B_1) \in M$, and the line segment $L := \{(\lambda_1, h_1(\lambda_1)) : \lambda_1 \in (\lambda_{11}(B_1), \lambda_1^o(B_1))\}$ is contained in U_1. Note that $\Pi_{\lambda_1}(\lambda, c\mu I) = 0$ for all $\lambda \in L$.

For $B_2 = 0$, we have $\Pi_{\lambda_2}(\lambda, c\mu I) = 0$ along the vertical $\lambda_1 = \lambda_{12}(B_1)$. If $\lambda_{12}(B_1) < \lambda_{11}(B_1)$, then h_1 and h_2 do not intersect in the positive orthant, and strategic idleness cannot be optimal. Increasing B_2 leaves the function $h_1(\lambda_1)$ and therefore the line segment L unchanged, but it changes the constraint $G(\lambda) = 0$ and the function $h_2(\lambda_1)$ as follows. By the implicit function theorem, we have

$$G(\lambda_1(B_2), \lambda_2; B_2) \equiv 0 \Rightarrow \frac{d\lambda_1(B_2)}{dB_2} = -\frac{\partial G/\partial B_2}{\partial G/\lambda_1} > 0 \text{ for } \lambda_2 > 0, \tag{119}$$

which implies that the line segment L remains in U_1. We have

$$\frac{\partial h_2(\lambda_1; B_2)}{\partial B_2} < 0 \text{ and } \lim_{B_2 \to \infty} h_2(\lambda_1; B_2) = 0 \text{ for } \lambda_1 > \lambda_{12}(B_1), \tag{120}$$

i.e., the line $\{(\lambda_1, h_2(\lambda_1)) : \lambda_1 \geq \lambda_{12}(B_1)\}$ "rotates clockwise" around the point $(\lambda_{12}(B_1), 0)$. As a result, there is a $B_2^o > 0$ such that $h_2(\lambda_1^o(B_1)) = h_1(\lambda_1^o(B_1))$, where B_2^o is a function of all parameters. (It corresponds to the parameters denoted by B_{21}, B_{22} and B_{23} for each of the triangles in the Proposition.) Note that $G(\lambda^o(B_1)) < 0$. So for B_1 and B_2^o, strategic idleness is optimal and $\lambda^o(B_1)$ is the unique revenue-maximizing vector. For all $B_2 > B_2^o$, $h_2(\lambda_1)$ intersects the line segment L exactly once in U_1.

2. We now provide conditions on the parameters A_1, A_2, c_1, c_2 and B_1 that satisfy (117).
Since $\lambda_{12}(B_1) < \lambda_{11}(B_1)$, $\Pi(\lambda, c\mu I)$ has no maximum on the boundary $\lambda_2 = 0$. Since $h_1'(\lambda_1) > 0$, it is equivalent to $h_1(\lambda_{12}(B_1)) < 0$. Solving for $\lambda_{12}(B_1)$ by setting $h_2(\lambda_{12}(B_1)) = 0$ yields $\lambda_{12}(B_1) = x/B_1$, where $x = A_1 - A_2 c_1/c_2$. Substituting in h_1 yields

$$h_1\left(\frac{x}{B_1}\right) < 0 \Leftrightarrow \frac{c_1}{\left(1 - \frac{x}{B_1}\right)^2} < A_1 - 2x, \tag{121}$$

where $x/B_1 < 1$ is required to satisfy the capacity constraint. This implies for $B_1 > 0$:

$$1 > \left(1 - \frac{x}{B_1}\right)^2 > \frac{c_1}{A_1 - 2x} > 0 \Leftrightarrow 1 - \sqrt{\frac{c_1}{A_1 - 2x}} > \frac{x}{B_1}. \tag{122}$$

If $x \leq 0$, this holds for all $B_1 > 0$. Otherwise, it holds if and only if

$$\frac{A_2 c_1}{A_1} < c_2 < \frac{2A_2 c_1}{A_1 + c_1} \text{ and } B_1 > B_1^o := \frac{A_1 c_2 - A_2 c_1}{c_2\left(1 - \sqrt{\frac{c_1 c_2}{2A_2 c_1 - A_1 c_2}}\right)}. \tag{123}$$

For $c_2 \geq \frac{2A_2 c_1}{A_1 + c_1}$, we have $\lambda_{12}(B_1) \geq \lambda_{11}(B_1)$ for all B_1.
Since $\lambda_{11}(B_1) < \lambda_{1G}(B_1)$, the line segment L of h_1 lies in U_1 for all B_2. We have

$$h_1(\lambda_{11}(B_1), 0) = 0 \Leftrightarrow A_1 - 2B_1 \cdot \lambda_{11}(B_1) = \frac{c_2}{(1 - \lambda_{11}(B_1))^2} \tag{124}$$

$$G(\lambda_{1G}(B_1), 0) = 0 \Leftrightarrow \frac{A_1 - A_2 - B_1 \cdot \lambda_{1G}(B_1)}{c_1 - c_2} = \frac{1}{(1 - \lambda_{1G}(B_1))^2}. \tag{125}$$

Noting that $G_{\lambda_1}(\lambda) < 0$ and $\Pi_{\lambda_1 \lambda_1} < 0$, we have $\lambda_{11}(B_1) < \lambda_{1G}(B_1)$ if and only if

$$\frac{A_1 - 2B_1 \cdot \lambda_{1G}(B_1)}{c_1} < \frac{A_1 - A_2 - B_1 \cdot \lambda_{1G}(B_1)}{c_1 - c_2} = \frac{1}{(1 - \lambda_{1G}(B_1))^2}. \tag{126}$$

The inequality is satisfied for all $\lambda_1 > 0$ if and only if

$$LHS(\lambda_1) := \lambda_1 \cdot B_1(2c_2 - c_1) < A_1\left(c_2 - \frac{A_2}{A_1} c_1\right) := RHS(\lambda_1). \tag{127}$$

Fix A_1 and c_1. Condition (127) can hold as follows.
(i) For $c_2 = \frac{c_1}{2}$, (127) holds for all B_1 if and only if $c_2 > \frac{A_2 c_1}{A_1} \Rightarrow A_2 < \frac{A_1}{2}$.

(*ii*) For $c_2 < \frac{c_1}{2}$: if $c_2 \geq \frac{A_2 c_1}{A_1}$, then $RHS(\lambda_1)$ is nonnegative and (126) holds for all B_1. Conversely, if $c_2 < \frac{A_2 c_1}{A_1}$, then $RHS(\lambda_1)$ is negative and (126) does not hold at $\lambda_1 = 0$, but LHS and RHS cross at

$$\lambda_1^{**}(B_1) = \frac{1}{B_1} \frac{A_1 c_2 - A_2 c_1}{2c_2 - c_1}, \tag{128}$$

where their value equals

$$\frac{A_1 - 2B_1 \cdot \lambda_1^{**}(B_1)}{c_1} = \frac{A_1 - A_2 - B_1 \cdot \lambda_1^{**}(B_1)}{c_1 - c_2} = \frac{2A_2 - A_1}{2c_2 - c_1}. \tag{129}$$

If this value is less than one, i.e.,

$$\frac{2A_2 - A_1}{2c_2 - c_1} \leq 1 \Leftrightarrow c_2 \leq \frac{c_1}{2} + \left(A_2 - \frac{A_1}{2}\right), \tag{130}$$

then (126) holds for no B_1, since the right-hand side of (126) is strictly larger than one. If

$$\frac{c_1}{2} + \left(A_2 - \frac{A_1}{2}\right) < c_2 < \frac{A_2}{A_1} c_1, \tag{131}$$

then $\lambda_{11}(B_1) < \lambda_1^{**}(B_1)$ if and only if $B_1 > B_1^*$, where B_1^* is the solution of

$$\frac{1}{(1 - \lambda_1^{**}(B_1))^2} = \frac{2A_2 - A_1}{2c_2 - c_1}. \tag{132}$$

(*iii*) For $c_2 > \frac{c_1}{2}$: if $c_2 \leq \frac{A_2 c_1}{A_1}$, then $RHS(\lambda_1)$ is nonpositive and (126) holds for no B_1. Conversely, if $c_2 > \frac{A_2 c_1}{A_1}$, $RHS(\lambda_1)$ is positive and (126) holds at $\lambda_1 = 0$. In this case, (126) holds for all B_1 if

$$\frac{2A_2 - A_1}{2c_2 - c_1} \leq 1 \Leftrightarrow c_2 \geq \frac{c_1}{2} + \left(A_2 - \frac{A_1}{2}\right), \tag{133}$$

but if and only if $B_1 < B_1^*$ in case c_2 satisfies

$$\frac{A_2}{A_1} c_1 < c_2 < \frac{c_1}{2} + \left(A_2 - \frac{A_1}{2}\right). \tag{134}$$

Notice that $A_1 - c_1 > A_2 - c_2$ requires $A_2 < A_1$ since $c_1 > c_2$. Combining the conditions yields the triangles T_1, T_2 and T_3. The fact that $\lambda_{11}(B_1) < \lambda_{1G}(B_1)$ if $\lambda_{12}(B_1) = \lambda_{11}(B_1)$ establishes that the interval (B_1^o, B_1^*) is nonempty for (A_2, c_2) in triangle T_3.□

6.7 Proof of Proposition 4

We focus on case 3. (The proofs for 1. and 2. are similar to that of Prop. 1.) For fixed $\lambda \in M$, problem (82) is equivalent to the delay cost minimization problem:

$$\min_W \sum_{i=1}^{2} \lambda_i \cdot c_i \cdot W_i, \tag{135}$$

subject to (77) and (67)-(69). Case (a): if $\lambda \in U_{c\mu}$, then the $c\mu$ rule is optimal: it satisfies (77) since $c_1 < c_2$, and it solves (135) by Lemma 11. Cases (b) and (c): if the expected class 2 delay under the $c\mu$ rule violates (77), the $c\mu$ rule is not IC and class 2 customers must be given faster service to restore IC. This is feasible if and only if (77) holds at the lower bound for $W_{2,2}(\lambda, r)$

over all nonpreemptive admissible policies $r \in \mathcal{A}^{NP}$, which the $Rc\mu$ rule attains. Thus, there are admissible IC policies for $\lambda \in U_{c\mu-Rc\mu}$ and none at $\lambda \in U_{No}$. If $\lambda \in U_{c\mu-Rc\mu}$, it is optimal to reduce W_2 only up to the point where it equals (87) by Lemma 11. The conservation law (67) implies that W_1 increases and must satisfy (86). By Lemma 9 there is an admissible (α, d) policy that attains (86-87), where $d = 0$ since $W_1^*(\lambda)$ and $W_2^*(\lambda)$ satisfy (67) with equality.□

References

Afeche, P., H. Mendelson. 2004. Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure. *Man. Sci.,* to appear.

Balachandran, K. R. 1972. Purchasing Priorities in Queues. *Man. Sci.* 18 (5) 319-326.

Cachon, G., P. Harker. 2002. Competition and outsourcing with scale economies. *Man. Sci.* 48 (10) 1314-1333.

Caldentey R.M., L.M. Wein. 2003. Revenue Management of a Make-to-Stock Queue. Working paper, New York University, New York.

Coffman, E.G. Jr., I. Mitrani. 1980. A Characterization of Waiting Time Performance Realizable by Single-Server Queues. *Oper. Res.* 28 (3) 810-821.

Conway, R.W., W.L. Maxwell, L. Miller. 1967. *Theory of Scheduling.* Add.-Wesley, Mass.

Cox, D., W. Smith. 1961. *Queues.* Methuen, London.

Dewan, S., H. Mendelson. 1990. User Delay Costs and Internal Pricing for a Service Facility. *Man. Sci.* 36 (12) 1502-1517.

Dolan, R.J. 1978. Incentive Mechanisms for Priority Queuing Problems. *Bell J. of Econ.* 9 (2) 421-436.

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

Gelenbe E., I. Mitrani. 1980. *Analysis and Synthesis of Computer Systems.* Acad. Press, London.

Ghanem, S.B. 1975. Computing central optimization by a pricing priority policy. *IBM Sys. J.* 14 272-292.

Gupta, A., D.O. Stahl, A.B. Whinston. 1996. An economic approach to networked computing with priority classes. *J. of Organizational Computing*, 6 71-95.

Ha, A.Y. 2001. Optimal Pricing That Coordinates Queues with Customer-Chosen Service Requirements. *Man. Sci.* 47 (7) 915-930.

Hassin, R. 1995. Decentralized regulation of a queue. *Man. Sci.* 41 (1) 163-173.

Hassin R., M. Haviv. 2003. *To Queue or not to Queue.* Kluwer, Boston.

Jehiel, P., B. Moldovanu, E. Stacchetti. 1996. How (Not) to Sell Nuclear Weapons. *The American Economic Review* 86 (4) 814-829.

Kakalik, J.S. 1969. Optimal Dynamic Operating Policies for a Service Facility. Technical Report OR Center, MIT, Boston.

Kalai, E., Kamien, M.I., M. Rubinovitch. 1992. Optimal Service Speeds in a Competitive Environment. *Man. Sci.* 38 (8) 1154-1163.

Kleinrock, L. 1967. Optimum Bribing for Queue Position. *Oper. Res.* 15 (2) 304-318.

Lederer, P.J., L. Li. 1997. Pricing, Production, Scheduling and Delivery-Time Competition. *Oper. Res.* 45 (3) 407-420.

Li, L., Y.S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Man. Sci.* 40 (5) 633-646.

Lippman, S.A., S. Stidham, Jr. 1977. Individual versus Social Optimization in Exponential Congestion Systems. *Oper. Res.* 25 (2) 233-247.

Maglaras, C., A. Zeevi. 2003. Pricing and Design of Differentiated Services: Approximate Analysis and Structural Insights. Working paper, Columbia University, New York.

Marchand, M.G. 1974. Priority Pricing. *Man. Sci.* 20 (7) 1131-1140.

Mendelson, H. 1985. Pricing Computer Services: Queueing Effects. *Comm. ACM* 28 (3) 312-321.

Mendelson, H., S.Whang. 1990. Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. *Oper. Res.* 38 (5) 870-883.

Myerson, R.B. 1981. Optimal Auction Design. *Math. of OR.* 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.,* to appear.

Rao, S., E.R. Petersen. 1998. Optimal Pricing of Priority Services. *Oper. Res.* 46 (1) 46-56.

Rubinovitch, M. 1985. The Slow Server Problem: A Queue with Stalling. *J. App. Prob.* 22 879-892.

Shaked, M., J.G. Shanthikumar. 1994. *Stochastic Orders and their Applications.* Acad. Press, San Diego.

Shanthikumar, J.G., D.D. Yao. 1992. Multiclass Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control. *Oper. Res.* 40 (2) S293-S299.

Shumsky, R.A., E.J. Pinker. 2003. Gatekeepers and Referrals in Services. *Man. Sci.* 49 (7) 839-856.

Stidham, S. Jr. 2002. Analysis, Desing and Control of Queueing Systems. *Oper. Res.*, 50 (1) 197-216.

Van Mieghem, J.A. 1995. Dynamic Scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. App. Prob,* 5(3) 809-833.

Van Mieghem, J.A.. 2000. Price and Service Discrimination in Queuing Systems: Incentive Compatibility of G$c\mu$ Scheduling. *Management Sci.* 46 (9) 1249-1267.

Wolff, R.W. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Englewood Cliffs.

Yechiali, U. 1971. On Optimal Balking Rules and Toll Charges in the GI/M/1 Queuing Process. *Oper. Res.* 19 (2) 348-370.