Optimal Price and Delay Differentiation in Queueing Systems

Costis Maglaras, John Yao, Assaf Zeevi

Graduate School of Business, Columbia University, New York, NY 10027,
c.maglaras@gsb.columbia.edu, jyao14@gsb.columbia.edu, assaf@gsb.columbia.edu

June 2013

We study a multi-server queueing model of a revenue-maximizing firm providing a service to a market of heterogeneous price- and delay-sensitive customers with private individual preferences. The firm may offer a selection of service classes that are differentiated in prices and delays. Using a deterministic relaxation, which highlights the first-order economic structure of the problem, we construct a solution that is incentive compatible and near-optimal in systems with large capacity and market potential. Our approach provides several new insights for large-scale systems: i) the tractable first-order analysis characterizes essentially all salient features of the optimal solution; ii) service differentiation is optimal when the less delay-sensitive market segment is sufficiently elastic; iii) depending on system capacity and market heterogeneity, "intentional delay" (whereby delay is artificially added) in cheaper service classes may be used to justify price premiums in the more expensive service classes, akin to the role of "damaged goods" in the economics literature; and iv) connecting economic optimization to queueing theory, the revenue-optimized system has the premium class operating in a "quality-driven" regime and the lower-tier service classes operating in an "efficiency-driven" regime (i.e., with noticeable delays that arise either endogenously or due to the injection of intentional delay by the service provider).

---

1 Introduction

1.1 Main Objectives and Overview of Results

Background and the main objectives. Price discrimination based on the "speed" at which a service is delivered has become a prevalent business practice. Standard examples include: parcel delivery services such as FedEx and UPS that offer overnight delivery at substantially higher prices than standard ground shipping; airport security screening whereby any economy class ticket holder, regardless of frequent flyer status, can purchase access to a priority lane; and various government services, e.g., passport issuance and renewals, that can be expedited by paying additional fees. A more recent example is the debate on "network neutrality" which questions whether Internet

service providers should be allowed to charge higher prices to certain content providers for faster data transmission rates. In all of the above, an essentially identical service is provisioned at varying quality grades (here, processing speed) and segments the market in a way that enables the firm to provide faster processing for impatient customers and shift system congestion to more patient customers. For revenue-maximizing firms, this service differentiation is driven by the potential to extract further revenues from the less-patient customer base, while non-profit providers can use service differentiation to better allocate resources and increase social welfare. Roughly speaking, the high-level problem faced by the service provider is how to optimally differentiate services by creating and appropriately implementing a suitable price-quality menu. In this paper we study a stylized queueing model of this service differentiation problem with the intention of shedding further light on the economics and operations considerations underlying this practice.

We consider a monopolistic revenue-maximizing firm (service provider) that offers a single service to a market of heterogeneous price- and delay-sensitive customers. The system is modeled as a multi-server queueing model, and the service may be offered at different grade levels that are differentiated in terms of price and delay; we will refer to these as "service classes." Each individual customer gains some positive utility from the service, but suffers negative utility for each unit of time he spends waiting in the system. Upon arriving at the system, he chooses one of these service classes (or opts out) so as to maximize his net utility (this utility being a linear function of the mean processing delay and price). In this manner, the set of price and delay combinations affects the demand rates into each service class, which in turn determines the congestion experienced there, and vice versa. An optimal solution will specify a menu of service classes together with a sequencing rule that maximize the expected revenue rate.

The market is composed of distinct customer segments or "types." All customers of a particular type have the same linear delay sensitivity and a random service valuation (or willingness-to-pay) drawn from a common distribution. A key assumption is that the type (and hence delay sensitivity) as well as the willingness-to-pay of an individual customer is *private information* and thus unknown to the service provider. Therefore, if the service provider chooses to offer different service levels, possibly at different prices, then he also needs a mechanism to ensure that the customers' self-optimizing choices are aligned with the service provider's revenue objective. The service provider's revenue maximization problem can be cast as a mechanism design problem.

The socially optimal menu for the above model is known and fairly straightforward to characterize and implement, stemming from the observations that it is optimal to set prices equal to the externality costs and to allocate servers so as to minimize aggregate delay costs; see further discussion in §1.2. For revenue maximization, however, both of these insights no longer hold and the firm's problem becomes more complex and only partially understood. This paper proposes an

approximate analysis, which is justified rigorously for systems with large capacity and large market potential, that offers significant insights into the structure of the optimal solution.

Main findings. The paper's findings contribute along three dimensions. First, we show that a *deterministic relaxation* of the service provider's revenue maximization problem gives rise to an intuitive price-delay menu which, in conjunction with a simple priority sequencing rule, achieves essentially optimal revenue performance when implemented in the stochastic system. This is described in §3 for a market with two customer types and is extended to more than two types in §5. In so doing, the paper demonstrates that a complex mechanism design problem can be made tractable by invoking large-scale analysis ideas that are commonplace in queueing theory. A key component of this approach is proving that incentive compatibility conditions are satisfied for systems of suitable size.

The paper's second contribution lies in the economic structural insights that are teased out of the deterministic relaxation: the service provider will determine whether to offer differentiated services depending on the extent to which the customer types' characteristics give rise to elastic or inelastic demand. In particular, increasing delay in certain service classes creates incentives for impatient customers to pay a premium for better (faster) service. In order to achieve this, it may be optimal to offer a form of "damaged goods," in which the service provider artificially delays the completion of service in some number of classes, in order to increase overall revenues. While strategically delaying some of the service classes may seem plausible when the system has slack processing capacity, one may question whether this insight persists when the resources in the system are more heavily utilized, since significant delays may then arise endogenously. Indeed, we show in §3-4, which focus on two customer types, that endogenous delays are sufficient and the use of damaged goods is unnecessary for systems with no "excess" capacity. However, as seen in the more general case discussed in §5, the damaged goods insight is more fundamental: our results suggest that purposeful delay of certain service classes is *necessary* in order to achieve optimal revenue performance *even* if there is no "slack" in systems with three or more service classes. In stark contrast, if the economic objective is social welfare optimization (as opposed to revenue maximization), the deterministic relaxation prescribes a non-differentiated solution; namely, only a single service class is needed to essentially achieve the optimal value of the objective function. To those readers familiar with the work of Mendelson and Whang (1990), this observation may seem puzzling at first glance; we provide an explanation in §6.

Finally, the paper also contributes to the literature on heavy-traffic analysis of queueing systems. Roughly speaking, we show that classical operating regimes, such as the so-called *efficiency-driven* (ED) and *quality-driven* (QD), may arise endogenously as a result of the revenue-maximizing mechanism design problem; specifically, the high priority class operates in the underloaded quality-driven

regime while the low priority class operates in the heavily utilized efficiency-driven regime. This complements earlier results by Maglaras and Zeevi (2003a), where it was first shown that, when customers are *homogenous* in their delay costs, the *quality and efficiency-driven* operating regime (QED) arises endogenously as a result of revenue maximization; this can be viewed as an intermediate case relative to the ones shown to be optimal for heterogeneous delay preferences.

1.2 Related Literature

The work on strategic customers in queues – where arrivals depend on system congestion – is extensive, dating back to the seminal study of Naor (1969); a survey of the topic area can be found in Hassin and Haviv (2003). Of particular note are Mendelson (1985) and Mendelson and Whang (1990), both of which model the system as an $M/M/1$ queue. The former considers a single customer type and formalizes the atomistic, utility-maximizing customer behavior in queues. The latter extends the welfare optimization analysis to multiple customer types and shows that prices that are set to externality costs are incentive compatible, and that delay cost minimization is optimal.

The closest paper in the literature is Afèche (2013), which addresses the revenue maximization problem in a single-server queueing system facing a market with two customer types (analogous to §3-4 in this work). He formulates the problem in a mechanism design framework, using ideas from the seminal work of Myerson (1979, 1981), and highlights the fact that externality pricing and delay cost minimization are no longer optimal in the revenue maximization setting. Moreover, he establishes that the optimal solution may even include so-called "strategic delay," in which the service provider chooses to artificially delay some customers beyond what is caused by system congestion alone. Our work adopts the mechanism design framework (which, among other things, allows for strategic delay), but our method of analysis is quite different and our main insights substantiate and extend Afèche (2013) in a more general setting, and separates in some sense the "first-order" effects from "lower-order" phenomena; we will return to discuss this in more detail in §5 (see Remark 3) after expounding on our main results. In particular, it reveals when and why service differentiation is revenue maximizing, how the firm should implement service differentiation, and precisely illustrates the role of strategic delay in this context.

A parallel stream of work analyzes multi-server queueing systems and leverages asymptotic analysis to gain insight into the optimal prices and policies. The work of Maglaras and Zeevi (2003a) considers a single-class $M/M/n$ system with price- and congestion-sensitive customers. Their work characterizes the asymptotic equilibrium operating point, and shows that when demand is elastic, the revenue-maximizing price induces customer arrivals that result in the QED regime; an operating regime where the probability of a customer experiencing delay is strictly positive but below one.

Maglaras and Zeevi (2005) extend this argument to a two-class system where aggregate demand into each class is affected by price and congestion, again linking economic optimality to the QED regime. Methodologically, this paper also relies on a suitable "deterministic relaxation" of the original optimization problem, but unlike Maglaras and Zeevi (2005), it allows for full substitution of services, whereas the former considers partially substitutable services where atomistic choice is not captured and incentive compatibility considerations are not present.

The "strategic delay" that features in the revenue-maximization setting can be viewed as the queueing system manifestation of damaged goods. This concept, at least within the economics and marketing literature, refers to a practice by which firms introduce a lower-price lower-quality version of a good with equal (or even higher) production cost, in order to segment the customer market and price discriminate. A number of examples of such cases can be found in Deneckere and McAfee (1996); see also McAfee (2007) who derives sufficient conditions for such practice in terms of marginal revenues. More recently, Anderson and Dana (2009) provide necessary conditions for a monopolist firm to increase profits by engaging in price discrimination, which may include offering damaged goods. However, the marketing and economics literature disregards the operational considerations of the service system, and the inherent conflict between price discrimination and efficient resource utilization (though Aféche (2013) takes a first step towards incorporating some of these aspects). Our model differs in two important aspects. We consider a market where customer valuations are continuous within discrete types of quality sensitivity. In the aforementioned works, each customer segment has a single valuation and a single quality sensitivity parameter, with a finite number of discrete segments or a continuum of segments. More importantly, we consider a system that is subject to congestion, so quality degrades as more customers purchase the service, and the service provider only has a partial (deliberate delay) or indirect (pricing and sequencing) influence on quality. For these reasons, while some of our results have a similar flavor to the ones mentioned above, they are neither a special case nor a generalization thereof.

Finally, there is a growing body of work that studies economic problems in queueing systems under different modeling assumptions, for example, allowing for non-linear delay cost functions, state-dependent pricing or delay notification, different user risk preferences, settings with uncertain model primitives and different degrees of learning, etc. We will not review this literature herein, but remark that the tractable analytical framework proposed in our paper has the potential to be applicable in those settings as well.

2 Model and Problem Formulation

We first describe the queueing system used to model the firm, then define a customer choice model, and finally formulate the optimization problem.

System model. The service provider (SP) operates s processing resources used to offer a particular service to a market of heterogeneous customers. The SP may offer k versions (or "classes") of the service that are differentiated by price and delay. We model this as a multi-class, multi-server queueing system. Customer requests arrive into each service class $j = 1, \ldots, k$ according to an independent Poisson process with rate λ_j. Each service class has an infinite-capacity buffer, and customers in that class wait in a queue according to their order of arrival until they are allocated a server. This allocation is determined by a control policy π to be discussed later on. The *delay* experienced by a customer in a given service class is the time he spends in the system minus the time spent in service.

Customers, irrespective of service class, have random processing requirements that are modeled as independent and identically distributed (i.i.d.) draws from an exponential distribution with mean $1/\mu$; in that way, the *mean* processing requirements are homogeneous across customers. The control policy π may not depend on the realized service times of customers. Each server may only work on one customer at a time and servers may idle with customers waiting in their queues; i.e., we allow non-work-conserving policies. Service for any customer may be interrupted without penalty and resumed without restarting service. We *do not* require the service discipline to be first-come-first-served within a service class. The control policy is represented as an allocation process $\pi(t) : [0, \infty) \to \mathbb{Z}_+^k$, where $\pi_j(t)$ is the number of servers processing class j customers at time t. We require $\pi_j(t)$ to be right continuous with left limits and Lebesgue integrable; further structural assumptions will be advanced shortly.

To define the system dynamics, consider $2k$ mutually independent unit-rate Poisson processes, $N_j^{(a)}(t)$ and $N_j^{(s)}(t)$ for $j = 1, \ldots, k$. Fix for now an arrival rate vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ where λ_j is the arrival rate into class j, which will be further detailed later in this section. Define $N_j^{(a)}(\lambda_j t)$ to be the number of customers that have arrived into class j by time t and $N_j^{(s)}\left(\int_0^t \mu \pi_j(s)\, ds\right)$ the number of class j customers that have completed service by time t. It is useful to describe the system in terms of the "headcount process" $((Z_1(t), \ldots, Z_k(t)) : 0 \le t < \infty)$ where $Z_j(t)$ is the number of class j customers in the system at time t, and the "queue length process" $((Q_1(t), \ldots, Q_k(t)) : 0 \le t < \infty)$ where $Q_j(t)$ is the number of class j customers in queue at time t. These processes must jointly satisfy the following conditions:

$$\sum_{j=1}^{k} \pi_j(t) \le s, \tag{1}$$

$$Q_j(t) = Z_j(t) - \pi_j(t) \ge 0 \quad \text{for } j = 1, \ldots, k, \tag{2}$$

$$Z_j(t) = N_j^{(a)}\left(\lambda_j t\right) - N_j^{(s)}\left(\int_0^t \mu \pi_j(s)\, ds\right) \ge 0 \quad \text{for } j = 1, \ldots, k. \tag{3}$$

Condition (1) limits the total number of servers working at any time to be at most s, but does not preclude servers from idling while there is work in the system. Condition (2) restricts the number of servers working on class j customers to be at most the number of class j customers in the system at that time. Condition (3) describes the system dynamics. We require that the control π, $(\pi_1(t), \ldots, \pi_k(t))$ be adapted to the filtration generated by $(Z_1(t), \ldots, Z_k(t))$.

As an example of an allocation process, consider a strict preemptive priority policy, with highest priority given to class 1 and lowest given to class k. Under such a policy, an arriving class j customer interrupts any lower-priority customer in service, from classes $j+1, \ldots, k$. If all servers are serving higher- or equal-priority customers, the arriving customer waits in queue. As long as the queues of all higher-priority classes are empty, then idle servers may resume interrupted lower-priority customers (from highest to lowest priority and in the order that they were interrupted) and start working on customers from the highest-priority non-empty queue. In other words, all processing capacity is first applied to class 1 and any remaining capacity is then successively applied to class 2, then to class 3, and so on. Such a policy can be expressed as follows:

$$\pi_1(t) = \min\{s, Z_1(t)\} \qquad \pi_j(t) = \min\{(s - Z_1(t) - \cdots - Z_{j-1}(t))^+, Z_j(t)\}, \quad j = 2, \ldots, k. \tag{4}$$

(In later sections, we see that this class of simple policies will largely suffice for our purposes.)

We say that a policy π is an "admissible control" for a given arrival rate vector λ if it satisfies the above conditions and there exists a unique stationary distribution for the headcount process under this policy. For an arrival rate vector λ and admissible control π, we define $\mathbb{E}D_j(\lambda, \pi)$ to be the expected time in queue for class j customers under the stationary distribution. (Expected values are all taken under the stationary distribution generated by a specified arrival rate vector λ and admissible control π.) Since customers are sensitive to delay, it is useful to think in terms of achievable delays instead of admissible controls. For an arrival rate vector λ, the set of *achievable delay vectors* is

$$\mathcal{D}(\lambda) = \{(d_1, \ldots, d_k) : d_j \geq \mathbb{E}D_j(\lambda, \pi),\ \pi \text{ is an admissible control}, j = 1, \ldots, k\}. \tag{5}$$

Note that we allow $d_j > \mathbb{E}D_j(\lambda, \pi)$, so a given service class may experience an overall delay *greater* than what can be attributed to system congestion alone. (The importance of this extra degree of freedom was first expounded on in Afèche (2013).) To operationalize this, we assume that the SP may "inject delay" after the service has been completed, so the server may begin processing a new customer while the injected delay is being imposed on the customer whose processing has just been completed. For example, a class j customer may be sent to an infinite-capacity "delay node" following service completion, where he is held for δ_j units of time and is then released from

the system; this will be referred to as "injected delay." Since customers perceive delay as time in system minus time in service, this delay node adds exactly δ_j to the *queueing delay*, $\mathbb{E}D_j$, for an "overall delay" of $d_j = \mathbb{E}D_j + \delta_j$. Note that with this convention, the headcount process $Z_j(t)$ is the number of class j customers in the system, *excluding* the delay node. As an aside, we note that this structure is most easily implemented in systems where service is not directly observed by the customer. Other methods to realize this "injected delay" include slowing down the processing rate, introducing appropriately timed server idleness, or adding a suitable delay node in the input queues.

Customer choice model and information structure. To communicate the key ideas of the paper in the simplest manner, the remainder of this section and §3-4 will consider a market comprised of two distinct customer segments (types), indexed by $i = 1, 2$; §5-6 will derive additional important insights that pertain to a market with $N > 2$ customer types. Customers of type i arrive at the system according to an independent Poisson process with rate Λ_i and may choose a service class to purchase or leave the system without service. Each arriving customer of type i has a willingness-to-pay V_i which is an i.i.d. draw from a distribution F_i. We assume that for each $i = 1, 2$ the (cumulative) distribution function F_i is concave, has a continuous density, an increasing generalized failure rate (IGFR), and a finite mean. Each type i customer incurs an additive linear delay cost of $\$c_i$ per unit of time spent waiting, and the parameter c_i is common across all type i customers.

Each customer seeks to maximize his individual utility. Upon arrival, a customer is informed of the k service classes, each having a per-access fee p_j and overall delay d_j. We assume that the queues themselves are *unobservable* to the customers. A customer of type i who is willing to pay V_i for completing service, computes his net utility for service class j as

$$U_i(j) = V_i - (p_j + c_i d_j), \tag{6}$$

and enters the service class that maximizes that utility, namely,

$$j^* = \text{argmax}_j\{U_i(j) : U_i(j) \geq 0,\ j = 1, \ldots, k\} \text{ with } j^* = 0 \text{ if } U_i(j) < 0 \text{ for all } j = 1, \ldots, k;$$

that is, $j = 0$ represents the no-purchase option. This type of behavior arises when service requests originate with atomistic and self-interested customers, who rationally decide how (and whether) to use the system. Customers who choose not to enter the system are lost and do not return.

We assume that the characteristics of each customer segment (Λ_i, c_i, F_i, and μ) are known to the SP, while the type $i \in \{1, 2\}$ and random valuation $V_i \sim F_i$ of any individual customer are *private information*, and thus unknown to the SP. Since the SP is unable to distinguish between customer types, he offers the same set of service classes to all customers.

Number of service classes offered. Observe that a customer of type i will select the service class with the minimum "full cost," given by $p_j + c_i d_j$, irrespective of his individual willingness-to-pay V_i. Therefore, all customers of type i will select the same service class. In a market with two customer types, the SP need only offer up to two service classes ($k \leq 2$) (or more generally, at most N classes if there are N customer segments). The resulting mean demand rate, governed by the *customer demand model*, for each of the two service classes is given by

$$\begin{aligned}\lambda_1(p_1,p_2,d_1,d_2) &= \Lambda_1\bar{F}_1(p_1+c_1d_1)\mathbf{1}\{p_1+c_1d_1 \leq p_2+c_1d_2\} \\ &\quad + \Lambda_2\bar{F}_2(p_1+c_2d_1)\mathbf{1}\{p_1+c_2d_1 < p_2+c_2d_2\}, \qquad (7)\\ \lambda_2(p_1,p_2,d_1,d_2) &= \Lambda_1\bar{F}_1(p_2+c_1d_2)\mathbf{1}\{p_2+c_1d_2 < p_1+c_1d_1\} \\ &\quad + \Lambda_2\bar{F}_2(p_2+c_2d_2)\mathbf{1}\{p_2+c_2d_2 \leq p_1+c_2d_1\}, \qquad (8)\end{aligned}$$

where $\bar{F}_i(\cdot) := 1 - F_i(\cdot)$ and $\mathbf{1}\{\cdot\}$ is the indicator function. We assume that if a customer of type i is indifferent between the two service classes, he will choose service class $j = i$. We note that the arrival process of customers into each service class is Poisson by the thinning property of Poisson processes.

System equilibrium. The queueing delays $(\mathbb{E}D_1, \mathbb{E}D_2)$ depend on the demand rates (λ_1, λ_2) and admissible control π, and, in turn, these demand rates depend, in part, on the queueing delays. An *equilibrium* for the system is an operating point where, for fixed prices, control policy, injected delays, and demand model, the congestion delays induce precisely the demand rates, and these in turn induce said delays.

Definition 1 (Equilibrium). Fix prices (p_1, p_2), a control policy π, injected delays (δ_1, δ_2), and a customer demand model $(\lambda_1, \lambda_2) = (\lambda_1(p_1, p_2, d_1, d_2), \lambda_2(p_1, p_2, d_1, d_2))$. The system admits an equilibrium if there exists a stationary probability distribution for the headcount process Z such that

$$d_j = \mathbb{E}D_j(\lambda_1, \lambda_2, \pi) + \delta_j \qquad j = 1, 2. \qquad (9)$$

Remark 1. We *do not* provide general conditions under which an equilibrium exists, but rather show in §4 that a unique equilibrium exists for the specific solution we propose to the following economic optimization problem.

Revenue maximization problem. The SP's problem is to find prices (p_1, p_2), a control policy π, and injected delays (δ_1, δ_2) to maximize the equilibrium revenue rate given by

$$R(\pi, p_1, p_2, \delta_1, \delta_2) = \sum_{j=1}^{2} p_j \lambda_j(p_1, p_2, d_1, d_2), \qquad (10)$$

where (d_1, d_2) are the overall delays in equilibrium (assuming that such an equilibrium exists), given in (9), and the customer demand model $\lambda_j(\cdot)$, $j = 1, 2$, is given in (7) and (8).

Afèche (2013) formulates the above as a mechanism design problem. Adopting Myerson's *revelation principle* (Myerson (1981)), it suffices to consider a *direct mechanism* where all customers report their private information (their type i and valuation V_i) to the SP. The SP then uses that information to determine which service class the customer purchases, if any. In order for arriving customers to truthfully report their types and valuations, the SP's mechanism needs to satisfy two sets of constraints:

- Incentive Compatibility: $p_i + c_i d_i \leq p_j + c_i d_j$ for all $j \neq i$,
- Individual Rationality: $\lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i)$ for $i = 1, 2$.

Note that this labeling assumes, without loss of generality, that type i customers are assigned to service class i or turned away. Myerson's revelation principle states that if the mechanism satisfies the incentive compatibility and individual rationality conditions above, then it is a *Nash equilibrium* for players to truthfully report their types and valuations. Moreover, there is no loss of generality in restricting our attention to direct mechanisms.

Adopting this mechanism design approach, the revenue maximization problem can be recast as follows. Find prices $p = (p_1, p_2)$ and control policy π to:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{2} p_i \lambda_i \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, 2 \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, 2 \\ & (d_1, d_2) \in \mathcal{D}(\lambda). \end{aligned} \tag{11}$$

The increasing generalized failure rate and finite mean properties of F_i ensure that an infinite price is not optimal (Lariviere (2006)). (This is a common assumption in the revenue management literature, but we note that weaker assumptions, e.g., that the functions $p\bar{F}_i(p)$ for $i = 1, 2$ are coercive, also suffice.) The reader should note that optimizing (11) over the space of admissible controls while taking into account the resulting equilibrium of the multi-server system is reasonably challenging, and a head-on treatment will likely require a brute-force computational analysis with limited insight. We denote the *supremum* of achievable revenues[1] over the feasible set of (11), by R_*. This can be viewed as the output of an oracle, and will serve as a benchmark against which we will compare the performance of an approximate solution to (11) that will be developed in the next section.

Discussion of modeling assumptions. Note that in the mechanism design formulation (11), the SP is not forced to offer two distinct service classes; the optimization problem allows both

[1] While it is possible to show that the supremum R_* may be achieved by a feasible solution, and hence an optimal solution exists, our subsequent analysis does not require this technical result.

classes to offer the same level of service, e.g., by pricing the "two options" equally and sequencing all customers through one queue that is served under a FIFO discipline. Additionally, in this problem setting the SP can only separate the customers in terms of their delay sensitivity preferences, not their willingness-to-pay; in other words the price may differ based on the customer's type i but not their willingness-to-pay V_i. This is a consequence of the structure of the underlying "products" that the SP can offer and the additive nature of the delay costs in each customer's net utility (cf. discussion before (7)-(8)); linearity of the delay cost is not required. We note that the formulation given in (11) can be extended to allow for multiple but distinct customer types, which is the setting of §5.

3 Deterministic Analysis

Our first step towards solving the mechanism design problem (11) is to solve a carefully chosen deterministic relaxation ("DR") of the latter. The DR preserves the essential economic and operational considerations of the SP's problem while ignoring the complications presented by the queueing dynamics and resulting equilibrium. The optimal solution to the DR will allow us to construct an approximate solution for the original, stochastic problem (11).

3.1 Deterministic Relaxation

The DR seeks prices (p_1, p_2) and delays (d_1, d_2) that

$$\begin{aligned} \text{maximize} \quad & p_1\lambda_1 + p_2\lambda_2 \\ \text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i,j = 1,2 \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,2 \\ & \lambda_1 + \lambda_2 \le s\mu \\ & d_1 \ge 0, d_2 \ge 0. \end{aligned} \tag{12}$$

That is, unlike in (11) where delays are endogenous, here they are *decision variables*. We only require that delays be non-negative and that total demand does not exceed the system capacity. It is in that precise sense that (12) is a (deterministic) relaxation of (11). (The objective and the incentive compatibility and individual rationality constraints remain as in (11).) This may appear overly simplistic, but, as we will see in the remainder of this paper, it captures the essential features of a near-optimal solution to the stochastic problem in (11). Observe that an optimal solution to (12) exists since the objective function is coercive and the feasible set is closed.

We will denote the optimal solution to (12) as $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ and set $\bar{\lambda}_i = \Lambda_i \bar{F}_i(\bar{p}_i + c_i\bar{d}_i)$, $i = 1, 2$. We also define the *relative workload contribution* in each class at the optimal solution as

$$\bar{\kappa}_i = \frac{\bar{\lambda}_i}{s\mu} \qquad i = 1, 2. \tag{13}$$

This is the fraction of "processing capacity" consumed by class i in the DR solution.

Table 1: Categorization of DR solutions.

	capacitated	uncapacitated
undifferentiated	$\bar{p}_1 = \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$	$\bar{p}_1 = \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$
differentiated	$\bar{p}_1 > \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$	$\bar{p}_1 > \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$

3.2 Characterization of the DR Solution

Our first structural result states that within the DR setting, it is best to have type 1 customers wait "as little as possible," i.e., $\bar{d}_1 = 0$, and to make type 2 customers wait "only long enough" to satisfy incentive compatibility, i.e., $\bar{p}_1 = \bar{p}_2 + c_1\bar{d}_2$.

PROPOSITION 1 **(Structure of the DR solution)**. *Let* $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ *be the optimal solution to the deterministic relaxation* (12)*. Then*

(a) $\bar{d}_1 = 0$*, and*

(b) $\bar{p}_1 = \bar{p}_2 + c_1\bar{d}_2$.

The main intuition here is that the SP earns revenue from fees but not delays. Therefore, a feasible solution (p_1, p_2, d_1, d_2) to the DR cannot be optimal if it is possible to maintain the same full cost in a service class while reducing the delay and increasing the price, since this would increase revenues while ensuring feasibility. This suggests that the sole purpose of imposing non-zero delay in class 2 is to segment the market.

We propose the following categorization and nomenclature for the DR solution, summarized in Table 1. If $\bar{p}_1 = \bar{p}_2$ we say that the solution is "undifferentiated," and if $\bar{p}_1 > \bar{p}_2$ we say it is "differentiated."[2] If $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ we say that the solution is "capacitated," and if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ we say it is "uncapacitated" (since the two cases refer to the DR solutions for which the capacity constraint in (12) is either binding or not).

With this in mind, we first answer the question of when the DR solution is differentiated. Consider the following deterministic relaxation of the "single-product problem," in which the SP is constrained to offering only one service class:

$$\max_p \left\{ p(\Lambda_1 + \Lambda_2)\bar{G}(p) : (\Lambda_1 + \Lambda_2)\bar{G}(p) \le s\mu \right\}, \tag{14}$$

where $\bar{G}(p) = 1 - G(p)$, and $G(p)$ is the *aggregate* willingness-to-pay distribution with density $g(p)$,

$$G(p) := \frac{\Lambda_1 F_1(p) + \Lambda_2 F_2(p)}{\Lambda_1 + \Lambda_2}, \qquad g(p) := \frac{\Lambda_1 f_1(p) + \Lambda_2 f_2(p)}{\Lambda_1 + \Lambda_2}. \tag{15}$$

[2]Note that if $\bar{p}_1 > \bar{p}_2$ and $\bar{\kappa}_2 = 0$, then $(\bar{p}_1, \bar{p}_1)$ is also a solution to the DR, and so the problem essentially reduces to a single product with a single market segment. Therefore we assume that any solution with $\bar{\kappa}_2 = 0$ is also "undifferentiated."

We assume that $G(\cdot)$ is strictly IGFR and therefore there is a unique maximizer of the single-product problem, which we denote by $\hat{p}$. Observe that if the optimal solution to the DR (12) is undifferentiated ($\bar{p}_1 = \bar{p}_2$), then the optimal solution to the single-product problem (14) must be $\hat{p} = \bar{p}_1 = \bar{p}_2$. In that case, no revenue is lost in restricting the SP to a single service class in the DR setting.

In Proposition 2 below we provide a necessary and sufficient condition for a differentiated solution, expressed in terms of demand elasticity[3]at the single-product optimal price $\hat{p}$. Let $\epsilon_i(p_1, p_2)$ be the elasticity of type i demand for class i service at prices (p_1, p_2), $i = 1, 2$, and let $\epsilon_g(p)$ be the elasticity of the aggregate demand for a single service class at price p:

$$\epsilon_1(p_1, p_2) = \frac{p_1 f_1(p_1)}{\bar{F}_1(p_1)}, \qquad \epsilon_2(p_1, p_2) = (1-c)\,\frac{p_2 f_2(cp_1 + (1-c)p_2)}{\bar{F}_2(cp_1 + (1-c)p_2)}, \qquad \epsilon_g(p) = \frac{pg(p)}{\bar{G}(p)}, \tag{16}$$

where $c := c_2/c_1 < 1$.

PROPOSITION 2 **(Conditions for service differentiation)**. *Assume that $G(\cdot)$ is strictly IGFR. Let $\hat{p}$ be the optimal solution of the single-product problem* (14)*, and let $\bar{p}_1, \bar{p}_2$ be the optimal prices of the deterministic relaxation* (12)*. Then*

$$\bar{p}_1 > \bar{p}_2 \quad \textit{if and only if} \quad \epsilon_2(\hat{p}, \hat{p}) > \epsilon_g(\hat{p}). \tag{17}$$

Differentiated services should be offered *if and only if* the demand elasticity for type 2 (delay-insensitive) customers at $\hat{p}$ is greater than the aggregate demand elasticity at that price. In that case, the SP may increase revenues by lowering the price for type 2 customers. Note that it must be elastic relative to the aggregate demand (as opposed to simply having an elasticity which is greater than 1), to account for the fact that any reduction in price must be matched by an increase in delays, in order to maintain incentive compatibility.

3.3 Translating the DR Solution

The DR solution provides fundamental insight on how to specify the number of service classes k, the prices for each, the control policy π, and how much injected delay (δ_1, δ_2), if any, is needed. This is summarized in Figure 1 and further interpreted below. The number of services classes and their respective prices are taken from the DR solution itself, while Proposition 1 indicates how to form the control policy, and when/whether to inject delay. When two service classes are offered, the prescribed solution gives strict preemptive priority to class 1, capturing the intuition that class 1 delays are targeted to be as small as possible.

[3]Recall that the demand elasticity at a price $p = (p_1, p_2)$ is the proportional change in demand due to a change in price:

$$\epsilon_i(p) = -\frac{p_i}{\lambda_i}\frac{\partial \lambda_i}{\partial p_i}.$$

Demand is *elastic* at p if $\epsilon(p) > 1$ in which case reducing the price will increase revenue; demand is *inelastic* at p if $\epsilon(p) < 1$ in which case increasing the price will increase revenue.

Figure 1: Prescribed stochastic solution based on DR for $N = 2$ customer types.

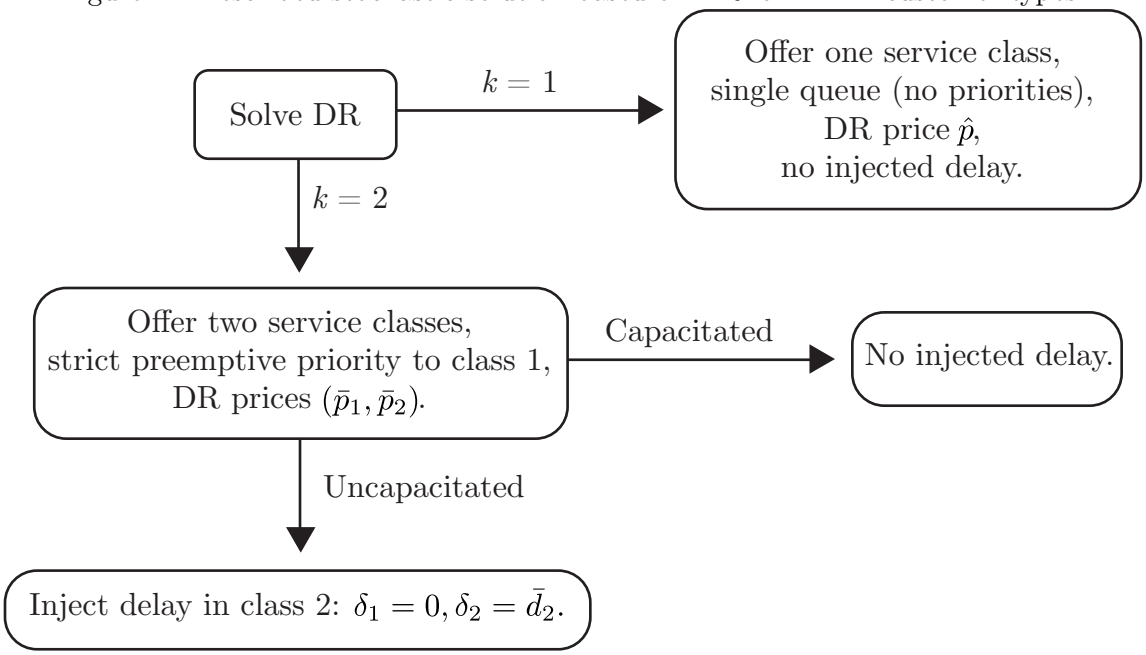


Figure 1 contains an important insight that will be justified in the stochastic analysis in the next section. Namely, injected delay is applied to class 2 only out of necessity, when the DR solution is differentiated and uncapacitated. If the DR solution is capacitated we expect, and indeed show in Theorem 1, that the natural congestion in the system will cause sufficient queueing delays in class 2 to satisfy the incentive compatibility condition. If the DR solution is uncapacitated, class 2 will also face a system operating at a low utilization rate and experience insignificant queueing delay. In that case, the SP injects delay to ensure that class 2 experiences $\bar{d}_2$ delay, needed to optimally segment the market. This is the key to ensuring that type 1 (delay-sensitive) customers have an incentive to pay a premium for high-priority service.

Henceforth, we will explicitly distinguish between the "DR solution" to (12) and its implementation in the stochastic system, described in Figure 1, which will be referred to as the "stochastic solution." We will also port the nomenclature in Table 1 to the stochastic setting. We call the stochastic solution "differentiated" if it offers two service classes and "undifferentiated" if it offers a single service class. With some abuse of terminology, we call the queueing system operating under the stochastic solution "capacitated" ("uncapacitated") if the underlying DR solution is capacitated, $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ (uncapacitated, $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$). Of course, the equilibrium traffic intensity in the queueing system under the stochastic solution is always less than 1.

4 Performance Analysis

4.1 Preliminaries

This section proves that the stochastic solution prescribed above achieves near-optimal performance in the stochastic system, and induces an equilibrium and operating regime that is consistent with the DR solution. To make this rigorous, we will focus on large-scale systems that are characterized by large processing capacities and large market potential. Specifically, we will consider a sequence of systems with increasing capacity and market potential, indexed by n:

$$\begin{aligned} s^n &:= n, \\ \Lambda_i^n &:= n\hat{\Lambda}_i, \quad i = 1, 2, \end{aligned} \tag{18}$$

with $\hat{\Lambda}_i := \Lambda_i/s$; note that at $n = s$ we recover the system with s servers and market potential Λ_i of original interest. Note that the size of each customer segment Λ_i^n scales with capacity, but the valuation distribution $F_i(\cdot)$ is held fixed. We will use a superscript n to index quantities that depend on the size of the system. Under this scaling, the nth system may be described by the queue length and headcount processes:

$$\begin{aligned} Q_j^n(t) &= Z_j^n(t) - \pi_j^n(t) && \text{for } j = 1, \ldots, k, \\ Z_j^n(t) &= N_j^{(a)}\left(\lambda_j^n t\right) - N_j^{(s)}\left(\int_0^t \mu \pi_j^n(s)\, ds\right) && \text{for } j = 1, \ldots, k, \end{aligned}$$

defined for a given control policy $\pi^n(t) = (\pi_1^n(t), \ldots, \pi_k^n(t))$ and arrival rate vector $\lambda^n = (\lambda_1^n, \ldots, \lambda_k^n)$.

For the nth system in the sequence, we formulate a revenue maximization problem, analogous to (11), where the quantities with superscript n replace their counterparts in (11); that is, we are analyzing the mechanism design optimization problem (11), but for a system with capacity and market potential scaled up proportionally by a factor n. We will denote by R_*^n the supremum of (11) for the nth problem. The policy prescribed in §3.3 can be applied to each system of size n as follows.

Undifferentiated DR solution (single class). If $\bar{p}_1 = \bar{p}_2 = \hat{p}$, offer a single service class ($k = 1$) at price $\hat{p}$ with no injected delay. The arrival rate into the single class is

$$\lambda^n = \Lambda_1^n \bar{F}_1(\hat{p} + c_1 d^n) + \Lambda_2^n \bar{F}_2(\hat{p} + c_2 d^n),$$

where d^n is the overall delay under the control policy $\pi^n(t) = \min\{s^n, Z^n(t)\}$. (The single-class problem is addressed in Maglaras and Zeevi (2003a,b).)

Differentiated DR solution (two classes). For the remainder of this section, we will focus on the case where the DR solution is differentiated, when necessary distinguishing between the capacitated and uncapacitated cases. If $\bar{p}_1 > \bar{p}_2$, the stochastic solution has two service classes ($k = 2$) at prices $(\bar{p}_1, \bar{p}_2)$ with injected delays $(0, \delta_2)$, where $\delta_2 = \bar{d}_2$ if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$, and $\delta_2 = 0$ otherwise.

Our first result studies a simplified setting where the arrival rate into each class is

$$\lambda_j^n = \Lambda_j^n \bar{F}_j(\bar{p}_j + c_j d_j^n), \quad \text{for } j = 1, 2, \tag{19}$$

and d_j^n is the overall delay in class j, under the control policy

$$\pi_1^n(t) = \min\{s^n, Z_1^n(t)\}, \quad \pi_2^n(t) = \min\{(s^n - Z_1^n(t))^+, Z_2^n(t)\}.$$

We denote by $\rho_j^n = \lambda_j^n / n\mu$ the *traffic intensity* in class $j = 1, 2$.

In (19) we explicitly *assume* that customers choose the "correct" service class, or equivalently, report their type truthfully. The following proposition shows that, under this assumption, the prescribed solution yields a unique equilibrium for each system in the sequence. Furthermore, the sequence of equilibria (i.e., the traffic intensities (ρ_1^n, ρ_2^n) and overall delays (d_1^n, d_2^n) induced by these prices, priority rule, and injected delays) converges to the DR solution.

PROPOSITION 3 **(System equilibrium)**. *Assume the scaling in* (18)*. Set the stochastic solution to be prices* $(\bar{p}_1, \bar{p}_2)$ *and injected delays* (δ_1, δ_2)*, together with the sequencing rule* π *prescribed in §3.3. Then, assuming* (19) *the following are true:*

(a) for every n, there exists a unique system equilibrium $(\rho_1^n, \rho_2^n, d_1^n, d_2^n)$*;*

(b) as $n \to \infty$*,* $\rho_j^n \to \bar{\kappa}_j$ *and* $d_j^n \to \bar{d}_j$*, for* $j = 1, 2$*.*

4.2 Incentive Compatibility and Revenue Optimality

Our next two theorems show that, in large systems, the stochastic solution derived in §3.3 is incentive compatible, i.e., it is a Nash equilibrium strategy for the customers to indeed choose the "correct" service classes as prescribed by the solution of the DR. Moreover, the proposed stochastic solution achieves near-optimal revenues. Note that Theorems 1-3 *do not* assume (19), but rather that the arrival rates λ_1^n and λ_2^n are determined by the atomistic customer choice models described by (7)-(8), under the scaling (18) and the stochastic solution composed of prices $(\bar{p}_1, \bar{p}_2)$, injected delays (δ_1, δ_2), and sequencing rule π prescribed in §3.3. Let $R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2)$ be the revenue rate in the nth system generated by this solution, namely,

$$R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2) = \bar{p}_1 \lambda_1^n + \bar{p}_2 \lambda_2^n.$$

THEOREM 1 **(Incentive compatibility)**. *Assume the scaling in* (18)*. Then, there exists a finite* N_{ic} *such that for all* $n \geq N_{ic}$*, the stochastic solution composed of prices* $(\bar{p}_1, \bar{p}_2)$*, injected delays* (δ_1, δ_2)*, and sequencing rule* π *prescribed in §3.3 is incentive compatible, namely*

$$\bar{p}_i + c_i d_i^n \leq \bar{p}_j + c_i d_j^n, \qquad i, j = 1, 2 \text{ and } i \neq j,$$

where d_j^n*,* $j = 1, 2$*, are the overall delays arising in the nth system in equilibrium.*

We emphasize that incentive compatibility is achieved for a *finite* sized system, i.e., for all systems in the sequence above the threshold N_{ic}, customers will choose the correct service class (in equilibrium).

THEOREM 2 **(Revenue optimality)**. *Assume the scaling in* (18)*. Then, the revenue rate* $R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2)$ *generated by the stochastic solution composed of prices* $(\bar{p}_1, \bar{p}_2)$*, injected delays* (δ_1, δ_2)*, and sequencing rule* π *prescribed in §3.3, satisfies*

$$R^n_* - R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2) \leq M, \qquad \text{for all } n \geq N_{ic},$$

for some finite positive constant M*, and* N_{ic} *as in Theorem 1. (*R^n_* *is the supremum revenue of the original mechanism design problem* (11) *for the scale-n system.)*

Theorem 2 is an uncharacteristically strong optimality result. Given that the DR is, in some sense, only a fairly crude (first-order) approximation of the mechanism design problem (11), we expect that the policy predicated on the DR would lead to a performance gap, in terms of revenue, that increases with system size. This asymptotic gap for policies based on deterministic analysis often grows proportionally to $\sqrt{n}$, which is the magnitude of the stochastic fluctuations not captured by the DR. Moreover, the optimality gap in settings where the $\sqrt{n}$ behavior has also been optimized will typically diverge with n, but at a slower rate. Surprisingly, our result shows that the optimality gap of the policy derived via the static DR *remains bounded*, regardless of the volume of workflow and scale of revenues. The DR solution has thus essentially optimized the original (stochastic) mechanism-design problem.

The result of Theorem 2 can be partially explained by Proposition 3 part (b). To fully deconstruct what underlies the strength of Theorem 2 requires a more careful examination of the *rate of convergence* of traffic intensities and delays, which is provided in §4.3.

4.3 System Operating Regime and Its Implications

The operating regime of a single-class multi-server queue can be naturally characterized by focusing on the probability that an arriving customer will have to wait prior to commencing service:

- $\mathbb{P}(\text{waiting time} > 0) \approx 0$: "quality driven" (QD) regime (focus on providing high-quality service).
- $\mathbb{P}(\text{waiting time} > 0) \approx 1$: "efficiency driven" (ED) regime (focus on efficient use of resources).
- $\mathbb{P}(\text{waiting time} > 0) \approx \nu \in (0, 1)$: "quality and efficiency driven" (QED) regime.

The celebrated work of Halfin and Whitt (1981) showed that these regimes are equivalently characterized by the system's traffic intensity. Specifically, the QED regime, where the probability of having to wait for service is modest, i.e., neither "never" nor "always," arises if and only if

$\rho^n = 1 - \beta/\sqrt{n}$ for some $0 < \beta < \infty$. This corresponds to the well-known "heavy-traffic" regime that has been studied extensively in the queueing literature. The ED regime operates at still higher asymptotic utilization rates, while the QD regime corresponds to lower utilization rates. The next theorem characterizes the operating regime that arises in our context as a consequence of the economic objectives in (11).

THEOREM 3 **(System operating regimes)**. *Assume the scaling in* (18)*, and consider the stochastic solution composed of prices* $(\bar{p}_1, \bar{p}_2)$*, injected delays* (δ_1, δ_2) *and sequencing rule* π *prescribed in §3.3. Then,*

(a) if the DR solution in (12) *is capacitated,* $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$*, then the traffic intensity in the stochastic system is*

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \textit{and} \quad \rho_2^n = \bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n),$$

and the system operates in the ED regime, namely,

$$\rho_1^n + \rho_2^n = 1 - \frac{\alpha}{n} + o(1/n),$$

where α *is a finite positive constant that depends on model primitives;*

(b) if the DR solution in (12) *is uncapacitated,* $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$*, then*

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \textit{and} \quad \rho_2^n = \bar{\kappa}_2 + o(1/n),$$

and the system operates in the QD regime.

Note that even if the system is capacitated, class 1 never experiences any significant delay since they receive static priority, and $\bar{\kappa}_1 < 1$ (that class is effectively facing an underutilized system operating in the QD regime). Class 2, on the other hand, experiences a system whose resources operate in the so-called ED regime, which results in significant congestion delays for the low priority class. This congestion is experienced entirely by class 2 customers, thereby achieving the necessary service differentiation. If the system is uncapacitated, delay is injected since there is essentially "not enough" endogenous congestion in the system to give rise to non-vanishing delay in class 2.

Discussion. Unlike the bulk of the literature on asymptotic analysis of queueing systems, where the operating regime is *imposed a priori* for analysis purposes, in our work this regime arises as a *consequence* of economic optimization considerations. In fact, as evidenced from our results, if the traditional QED-type regime were imposed, it would be strictly sub-optimal in the context of our problem. To put our results in further perspective, let us first contrast them with Maglaras and Zeevi (2003a) who showed that the QED regime emerges as a direct consequence of economic optimization (revenues or welfare) when all customers have the same delay sensitivity parameter

(single type). Our results are complementary and, taken together, show that each of the three asymptotic operating regimes outlined above may arise when we allow for multiple service classes: i) in a capacitated system, a single-class stochastic solution gives rise to the QED regime; ii) a two-class stochastic solution in a capacitated system places class 1 in the QD regime and class 2 in the ED regime; and iii) in the uncapacitated case all classes operate in the QD regime and injected delay is required to differentiate the two service classes. Emphasizing the last point, injected delay plays a fundamental role in a large scale system only if it has ample capacity, and, as such, suggests that in systems where capacity has been optimized, injected delay will have a vanishing effect.

Theorem 3 also provides the key analytical foundations that support Theorem 2 (revenue optimality). Note that in the capacitated case

$$\begin{aligned} R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2) &= \bar{p}_1\lambda_1^n + \bar{p}_2\lambda_2^n = n\mu\left(\bar{p}_1\rho_1^n + \bar{p}_2\rho_2^n\right), \\ &= n\mu\left(\bar{p}_1(\bar{\kappa}_1 + o(1/n)) + \bar{p}_2\left(\bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n)\right)\right), \\ &= n\mu(\bar{p}_1\bar{\kappa}_1 + \bar{p}_2\bar{\kappa}_2) + n\mu\left(\bar{p}_1 o(1/n) - \bar{p}_2\frac{\alpha}{n} + \bar{p}_2 o(1/n)\right), \\ &= n\bar{R} - \mu\bar{p}_2\alpha + o(1), \end{aligned} \tag{20}$$

where $\bar{R}$ is the value of the DR (12) under the optimal solution $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$. Since the DR (12) is a relaxation of the original mechanism design problem (11), the supremum revenue in the DR dominates the supremum revenue in the original problem. Moreover, the DR has an optimal solution (i.e., achieves its supremum), and thus $n\bar{R} \geq R_*^n$, where R_*^n is the supremum revenue for the nth system. In the uncapacitated case, ρ_2^n converges at rate $o(1/n)$ in the QD regime, so the stochastic solution will provide revenues that are close, in absolute dollars, to the optimum.

Remark 2 (The single-server system model). The deterministic relaxation (12) applies to a single-server system model, and the stochastic solution outlined in §3.3 is well-defined in that setting as well. The scaling in that model would increase the speed of the single server $\mu^n = n\mu$ together with potential demand Λ_i^n, $i = 1, 2$, defined as in (18). If the stochastic solution is differentiated, one can show that delays in class 1 and class 2 will still converge to $(0, \bar{d}_2)$, but a different analysis is needed to establish the eventual satisfaction of the incentive compatibility property.

5 The Essential Role of Injected Delay

We now consider the more general problem with $N \geq 3$ customer types. Apart from providing an important technical extension, which seems intractable using direct analysis but is possible within our framework, it turns out that the analysis of the multi-type model offers new insights that are obscured in the simpler two-type setting, in particular, concerning the importance of injected delay.

5.1 Analysis of the Deterministic Relaxation

The problem formulation in §2 is easily extended to N customer types with linear delay costs $c_1 > c_2 > \cdots > c_N$, valuation distributions $F_i(\cdot)$, and potential demand Λ_i, $i = 1, \ldots, N$. The mechanism design problem is then to find prices $(p_1, \ldots, p_N)$, a control policy π, and the injected delay prescription $(\delta_1, \ldots, \delta_N)$ that maximize revenues. We start by solving the following DR, which is the analogue of (12):

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} p_i \lambda_i \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, \ldots, N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, \ldots, N \\ & \sum_{i=1}^{N} \lambda_i \leq s\mu \\ & d_i \geq 0 \quad i = 1, \ldots, N. \end{aligned} \tag{21}$$

With two customer types, the solution to (12) either assigned both types to a single service class, or each type to its own service class. In the more general setting with N types, the solution to (21) may give rise to $k \leq N$ distinct service classes. With this in mind, we denote the optimal solution to (21), indexed by customer type, $\bar{p} = (\bar{p}_1, \ldots, \bar{p}_N)$ and $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_N)$. Equivalently, we may write the DR solution in terms of the distinct service classes, namely $\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$ and $\hat{d} = (\hat{d}_{(1)}, \ldots, \hat{d}_{(k)})$, along with k sets $\{A_{(1)}, \ldots, A_{(k)}\}$, where $A_{(j)}$ is the set of all customer types that choose class j. That is, $\bar{p}_i = \hat{p}_{(j)}$ for all $i \in A_{(j)}$. We will call the sets $A_{(j)}$, $j = 1, \ldots, k$, "market segments."

The generalization of Proposition 1, which describes the structure of the optimal solution of the DR in the two-type problem, is given by the following result.

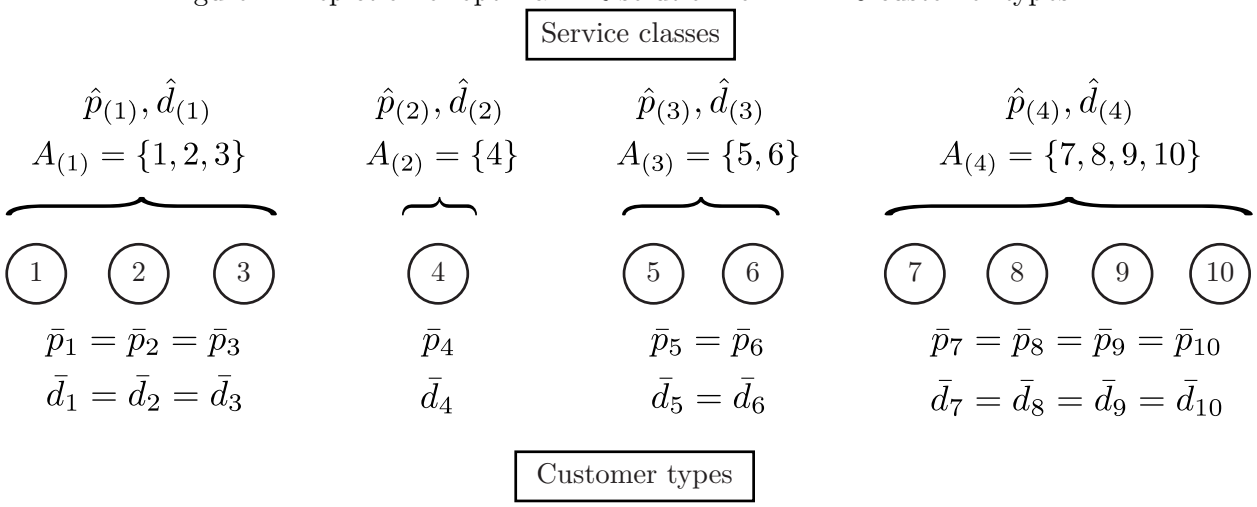
PROPOSITION 4 **(Structure of the multi-type DR solution)**. *Let $\bar{p}$, $\bar{d}$ be an optimal solution to the DR* (21)*. Then*

(a) $\bar{d}_1 = 0$ *and* $\bar{p}_i + c_i \bar{d}_i = \bar{p}_{i+1} + c_i \bar{d}_{i+1}$*, for* $i = 1, \ldots, N-1$.

(b) Recall that types are labelled in decreasing order of their delay sensitivity parameters, i.e., $c_1 > c_2 > \cdots > c_N$*. The market segments* $A_{(j)}$*,* $j = 1, \ldots, k$ *are contiguous in the following sense*

$$\begin{aligned} A_{(1)} &= \left\{1, \ldots, |A_{(1)}|\right\}, \\ A_{(2)} &= \left\{|A_{(1)}| + 1, \ldots, |A_{(1)}| + |A_{(2)}|\right\}, \\ &\vdots \\ A_{(k)} &= \left\{\textstyle\sum_{j=1}^{k-1} |A_{(j)}| + 1, \ldots, N\right\}. \end{aligned}$$

Figure 2: Depiction of optimal DR solution for $N = 10$ customer types.



Note. This DR solution specifies $k = 4$ service classes, where $\hat{p}_{(j)}$ and $\hat{d}_{(j)}$ denote the price and delay, respectively, of service class j and $A_{(j)}$ denotes the segment of customer types that choose service class j.

Part (a) suggests that delays should be large enough to satisfy incentive compatibility, but not larger. Part (b) shows that the market segment $A_{(j)}$, is composed of *consecutive* customer types. An example with $N = 10$ customer types and $k = 4$ service classes, along with the associated DR solution $\bar{p}, \bar{d}$ and $\hat{p}, \hat{d}, \{A_{(1)}, \ldots, A_{(4)}\}$ is shown in Figure 2. (In what follows, we assume access to the solution of (21).)

5.2 Prescribed Solution for the Stochastic System

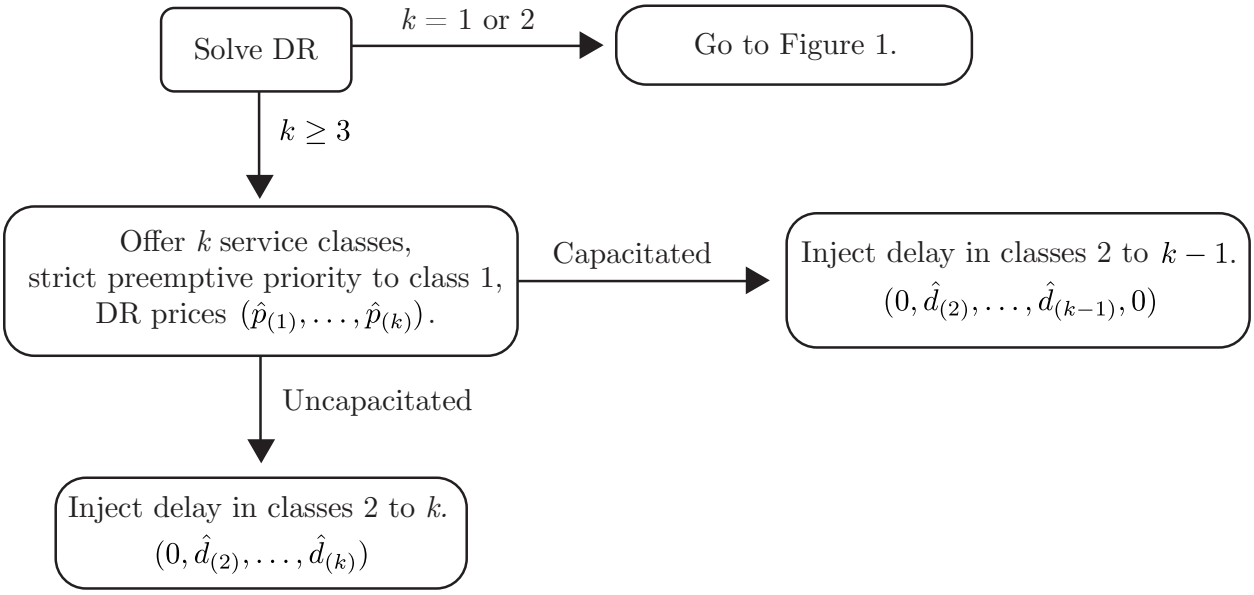
Assume that the optimal solution to the DR (21) offers k service classes at prices $\hat{p}_{(1)} > \hat{p}_{(2)} > \cdots > \hat{p}_{(k)}$ and delays $\hat{d}_{(k)} > \cdots > \hat{d}_{(2)} > \hat{d}_{(1)} = 0$, with market segments $A_{(1)}, \ldots, A_{(k)}$. For the DR solution, we define the relative workload contribution from class j to be

$$\hat{\kappa}_{(j)} := \frac{\sum_{i \in A_{(j)}} \Lambda_i \bar{F}_i(\hat{p}_{(j)} + c_i \hat{d}_{(j)})}{s\mu}$$

and, following terminology established in §3, we say that the DR solution is *capacitated* if $\sum_{j=1}^{k} \hat{\kappa}_{(j)} = 1$ and *uncapacitated* otherwise.

We start by specifying the prescribed stochastic solution in the case $k \geq 3$; there are k classes of service and our prescription sets prices $\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$. Classes are served using a strict preemptive priority rule, giving highest priority to class 1 and lowest to class k. Injected delays are given by $\delta = (\delta_{(1)}, \ldots, \delta_{(k)})$, where: $\delta_{(1)} = 0$; $\delta_{(j)} = \hat{d}_{(j)}$ for $j = 2, \ldots, k-1$; $\delta_{(k)} = \hat{d}_{(k)}$ if the system is uncapacitated; and $\delta_{(k)} = 0$ otherwise. If $k = 1$, there is only a single class priced at $\hat{p}_{(1)}$; no priorities or injected delays are needed. If $k = 2$, there are two service classes with prices

Figure 3: Prescribed stochastic solution based on DR for $N > 2$ customer types.



$(\hat{p}_{(1)}, \hat{p}_{(2)})$; service is sequenced according to a strict preemptive priority rule; and injected delays are given by $\delta_{(1)} = 0$, and $\delta_{(2)} = \hat{d}_{(2)}$ if the system is uncapacitated or $\delta_{(2)} = 0$ otherwise.

Necessity of injected delay. Note that if $k \geq 3$ the prescribed stochastic solution will *always inject delay* in some of the service classes, *irrespective* of whether the system is capacitated or uncapacitated. As will be shown in Theorem 4, this results in near-optimal performance. In contrast, when $k = 2$ no injected delay is needed in a capacitated system, since delays arise endogenously as a result of congestion. In the multi-class setting all classes outside the lowest priority do not experience measurable delays as a result of congestion, hence the necessity of injected delay to optimize the service offering.

To move forward with the stochastic analysis of this solution, we first apply the scaling in (18) to all customer types $i = 1, \ldots, N$. Then, in the nth system in the sequence, the demand for each class j is given by

$$\begin{aligned}\gamma^n_{(j)} = &\sum_{i \in A_{(j)}} \Lambda^n_i \bar{F}_i(\hat{p}_{(j)} + c_i d^n_{(j)}) \mathbf{1}\{\hat{p}_{(j)} + c_i d^n_{(j)} \leq \hat{p}_{(\ell)} + c_i d^n_{(\ell)} \ \text{ for all } \ell = 1, \ldots, k\} \\ &+ \sum_{i \notin A_{(j)}} \Lambda^n_i \bar{F}_i(\hat{p}_{(j)} + c_i d^n_{(j)}) \mathbf{1}\{\hat{p}_{(j)} + c_i d^n_{(j)} < \hat{p}_{(\ell)} + c_i d^n_{(\ell)} \ \text{ for all } \ell \neq j\},\end{aligned}$$

where d^n is the overall delay vector under the control policy $\pi^n(t) = (\pi^n_1(t), \ldots, \pi^n_k(t))$,

$$\pi^n_1(t) = \min\{s^n, Z^n_1(t)\}, \qquad \pi^n_j(t) = \min\left\{\left(s^n - \textstyle\sum_{\ell=1}^{j-1} Z^n_\ell(t)\right)^+, Z^n_j(t)\right\}, \ j = 2, \ldots, k.$$

The revenue earned in the nth system under our solution is

$$R^n(\pi, \hat{p}, \delta) = \bar{p}_{(1)}\gamma^n_{(1)} + \cdots + \bar{p}_{(k)}\gamma^n_{(k)},$$

and let R_*^n denote the the supremum over revenues earned under any feasible solution to the N-type mechanism design problem which generalizes (11).

THEOREM 4 **(Incentive compatibility and revenue optimality)**. *Under the scaling described above,*

(a) there exists a finite $N_{ic} > 0$ such that for all $n \geq N_{ic}$, the incentive compatibility conditions are satisfied

$$\hat{p}_{(j)} + c_i d^n_{(j)} \leq \hat{p}_{(\ell)} + c_i d^n_{(\ell)} \qquad \text{for all } i \in A_{(j)} \text{ and } j, \ell = 1, \ldots, k, \text{ with } \ell \neq j;$$

(b) if the stochastic solution is differentiated, $k \geq 2$, then there exists a finite positive constant M such that

$$R_*^n - R^n(\pi, \hat{p}, \delta) \leq M \qquad \text{for all } n \geq N_{ic}.$$

REMARK 3 (CONNECTION TO AFÈCHE (2013)). Our paper leverages the formulation of Afèche (2013), extending that framework to the more complex setting of the multi-server and multi-type systems. Moreover, our treatment strengthens prior findings and provides additional insight in several dimensions. In particular, we are able to separate the questions of when a revenue-maximizing SP should offer multiple service classes and how the system should be operated, including possibly injecting delay. We are also able analyze the revenue-maximization problem for more than two customer types, with the rather surprising insight that in the multi-class setting injected delays are needed to guarantee optimal revenues. In particular, for two customer types (Figure 1) and a "capacitated" system, injected delay is not a first-order effect and is insignificant in large systems. Since we expect that two-class systems where capacity has been optimized will never be uncapacitated, injected delay will play a minor role in such settings. When we extend our method to more than two customer types (Figure 3), it is revealed that injected delay may be a first-order effect that remains significant in large systems, even those with optimized capacity.

A partial extension to multiple customer types can also be found in Afèche and Pavlin (2011) and Katta and Sethuraman (2005), which show that some pooling of customer types may occur and, in the former, some form of injected delay may be necessary. However, the results in both of those works require strong, restrictive assumptions on the structure of customer valuations and delay costs. By contrast, the N-type model presented in this section is a seamless extension of the two-type setting and our results and insights generalize directly.

REMARK 4 (AN ALTERNATIVE IMPLEMENTATION). Is it possible to achieve the same degree of delay differentiation if $k \geq 3$ without the use of injected delay in a capacitated system? While the answer is affirmative, the resulting heuristic may not be desirable. For example, consider a structure with two priority lanes. Users that select the most expensive service class $\hat{p}_{(1)}$ get

assigned to the high priority queue and experience negligible delay. Users that select the cheapest class $\hat{p}_{(k)}$ get assigned the second (low) priority queue. Users that select any intermediate service class $\hat{p}_{(j)}$ get assigned to the high priority queue with probability $1 - \hat{d}^n_{(j)}/\hat{d}^n_{(k)}$ and to the low priority queue with probability $\hat{d}^n_{(j)}/\hat{d}^n_{(k)}$. One can verify that this policy is incentive compatible and results in near-optimal revenues, in the sense of Theorem 4. However, while the *average* delays in the intermediate service classes are asymptotically optimal, this policy would subject those customers to either very long delays or no delay at all, a quality that makes it less desirable from an operational standpoint. While this demonstrates that the solution to the DR may have multiple implementations in the stochastic setting, we believe that the one provided in §5.2 is the most natural and efficient interpretation of the DR solution.

6 Contrast with Mendelson-Whang's Socially Optimal Solution

In the welfare-maximization problem, the SP seeks to find prices $(p_1, \ldots, p_N)$ and a policy π that maximize the overall welfare in the system (net utility to customers plus revenue to the SP). As with the revenue maximization objective in (11), this can be reformulated as a mechanism design problem:

$$\begin{aligned} \text{maximize} \quad & W(p,d) = \sum_{i=1}^{N} \Lambda_i \left(\int_{p_i + c_i d_i}^{\infty} v f_i(v)\, dv - c_i d_i \bar{F}_i(p_i + c_i d_i) \right) \\ \text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i,j = 1,\ldots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,\ldots,N \\ & d \in \mathcal{D}(\lambda). \end{aligned} \tag{22}$$

Here, the objective is to maximize the sum of the total value generated per customer type, integrated over the corresponding subsets that purchase service, minus their total delay costs. Pricing transfers are "internal" in this formulation and do not affect welfare.

Mendelson and Whang (1990) offered a complete analysis of this problem for a system modeled as an $M/M/1$ queue. Their main insights were: i) the SP should offer N service classes, i.e., one for each customer type; ii) the optimal prices are equal to the externality costs for each class; and iii) resulting equilibrium delays arise naturally as the result of system congestion under a strict priority rule that strives to minimize the total delay costs (the "$c\mu$-rule"). A relatively simple variation of their arguments in the $M/M/1$ context can be applied in the multi-server setting of our paper to re-establish i)-iii).

First, consider the following deterministic relaxation (DR) of the social welfare optimization problem (22):

$$\begin{aligned} \text{maximize} \quad & W(p,d) \\ \text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i,j = 1,\dots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,\dots,N \\ & \sum_{i=1}^{N} \lambda_i \le s\mu \\ & p_i \ge 0, d_i \ge 0 \quad i = 1,\dots,N. \end{aligned} \tag{23}$$

Assume that the model primitives are such that an optimal solution to (23) satisfies $\lambda_i > 0$ for all $i = 1,\dots,N$. In this case, Proposition 5 below shows that the optimal solution is *unique* and *undifferentiated*.

PROPOSITION 5 **(Solution to the Social Welfare DR)**. *If $p_1, d_1, \dots, p_N, d_N$ is an optimal solution to the social-welfare DR* (23) *such that $\lambda_i > 0$ for all $i = 1,\dots,N$, then $p_1 = \cdots = p_N = \hat{p}_{soc}$, and $d_1 = \cdots = d_N = 0$, where*

$$\hat{p}_{soc} = \begin{cases} \bar{G}^{-1}\left(\frac{s\mu}{\sum_{i=1}^{N}\Lambda_i}\right), & \sum_{i=1}^{N}\Lambda_i > s\mu \\ 0, & \text{otherwise.} \end{cases}$$

This appears to contradict all three of the findings of Mendelson and Whang (1990)! To reconcile, we will argue shortly that as the system size grows large, and under the optimal strategy identified by the Mendelson-Whang solution, pricing decisions converge to a common price, delays approach zero and, moreover, delay differentiation becomes negligible. As a consequence, the differentiated N class menu asymptotically degenerates to a single class offering, as derived in the DR solution.

To be more precise, the Mendelson-Whang solution under the scaling (18), prescribes the vector of social welfare optimal prices in the nth system, $p^n_* = (p^n_{1*},\dots,p^n_{N*})$, to be

$$p^n_{j*} = \sum_{\ell=1}^{N} c_\ell \lambda^n_{\ell*} \frac{\partial \mathbb{E}D^n_\ell}{\partial \lambda^n_j}, \quad j = 1,\dots,N. \tag{24}$$

Here, $\lambda^n_{j*} = \Lambda^n_j \bar{F}(p^n_{j*} + c_j \mathbb{E}D^n_j)$ is the demand rate, and $\mathbb{E}D^n_j$ is the queueing delay in each class $j = 1,\dots,N$ under a strict preemptive priority policy

$$\pi^n_1(t) = \min\{s^n, Z^n_1(t)\}, \qquad \pi^n_j(t) = \min\left\{\left(s^n - \textstyle\sum_{\ell=1}^{j-1} Z^n_\ell(t)\right)^+, Z^n_j(t)\right\},\ j = 2,\dots,N.$$

Let $\rho^n_{j*} = \lambda^n_{j*}/n\mu$ denote the traffic intensity in class j in the nth system under this optimal solution.

PROPOSITION 6 **(Social welfare solution structure)**. *Assume the scaling in* (18) *and assume that $\bar{F}_i(\hat{p}_{soc}) > 0$ for all $i = 1,\dots,N$. Then as $n \to \infty$,*

(a) $p^n_{j*} \to \hat{p}_{soc}$ *for* $j = 1, \ldots, N$*;*

(b) $\sqrt{n}\left(1 - \sum_{j=1}^N \rho^n_{j*}\right) \to \beta$ *for some strictly positive, finite constant* β *that depends on model primitives.*

Part (a) of the above proposition asserts that the DR indeed captures the first order properties of the optimal solution for the original mechanism design problem (22), and that the exact analysis in Mendelson and Whang (1990) provides a lower order (and asymptotically vanishing) refinement around the DR solution (that may, of course, be significant in systems of modest size).

Part (b) of the above proposition asserts that the social-welfare optimized system must equilibrate in the QED regime, namely $\sum_{j=1}^N \rho^n_{j*} \approx 1 - \beta/\sqrt{n}$. This complements the analysis in Maglaras and Zeevi (2003a), who showed that the QED regime was welfare maximizing for a single customer type. That is, the socially optimal resource utilization rate is "similar" in a single-type and a multi-type system. In contrast, the revenue maximizing solution may lead to different resource utilization regimes in single-type versus multi-type settings. Additionally, the resource utilization findings imply that the socially optimal solution leads to almost negligible delay differentiation, whereas the revenue maximizing solution may prescribe significant delay differentiation and, as a result, charge significant price premiums for faster service. These asymptotic findings provide interesting contrasts between social welfare and revenue optimization that do not seem apparent via exact analysis.

A Proofs

This appendix contains the proofs of Propositions 2-3 and Theorems 1-4. We defer the proofs of Propositions 1, 4-6 along with a few side lemmas to Appendix B.

Proof of Proposition 2. We prove the equivalent statement: $\bar{p}_1 = \bar{p}_2 = \hat{p}$ if and only if $\epsilon_2(\hat{p},\hat{p}) \le \epsilon_g(\hat{p})$. Fix (p_1, p_2, d_1, d_2) to be a feasible solution to the DR (12) that additionally satisfies

$$d_1 = 0, \qquad d_2 = \frac{1}{c_1}(p_1 - p_2).$$

The full cost for each class at this solution is

$$p_1 + c_1 d_1 = p_1 \quad \text{and} \quad p_2 + c_2 d_2 = cp_1 + (1-c)p_2,$$

respectively, where $c := c_2/c_1$. Define the functions $\kappa_1(p_1)$ and $\kappa_2(p_1,p_2)$ to be the relative workload contributions by class 1 and class 2, respectively, at the price point (p_1,p_2):

$$\kappa_1(p_1) := \frac{\Lambda_1 \bar{F}_1(p_1)}{s\mu}, \qquad \kappa_2(p_1,p_2) := \frac{\Lambda_1 \bar{F}_2(cp_1 + (1-c)p_2)}{s\mu}. \tag{25}$$

The following result, specifically (26), proves the "only if" part of the above assertion.

LEMMA 1. *Let $\hat{p}$ be the optimal solution to the single-product problem* (14)*, and let $(\bar{p}_1,\bar{p}_2)$ be the optimal solution to the DR* (12)*. Then*

$$\bar{p}_1 = \bar{p}_2 = \hat{p} \quad \textit{implies} \quad \epsilon_2(\hat{p},\hat{p}) \le \epsilon_g(\hat{p}) \quad \textit{and} \tag{26}$$

$$\bar{p}_1 > \bar{p}_2 \quad \textit{implies} \quad \frac{\epsilon_1(\bar{p}_1)}{\bar{p}_1} < \left(1 - \frac{c}{1-c}\frac{\kappa_2(\bar{p}_1,\bar{p}_2)}{\kappa_1(\bar{p}_1)}\right)\frac{\epsilon_2(\bar{p}_1,\bar{p}_2)}{\bar{p}_2}, \tag{27}$$

where $\epsilon_1(p_1)$, $\epsilon_2(p_1,p_2)$ and $\epsilon_g(p)$ are the price elasticities defined in (16) *and $\kappa_1(p_1)$ and $\kappa_2(p_1,p_2)$ are defined in* (25)*.*

It remains to show that $\epsilon_2(\hat{p},\hat{p}) \le \epsilon_g(\hat{p})$ implies $\bar{p}_1 = \bar{p}_2 = \hat{p}$. Note that (27) is equivalent to the statement that $\bar{p}_1 = \bar{p}_2 = \hat{p}$, provided that

$$\frac{\epsilon_1(\bar{p}_1)}{\bar{p}_1} \ge \left(1 - \frac{c}{1-c}\frac{\kappa_2(\bar{p}_1,\bar{p}_2)}{\kappa_1(\bar{p}_1)}\right)\frac{\epsilon_2(\bar{p}_1,\bar{p}_2)}{\bar{p}_2}.$$

Hence, we have that

$$\epsilon_1(\hat{p}) \ge \left(1 - \frac{c}{1-c}\frac{\kappa_2(\hat{p},\hat{p})}{\kappa_1(\hat{p})}\right)\epsilon_2(\hat{p},\hat{p}),$$

which we rewrite in terms of f_i and $\bar{F}_i$,

$$\frac{\hat{p} f_1(\hat{p})}{\bar{F}_1(\hat{p})} \ge \left(1 - \frac{c}{1-c}\frac{\Lambda_2 \bar{F}_2(\hat{p})}{\Lambda_1 \bar{F}_1(\hat{p})}\right)(1-c)\frac{\hat{p} f_2(\hat{p})}{\bar{F}_2(\hat{p})}.$$

Some algebraic manipulation yields

$$\begin{aligned}
\Lambda_1 f_1(\hat{p}) &\ge \left((1-c)\Lambda_1 \bar{F}_1(\hat{p}) - c\Lambda_2 \bar{F}_2(\hat{p})\right)\frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})},\\
\frac{\Lambda_1 f_1(\hat{p}) + \Lambda_2 f_2(\hat{p})}{\Lambda_1 \bar{F}_1(\hat{p}) + \Lambda_2 \bar{F}_2(\hat{p})} &\ge (1-c)\frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})},\\
\epsilon_g(\hat{p}) &\ge \epsilon_2(\hat{p},\hat{p}),
\end{aligned}$$

and we deduce that $\epsilon_2(\hat{p},\hat{p}) \le \epsilon_g(\hat{p})$ implies $\bar{p}_1 = \bar{p}_2 = \hat{p}$. This concludes the proof. □

Proof of Proposition 3. Consider the sequence of systems under the scaling (18).

Proof of (a) (Existence and uniqueness of equilibrium.) Fix a positive integer n and put $s^n = n$. We make two trivial observations that substantially simplify our analysis.

Observation 1: Since the control is a strict preemptive priority, the number of class 1 customers in the system form a Markov process that is an $M/M/n$ queue with arrival rate λ_1^n and service rate μ; customers in class 2 are "invisible" to customers in class 1.

Observation 2: Since the service requirements of all customers are i.i.d. exponential with rate μ, the total number of customers in the system form a Markov process that is an $M/M/n$ queue with arrival rate $\lambda_1^n+\lambda_2^n$ and service rate μ.

For any arrival rate $0 \leq \lambda_1^n < n\mu$, we define, with some abuse of notation, $\mathbb{E}D_1^n(\lambda_1^n)$ to be the queueing delay in class 1 as an explicit function of the arrival rate in class 1. The expectation is taken with respect to the stationary distribution of the class 1 headcount process under the arrival rate λ_1^n and the sequencing rule $\pi_1(t)$. With Observation 1, standard queueing results show that such a stationary distribution exists and is unique as long as $\lambda_1^n < n\mu$.

For any arrival rate pair $(\lambda_1^n, \lambda_2^n)$, with $\lambda_1^n, \lambda_2^n \geq 0$ and $\lambda_1^n + \lambda_2^n < n\mu$, we define $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n)$ to be the queueing delay in class 2 as a function of arrival rates in both classes. The expectation is taken with respect to the stationary distribution of the headcount process under arrival rates $(\lambda_1^n, \lambda_2^n)$ and the sequencing rule $(\pi_1^n(t), \pi_2^n(t))$. With Observation 2, standard queueing results show that such a stationary distribution exists and is unique as long as $\lambda_1^n + \lambda_2^n < n\mu$. Note that $\mathbb{E}D_1^n(\lambda_1^n)$ is continuous and monotone increasing in λ_1^n. $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n)$ is continuous and monotone increasing in λ_1^n and in λ_2^n.

For each class $i = 1, 2$, we write the class i arrival rate in that class as an explicit function of the class i overall delay $d_i^n \geq 0$: $\lambda_i^n(d_i^n) = \Lambda_i^n \bar{F}_i(\bar{p}_i + c_i d_i^n)$, $i = 1, 2$. Note that $\lambda_i^n(d_i^n)$ is monotone non-inceasing in d_i^n. An equilibrium in the nth system is given by a delay pair (ξ_1^n, ξ_2^n) that jointly satisfies

$$\begin{aligned} \lambda_1^n(\xi_1^n) + \lambda_2^n(\delta_2 + \xi_2^n) &< n\mu, \\ \mathbb{E}D_1^n\left(\lambda_1^n(\xi_1^n)\right) &= \xi_1^n, \\ \mathbb{E}D_2^n\left(\lambda_1^n(\xi_1^n), \lambda_2^n(\delta_2 + \xi_2^n)\right) &= \xi_2^n. \end{aligned} \tag{28}$$

Since class 2 customers are "invisible" to class 1, we first show that a unique ξ_1 exists for class 1 and then, given ξ_1, we show that a unique ξ_2 exists for class 2.

Class 1: Define $h_1(x) := x - \mathbb{E}D_1^n(\lambda_1^n(x))$. Note that $h_1(\cdot)$ is continuous with $h_1(0) < 0$ and $h_1(\infty) > 0$ (since $\lambda_1^n(0) = \Lambda_1^n \bar{F}_1(\bar{p}_1) < n\mu$ and $\lambda_1^n(\infty) = 0$). Furthermore, $h_1(x)$ is monotone increasing in x since $\mathbb{E}D_1^n(\lambda_1^n(x))$ is monotone non-increasing in x. Therefore, there exists a unique ξ_1^n such that $h_1(\xi_1^n) = 0$.

Class 2: Fix $\lambda_1^n = \Lambda_1^n \bar{F}_1(\bar{p}_1 + c_1\xi_1^n)$ and note that $\lambda_1^n < n\mu\bar{\kappa}_1$. Define $h_2(x) := x - \delta_2 - \mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n(\delta_2 + x))$. Note that $h_2(\cdot)$ is continuous with $h_2(\infty) > 0$ since $\lambda_2^n(\delta_2 + \infty) = 0$. Furthermore, $h_2(x)$ is monotone increasing in x since $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n(\delta_2 + x))$ is monotone non-increasing in x. If $h(\underline{x}) < 0$ for some $\underline{x} \geq 0$ and $\lambda_2^n(\delta_2 + d_2^n) < n\mu - \lambda_1^n$ for $d_2^n > \underline{x}$, then there exists a unique ξ_2^n such that $h_2(\xi_1^n) = 0$.

There are two cases we need to discuss. First, for the uncapacitated case ($\bar{\kappa}_1 + \bar{\kappa}_2 < 1$, $\delta_2 = \bar{d}_2$), take $\underline{x} = 0$ since $\lambda_2^n(\bar{d}_2) = \Lambda_2^n \bar{F}_2(\bar{p}_2 + c_2\bar{d}_2) > 0$ and $\lambda_1^n + \lambda_2^n(\bar{d}_2) < n\mu$. Second, for the capacitated case ($\bar{\kappa}_1 + \bar{\kappa}_2 = 1$,

$\delta_2 = 0$), take $\underline{x}$ such that $\lambda_1^n + \lambda_2^n(\underline{x}) = n\mu$. Then as $x \to \underline{x}$, $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n(x)) \to \infty$ and, for some $\epsilon > 0$, $h(\underline{x} + \epsilon) < 0$ and $\lambda_1^n + \lambda_2^n(\underline{x} + \epsilon) < n\mu$.

We conclude that there exists a unique equilibrium for each n, which can be represented by the delay pair (ξ_1^n, ξ_2^n) satisfying (28), or equivalently the traffic intensity pair (ρ_1^n, ρ_2^n), where

$$\rho_i^n = \frac{\Lambda_i^n \bar{F}_i(\bar{p}_1 + c_i \xi_i^n)}{n\mu}, \qquad i = 1, 2.$$

Note that under this equilibrium, $\rho_1^n + \rho_2^n < 1$ and therefore a unique stationary distribution exists for every n.

Proof of (b) (Convergence of equilibria to DR solution). We prove part (b) in two steps. In Step 1 we show that a limit exists, $\rho_i^n \to \rho_i^\infty$, $i = 1, 2$. In Step 2 we show that the overall delays converge to the delays in the DR solution, $d_i^n \to \bar{d}_i$, $i = 1, 2$. From Step 2, it follows immediately, by the continuity of $F_i(\cdot)$, that $\rho_i^\infty = \bar{\kappa}_i$, $i = 1, 2$.

In what follows, let $\{\rho_i^n\}_{n=1}^\infty$ be the sequence of class i traffic intensities in equilibrium and let $\{\mathbb{E}D_i^n\}_{n=1}^\infty$ be the associated sequence of class i expected queueing delays, $i = 1, 2$. For each n,

$$\rho_1^n = \frac{\hat{\Lambda}_1}{\mu} \bar{F}_1(\bar{p}_1 + c_1 \mathbb{E}D_1^n),$$
$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu} \bar{F}_2(\bar{p}_2 + c_2\delta_2 + c_2 \mathbb{E}D_2^n),$$

where the expectation is taken with respect to the unique stationary distribution established in part (a).

Step 1. Proving that $\rho_i^n \to \rho_i^\infty$, $i = 1, 2$.

If $\rho_1^n = 0$ then $\mathbb{E}D_1^n = 0$ (since there are no class 1 customers in the system), but then $\rho_1^n = \bar{\kappa}_1 > 0$, in contradiction. Therefore, $\rho_1^n > 1$ for all n. Now, suppose there exist subsequences $\{n_k\}_{k=1}^\infty$ and $\{n_\ell\}_{\ell=1}^\infty$ such that

$$\lim_{k\to\infty} n_k(1 - \rho_1^{n_k}) = \overline{g} \quad \text{and} \quad \lim_{\ell\to\infty} n_\ell(1 - \rho_1^{n_\ell}) = \underline{g},$$

where $0 \le \underline{g} < \overline{g} \le \infty$.

LEMMA 2. *Given a sequence of single-class $M/M/n$ systems, indexed by n, with arrival rate λ^n and service rate μ, with $\lambda^n < n\mu$, let $\mathbb{E}D^n$ be the expected queueing delay with respect to the stationary distribution.*

1. *If $n(1 - \rho^n) \to 0$, then $\mathbb{E}D^n \to \infty$.*
2. *$n(1 - \rho^n) \to g \in (0, \infty)$ if and only if $\mathbb{E}D^n \to d = \frac{1}{\mu g} \in (0, \infty)$.*
3. *If $n(1 - \rho^n) \to \infty$, then $\mathbb{E}D^n \to 0$.*

Since $0 \le \underline{g} < \overline{g} \le \infty$, by Lemma 2, we have that

$$0 \le \lim_{k\to\infty} \mathbb{E}D_1^{n_k} < \lim_{\ell\to\infty} \mathbb{E}D_1^{n_\ell} \le \infty.$$

Noting that ρ_1^n is continuous and strictly decreasing in $\mathbb{E}D_1^n$,

$$0 \le \lim_{\ell\to\infty} \rho_1^{n_\ell} < \lim_{k\to\infty} \rho_1^{n_k} \le 1.$$

Since $\lim_{\ell\to\infty}\rho_1^{n_\ell}$ is strictly less than 1, we have

$$\lim_{\ell\to\infty} n_\ell(1-\rho_1^{n_\ell}) = \underline{g} = \infty$$

and therefore $\overline{g} \le \underline{g}$, contradicting our assumption. Therefore, all subsequences converge to a common limit, which we denote ρ_1^∞. The same argument shows that $\rho_1^n + \rho_2^n$ converges as $n\to\infty$. Therefore, $\rho_2^n \to \rho_2^\infty$.

Step 2. Proving that overall delays converge to the DR solution $d_i^n \to \bar{d}_i$, $i = 1, 2$.

We treat separately the capacitated and uncapacitated cases.

For the uncapacitated case ($\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ and $\delta_2 = \bar{d}_2$), note that

$$\begin{aligned}\rho_1^n + \rho_2^n &= \frac{\hat{\Lambda}_1}{\mu}\bar{F}_1(\bar{p}_1 + c_1\mathbb{E}D_1^n) + \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2(\bar{d}_2 + \mathbb{E}D_2^n))\\ &\le \bar{\kappa}_1 + \bar{\kappa}_2 < 1.\end{aligned}$$

Since $\rho_1^n + \rho_2^n$ is eventually strictly less than 1, $\mathbb{E}D_1^n \to 0$ and $\mathbb{E}D_2^n \to 0$, and we conclude that $d_1^n \to \bar{d}_1 = 0$ and $d_2^n \to \delta_2 = \bar{d}_2$. For the capacitated case ($\bar{\kappa}_1 + \bar{\kappa}_2 = 1$, $\bar{\kappa}_2 > 0$, and $\delta_2 = 0$), note that in class 1, $\rho_1^n \le \bar{\kappa}_1 < 1$ for all n so $\mathbb{E}D_1^n \to 0$ and $\rho_1^n \to \bar{\kappa}_1$.

In class 2, suppose $\lim_{n\to\infty}\mathbb{E}D_2^n < \bar{d}_2$. Then there exists $\epsilon > 0$ such that for all n sufficiently large

$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2\mathbb{E}D_2^n) \ge \bar{\kappa}_2 + \epsilon.$$

Since $\rho_1^n \to \bar{\kappa}_1$, we have that eventually $\rho_1^n + \rho_2^n > \bar{\kappa}_1 + \bar{\kappa}_2 = 1$, in contradiction.

Suppose $\lim_{n\to\infty}\mathbb{E}D_2^n > \bar{d}_2$. Then there exists $\epsilon > 0$ such that for all n sufficiently large

$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2\mathbb{E}D_2^n) \le \bar{\kappa}_2 - \epsilon.$$

Since $\rho_1^n \to \bar{\kappa}_1$, we have that eventually $\rho_1^n + \rho_2^n < 1$, which implies $\mathbb{E}D_2^n \to 0$, in contradiction. This completes the proof. □

LEMMA 3 **(Rates of convergence)**. *Assume the scaling in* (18). *Set the stochastic solution to prices* $(\bar{p}_1, \bar{p}_2)$ *and injected delays* (δ_1, δ_2), *together with the sequencing rule* π *prescribed in §3.3. Assume that customer types choose the "correct" service class, i.e.,*

$$\lambda_j^n = \Lambda_j^n\bar{F}_j(\bar{p}_j + c_j d_j^n), \qquad \textit{for } j = 1, 2.$$

If the DR solution is uncapacitated ($\bar{\kappa}_1 + \bar{\kappa}_2 < 1$),

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + o(1/n), \tag{29}$$

while if the DR solution is capacitated ($\bar{\kappa}_1 + \bar{\kappa}_2 = 1$),

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + \mathcal{O}(1/n). \tag{30}$$

Proof of Lemma 3. This will be central to the proof of Theorem 1. We prove this in three steps.

Step 1. We first prove (29). From part (b), $\rho_1^n \to \bar{\kappa}_1 < 1$ and therefore $\sqrt{n}(1-\rho_1^n) \to \infty$. The proof of Proposition 1 of Halfin and Whitt (1981) shows that for a single-class multi-server queue,

$$\sqrt{n}(1-\rho_1^n)\exp(n(1-\rho_1^n)^2/2)\nu(\rho_1^n) \to \frac{1}{1+\sqrt{2\pi}} \quad \text{as } n\to\infty.$$

Here, $\nu(\cdot)$ is the probability that a class 1 customer has a positive waiting time, as a function of traffic intensity. Therefore,

$$n^{3/2}\exp(n(1-\rho_1^n)^2/2)\mathbb{E}D_1^n \to \frac{1}{\mu(1-\bar{\kappa}_1)(1+\sqrt{2\pi})} \in (0,\infty) \quad \text{as } n\to\infty,$$

which yields $d_1^n = \mathcal{O}(n^{-3/2}e^{-bn}) = o(1/n)$ where $b = \frac{1}{2}(1-\bar{\kappa}_1)^2$. This also proves that $\mathbb{E}D_2^n = o(1/n)$, and therefore $d_2^n = \delta_2 + o(1/n) = \bar{d}_2 + o(1/n)$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$.

Step 2. We now show that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$. Since $F_1(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}_1^n \in [0, d_1^n]$ such that

$$n(\bar{\kappa}_1 - \rho_1) = nd_1^n \frac{c_1\hat{\Lambda}_1 f(\bar{p}_1 + c_1\tilde{d}_1^n)}{\mu}.$$

Since $nd_1^n \to 0$ as $n\to\infty$, we conclude that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$. This also proves $n(\bar{\kappa}_2 - \rho_2^n) \to 0$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$.

Step 3. We now show that $n(\bar{\kappa}_2 - \rho_2^n) \to \left(\mu\bar{\kappa}_2\bar{d}_2\right)^{-1} \in (0,\infty)$. Note that

$$\frac{\rho_1^n\mathbb{E}D_1^n + \rho_2^n\mathbb{E}D_2^n}{\rho_1^n + \rho_2^n} = \frac{\nu(\rho_1^n + \rho_2^n)}{n\mu(1-\rho_1^n-\rho_2^n)},$$

where $\nu(\cdot)$ is the probability that a customer in an $M/M/n$ queue has a positive waiting time as a function of traffic intensity. Applying part (b), we see that $\mathbb{E}D^n \to \bar{\kappa}_2\bar{d}_2$ and by Lemma 2(b) it must hold that

$$n(1-\rho_1^n-\rho_2^n) = n(\bar{\kappa}_1-\rho_1^n) + n(\bar{\kappa}_2-\rho_2^n) \to \frac{1}{\mu\bar{\kappa}_2\bar{d}_2} \in (0,\infty).$$

$F_2(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}_2^n$, where $|\tilde{d}_2^n - \bar{d}_2| \le |d_2^n - \bar{d}_2|$, such that

$$n(\bar{\kappa}_2 - \rho_2^n) = n(d_2^n - \bar{d}_2)\frac{c_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\tilde{d}_2^n)}{\mu},$$

and, by continuity, $f_2(\bar{p}_2 + c_2\tilde{d}_2^n) \to f_2(\bar{p}_2 + c_2\bar{d}_2) > 0$. Therefore

$$n(d_2^n - \bar{d}_2) \to \frac{1}{c_2\bar{\kappa}_2\bar{d}_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\bar{d}_2)} \in (0,\infty).$$

We conclude that $d_2^n = \bar{d}_2 + \mathcal{O}(1/n)$. This completes the proof. □

Proof of Theorem 1. It suffices to show that the delays (d_1^n, d_2^n) from Proposition 3 are incentive compatible for sufficiently large n. If incentive compatibility is satisfied, then by the revelation principle it is a Nash equilibrium for customers to truthfully report their types and valuations. This allows us to drop the *assumption* that customers choose the correct service class and thus define for any $n \ge N_{ic}$ a system where the customer demand model is given by (7)-(8), under which an equilibrium exists, and where the prices and equilibrium delays are incentive compatible.

Applying Proposition 1(b) to the incentive compatibility conditions, the delays (d_1^n, d_2^n) are incentive compatible if

$$\bar{d}_2 \le (d_2^n - d_1^n) \le \frac{c_1}{c_2}\bar{d}_2. \tag{31}$$

From Proposition 3(b) we have that $d_1^n \to 0$ and $d_2^n \to \bar{d}_2$ as $n\to\infty$ Since $c_1/c_2 > 1$, there exists some N_{ic}^1 such that for all $n \ge N_{ic}^1$, $d_2^n - d_1^n \le \frac{c_1}{c_2}\bar{d}_2$.

For the inequality $\bar{d}_2 \leq (d_2^n - d_1^n)$, we consider separately the uncapacitated and capacitated cases. In the uncapacitated case, $d_1^n = \mathbb{E}D_1^n$ and $d_2^n = \mathbb{E}D_2^n + \delta_2 = \mathbb{E}D_2^n + \bar{d}_2$. The strict priority rule implies that $\mathbb{E}D_1^n \leq \mathbb{E}D_2^n$ for all n. Therefore, for all $n \geq N_{ic}^2 = 1$,

$$d_2^n - d_1^n = \bar{d}_2 + \mathbb{E}D_2^n - \mathbb{E}D_1^n \geq \bar{d}_2.$$

In the capacitated case, $d_1^n = \mathbb{E}D_1^n$ and $d_2^n = \mathbb{E}D_2^n$ (no delay is injected), so

$$d_2^n - d_1^n = \bar{d}_2 + \frac{1}{n}\left(n(d_2^n - \bar{d}_2) - nd_1^n\right). \tag{32}$$

In the proof of Lemma 3 we showed that $n(d_2^n - \bar{d}_2) \to \beta > 0$ and $nd_1^n \to 0$, as $n \to \infty$. Therefore, there exists some N_{ic}^2 such that for all $n \geq N_{ic}^2$,

$$\left(n(d_2^n - \bar{d}_2) - nd_1^n\right) \geq 0,$$

so $d_2^n - d_1^n \geq \bar{d}_2$.

For all $n \geq N_{ic} = \max\{N_{ic}^1, N_{ic}^2\}$, the delays (d_1^n, d_2^n) are incentive compatible. This concludes the proof. □

Proof of Theorem 2. By Theorem 1, for any $n \geq N_{ic}$, the prescribed solution is incentive compatible and customers choose the "correct" service class. We write the revenues earned in the nth system as

$$R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2) = \bar{p}_1\lambda_1^n + \bar{p}_2\lambda_2^n = n\mu(\bar{p}_1\rho_1^n + \bar{p}_2\rho_2^n)$$

where $\lambda_i^n = \Lambda_i \bar{F}_i(\bar{p}_i + c_i d_i^n)$ and $\rho_i^n = \lambda_i^n/n\mu$. Therefore

$$\begin{aligned} R^n(\pi, \bar{p}_1, \bar{p}_2, \delta_1, \delta_2) &= n\mu(\bar{p}_1\bar{\kappa}_1 + \bar{p}_2\bar{\kappa}_2) - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n) \\ &= n\bar{R} - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n). \end{aligned} \tag{33}$$

From (29) and (30) we have that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$ while, if the DR solution is uncapacitated $n(\bar{\kappa}_2 - \rho_2^n) \to 0$ and if the DR solution is capacitated $n(\bar{\kappa}_2 - \rho_2^n) \to 1/\mu\bar{\kappa}_2\bar{d}_2$. Therefore, there exists a finite, positive constant M such that

$$n(\bar{\kappa}_1 - \rho_1^n) + n(\bar{\kappa}_2 - \rho_2^n) \leq M \qquad \text{for all } n \geq N_{ic}. \quad \square$$

Proof of Theorem 3. By Theorem 1, for any $n \geq N_{ic}$, the prescribed solution is incentive compatible and customers choose the "correct" service class. Therefore, all the assumptions of Proposition 3 and Lemma 3 are satisfied for the sequence of systems indexed by n, starting at N_{ic}, and the results of Proposition 3 and Lemma 3 hold. In particular, a unique sequence of equilibria exists, the equilibrium delays converges to the DR solution, and as $n \to \infty$, if the DR solution is uncapacitated,

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \text{and} \quad \rho_2^n = \bar{\kappa}_2 + o(1/n),$$

while if the DR solution is capacitated,

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \text{and} \quad \rho_2^n = \bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n).$$

where $\alpha = 1/\mu\bar{\kappa}_2\bar{d}_2$. This concludes the proof. □

Proof of Theorem 4. This is a direct extension of the two-type case, namely Theorems 1-2 and supporting results. The proof is essentially identical, and we will provide an outline here for $k > 2$ service classes, omitting further details.

Class 1, the highest priority service class, is the same as in the two-type, two-class case. Intermediate service classes, $j = 2, \ldots, k-1$, also operate in the QD regime, in terms of traffic intensity and queueing delays, but the SP adds injected delay to ensure that the IC constraints are satisfied since the system congestion for these classes is not sufficient (analogous to class 2 in the two-type uncapacitated system). Finally, the lowest priority service class k operates in the QD regime with injected delay if the system is uncapacitated and in the ED regime with no injected delay if the system is capacitated. This is analogous to class 2 in the two-type, two-class system.

B Supplementary Proofs (to be included in full technical report)

Proof of Proposition 1. Proposition 1 is a special case of Proposition 4(a). □

Proof of Proposition 4.

Proof of (a). Suppose each property does *not* hold for a feasible solution $(\bar{p}_1, \ldots, \bar{p}_N)$, $(\bar{d}_1, \ldots, \bar{d}_N)$. We construct an alternative solution $(\breve{p}_1, \ldots, \breve{p}_N)$, $(\breve{d}_1, \ldots, \breve{d}_N)$, which is feasible and strictly improves on the former.

Suppose $\bar{d}_1 > 0$. Take $\breve{p}_1 = p_1 + c_1 d_1$, $\breve{d}_1 = 0$, and $\breve{p}_i = p_i$, $\breve{d}_i = d_i$ for $i = 2, \ldots, N$.

Suppose $\bar{p}_i + c_i \bar{d}_i < \bar{p}_{i+1} + c_i \bar{d}_{i+1}$. Take

$$\breve{p}_{i+1} = \frac{c_i(\bar{p}_{i+1} + c_{i+1}\bar{d}_{i+1}) - c_{i+1}(\bar{p}_i + c_i \bar{d}_i)}{c_i - c_{i+1}} \qquad \breve{d}_{i+1} = \frac{\bar{p}_i + c_i \bar{d}_i - \bar{p}_{i+1} - c_{i+1}\bar{d}_{i+1}}{c_i - c_{i+1}}$$

and $\breve{p}_j = \bar{p}_j$, $\breve{d}_j = \bar{d}_j$ for $j \neq i+1$.

Proof of (b). We write the result of part (a) as

$$\bar{d}_i = \bar{d}_{i-1} + \frac{1}{c_{i-1}}(\bar{p}_{i-1} - \bar{p}_i).$$

Additionally, incentive compatibility requires

$$\bar{p}_i + c_i \bar{d}_i \leq \bar{p}_{i-1} + c_i \bar{d}_{i-1} \quad \text{for } i = 2, \ldots, N.$$

Therefore $\bar{p}_1 \geq \bar{p}_2 \geq \cdots \geq \bar{p}_N$ and the sets $\{A_{(1)}, \ldots, A_{(N)}\}$ must have the structure described. □

Proof of Lemma 1. Apply Proposition 1 to reduce the deterministic relaxation (12) to two variables p_1 and p_2, and set $c := \frac{c_2}{c_1} < 1$,

$$\begin{aligned} \text{maximize} \quad & \Lambda_1 p_1 \bar{F}_1(p_1) + \Lambda_2 p_2 \bar{F}_2(cp_1 + (1-c)p_2) \\ \text{subject to} \quad & p_1 \geq p_2 \\ & \Lambda_1 \bar{F}_1(p_1) + \Lambda_2 \bar{F}_2(cp_1 + (1-c)p_2) \leq s\mu. \end{aligned} \tag{34}$$

Equations (26) and (27) follow from the KKT necessary conditions of (34). □

Proof of Lemma 2. Lemma 2 follows immediately from Lemma 4 and the $M/M/n$ delay formula. □

LEMMA 4 **(Halfin and Whitt)**. *Given a sequence of single-class $M/M/n$ systems, indexed by n, with arrival rate λ^n and service rate μ, we define $\rho^n = \frac{\lambda^n}{n\mu}$ and $\nu^n = \mathbb{P}(Z^n \geq n)$, the probability that all servers are busy.*

(a) If $\sqrt{n}(1-\rho^n) \to 0$ then $\nu^n \to 1$.

(b) $\sqrt{n}(1-\rho^n) \to \beta \in (0,\infty)$ if and only if $\nu^n \to \nu \in (0,1)$.

(c) If $\sqrt{n}(1-\rho^n) \to \infty$ then $\nu^n \to 0$.

Proof of Proposition 5. Let $p_1, d_1, \ldots, p_N, d_N$ be an optimal solution to (23). If $\sum_{i=1}^N \Lambda_i \leq s\mu$ then $p_1 = \cdots = p_N = \hat{p}_{soc} = 0$ and $d_1 = \cdots = d_N = 0$ is trivially the optimal solution. Assume $\sum_{i=1}^N \Lambda_i > s\mu$ and thus $\sum_{i=1}^N \Lambda_i \bar{F}_i(\hat{p}_{soc}) = s\mu$. We prove that if $\lambda_i > 0$ then $p_i = \hat{p}_{soc}$ and $d_i = 0$, for any $i = 1, \ldots, N$. Proposition 5 follows immediately.

Apply the identity $\int_x^\infty v f_i(v)\, dv = x\bar{F}_i(x) + \int_x^\infty \bar{F}_i(v)\, dv$ to rewrite the social welfare at the optimal solution as

$$W(p_1, d_1, \ldots, p_N, d_N) = \sum_{i=1}^N p_i \Lambda_i \bar{F}_i(p_i + c_i d_i) + \sum_{i=1}^N \Lambda_i \int_{p_i + c_i d_i}^\infty \bar{F}_i(v)\, dv$$

and define W_{soc} to be the social welfare at the single-class solution $(\hat{p}_{soc}, 0)$

$$W_{soc} := \hat{p}_{soc} s\mu + \sum_{i=1}^N \Lambda_i \int_{\hat{p}_{soc}}^\infty \bar{F}_i(v)\, dv.$$

Consider the difference in social welfare between the two solutions

$$\begin{aligned} W(p_1, d_1, \ldots, p_N, d_N) - W_{soc} &= \sum_{i=1}^N \Lambda_i \left(p_i \bar{F}_i\left(p_i + c_i d_i\right) + \int_{p_i + c_i d_i}^\infty \bar{F}_i(v)\, dv \right) - \hat{p}_{soc} s\mu - \sum_{i=1}^N \Lambda_i \int_{\hat{p}_{soc}}^\infty \bar{F}_i(v)\, dv \\ &= \hat{p}_{soc} \underbrace{\left(\sum_{i=1}^N \Lambda_i \bar{F}_i(p_i + c_i d_i) - s\mu \right)}_{\leq 0} + \sum_{i=1}^N \Lambda_i \left((p_i - \hat{p}_{soc}) \bar{F}_i\left(p_i + c_i d_i\right) + \int_{p_i + c_i d_i}^{\hat{p}_{soc}} \bar{F}_i(v)\, dv \right) \\ &\leq \sum_{i=1}^N \Lambda_i \left((p_i - \hat{p}_{soc}) \bar{F}_i\left(p_i + c_i d_i\right) + \int_{p_i + c_i d_i}^{\hat{p}_{soc}} \bar{F}_i(v)\, dv \right). \end{aligned}$$

Since $\bar{F}_i(\cdot)$ is non-negative and non-increasing, we have for any $i = 1, \ldots, N$,

$$(p_i - \hat{p}_{soc}) \bar{F}_i\left(p_i + c_i d_i\right) + \int_{p_i + c_i d_i}^{\hat{p}_{soc}} \bar{F}_i(v)\, dv \leq 0. \tag{35}$$

Moreover, (35) holds with *strict* inequality for any i such that $\lambda_i > 0$ and $p_i \neq \hat{p}_{soc}$. (This is easily checked in each of the cases: i) $\hat{p}_{soc} < p_i$, ii) $p_i < \hat{p}_{soc} \leq p_i + c_i d_i$, and iii) $\hat{p}_{soc} > p_i + c_i d_i$.) Therefore, if $\lambda_i > 0$ and $p_i \neq \hat{p}_{soc}$ then $W(p_1, d_1, \ldots, p_N, d_N) < W_{soc}$ in contradiction. □

Proof of Proposition 6.

Proof of (a). Let $\mathbb{E}D^n_{j*}$ be the queueing delay for class j, $j = 1, \ldots, N$, in the nth system operating under the optimal solution; let W^n_* be the optimal social welfare:

$$W^n_* := \sum_{j=1}^N \Lambda^n_j \left(\int_{p^n_{j*}+c_j\mathbb{E}D^n_{j*}}^\infty v f_j(v)\, dv - c_j \mathbb{E}D^n_{j*} \bar{F}_j(p^n_{j*} + c_j \mathbb{E}D^n_{j*}) \right).$$

Let $\mathbb{E}D^n_{soc}$ be the queueing delay in the nth system operating with a single service class at price $\hat{p}_{soc}$ and let W^n_{soc} be the resulting social welfare:

$$W^n_{soc} := \sum_{j=1}^N \Lambda^n_j \left(\int_{\hat{p}_{soc}+c_j\mathbb{E}D^n_{soc}}^\infty v f_j(v)\, dv - c_j d_j \bar{F}_j(\hat{p}_{soc} + c_j \mathbb{E}D^n_{soc}) \right).$$

We first show that $\mathbb{E}D^n_{soc} \to 0$. Define ρ^n_{soc} to be the utilization in the nth system and note that

$$\rho^n_{soc} = \sum_{j=1}^N \frac{\Lambda^n_j}{n\mu} \bar{F}_j(\hat{p}^n_{soc} + c_j \mathbb{E}D^n_{soc}) < \sum_{j=1}^N \frac{\Lambda^n_j}{n\mu} \bar{F}_j(\hat{p}^n_{soc}) \le 1 \quad \text{for all } n.$$

If $\lim_{n\to\infty} \mathbb{E}D^n_{soc} > 0$ then $\lim_{n\to\infty} \rho^n_{soc} < 1$ implying that $\lim_{n\to\infty} \mathbb{E}D^n_{soc} = 0$, in contradiction. If $\lim_{n\to\infty} p^n_{j*} \neq \hat{p}_{soc}$ then $\frac{W^n_*}{W^n_{soc}} < 1$ for sufficiently large n, in contradiction.

Proof of (b). We can write the queueing delays in each class as

$$\mathbb{E}D^n_{1*} = \psi^n(\rho^n_{1*}), \qquad \text{and} \qquad \mathbb{E}D^n_{j*} = \frac{\omega^n_{j*}\psi^n(\omega^n_{j*})}{\rho^n_{j*}} - \frac{\omega^n_{(j-1)*}\psi^n(\omega^n_{(j-1)*})}{\rho^n_{j*}} \quad \text{for } j = 2, \ldots, N. \tag{36}$$

where $\omega^n_{j*} := \sum_{\ell=1}^j \rho^n_{\ell*}$ for $j = 1, \ldots, N$,

$$\nu^n(x) := \left(\sum_{j=0}^{n-1} \frac{(nx)^j}{j!} + \frac{(nx)^n}{n!(1-x)} \right)^{-1} \frac{(nx)^n}{n!(1-x)} \qquad \text{and} \qquad \psi^n(x) := \frac{\nu^n(x)}{n\mu(1-x)}. \tag{37}$$

Note that $\nu^n(x)$ is the formula for probability of delay and $\psi^n(x)$ is the formula for expected delay in a standard $M/M/n$ queue in stationarity, each as a function of traffic intensity $x \in [0, 1)$.

Define

$$\kappa_{j*} := \frac{\hat{\Lambda}_j \bar{F}_j(\hat{p}_{soc})}{\mu} \quad \text{for } j = 1, \ldots, N.$$

From part (a) we have $\rho^n_{j*} \to \kappa_{j*}$. Since $\sum_{j=1}^{N-1} \kappa_{j*} < 1$, we have that, $n(\kappa_{j*} - \rho^n_{j*}) \to 0$ for all $j = 1, \ldots, N-1$ (see Step 2 in the proof of Lemma 3) and therefore $\sqrt{n}(\kappa_{j*} - \rho^n_{j*}) \to 0$ for all $j = 1, \ldots, N-1$. It remains to show that $\sqrt{n}(\kappa_{N*} - \rho^n_{N*}) \to \beta \in (0, \infty)$.

Since $F_N(\cdot)$ is continuously differentiable, there exists some $\tilde{d}^n$ such that

$$(\kappa_{N*} - \rho^n_{N*}) = \mathbb{E}D^n_{N*} \frac{\hat{\Lambda}_N f_N(p^n_{N*} + c_N \tilde{d}^n)}{\mu}.$$

According to the formulas above, we can write

$$\begin{aligned}
\mathbb{E}D^n_{N*} &= \frac{\omega^n_{N*}}{\rho^n_{N*}} \frac{\nu^n(\omega^n_{N*})}{n\mu(1-\omega^n_{N*})} - \frac{\omega^n_{(N-1)*}}{\rho^n_{N*}} \frac{\nu^n(\omega^n_{(N-1)*})}{n\mu(1-\omega^n_{(N-1)*})} \\
n(1-\omega^n_{N*})\mathbb{E}D^n_{N*} &= \frac{\omega^n_{N*}}{\mu\rho^n_{N*}} \left(\nu^n(\omega^n_{N*}) - \frac{\omega^n_{(N-1)*}}{\omega^n_{N*}} \frac{(1-\omega^n_{N*})}{(1-\omega^n_{(N-1)*})} \nu^n(\omega^n_{(N-1)*}) \right) \\
\lim_{n\to\infty} n(1-\omega^n_{N*})\mathbb{E}D^n_{N*} &= \frac{1}{\mu\kappa_{N*}} \lim_{n\to\infty} \nu^n(\omega^n_{N*}).
\end{aligned}$$

Also, note that

$$n(1-\omega_{N*}^n)(\kappa_{N*}-\rho_{N*}^n) = \sum_{j=1}^{N-1} n\left(\kappa_{j*}^n-\rho_{j*}^n\right)(\kappa_{N*}-\rho_{N*}^n) + n(\kappa_{N*}-\rho_{N*}^n)^2$$
$$\lim_{n\to\infty} n(1-\omega_{N*}^n)(\kappa_{N*}-\rho_{N*}^n) = \lim_{n\to\infty} n(\kappa_{N*}-\rho_{N*}^n)^2.$$

Therefore, we have that

$$\left(\lim_{n\to\infty}\sqrt{n}(\kappa_{N*}-\rho_{N*}^n)\right)^2 = \frac{\hat{\Lambda}_N f_N(\hat{p}_{soc})}{\mu^2\kappa_{N*}}\lim_{n\to\infty}\nu^n(\omega_{N*}^n).$$

By Lemma 4, it must be that $\sqrt{n}(\kappa_{N*}-\rho_{N*}^n)\to\beta\in(0,\infty)$. $\square$

References

Afèche, Philipp. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* Forthcoming.

Afèche, Philipp, Michael Pavlin. 2011. Optimal price-lead time menus for queues with customer choice: Priorities, pooling & strategic delay. Working paper.

Anderson, Eric T., James D. Dana, Jr. 2009. When is price discrimination profitable? *Management Science* **55**(6) 980–989.

Deneckere, Raymond J., R. Preston McAfee. 1996. Damaged goods. *Journal of Economics & Management Strategy* **5**(2) 149–174.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.

Hassin, Refael, Moshe Haviv. 2003. *To Queue or Not To Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers.

Katta, Akshay-Kumar, Jay Sethuraman. 2005. Pricing strategies and service differentiation in queues – a profit maximization perspective. Tech. rep., Computational Optimization Research Center, Columbia University. TR-2005-04.

Lariviere, Martin A. 2006. A note on probability distributions with increasing generalized failure rates. *Operations Research* **54**(3) 602–604.

Maglaras, Costis, Assaf Zeevi. 2003a. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Maglaras, Costis, Assaf Zeevi. 2003b. Pricing and performance analysis for a system with differentiated services and customer choice. R. Srikant, P. Voulgaris, eds., *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*.

Maglaras, Costis, Assaf Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* **53**(2) 242–262.

McAfee, Preston. 2007. Pricing damaged goods. Economics Discussion Paper 2, Kiel Institute for the World Economy. URL `http://www.economics-ejournal.org/economics/discussionpapers/2007-2`.

Mendelson, Haim. 1985. Pricing computer services: Queueing effects. *Communications of the ACM* **28**(3) 312–321.

Mendelson, Haim, Seungjin Whang. 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research* **38**(5) 870–883.

Myerson, Roger B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* **47**(1) 61–74.

Myerson, Roger B. 1981. Optimal auction design. *Mathematics of Operations Research* **6**(1) 58–73.

Naor, Pinhas. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.