Submitted to *Management Science*
manuscript MS-13-00926.R3

Optimal Price and Delay Differentiation in Large-Scale Queueing Systems

(Authors' names blinded for peer review)

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 simplifies the problem by preserving the economic aspects of price-and-delay differentiation while ignoring queueing delays, we construct a solution to the fully stochastic problem 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 deterministic analysis captures all pricing, differentiation, and delay characteristics of the stochastic solution that are non-negligible at large scale; ii) service differentiation is optimal when the less delay-sensitive market segment is sufficiently elastic; iii) the use of "strategic delay" depends on system capacity and market heterogeneity – and contributes significant delay when the system capacity is under-utilized or when the firm offers three or more service classes; 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 with noticeable delays that arise either endogenously ("efficiency-driven" regime) or with the addition of strategic delay by the service provider.

Key words: service differentiation; pricing; revenue management; damaged goods; queueing games; many-server limits

1. Introduction

1.1. Motivation and Overview of Results

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. The on-going debate over network neutrality principles 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 levels (based on delay) 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. The high-level problem for the service provider is how to optimally design and implement a menu of price-delay service offerings in such settings. We study this service differentiation problem in the context of a large-scale stochastic service system that is prone to congestion due to queueing.

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 queue and may have multiple service classes that are differentiated in terms of price and delay. Demand for each service class consists of a stream of atomistic and rational customers. An individual customer gains positive utility from receiving service, but suffers negative utility for each unit of time spent waiting. Upon arrival, he chooses the service class (or opts out) that maximizes his net expected utility. In this manner, the set of price and delay combinations affects the demand for each service class, which in turn determines the congestion in each class, and so on. An optimal solution specifies a menu of service classes and 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. The type and valuation of any individual customer is private information and thus unknown to the service provider. Designing the service provider's revenue maximizing product menu, taking into account the effect of customers' self-optimizing choices, can be cast as a mechanism design problem. As a point of reference, the socially optimal menu for the above model is known and fairly straightforward to characterize and implement, based on the key insights 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.

Main findings. This paper proposes an approximate analysis, that applies to systems with large processing capacity operating in settings with large market potential. This greatly simplifies the study of the revenue-maximization problem, while preserving the significant insights into the structure of the optimal solution. Some of the key contributions are the following.

1. Solution via Deterministic Analysis. Setting aside queueing dynamics, we propose a *deterministic relaxation* of the revenue maximization problem and show that its solution yields an intuitive price-delay menu and suggests a simple priority sequencing rule. This translates to a solution in the stochastic setting that achieves near-optimal revenue performance in large-scale systems. We apply this framework to the setting with two customer types (§2-4) and show that it easily extends to multiple customer segments (§5), which is relevant to settings with significant market heterogeneity. Our deterministic analysis does not provide closed-form price prescriptions in terms of model primitives, but rather allows an easy-to-compute solution that is accurate in large-scale systems (see Remark 2).

2. Insights into Service Differentiation. Our approach shows that the first-order (non-vanishing at large scale) features of the stochastic solution can be immediately determined from the solution to the deterministic relaxation. Such features are prices, delays, level of differentiation, system utilization, sequencing of customers, and strategic delay (which was first analyzed in Afèche (2004)). For example, we identify conditions that imply first-order service differentiation is optimal. We also establish that in systems with two service classes, strategic delay is a first-order effect when there is ample capacity (some fraction of servers permanently idle at large scale), but second-order (vanishing at large scale) when there is not, including settings where the service provider decides to set capacity at a level that avoids permanently idle servers. In systems where it is optimal to offer three or more service classes, strategic delay is always a first-order consideration. These results do not rely on restrictive assumptions on the market primitives, such as uniform or exponential valuation distributions.

3. Connection to Asymptotic Queueing Regimes. The paper also contributes to the literature on heavy-traffic analysis of queueing systems. We believe that this is the first work to show that classical operating regimes, such as the so-called *efficiency-driven* (ED) and *quality-driven* (QD), may arise endogenously as a result of price discrimination and service differentiation; specifically, the high priority class operates in the QD regime, experiencing an underloaded and uncongested system, while the lowest priority class operates in the heavily utilized ED regime, experiencing a system that is always congested. This complements earlier results by Maglaras and Zeevi (2003a) that first showed that the *quality and efficiency-driven* (QED) operating regime arises endogenously as a result of revenue maximization when customers are homogenous in their delay costs.

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). Two early references that are relevant to our work are Mendelson (1985) and Mendelson and Whang (1990), which introduced the atomistic, utility-maximizing customer behavior model in queues with single and multi-type markets, respectively; the latter focused on welfare maximization.

The revenue maximization problem that we consider is most closely related to Afèche (2013), who analyzed a single-server queueing system facing a market with two customer types (analogous to §3-4 in this work), and made three important and related contributions. First, he formulated the problem in a mechanism design framework, and, second, showed that externality pricing and delay cost minimization are no longer optimal in the revenue maximization setting. Third, he established necessary and sufficient conditions for the optimal solution to include strategic delay, in which the service provider chooses to artificially delay some customers beyond what is caused by system congestion alone. His study provides an exact analysis of the two-type case and partial extensions of this approach to multiple (more than two) customer types in a $M/M/1$ setting can be found in Afèche and Pavlin (2011) and Katta and Sethuraman (2005). These partial extensions require more restrictive assumptions on the market primitives – specifically, all customers of a given type (common delay cost) share a common service valuation, and there is a monotone relationship between delay cost and service valuation. Our work adopts the mechanism design formulation (which allows for strategic delay) introduced in Afèche (2013), applied to a multi-server setting. More importantly, our method of analysis and the focus of our results are different. Unlike the above papers, we undertake an approximate rather than exact analysis approach, which provides new and complementary insights. The exact analysis papers describe features of the optimal solution directly in terms of model primitives, while we formulate an approximate solution, which in turn is based on the solution to a deterministic relaxation. In particular, the deterministic relaxation solution captures the first-order features of the optimal solution and ignores those that become vanishingly small in large systems. We note that our proposed framework extends to the multi-type setting without further restrictions. Another example of interest that can be handled within our framework and is of interest to service systems and information service networks is the parallel multi-pool, multi-server system. We provide a detailed comparison with Afèche (2013) in §5 (see Remark 4).

The above references and the closely related literature uses exact analysis for single-server queueing systems. There is a parallel stream of work that, like this paper, considers multi-server systems and leverages asymptotic analysis to gain insight into the optimal prices and policies. Maglaras and Zeevi (2003a) consider a single-class system, characterize the asymptotic equilibrium operating point, and show that, when demand is elastic, the revenue-maximizing price places the system in the QED regime. Maglaras and Zeevi (2005) introduces the use of a deterministic relaxation for a two class system, where choice is captured via an aggregated demand function in a setting with partially substitutable products; atomistic choice, incentive compatibility, and delay preference heterogeneity were not considered.

The terminology describing the operating regimes of large-scale, multi-server systems is due to Borst et al. (2004). In that paper and much of the work in capacity sizing and optimal control of multi-server systems (typically motivated by call center applications), demand is exogenous. By contrast, demand in our model is delay-sensitive and therefore endogenously determined via a game-theoretic equilibrium, which captures the complex interaction between individual, utility-maximizing customers and a revenue or social-welfare maximizing service provider. While there is a significant body of work in which asymptotic operating regimes arise from endogenous demand, including Maglaras and Zeevi (2003a,b, 2005), Whitt (2003), Armony and Maglaras (2004a,b), and Plambeck and Ward (2006), most consider problems in which large-scale delay differentiation is absent and find that the QED regime is economically optimal.

Strategic delay can be viewed as the queueing system manifestation of damaged goods, a concept from the economics and marketing literature, which refers to the practice of introducing a low-price low-quality version of a good, despite equal (or greater) production costs, that serves to segment the customer market and price discriminate. A number of examples of such cases can be found in Deneckere and McAfee (1996), while McAfee (2007) derives sufficient conditions this practice to be optimal. 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. A significant difference between our work and these is that 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. The marketing and economics literature generally disregards the operational considerations of the service system, and the inherent conflict between price discrimination and efficient resource utilization that gives rise to congestion effects.

2. Model and Problem Formulation

System model. The service provider (SP) operates s servers, which are used to offer k classes of service that are differentiated by price and delay. Arrivals into a service class $j \in \{1,\dots,k\}$ form an independent Poisson process with rate λ_j, which is determined by the customer choice model specified below. Each service class has an infinite-capacity buffer and customers in that class wait in a queue until they are allocated a server. 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.[1] All customers have

[1] All results hold if delay is defined to be the sojourn time, with only trivial changes to the proofs.

random processing requirements that are independent and identically distributed (i.i.d.) draws from an exponential distribution with mean $1/\mu$.

The allocation of servers to customers is determined by a control policy π, which satisfies the following assumptions: i) each server may only work on one customer at a time; ii) service for any customer may be interrupted without penalty and resumed without restarting service (allow preempt/resume); iii) the policy does not depend on the realized service times of customers; iv) servers may not idle if there are any customers waiting in queue.

Assumption i) is for ease of exposition – all major results hold if processor sharing is allowed. Assumption ii) simplifies many of the proofs; if preemption is not allowed, the asymptotic results are the same in the limit, but the rates of convergence may differ – see Remark 3. Assumptions iii)-iv) are standard work-conservation assumptions. A formal description of these queueing dynamics is provided in Appendix B. We allow for strategic delay by assuming that customers are sent to an infinite-capacity "delay node" *following service completion*, where a customer from service class j is held for $\delta_j \geq 0$ units of time and then released from the system. This is one of several ways to add strategic delay (see §3.2 and §7 of Afèche (2013)), and can achieve the expected delays obtained under any alternative implementation.

Given a control policy π and an arrival rate vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ that satisfies $\sum_{j=1}^{k} \lambda_j < s\mu$, standard queueing results (e.g., Saaty (1961) and references therein) show that there exists a unique stationary distribution for the number of customers for each service class that are in queue or in service, but not in the delay node (sometimes called the "headcount process"). Define $\mathbb{E}D_j(\lambda, \pi)$ to be the expected time in queue for class j customers under this stationary distribution. The overall delay experienced by a customer in class j is therefore $\mathbb{E}D_j(\lambda, \pi) + \delta_j$. (Expected values are always with respect to the stationary distribution generated by a specified arrival rate vector λ and admissible control π.)

Customer choice model. Customers of type $i = 1, 2$ 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 type i customer 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 the cumulative distribution function F_i is strictly increasing on its support, has a continuous density f_i, an increasing generalized failure rate (IGFR), and a finite mean. The IGFR and finite mean assumptions ensure that an infinite price is not optimal (Lariviere (2006)). (This is a common condition in the revenue management literature, but weaker assumptions, e.g., that the functions $p\bar{F}_i(p)$ for $i = 1, 2$ are coercive, also suffice.) Each type i customer incurs an additive linear delay cost of c_i per unit time spent waiting, where c_i is common across all type i customers. We assume, without loss of generality, that $c_1 > c_2$, so type 1 customers are more delay sensitive than type 2 customers.

A type i customer with willingness-to-pay V_i, who arrives at a system offering k service classes with prices p_j and overall delays d_j, $j=1,\ldots,k$, calculates his net utility for each service class j,

$$U_i(j) = V_i - (p_j + c_i d_j), \tag{1}$$

and chooses the option that maximizes his net utility,

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

where $j=0$ represents the no-purchase option. Customers who choose not to enter the system are lost and do not return.

Information structure. 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 valuation V_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. We also assume that the queues are *unobservable* so customers make their choice based on the announced prices and delays (which we require to be credible).

Number of service classes offered. Observe that all customers of type i will select the same service class, because any individual type i customer selects the service class j with the minimum "full cost," $p_j + c_i d_j$, irrespective of his individual willingness-to-pay V_i. In a market with N customer types, the SP need only offer up to N service classes ($k \leq N$). For $N=2$, the resulting mean demand rate for each service class is given by

$$\begin{aligned}\lambda_1(p_1,p_2,d_1,d_2) &= \Lambda_1 \bar{F}_1(p_1+c_1d_1)\mathbf{1}\{p_1+c_1d_1 \leq p_2+c_1d_2\} \\ &\quad + \Lambda_2 \bar{F}_2(p_1+c_2d_1)\mathbf{1}\{p_1+c_2d_1 < p_2+c_2d_2\}, \qquad (2)\\ \lambda_2(p_1,p_2,d_1,d_2) &= \Lambda_1 \bar{F}_1(p_2+c_1d_2)\mathbf{1}\{p_2+c_1d_2 < p_1+c_1d_1\} \\ &\quad + \Lambda_2 \bar{F}_2(p_2+c_2d_2)\mathbf{1}\{p_2+c_2d_2 \leq p_1+c_2d_1\}, \qquad (3)\end{aligned}$$

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 $j=i$. By the Poisson thinning property, the arrival process into each service class is itself Poisson.

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

Definition 1 (Equilibrium). Fix prices (p_1, p_2), a control policy π, strategic 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 $\lambda_1 + \lambda_2 < s\mu$ and

$$d_j = \mathbb{E}D_j(\lambda_1, \lambda_2, \pi) + \delta_j \qquad j = 1, 2. \tag{4}$$

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 strategic delays (δ_1, δ_2) to maximize the equilibrium revenue rate given by

$$R(\pi, p_1, p_2, \delta_1, \delta_2) = \sum_{j=1}^{2} p_j \lambda_j(p_1, p_2, d_1, d_2), \tag{5}$$

where (d_1, d_2) are the overall delays in equilibrium (assuming it exists), given in (4), and the customer demand model $\lambda_j(\cdot)$, $j = 1, 2$, is given in (2) and (3).

We adopt the formulation of Afèche (2013), which states the above as a mechanism design problem. Applying the revelation principle (Myerson (1979)), we consider, without loss of generality, only *direct mechanisms* that satisfy incentive compatibility and individual rationality.

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

In a direct mechanism, each customer reports their private information (type i and valuation V_i) to the SP, who then uses that information to determine which service class the customer purchases, if any. If such a mechanism satisfies the incentive compatibility and individual rationality conditions, then it is a Nash equilibrium for customers to truthfully report their types and valuations. Under this labeling, type i customers are either assigned to service class i or turned away.

The revenue maximization problem is to find prices (p_1, p_2), a control policy π, and strategic delays (δ_1, δ_2) to:

$$\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{2} p_i \lambda_i & (6)\\
\text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i, j = 1, 2 \text{ and } i \ne j \\
& \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, 2 \\
& \lambda_1 + \lambda_2 < s\mu \\
& d_i = \mathbb{E}D_i(\lambda_1, \lambda_2, \pi) + \delta_i \quad i = 1, 2 \\
& \delta_i \ge 0 \quad i = 1, 2.
\end{aligned}$$

The solution to (6) does not necessarily have two *distinct* service classes; the optimization problem allows both 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. We consider such solutions to be single-class. The ability of the SP to segment the market by delay sensitivity, but not valuation, is a consequence of additive delay costs; linearity of the delay cost is not required.

3. Deterministic Analysis

Our proposed analysis framework relies on a deterministic relaxation ("DR"), which preserves the essential economic considerations and the capacity constraint of the original problem (6) while ignoring the complications presented by the queueing dynamics and resulting equilibrium. We then use the optimal solution to the DR to construct an approximate solution to the original problem, which achieves near-optimal performance in large systems in a way we make precise in the next section.

3.1. Deterministic Relaxation

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

$$\begin{aligned} \text{maximize} \quad & p_1\lambda_1 + p_2\lambda_2 \\ \text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i,j = 1,2 \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,2 \\ & \lambda_1 + \lambda_2 \le s\mu \\ & d_1 \ge 0, d_2 \ge 0. \end{aligned} \tag{7}$$

The delays are treated as "free" *decision variables*, only constrained to be non-negative and to satisfy the system-wide capacity constraint; they *do not* need to correspond to an achievable pair of equilibrium delays in the queueing system as required in (6). In this precise sense, (7) is a (deterministic) relaxation of (6).

An optimal solution to (7), which we call the "DR solution," exists since the objective function is coercive and the feasible set is closed. We denote the DR solution $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ and set $\bar{\lambda}_i = \Lambda_i \bar{F}_i(\bar{p}_i + c_i \bar{d}_i)$, $i = 1,2$. We also denote by $\bar{\kappa}_i$ the fraction of system capacity consumed by class i in the DR solution

$$\bar{\kappa}_i = \frac{\bar{\lambda}_i}{s\mu} \qquad i = 1,2. \tag{8}$$

Remark 2. Note that while we guarantee the existence of a DR solution and describe some of its properties that are useful in constructing a stochastic solution, we do not provide closed-form expressions for the DR solution. By treating delays as decision variables, computing the DR solution to (7) is substantially easier than directly solving (6), both of which, in general, may require numerical methods. We do not discuss numerical methods in this paper and assume that the solution to the deterministic optimization problem (7) is accessible.

Since (7) is a relaxation of (6), the optimal revenue rate in the DR setting,

$$\bar{R} = \bar{p}_1 \bar{\lambda}_1 + \bar{p}_2 \bar{\lambda}_2,$$

is an upper bound on the optimal revenue rate in (6). In later sections, we prove asymptotic optimality of approximate solutions by demonstrating that their revenues converge to this upper bound.

3.2. Characterization of the DR Solution

The SP earns revenue from fees but not delays. Therefore, a feasible DR solution (p_1, p_2, d_1, d_2) cannot be optimal if it is possible to maintain the same full cost in a service class while reducing its delay and increasing its price, since this would increase revenues and maintain feasibility.

PROPOSITION 1 **(Structure of the DR solution)**. *It suffices to consider solutions* (p_1, p_2, d_1, d_2) *that satisfy*

(a) $d_1 = 0$*, and*

(b) $p_1 = p_2 + c_1 d_2$.

Recall that $c_1 > c_2$. At the optimal solution $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$, type 1 customers do not wait; type 2 customers wait "only long enough" to satisfy incentive compatibility, i.e., $\bar{p}_1 = \bar{p}_2 + c_1 \bar{d}_2$, and 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 DR solution is "undifferentiated," and if $\bar{p}_1 > \bar{p}_2$ it is "differentiated."[2] If $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ it is "capacitated," and if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ it is "uncapacitated" (since the two cases refer to the DR solutions for which the capacity constraint in (7) is either binding or not). With this in mind, we first answer the question of when the DR solution is differentiated.

Consider the following "single-product problem," in which the SP is constrained to offering only one service class:

$$\max_p \left\{ p(\Lambda_1 + \Lambda_2)\bar{G}(p) : (\Lambda_1 + \Lambda_2)\bar{G}(p) \le s\mu \right\}, \tag{9}$$

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

$$G(p) := \frac{\Lambda_1 F_1(p) + \Lambda_2 F_2(p)}{\Lambda_1 + \Lambda_2}, \qquad g(p) := \frac{\Lambda_1 f_1(p) + \Lambda_2 f_2(p)}{\Lambda_1 + \Lambda_2}. \tag{10}$$

[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."

Table 1 **Categorization of DR solutions** ($N = 2$).

	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	$\bar{p}_1 > \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$	$\bar{p}_1 > \bar{p}_2$ $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$

We assume that there is a unique maximizer of the single-product problem, which we denote by $\hat{p}$.[3] Observe that if the optimal solution to the DR (7) is undifferentiated ($\bar{p}_1 = \bar{p}_2$), then the optimal solution to the single-product problem (9) 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[4] at the single-product optimal price $\hat{p}$. Let $\epsilon_i(p_i, d_i)$ be the demand elasticity for service class i at price p_i and delay d_i, for $i = 1, 2$, and let $\epsilon_g(p)$ be the elasticity of the aggregate demand for a single service class at price p:

$$\epsilon_i(p_i, d_i) = \frac{p_i f_i(p_i + c_i d_i)}{\bar{F}_i(p_i + c_i d_i)}, \qquad \epsilon_g(p) = \frac{p g(p)}{\bar{G}(p)}. \tag{11}$$

Proposition 2 **(Conditions for service differentiation)**. *Assume that the optimal solution of the single-product problem* (9) *has a unique solution,* $\hat{p}$*, and assume that* $\bar{F}_2(\hat{p}) > 0$*. Let* $\bar{p}_1, \bar{p}_2$ *be the optimal prices of the deterministic relaxation* (7)*. Then*

$$\bar{p}_1 > \bar{p}_2 \quad \textit{if and only if} \quad \left(1 - \frac{c_2}{c_1}\right) \epsilon_2(\hat{p}, 0) > \epsilon_g(\hat{p}). \tag{12}$$

We assume that $\bar{F}_2(\hat{p}) > 0$, so that $\epsilon_2(\hat{p}, 0)$ is well-defined.[5] Differentiated services should be offered *if and only if* the demand for type 2 (delay-insensitive) customers at $\hat{p}$ is sufficiently more elastic than the aggregate demand at that price. In that case, the SP may increase revenues by lowering the price for type 2 customers. Elasticity relative to the aggregate demand (as opposed to simply having an elasticity which is greater than 1) allows for the single-product solution to be capacitated. The factor of $(1 - c_2/c_1)$ accounts for the fact that any reduction in class 2 price must be matched by an increase in delays, in order to maintain incentive compatibility.

[3] It is straightforward to extend Proposition 2 to the case of multiple solutions to (9) by requiring that the condition (12) hold for *all* single-product optimal prices. Moreover, uniqueness of $\hat{p}$ is guaranteed if, for example, G is strictly IGFR, but this is an additional assumption and does not follow from IGFR assumptions on individual demand distributions F_1 and F_2.

[4] In general, the demand elasticity at a price p is the proportional change in demand due to a change in price:

$$\epsilon(p) = -\frac{p}{\lambda}\frac{\partial \lambda}{\partial p}.$$

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.

[5] If $\bar{F}_2(\hat{p}) = 0$, it can be shown that a sufficient condition for service differentiation is $\bar{F}_2\left(\hat{p}\left(1 - (1 - c)/\epsilon_g(\hat{p})\right)\right) > 0$.

3.3. Translating the DR Solution

We construct a solution to the stochastic problem (6) based on the results of §3.1-3.2, thereby translating the DR solution into a stochastic solution. The number of services classes k and their respective prices $\bar{p}_1$, $\bar{p}_2$ are taken directly from the DR solution. For $k=1$, this fully specifies the solution (of course, no strategic delay is added to a single class). When two service classes are offered, $k=2$ with $\bar{p}_1 > \bar{p}_2$, the control policy π gives strict preemptive priority to class 1 and strategic delay δ_2 is added to class 2 as needed to discourage type 1 customers (no strategic delay in class 1, $\delta_1 = 0$).

$$\delta_2 = \max(0, \bar{d}_2 - (\mathbb{E}D_2 - \mathbb{E}D_1)).$$

This captures the intuition, from Proposition 1, that class 1 delays should be as small as possible and class 2 delays should be only large enough to guarantee type 1 incentive compatibility.

Henceforth, we will explicitly distinguish between the "DR solution" to (7) and its interpretation in the stochastic system, 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. Asymptotic Performance Analysis

4.1. Preliminaries

We now prove that the stochastic solution prescribed above is asymptotically optimal in the stochastic setting, and induces an equilibrium and operating regime that is consistent with the DR solution. Consider a sequence of systems with increasing capacity and market potential, indexed by n:

$$\begin{aligned} s^n &:= n, \\ \Lambda_i^n &:= n\hat{\Lambda}_i, \quad i = 1, 2, \end{aligned} \tag{13}$$

with $\hat{\Lambda}_i := \Lambda_i / s$, and Λ_i and s are the parameters of the system of original interest. With this definition in place, when $n = s$, the corresponding system in that sequence matches the original system. While the size of each customer segment Λ_i^n scales with capacity, the valuation distribution $F_i(\cdot)$ and delay cost parameter c_i are held fixed. In this way, the customer population grows large, but the characteristics and behavior of *individual* customers remain the same. We use a superscript n to index quantities that depend on the size of the system.

For the nth system in the sequence, the revenue maximization problem is analogous to (6) with quantities having a superscript n replacing their counterparts. The scaled DR revenue rate $n\bar{R}/s$ is again an upper bound on the optimal revenue rate earned in the nth system. The stochastic solution constructed 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 strategic 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 simply the queueing delay $\mathbb{E}D^n$ under the work-conserving control policy π^n. The single-class problem is largely addressed in Maglaras and Zeevi (2003a), whose results easily extend to a heterogenous market of customers that are offered a single service class. In particular, their Theorems 1 and 2 can prove that $\hat{p}$ is asymptotically optimal and the resulting system operates in the QED regime (in the capacitated case).

Differentiated DR solution (two classes). If $\bar{p}_1 > \bar{p}_2$, offer two service classes ($k = 2$) at prices $(\bar{p}_1, \bar{p}_2)$ and add strategic delays $(0, \delta_2^n)$, where $\delta_2^n = \max(0, \bar{d}_2 - (\mathbb{E}D_2^n - \mathbb{E}D_1^n))$. The control policy π^n gives class 1 strict preemptive priority over class 2. For the remainder of this section, we focus on this differentiated case, when necessary distinguishing between the capacitated and uncapacitated cases.

Our first result shows that the stochastic solution yields a unique equilibrium for each system in the sequence, under a *simplified* customer choice model,

$$\lambda_j^n = \Lambda_j^n \bar{F}_j(\bar{p}_j + c_j d_j^n), \quad \text{for } j = 1, 2. \tag{14}$$

In contrast to the demand model described in (2)-(3), (14) explicitly *assumes* that customers choose the "correct" service class, or equivalently, report their type truthfully. We denote by $\rho_j^n = \lambda_j^n / n\mu$ the *traffic intensity* in class $j = 1, 2$. 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 the stochastic solution) converges to the DR solution.

Proposition 3 **(System equilibrium)**. *Assume the scaling in* (13) *and the customer choice model in* (14). *Under the stochastic solution consisting of prices* $(\bar{p}_1, \bar{p}_2)$, *strategic delays* (δ_1^n, δ_2^n), *and priority rule* π^n *described above:*

(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* $j = 1, 2$*;*

(c) as $n \to \infty$, *if the DR solution in* (7) *is capacitated,* $\bar{\kappa}_i + \bar{\kappa}_2 = 1$, *then* $\delta_2^n \to 0$*; and if it is uncapacitated,* $\bar{\kappa}_i + \bar{\kappa}_2 < 1$, *then* $\delta_2^n \to \bar{d}_2$.

4.2. Incentive Compatibility and Revenue Optimality

Proposition 3 establishes the asymptotic system behavior under the assumption that customers make the "correct" choices. Theorem 1 establishes that the stochastic solution becomes incentive compatible in large systems, which implies it is a Nash equilibrium strategy for customers to choose the "correct" service classes (or equivalently to truthfully report their type and valuation).

Theorem 1 **(Large-scale incentive compatibility)**. *Assume the scaling in* (13)*. Then, there exists a finite N_{ic} such that for all $n \geq N_{ic}$, the stochastic solution composed of prices $(\bar{p}_1, \bar{p}_2)$, strategic delays (δ_1^n, δ_2^n), and priority rule π^n described in §4.1 is incentive compatible, namely*

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

Moreover, if the solution is capacitated, $\bar{\lambda}_1 + \bar{\lambda}_2 = s\mu$, then $\delta_2^n = 0$ for all n sufficiently large.

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). So, one does not need to assume that customers make the right choices through (14), as in Proposition 3, but rather the atomistic, utility maximizing behavior of customers described in (2)-(3) guarantee the desired behavior in large systems. If the solution is capacitated, the system congestion creates sufficient queueing delay in class 2 to satisfy the incentive compatibility condition and strategic delay becomes vanishingly small in large systems; if the solution is uncapacitated, queueing delays in both classes will become negligible, in which case, the SP adds strategic delay to class 2 in order to optimally segment the market and ensure that delay-sensitive customers have an incentive to pay a premium for high-priority service (cf. Proposition 3(c)).

We define

$$R^n = \bar{p}_1 \lambda_1^n + \bar{p}_2 \lambda_2^n$$

to be the revenue rate in the nth system generated by this solution.

Theorem 2 **(Asymptotic revenue optimality)**. *Assume the scaling in* (13)*. Then, the revenue rate R^n generated by the stochastic solution composed of prices $(\bar{p}_1, \bar{p}_2)$, strategic delays (δ_1^n, δ_2^n), and priority rule π^n described in §4.1, satisfies*

$$\frac{n\bar{R}}{s} - R^n \leq M, \qquad \text{for all } n \geq N_{ic},$$

for some finite positive constant M, and N_{ic} as in Theorem 1. (Note that $n\bar{R}/s$ is an upper bound on the optimal revenue of the original mechanism design problem (6) *for the scale-n system.)*

Theorem 2 is an unusually strong optimality result. Given that the DR is, in some sense, a fairly crude (first-order) approximation of the mechanism design problem (6), one might expect that the policy predicated on the DR would lead to a performance gap, in terms of revenue, that increases with system size. Indeed, it is typical that system design optimized via a deterministic analysis may result in a asymptotic optimality gap that grows proportionally to $\sqrt{n}$, and that even systems where the "second-order" behavior has been optimized will still have an asymptotic gap that is $o(\sqrt{n})$, but still diverges with n. Indeed, in Maglaras and Zeevi (2003a, 2005) 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. They further optimized the $\sqrt{n}$ behavior so the gap is then $o(\sqrt{n})$, but still diverges with n. Theorem 2 shows that the optimality gap of the policy derived via the static DR *remains bounded*, regardless of the volume of workflow and scale of the resulting revenues. This type of bounded error result is also featured in Randhawa (2013). The underlying driver is that the fluid-optimal solution describes a critically loaded system with non-degenerate delays, which is uniquely determined by the ED regime, and, in turn, guarantees $O(1)$ accuracy of the fluid model. We discuss this in detail in the following section.

4.3. System Operating Regime and Its Implications

The asymptotic 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, $\sqrt{n}(1-\rho^n) \to 0$, implying that arriving customers always have to wait. The QD regime corresponds to lower asymptotic utilization rates where arriving customers never wait, $\sqrt{n}(1-\rho^n) \to \infty$. The next theorem characterizes the operating regime that arises as a consequence of the economic objectives in (6).

Theorem 3 **(System operating regimes)**. *Assume the scaling in* (13)*, and consider the stochastic solution composed of prices* $(\bar{p}_1, \bar{p}_2)$*, strategic delays* (δ_1^n, δ_2^n) *and priority rule* π^n *described in §4.1. Then,* *(a) if the DR solution in* (7) *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 (7) *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.

Relating back to Proposition 3 and Theorem 1, if the DR solution is capacitated, then the resulting equilibrium converges to the ED regime in which the delay of the low priority class emerges due to significant congestion effects (strategic delay vanishes in those cases). The high priority class 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).

The system operating regimes characterized above are the result of economic optimization, and are not *imposed a priori* for analysis purposes. To summarize, i) in a capacitated system, a single-class stochastic solution gives rise to the QED regime (cf. Maglaras and Zeevi (2003a)); 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 strategic delay is required to differentiate the two service classes. Therefore, we show that strategic delay is a first-order effect in the two-class system only in the uncapacitated case, when some fraction of servers are asymptotically always idle. In a system where the service provider sets capacity, with an associated positive cost (e.g. analogous to the setting of §5 in Maglaras and Zeevi (2003a)), this suggests an optimized capacity level avoids permanently idle servers and thus strategic delay will be of second-order importance – i.e., approaches zero as the system grows large. In finite sized systems, the optimal solution may include non-zero strategic delay even when the service provider optimizes capacity.

The $O(1/n)$ convergence characterized by the ED regime also explains the bounded revenue optimality gap in Theorem 2. Note that in the capacitated case

$$\begin{aligned} R^n &= \bar{p}_1\lambda_1^n+\bar{p}_2\lambda_2^n = n\mu\left(\bar{p}_1\rho_1^n+\bar{p}_2\rho_2^n\right), \\ &= n\mu\left(\bar{p}_1(\bar{\kappa}_1+o(1/n))+\bar{p}_2\left(\bar{\kappa}_2-\frac{\alpha}{n}+o(1/n)\right)\right), \\ &= n\mu(\bar{p}_1\bar{\kappa}_1+\bar{p}_2\bar{\kappa}_2)+n\mu\left(\bar{p}_1 o(1/n)-\bar{p}_2\frac{\alpha}{n}+\bar{p}_2 o(1/n)\right), \\ &= \frac{n\bar{R}}{s}-\mu\bar{p}_2\alpha+o(1). \end{aligned} \tag{15}$$

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 3 (Non-preemption). If we restricted our control policy π to non-preemptive priorities, much of this analysis would carry through directly. Class 1 would get strict *non-preemptive* priority in the differentiated case, and prices and strategic delays would remain unchanged. (A different proof would be required to extend Proposition 3(a), which establishes equilibrium delays.) In this setting, both class 1 and class 2 delays will converge to their respective limits at rate $O(1/n)$, and the incentive compatibility and revenue optimality results would carry through. (This is also true, for example, in the appropriately scaled $M/M/1$ system). In contrast, class 1 delay converged exponentially fast to zero in the preemptive case.

Finally, the assumptions on $F_i(\cdot)$, $i=1,2$, can be substantially weakened as long as the DR solution to (7) is guaranteed and accessible. In that case, the results and intuition of Propositions 1 and 3 as well as Theorems 1-3 still hold under much weaker assumptions, for example the functions $F_i(\cdot)$ are only required to be strictly increasing and continuously differentiable in a neighborhood of the DR solution.

5. The Essential Role of Injected Delay

The analysis of the two-type model of the preceding sections establishes that strategic delay becomes asymptotically negligible in large-scale capacitated systems. This sharp insight turns out to hinge crucially on the restrictive assumption of a market with only two segments. In this section we study a market with multiple types ($N \geq 3$) and demonstrate that strategic delay is a first-order effect that is needed to allow differentiation into three or more service classes, regardless of system capacity. The problem formulation and methodology described in §2-4 is readily extended to the multi-type setting. We focus on highlighting additional insights rather than the straightforward extensions of Propositions 1 and 3 or Theorems 1-3.

5.1. Analysis of the Deterministic Relaxation

We consider 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 strategic delay prescription $(\delta_1,\ldots,\delta_N)$ that maximize revenues. The following DR is the analogue of (7):

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} p_i \lambda_i \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i,j=1,\ldots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i=1,\ldots,N \end{aligned} \tag{16}$$

$$\sum_{i=1}^{N} \lambda_i \le s\mu$$
$$d_i \ge 0 \quad i = 1, \ldots, N.$$

The optimal solution to (16), indexed by customer type, is denoted $\bar{p} = (\bar{p}_1, \ldots, \bar{p}_N)$ and $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_N)$, where two or more customer types may have the same price and delay offering. (In the two-type setting, this corresponded to the undifferentiated solution.) The solution to (16) can be expressed with respect to *distinct* service classes, denoted by $\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$ 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 prefer class j to any other service class (i.e., $\bar{p}_i = \hat{p}_{(j)}$ and $\bar{d}_i = \hat{d}_{(j)}$ for all $i \in A_{(j)}$). We will call the sets $A_{(j)}$, $j = 1, \ldots, k$, "market segments." Note that a customer *prefers* one service class over others but may still *choose* the no-purchase option. Therefore Lemma 1 does not claim that it is optimal to *serve* consecutive types and the optimal solution to (16) may satisfy (17) and still price out intermediate customers types. More technically, these market segments reflect the structure of the incentive compatibility conditions, but not individual rationality conditions.

Generalizing Proposition 1, it suffices to consider solutions that satisfy

$$d_1 = 0 \quad \text{and} \quad p_i + c_i d_i = p_{i+1} + c_i d_{i+1} \text{ for } i = 1, \ldots, N-1. \tag{17}$$

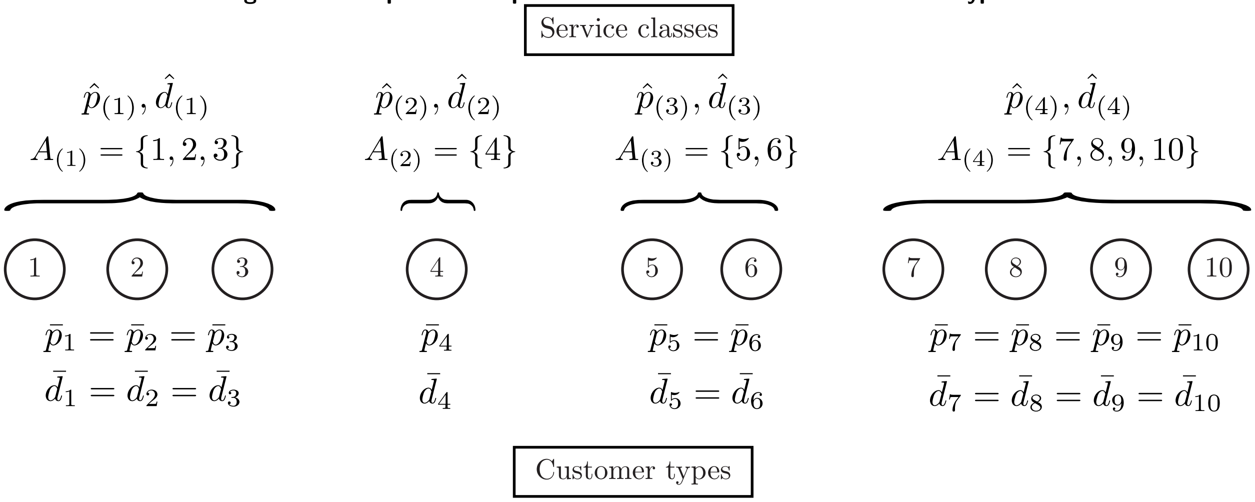
In the multi-type setting, this structure describes the optimal pooling of customer types in the DR.

LEMMA 1. *For any feasible solution to* (16) $(p_1, \ldots, p_N)$, $(d_1, \ldots, d_N)$ *that satisfies the conditions* (17)*, the market segments* $A_{(j)}$*,* $j = 1, \ldots, k$ *are contiguous in the following sense*

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

Lemma 1 shows that the market segments $A_{(j)}$, $j = 1, \ldots, k$, consist of *consecutive* customer types (recall that customer types are ordered by their delay sensitivity $c_1 > c_2 > \cdots > c_N$). 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 1. We note that a partial extension to Proposition 2 may be derived, but it adds little insight.

Figure 1 **Depiction of optimal DR solution for $N = 10$ customer types.**


Service classes
$\hat{p}_{(1)}, \hat{d}_{(1)}$
$A_{(1)} = \{1,2,3\}$
$\hat{p}_{(2)}, \hat{d}_{(2)}$
$A_{(2)} = \{4\}$
$\hat{p}_{(3)}, \hat{d}_{(3)}$
$A_{(3)} = \{5,6\}$
$\hat{p}_{(4)}, \hat{d}_{(4)}$
$A_{(4)} = \{7,8,9,10\}$
1
2
3
4
5
6
7
8
9
10
$\bar{p}_1 = \bar{p}_2 = \bar{p}_3$
$\bar{d}_1 = \bar{d}_2 = \bar{d}_3$
$\bar{p}_4$
$\bar{d}_4$
$\bar{p}_5 = \bar{p}_6$
$\bar{d}_5 = \bar{d}_6$
$\bar{p}_7 = \bar{p}_8 = \bar{p}_9 = \bar{p}_{10}$
$\bar{d}_7 = \bar{d}_8 = \bar{d}_9 = \bar{d}_{10}$
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.

5.2. Prescribed Solution for the Stochastic System

Suppose the DR solution to (16) offers k distinct 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 $A_{(1)}, \ldots, A_{(k)}$. At the DR solution, we define the relative workload contribution from class $j \in \{1, \ldots, k\}$ 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 again specify a stochastic solution with the same number of service classes and prices as the DR, combined with strict preemptive priorities and strategic delays that are added only if queueing delays are insufficient. If $k = 1$, there is only a single class priced at $\hat{p}_{(1)}$; no priorities or strategic delays are needed. If $k \geq 2$ there are k service classes with prices $\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$, served with a strict preemptive priority rule, with highest priority given to class 1 and lowest to class k. Strategic delay is given by $\delta = (\delta_{(1)}, \ldots, \delta_{(k)})$, where: $\delta_{(1)} = 0$ and

$$\delta_{(j)} = \max(0, \hat{d}_{(j)} - (\mathbb{E}D_{(j)} - \mathbb{E}D_{(j-1)})) \quad \text{for } j = 2, \ldots, k.$$

Applying the scaling in (13) to all customer types $i = 1, \ldots, N$, the demand for each class j in the nth system in the sequence is given by

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

where $d^n_{(j)} = \mathbb{E}D^n_{(j)} + \delta^n_{(j)}$ is the overall delay. The revenue earned in the nth system under our solution is $R^n = \sum_{j=1}^k \hat{p}_{(j)} \gamma^n_{(j)}$.

Necessity of strategic delay. Proposition 3 and Theorems 1-3 all generalize in the multi-class case. Focusing on the intermediate classes $j = 2, \ldots, k-1$, i.e., excluding the highest and lowest priority classes, the strategic delay added to an intermediate class j is non-vanishing in large systems,

$$\delta^n_{(j)} \to \hat{d}_{(j)} \qquad \text{as } n \to \infty,$$

irrespective of capacity utilization. The limiting amount of strategic delay added to the lowest priority class k depends on the capacity constraint, as it did in the two-class setting. Essentially, the priority rule causes all congestion to be experienced in only the lowest priority class, so first-order strategic delay must be added to differentiate intermediate service classes.

Remark 4 (Connection to Afèche (2013)). Afèche (2013) introduced a mechanism design (incentive-compatible) formulation of revenue maximization problems in queueing systems, where he was the first to demonstrate the use of strategic delay in the context of revenue maximization in a queueing system, highlight the use of delay in the low priority class to achieve incentive compatibility, the importance of capacity, and obtain parameter conditions that favor differentiation. His study focused on a two-type market served by an $M/M/1$ system and used exact analysis, and some of his results and conditions imposed further restrictions on the valuation distributions. Some of his results may be extended to service systems in which the achievable region of delays is explicitly and tractably characterized, including a two-class multi-server queue. As pointed out in § 7 of Afèche (2013), the exact analysis approach based on the achievable region may become intractable in queueing systems of increasing complexity, including multi-type and multi-class queues, whereat progress is made by imposing additional restrictions on the customer market. Our analysis leverages Afèche's formulation but uses a more tractable framework that relies on the solution of a much simpler deterministic relaxation and asymptotic approximations. Such model approximations are justified via asymptotic limits in large-scale systems, and offer a framework that generates strong insights regarding first-order drivers of optimized system performance and allows the treatment of systems that may not be amenable to exact analysis. The latter is underscored by the analysis of a market with multiple ($N \geq 3$) types. As previously mentioned in Remark 2, the DR may not, in the generality presented here, yield closed-form expressions for optimal prices. However, when numerical computation is required, the DR solution is likely considerably easier to compute than the exact solution, which additionally depends on the queueing delay equilibrium. Moreover, once a DR solution is found, all of its features (price, service differentiation, and insight into operational considerations) carry over as first-order drivers of system performance in an asymptotically optimal solution to the stochastic problem. The insights gleaned from model approximations become accurate in systems and application settings characterized by large processing capacity and large market potential. For example, while the exact analysis of Afèche (2013) simply shows that the two customer types are always offered distinct service classes (if both types are present in the system), our asymptotic analysis suggests that this distinction may become negligible in large systems, in particular when type 2 demand is sufficiently inelastic (in the sense of Proposition 2). An even more extreme example of asymptotically negligible differentiation is detailed in the next section.

Remark 5 (An alternative implementation). Is it possible to achieve the same degree of delay differentiation if $k \geq 3$ without the use of strategic delay in a capacitated system? While the answer is affirmative, the resulting heuristic may not be desirable. For example, suppose $k = 3$ and 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 $\hat{p}_{(3)}$ get assigned the second (low) priority queue. Users that select the intermediate service class $\hat{p}_{(2)}$ get assigned to the high priority queue with probability $1 - \hat{d}_{(2)}/\hat{d}_{(3)}$ and to the low priority queue with probability $\hat{d}_{(2)}/\hat{d}_{(3)}$, which results in an average delay that converges to $\hat{d}_{(2)}$. One can verify that this policy is incentive compatible and results in near-optimal revenues. 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 (6), this can be reformulated as a mechanism design problem:

$$\begin{aligned} \text{maximize} \quad & W(p,d) = \sum_{i=1}^{N} \Lambda_i \left(\int_{p_i + c_i d_i}^{\infty} v f_i(v)\, dv - c_i d_i \bar{F}_i(p_i + c_i d_i) \right) \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i,j = 1,\ldots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,\ldots,N \\ & \sum_{i=1}^{N} \lambda_i < s\mu \end{aligned} \tag{18}$$

Money transfers from customers to the SP are "internal" and are not reflected in the welfare objective.

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 (18):

$$\begin{aligned} \text{maximize} \quad & W(p,d) \\ \text{subject to} \quad & p_i + c_i d_i \le p_j + c_i d_j \quad i,j = 1,\ldots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,\ldots,N \\ & \sum_{i=1}^{N} \lambda_i \le s\mu \\ & p_i \ge 0, d_i \ge 0 \quad i = 1,\ldots,N. \end{aligned} \tag{19}$$

The social-welfare objective is equivalent to delay-cost minimization and so, in the DR setting (19), the optimal solution is unique and undifferentiated[6] with zero delay and optimal price $\hat{p}_{soc}$,

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

Since we expect the DR to be asymptotically optimal in large systems, this suggests that as the system size grows large, the optimal strategy identified by the Mendelson-Whang solution degenerates to a single-class offering. That would imply that delay differentiation is always asymptotically negligible in the social welfare setting.

To be more precise, the Mendelson-Whang solution under the scaling (13), prescribes the vector of social welfare optimal prices in the nth system, $p^n = (p_1^n, \ldots, p_N^n)$, to be

$$p_j^n = \sum_{\ell=1}^{N} c_\ell \lambda_\ell^n \frac{\partial \mathbb{E}D_\ell^n}{\partial \lambda_j^n}, \quad j = 1,\ldots,N. \tag{20}$$

Here, $\lambda_j^n = \Lambda_j^n \bar{F}(p_j^n + c_j \mathbb{E}D_j^n)$ is the demand rate, and $\mathbb{E}D_j^n$ is the queueing delay in each class $j = 1,\ldots,N$ under a strict preemptive priority policy π^n that gives class j priority over class $j+1$. Let $\rho_j^n = \lambda_j^n / n\mu$ denote the traffic intensity in class j in the nth system under this optimal solution.

[6] This assumes that the model primitives are such that $\bar{F}_N(\hat{p}_{soc}) > 0$ to rule out meaningless "differentiation" for the Nth type, such as $p_N = 0$, $d_N = \hat{p}_{soc}/c_N$.

Proposition 4 **(Social welfare solution structure)**. *Assume the scaling in* (13) *and assume that* $\bar{F}_N(\hat{p}_{soc}) > 0$. *Then as* $n \to \infty$,

(a) $p^n_{j*} \to \hat{p}_{soc}$ *and* $\mathbb{E}D^n_j \to 0$ *for* $j = 1, \ldots, N$;

(b) if $\hat{p}_{soc} > 0$ *then* $\sqrt{n}\left(1 - \sum_{j=1}^N \rho^n_{j*}\right) \to \beta$ *for some strictly positive, finite constant* β *that depends on model primitives.*

Part (a) asserts that the DR indeed captures the first order properties of the optimal solution for the original mechanism design problem (18), 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) asserts that a capacitated 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 in a market with a single customer type. In contrast, revenue maximization requires significant delay differentiation to extract, in return, significant price premia, and this leads the system to operate in the ED regime that is accompanied by higher resource utilization rates.

Appendix A: Proofs

This appendix contains the proofs of Propositions 2-3 and Theorems 1-2. We defer the proofs of Proposition 1, Lemma 1, and Proposition 4 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 $(1 - c_2/c_1)\epsilon_2(\hat{p}, 0) \leq \epsilon_g(\hat{p})$.

Fix (p_1, p_2, d_1, d_2) to be a feasible solution to the DR (7) 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

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

respectively, where $c := c_2/c_1$. Define the functions $\kappa_1(p_1, d_1)$ and $\kappa_2(p_2, d_2)$ to be the relative workload contributions by class 1 and class 2, respectively, at the price point (p_1, d_1, p_2, d_2):

$$\kappa_1(p_1, d_1) := \frac{\Lambda_1 \bar{F}_1(p_1 + c_1 d_1)}{s\mu}, \qquad \kappa_2(p_1, d_2) := \frac{\Lambda_1 \bar{F}_2(p_2 + c_2 d_2)}{s\mu}. \tag{21}$$

The following result, specifically (22), proves the "only if" part of the above assertion.

Lemma 2. *Let* $\hat{p}$ *be the optimal solution to the single-product problem* (9), *and let* $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ *be the optimal solution to the DR* (7). *Then*

$$\bar{p}_1 = \bar{p}_2 = \hat{p} \quad \textit{implies} \quad (1 - c)\epsilon_2(\hat{p}, 0) \leq \epsilon_g(\hat{p}) \quad \textit{and} \tag{22}$$

$$\bar{p}_1 > \bar{p}_2 \quad \textit{implies} \quad \frac{\epsilon_1(\bar{p}_1, 0)}{\bar{p}_1} < \left(1 - \frac{c}{1 - c}\frac{\kappa_2(\bar{p}_2, \bar{d}_2)}{\kappa_1(\bar{p}_1, 0)}\right)(1 - c)\frac{\epsilon_2(\bar{p}_2, \bar{d}_2)}{\bar{p}_2}, \tag{23}$$

where $\epsilon_i(p_i, d_i)$, $i = 1, 2$ *and* $\epsilon_g(p)$ *are the price elasticities defined in* (11) *and* $\kappa_i(p_i, d_i)$, $i = 1, 2$, *are defined in* (21).

It remains to show that $\epsilon_2(\hat{p},0) \leq \epsilon_g(\hat{p})$ implies $\bar{p}_1 = \bar{p}_2 = \hat{p}$. Note that (23) is equivalent to the statement that $\bar{p}_1 = \bar{p}_2 = \hat{p}$, provided that

$$\frac{\epsilon_1(\bar{p}_1,0)}{\bar{p}_1} \geq \left(1 - \frac{c}{1-c}\frac{\kappa_2(\bar{p}_2,\bar{d}_2)}{\kappa_1(\bar{p}_1,0)}\right)(1-c)\frac{\epsilon_2(\bar{p}_2,\bar{d}_2)}{\bar{p}_2}.$$

Also, if $\bar{p}_1 = \bar{p}_2 = \hat{p}$ then $\bar{d}_2 = 0$, and hence

$$\epsilon_1(\hat{p},0) \geq \left(1 - \frac{c}{1-c}\frac{\kappa_2(\hat{p},0)}{\kappa_1(\hat{p},0)}\right)(1-c)\epsilon_2(\hat{p},0),$$

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

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

Some algebraic manipulation yields

$$\begin{aligned}\Lambda_1 f_1(\hat{p}) &\geq \left((1-c)\Lambda_1\bar{F}_1(\hat{p}) - c\Lambda_2\bar{F}_2(\hat{p})\right)\frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})},\\ \frac{\Lambda_1 f_1(\hat{p}) + \Lambda_2 f_2(\hat{p})}{\Lambda_1\bar{F}_1(\hat{p}) + \Lambda_2\bar{F}_2(\hat{p})} &\geq (1-c)\frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})},\\ \epsilon_g(\hat{p}) &\geq (1-c)\epsilon_2(\hat{p},0),\end{aligned}$$

and we deduce that $(1-c)\epsilon_2(\hat{p},\hat{p}) \leq \epsilon_g(\hat{p})$ 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 (13).

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 π^n. 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 π^n. 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 $d_i^n \geq 0$: $\lambda_i^n(d_i^n) = \Lambda_i^n \bar{F}_i(\bar{p}_i + c_i d_i^n)$, $i = 1,2$. In class 2, strategic delay is added such that the overall delay $d_2^n = \delta_2^n + \xi_2^n = \max\{\xi_2^n, \xi_1^n + \bar{d}_2\}$. Note that $\lambda_i^n(d_i^n)$ is monotone non-increasing 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

$$\begin{aligned} \lambda_1^n(\xi_1^n) + \lambda_2^n(\delta_2^n + \xi_2^n) &< n\mu, \\ \mathbb{E}D_1^n\left(\lambda_1^n(\xi_1^n)\right) &= \xi_1^n, \\ \mathbb{E}D_2^n\left(\lambda_1^n(\xi_1^n), \lambda_2^n(\delta_2^n + \xi_2^n)\right) &= \xi_2^n. \end{aligned} \tag{24}$$

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 $h_1(x)$ exists for all $x \geq 0$, since $\lambda_1^n(0) = \Lambda_1^n \bar{F}_1(\bar{p}_1) < n\mu$, and is continuous with $h_1(0) < 0$ and $h_1(\infty) > 0$ (since $\lambda_1^n(\infty) = 0$). 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

$$h_2(x) := x - \max\{\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n(\max\{x, \xi_1^n + \bar{d}_2\})), \xi_1^n + \bar{d}_2\}.$$

Note that $h_2(x)$ exists for all $x \geq 0$, since $\lambda_2^n(\xi_1^n + \bar{d}_2) < n\mu - \lambda_1^n$, and is continuous with $h_2(0) < 0$ and $h_2(\infty) > 0$ (since $\lambda_2^n(\infty) = 0$). Furthermore, $h_2(x)$ is monotone increasing in x since the second term is monotone non-increasing in x. Therefore, there exists a unique ξ_2^n such that $h_2(\xi_2^n) = 0$.

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 (24), or equivalently the traffic intensity pair (ρ_1^n, ρ_2^n), where

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

$d_1^n = \xi_1^n$, and $d_2^n = \max\{\xi_2^n, \xi_1^n + \bar{d}_2\}$. 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,

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

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 $0 \leq \underline{g} < \overline{g} \leq \infty$.

LEMMA 3. *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 3, 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 $d_i^n \to \bar{d}_i$, $i = 1,2$.

First, observe that $d_1^n = \mathbb{E}D_1^n > \bar{d}_1 = 0$ and $d_2^n = \max\{\mathbb{E}D_2^n, \bar{d}_2 + \mathbb{E}D_1^n\} > \bar{d}_2$. Therefore,

$$\rho_1^n + \rho_2^n = \frac{\hat{\Lambda}_1}{\mu}\bar{F}_1(\bar{p}_1 + c_1 d_1^n) + \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2 d_2^n) < \bar{\kappa}_1 + \bar{\kappa}_2 \le 1.$$

In the uncapacitated case, $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ so $\rho_1^n + \rho_2^n$ is bounded away from 1 so $\mathbb{E}D_1^n \to 0$ and $\mathbb{E}D_2^n \to 0$, and we conclude that $d_1^n \to 0$, $d_2^n \to \bar{d}_2$, and $\delta_2^n \to \bar{d}_2$.

In the capacitated case, $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ ($\bar{\kappa}_2 > 0$), $\rho_1^n < \bar{\kappa}_1 < 1$ is bounded away from 1 so $\mathbb{E}D_1^n \to 0$ and therefore $d_1^n \to 0$. Since $\bar{F}_1$ is continuous, this implies that $\rho_1^n \to \bar{\kappa}_1$.

For class 2, suppose $\lim_{n\to\infty} d_2^n > \bar{d}_2$, then there exists $\epsilon > 0$ such that for all n sufficiently large

$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2 d_2^n) \le \bar{\kappa}_2 - \epsilon.$$

since $\bar{F}_2(\cdot)$ is strictly decreasing. Therefore, eventually $\rho_1^n + \rho_2^n < 1$, which implies $\mathbb{E}D_2^n \to 0$, in contradiction. Since $d_2^n > \bar{d}_2$ for all n, we conclude that $d_2^n \to \bar{d}_2$ and, by continuity of $\bar{F}_1(\cdot)$, $\rho_2^n \to \bar{\kappa}_2$.

Proof of (c) (Strategic delay). For the uncapacitated case, since $\mathbb{E}D_2^n \to 0$ and $d_2^n \to \bar{d}_2$, it must be that $\delta_2^n \to \bar{d}_2$. For the capacitated case, we defer to the proof of Lemma 4, where it is shown that $\mathbb{E}D_2^n \to \bar{d}_2$.

This completes the proof. □

The following Lemma is central to the proof of Theorem 1-3.

LEMMA 4 **(Rates of convergence)**. *Assume the scaling in* (13). *Set the stochastic solution to prices $(\bar{p}_1, \bar{p}_2)$, strategic delays (δ_1^n, δ_2^n), and priority rule π^n described in §4.1. 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 ($\bar{\kappa}_1 + \bar{\kappa}_2 < 1$),

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + o(1/n), \tag{25}$$

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

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + \mathcal{O}(1/n). \tag{26}$$

Proof of Lemma 4. We prove this in three steps.

Step 1. We first prove that $d_1^n = o(1/n)$ in both the capacitated and uncapacitated cases. From Proposition 3(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 $d_2^n = \bar{d}_2 + o(1/n)$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$, so we have proven (25).

Step 2. We now provide an intermediate step showing that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$ in both the capacitated and uncapacitated cases. Since $F_1(\cdot)$ is continuously differentiable, the mean value theorem ensures that there exists some $\tilde{d}_1^n \in [0, d_1^n]$ such that

$$\rho_1^n = \frac{\hat{\Lambda}_1\bar{F}_1(\bar{p}_1 + c_1 d_1^n)}{\mu} = \underbrace{\frac{\hat{\Lambda}_1\bar{F}_1(\bar{p}_1)}{\mu}}_{=\bar{\kappa}_1} - d_1^n\frac{c_1\hat{\Lambda}_1 f_1(\bar{p}_1 + c_1\tilde{d}_1^n)}{\mu}$$

and therefore

$$n(\bar{\kappa}_1 - \rho_1^n) = nd_1^n\frac{c_1\hat{\Lambda}_1 f_1(\bar{p}_1 + c_1\tilde{d}_1^n)}{\mu}.$$

Since $nd_1^n \to 0$ as $n \to \infty$ and $\tilde{d}_1^n \le d_1^n$ we conclude that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$. (A nearly identical argument also proves $n(\bar{\kappa}_2 - \rho_2^n) \to 0$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$.)

Step 3. This step proves the d_2^n rate of convergence in the capacitated case, (26). $F_2(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}_2^n \in [\bar{d}_2, d_2^n]$ 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 $f_2(\bar{p}_2 + c_2\tilde{d}_2^n) \to f_2(\bar{p}_2 + c_2\bar{d}_2) > 0$. Note that $(1-\rho^n) = (\bar{\kappa}_1 - \rho_1^n) + (\bar{\kappa}_2 - \rho_2^n)$, which combined with the result of Step 2, gives us

$$\lim_{n\to\infty} n(1-\rho^n) = \lim_{n\to\infty} n(\bar{\kappa}_2 - \rho_2^n) = \frac{c_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\bar{d}_2)}{\mu}\lim_{n\to\infty} n(d_2^n - \bar{d}_2).$$

Recall that $d_2^n - \bar{d}_2 = \max\{\mathbb{E}D_2^n - \bar{d}_2, \mathbb{E}D_1^n\}$ and therefore

$$\lim_{n\to\infty} n(d_2^n - \bar{d}_2) = \max\left\{\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2), 0\right\}.$$

If $\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) \leq 0$ then $n(1-\rho^n) \to 0$ and, by Lemma 3, $\mathbb{E}D_2^n \geq \mathbb{E}D^n \to \infty$, a contradiction. Similarly, if $\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) = \infty$ then $\mathbb{E}D^n \to 0$ and therefore $\mathbb{E}D_2^n \to 0$, again a contradiction. Therefore, it must be that

$$\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) = \frac{1}{c_2\bar{\kappa}_2\bar{d}_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\bar{d}_2)} \in (0,\infty)$$

since $\rho_1^n\mathbb{E}D_1^n + \rho_2^n\mathbb{E}D_2^n = (\rho_1^n + \rho_2^n)\mathbb{E}D^n$ implying $\mathbb{E}D^n \to \bar{\kappa}\bar{d}_2$ and $n(1-\rho^n) \to 1/\mu\bar{\kappa}_2\bar{d}_2$. Therefore $d_2^n = \bar{d}_2 + \mathcal{O}(1/n)$, proving the remainder of (26). □

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 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 $n \geq N_{ic}$, a system where the customer demand model is given by (2)-(3), 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

$$\bar{d}_2 \leq (d_2^n - d_1^n) \leq \frac{c_1}{c_2}\bar{d}_2. \tag{27}$$

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} such that for all $n \geq N_{ic}$, $d_2^n - d_1^n \leq \frac{c_1}{c_2}\bar{d}_2$. Strategic delay δ_2^n ensures that the left hand inequality is satisfied for all n.

In the capacitated case, the results of Lemma 4 show that

$$d_2^n = \max\{\mathbb{E}D_2^n, \bar{d}_2 + \mathbb{E}D_1^n\} = \max\{\bar{d}_2 + \mathcal{O}(1/n), \bar{d}_2 + o(1/n)\}$$

and therefore $\mathbb{E}D_2^n > \bar{d}_2 + \mathbb{E}D_1^n$ and $\delta_2^n = 0$ for all n sufficiently large (this may be larger than N_{ic}).

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

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

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

$$\begin{aligned} R^n &= n\mu(\bar{p}_1\bar{\kappa}_1 + \bar{p}_2\bar{\kappa}_2) - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n) \\ &= \frac{n\bar{R}}{s} - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n). \end{aligned}$$

From (25) and (26) 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 4 are satisfied for the sequence of systems indexed by n, starting at N_{ic}, and the results of Proposition 3 and Lemma 4 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. □

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

Queueing Dynamics

We represent the control policy π 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. As an example, 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:

$$\pi_1(t) = \min\{s, Z_1(t)\} \qquad \pi_j(t) = \min\{(s - Z_1(t) - \cdots - Z_{j-1}(t))^+, Z_j(t)\}, \quad j = 2,\ldots,k, \tag{28}$$

where $(Z_1(t),\ldots,Z_k(t))$ is the headcount process defined below.

We now define the system dynamics for fixed arrival rate vector $\lambda = (\lambda_1,\ldots,\lambda_k)$ and control policy $\pi(t)$. Consider $2k$ mutually independent unit-rate Poisson processes, $N_j^{(a)}(t)$ and $N_j^{(s)}(t)$ for $j = 1,\ldots,k$. $N_j^{(a)}(\lambda_j t)$ is the number of customers that have arrived into class j by time t and $N_j^{(s)}\left(\int_0^t \mu\pi_j(s)\,ds\right)$ is the number of class j customers that have completed service by time t. The system may be described in terms of the "headcount process" $((Z_1(t),\ldots,Z_k(t)) : 0 \leq t < \infty)$ where $Z_j(t)$ is the number of class j customers in the system *excluding the delay node* at time t, and the "queue length process" $((Q_1(t),\ldots,Q_k(t)) : 0 \leq t < \infty)$ where $Q_j(t)$ is the number of class j customers in queue at time t. These processes must jointly satisfy the following conditions:

$$\sum_{j=1}^k \pi_j(t) = \min\left\{s, \sum_{j=1}^k Z_j(t)\right\}, \tag{29}$$

$$Q_j(t) = Z_j(t) - \pi_j(t) \geq 0 \quad \text{for } j = 1,\ldots,k, \tag{30}$$

$$Z_j(t) = N_j^{(a)}(\lambda_j t) - N_j^{(s)}\left(\int_0^t \mu\pi_j(s)\,ds\right) \geq 0 \quad \text{for } j = 1,\ldots,k. \tag{31}$$

Condition (29) ensures the total number of servers working at any time does not exceed s, and that no servers idle while there are customers waiting in the queue. Condition (30) 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 (31) describes the system dynamics. We require that the control π, $(\pi_1(t), \ldots, \pi_k(t))$ be adapted to the filtration generated by $(Z_1(t), \ldots, Z_k(t))$.

Proof of Proposition 1. We prove the general N-type case stated in (17). Note that in the case of additive, linear delay costs, local incentive compatibility implies global incentive compatibility. This is also shown in Lemma 2 of Katta and Sethuraman (2005) although we clarify that the assumption $d_1 \leq d_2 \leq \ldots d_N$ is redundant.

Lemma 5 **(Local incentive compatibility implies global incentive compatibility.)**.

$$\begin{aligned} p_i + c_i d_i &\leq p_{i+1} + c_i d_{i+1} \quad \textit{for } i = 1, \ldots, N-1 \\ p_i + c_i d_i &\leq p_{i-1} + c_i d_{i-1} \quad \textit{for } i = 2, \ldots, N \end{aligned}$$

implies

$$p_i + c_i d_i \leq p_j + c_i d_j \quad \textit{for all } i, j = 1, \ldots, N.$$

Proof of Lemma 5. First, we note that local incentive compatibility is equivalent to

$$c_{i+1}(d_{i+1} - d_i) \leq p_i - p_{i+1} \leq c_i(d_{i+1} - d_i)$$

and since $c_i > c_{i+1}$ this implies that $d_{i+1} \geq d_i$ and $p_i \geq p_{i+1}$. We now prove by induction.

Fix $i \in 1, \ldots, N-2$. For $j > i$, assume $p_i + c_i d_i \leq p_j + c_i d_j$ (the base case $j = i+1$ is true by local incentive compatibility).

$$\begin{aligned} p_{j+1} + c_i d_{j+1} &= p_{j+1} + c_j d_{j+1} + (c_i - c_j) d_{j+1} \\ &\geq p_j + c_j d_j + (c_i - c_j) d_{j+1} \\ &= p_j + c_i d_j + (c_i - c_j)(d_{j+1} - d_j) \\ &\geq p_i + c_i d_i \end{aligned}$$

Fix $i \in 3, \ldots, N$. For $j < i$, assume $p_i + c_i d_i \leq p_j + c_i d_j$ (the base case $j = i-1$ is true by local incentive compatibility).

$$\begin{aligned} p_{j-1} + c_i d_{j-1} &= p_{j-1} + c_j d_{j-1} - (c_j - c_i) d_{j-1} \\ &\geq p_j + c_j d_j - (c_j - c_i) d_{j-1} \\ &= p_j + c_i d_j + (c_j - c_i)(d_j - d_{j-1}) \\ &\geq p_i + c_i d_i \end{aligned}$$

This concludes the proof of Lemma 5. □

Supposing each property does *not* hold for a feasible solution $(p_1,\dots,p_N)$, $(d_1,\dots,d_N)$, we construct an alternative solution $(\breve{p}_1,\dots,\breve{p}_N)$, $(\breve{d}_1,\dots,\breve{d}_N)$, that satisfies the property, is feasible, and achieves at least as high a revenue rate. In particular, the alternative solution is constructed to satisfy $\breve{p}_i + c_i\breve{d}_i = p_i + c_i d_i$ for all $i = 1,\dots,N$, which guarantees that the capacity constraint is satisfied, and it is trivial to check local incentive compatibility and therefore global incentive compatibility.

Proof of (a). Suppose $d_1 > 0$. 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 $i = 2,\dots,N$. Note that if $\bar{F}_1(p_1 + c_1 d_1) > 0$ then revenues are *strictly* improved.

Proof of (b). Suppose $p_i + c_i d_i < p_{i+1} + c_i d_{i+1}$. Take

$$\breve{p}_{i+1} = \frac{c_i(p_{i+1} + c_{i+1}d_{i+1}) - c_{i+1}(p_i + c_i d_i)}{c_i - c_{i+1}} \qquad \breve{d}_{i+1} = \frac{p_i + c_i d_i - p_{i+1} - c_{i+1}d_{i+1}}{c_i - c_{i+1}}$$

and $\breve{p}_j = p_j$, $\breve{d}_j = d_j$ for $j \neq i+1$. Note that if $\bar{F}_{i+1}(p_{i+1} + c_{i+1}d_{i+1}) > 0$ then revenues are *strictly* improved.

Proof of Proposition 1. A feasible solution that satisfies (17) implies

$$d_i = d_{i-1} + \frac{1}{c_{i-1}}(p_{i-1} - p_i).$$

Since incentive compatibility implies $p_1 \geq p_2 \geq \cdots \geq p_N$, we see that if $p_i = p_j$ for some $i > j$ then $p_i = p_{i+1} = \cdots = p_j$ and $d_i = d_{i+1} = \cdots = d_j$. Therefore the sets $\{A_{(1)},\dots,A_{(N)}\}$ must have the structure described. Note that it is possible for $i \in A_{(j)}$ and $\bar{F}_i(p_{(j)} + c_i d_{(j)}) = 0$, in which case no type i customers will purchase service. However, the solution will still segment the market such that type i customers are in the jth segment. □

Proof of Lemma 2. Apply Proposition 1 to reduce the deterministic relaxation (7) to two variables p_1 and p_2, and set $c := \frac{c_2}{c_1} < 1$,

$$\begin{aligned} \text{maximize} \quad & \Lambda_1 p_1 \bar{F}_1(p_1) + \Lambda_2 p_2 \bar{F}_2(cp_1 + (1-c)p_2) \\ \text{subject to} \quad & p_1 \geq p_2 \\ & \Lambda_1 \bar{F}_1(p_1) + \Lambda_2 \bar{F}_2(cp_1 + (1-c)p_2) \leq s\mu. \end{aligned} \tag{32}$$

Equations (22) and (23) follow from the KKT necessary conditions of (32). □

Proof of Lemma 3. Lemma 3 follows immediately from Lemma 6 and the $M/M/n$ delay formula. □

LEMMA 6 **(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 4.

Proof of (a). Let $\mathbb{E}D^n_{j*}$ be the queueing delay for class j, $j = 1,\dots,N$, in the nth system operating under the optimal prices p^n_{j*} and a strict priority rule. Let W^n_* be the optimal social welfare under this solution.

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. 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

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

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

$$\mathbb{E}D^n_{1*} = \psi^n(\rho^n_{1*}), \qquad \text{and} \qquad \mathbb{E}D^n_{j*} = \frac{\omega^n_{j*}\psi^n(\omega^n_{j*})}{\rho^n_{j*}} - \frac{\omega^n_{(j-1)*}\psi^n(\omega^n_{(j-1)*})}{\rho^n_{j*}} \quad \text{for } j = 2, \ldots, N. \tag{33}$$

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

$$\nu^n(x) := \left(\sum_{j=0}^{n-1} \frac{(nx)^j}{j!} + \frac{(nx)^n}{n!(1-x)} \right)^{-1} \frac{(nx)^n}{n!(1-x)} \qquad \text{and} \qquad \psi^n(x) := \frac{\nu^n(x)}{n\mu(1-x)}. \tag{34}$$

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 4) and therefore $\sqrt{n}(\kappa_{j*} - \rho^n_{j*}) \to 0$ for all $j = 1, \ldots, N-1$. It remains to show that $\sqrt{n}(\kappa_{N*} - \rho^n_{N*}) \to \beta \in (0, \infty)$.

$F_N(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}^n \in [0, \mathbb{E}D^n_{N*}]$ 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

$$\begin{aligned}
n(1-\omega^n_{N*})(\kappa_{N*} - \rho^n_{N*}) &= \sum_{j=1}^{N-1} n\left(\kappa^n_{j*} - \rho^n_{j*}\right)(\kappa_{N*} - \rho^n_{N*}) + n(\kappa_{N*} - \rho^n_{N*})^2 \\
\lim_{n\to\infty} n(1-\omega^n_{N*})(\kappa_{N*} - \rho^n_{N*}) &= \lim_{n\to\infty} n(\kappa_{N*} - \rho^n_{N*})^2.
\end{aligned}$$

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 6, it must be that $\sqrt{n}(\kappa_{N*} - \rho^n_{N*}) \to \beta \in (0, \infty)$. □

References

Afèche, Philipp. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delay tactics. Working paper.

Afèche, Philipp. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.

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.

Armony, Mor, Costis Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.

Armony, Mor, Costis Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* **52**(2) 271–292.

Borst, Sem, Avi Mandelbaum, Martin I. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

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.

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

Plambeck, Erica L., Amy R. Ward. 2006. Optimal control of a high-volume assemble-to-order system. *Mathematics of Operations Research* **31**(3) 453–477.

Randhawa, Ramandeep S. 2013. Accuracy of fluid approximations for queueing systems with congestion-sensitive demand and implications for capacity sizing. *Operations Research Letters* **41**(1) 27–31.

Saaty, Thomas L. 1961. *Elements of Queueing Theory with Applications*. McGraw-Hill.

Whitt, Ward. 2003. How multiserver queues scale with growing congestion-dependent demand. *Operations Research* **51**(4) 531–542.