Pricing and Design of Differentiated Services: Approximate Analysis and Structural Insights

Constantinos Maglaras Assaf Zeevi[*]

Submitted: April 2003 Revised: November 2003; March 2004

To appear in *Operation Research, 2005*

Abstract

We consider a model of a service system that delivers two non-substitutable services to a market of heterogenous users. The first service is delivered subject to a "guaranteed" (G) processing rate, and the second is a "best-effort" (BE) type service in which residual capacity not allocated to the guaranteed class is *shared* among BE-users. Users, in turn, are sensitive to both price and congestion-related effects. The service provider's objective is to optimally design the system so as to extract maximum revenues. The design variables in this problem consist of a pair of static prices for the two services, a policy that controls admission of G-users into the system, and the mechanism by which users are informed of the state of congestion in the system. Since these objectives are difficult to address using exact analysis, we pursue approximations that are tractable and lead to structural insights. Specifically, we first solve a deterministic relaxation of the original objective to obtain a "fluid-optimal" solution which is subsequently evaluated and refined to account for stochastic fluctuations. Using diffusion limits, we derive approximations that yield the following structural results: **(i) pricing** rules derived from the deterministic analysis are "almost" optimal; **(ii)** the **optimal operational regime** for the system is close to heavy-traffic, and; **(iii) real-time congestion notification** results in increased revenues. Numerical results illustrate the accuracy of the proposed approximations and validate the aforementioned structural insights.

Short Title: Design of differentiated services

Keywords: congestion notification, diffusion approximations, economics, Halfin-Whitt regime, many server limits, pricing, queueing, revenue management, service differentiation

[*] Both authors are with Graduate School of Business, Columbia University, 3022 Broadway, New York, NY 10027. Email: `c.maglaras@columbia.edu` and `assaf@gsb.columbia.edu`

1 Introduction

Recent years have witnessed an explosive growth in services offered over the Internet via the world-wide-web. These web-based services include electronic commerce, internet telephony, streaming audio and video, e-mail and information retrieval, to name but a few examples. In an effort to address the processing requirements of these diverse applications and better segment the market of potential users, service providers are attempting to offer multiple grades of service so that users are differentiated according to their quality-of-service (QoS) requirements and willingness to pay. Inspired by these recent developments, in particular the emergence of *information services*, this paper introduces a simple stylized model of *differentiated services*, and addresses questions of optimal system design.

The goal of our model is to capture some of the stylized features that characterize information services. The first feature is that congestion in such services typically manifests itself as a degradation of the processing rate which in turn leads to delays. This should be contrasted with more traditional service operations where delays are driven by queueing effects. The second is the QoS levels that are common in the delivery of these type of services. In particular, in many instances a service provider may offer "real-time" applications that necessitate a guaranteed performance (e.g., software on demand), and "low QoS" applications that may be delivered subject to rate degradations (e.g., on-line help desk or database searches).

Motivated by such QoS provisioning, this paper considers a system that delivers two *non-substitutable* services or application classes. The first service is delivered subject to a "guaranteed" (G) processing rate, and the second is a "best-effort" (BE) type service in which residual capacity not allocated to the G-class is *shared* among BE-users. An important feature of this model is that both services are delivered using common processing resources, i.e., capacity is not split in such a way that a fraction is dedicated to each service class. Demand for service in each application class is determined by the total cost faced by its users, this being comprised of a class-specific usage fee and a congestion-related cost.

In terms of probabilistic primitives, we assume that nominal connection requests arrive according to independent Poisson processes, and the processing requirements of the two services are exponentially distributed with potentially different rates. With these assumptions in place, the system dynamics are Markovian. We note that our formulation and analysis focuses on the overall demand induced by a given price and congestion level. That is, we do not attempt to model the flow or packet level dynamics that characterize the means of delivering information services.

The service provider's objective is to extract maximum revenues by: (i) optimally pricing the two service classes; (ii) choosing an admission control policy for guaranteed-rate requests; and (iii.)

selecting the mechanism by which users "learn" about the state of congestion in the system. In achieving these goals, the service provider is assumed to possess full information on the customer types. Note that the congestion notification mechanism alluded to above introduces feedback: delay-averse users may be more reluctant to connect to the system when congestion is high, and this in turn reduces congestion thus inducing more users to connect. This process then continues until an *equilibrium* is reached, a notion that is central to our analysis. In addition to the design objectives stated above, our analysis strives to illuminate various other aspects that characterize the performance of the system, e.g., magnitude of congestion-related effects, the nominal operating point for the system and other equilibrium properties.

The objectives mentioned above are difficult to address directly, even under simplifying Markovian assumptions. In particular, the stochastic modulation of capacity available to BE-users and the feedback mechanism that is introduced by congestion notification, render the above design problems intractable as far as exact analysis is concerned. Instead, we propose an approximate analysis that gives rise to important structural insights and supports simple computations. In hindsight, this approach is seen to be quite accurate in large capacity systems. The first step of this hierarchical analysis consists of formulating a deterministic relaxation of the original optimization problem. The solution to this problem yields "fluid-optimal" per-access prices for the two services, and suggests an admission policy for G-users. The latter amounts to giving "high priority" to the service class that generates more revenue per unit of capacity per unit time. Since it is natural to think of this value being higher for services that require strict performance guarantees, we will hereafter assume that indeed the guaranteed service is given "high priority" in the sense that its users are always admitted when capacity is available. (As will be argued in what follows, the main structural insights that arise in this setting are essentially preserved when this priority is switched to the BE-class.) The second step of this analysis examines the performance of the system under the "fluid-optimal" solution, assessing the effects of stochastic fluctuations. Subsequently, in the final step, the fluid-optimal solution is refined to account for stochastic fluctuations so as to further optimize system performance and extract additional revenues. In terms of methodology, the approximate analysis described above hinges to a large extent on diffusion limits. This machinery enables us to pursue several objectives that would otherwise not be tractable via exact analysis.

The main contributions of this paper are the following.

i.) **Pricing and admission rules derived via deterministic analysis.** The fluid-optimal prices and the associated admission control policy for G-users turn out to be "almost optimal." Namely, the revenues extracted by these choices when implemented in the stochastic system are very close to those generated by the optimal rule. (See Theorem 2.)

ii.) **System operational regime and structural insights.** Under nominal assumptions on the revenue functions (essentially, concave increasing), fluid-optimal prices derived via the deterministic analysis (see Proposition 1) induce an equilibrium operating point where utilization is high, congestion effects are "small," and stochastic fluctuations are of order square root of the system "size." (See (i)-(iv) in Theorem 1.) When the mean service requirement is identical in the two classes, the system equilibrium can be approximated with high accuracy via a solution to a simple fixed point equation. (See (10) in Theorem 1).

iii.) **Performance analysis when classes are differentiated with respect to their service requirements.** A simple approximation that relies on underlying diffusion limits is proposed, so as to derive closed-form approximations as in Theorem 1. (See Section 6.)

iv.) **Second order price correction.** The structural insights that follow from the equilibrium analysis give rise to a simple "second-order" price correction that refines the fluid-optimal price so as to extracts higher revenues. (See Section 7.)

v.) **The value of real-time congestion notification.** A system that informs users of "real-time" congestion generates more revenue than one that provides static congestion information; the magnitude of this contribution is seen to be "second-order." (See Theorem 3.) These results are established with the aid of diffusion limits. (See Proposition 4).

Numerical results validate the structural insights and illustrate the accuracy of the approximations discussed above.

The remainder of the paper is structured as follows. This section concludes with a review of the literature, while Section 2 describes the system model and design objectives. Section 3 pursues a deterministic analysis and Section 4 derives the system behavior under the deterministic solution. Section 5 discusses some of the qualitative insights extracted from the analysis in Sections 3 and 4. Sections 6– 8 focus on extensions and refinements of the previous analysis focusing on non-identical service rates, second order optimization, and the economic value of real-time congestion notification, respectively. Finally, there are two appendices: Appendix A contains background material on diffusion limits; and Appendix B contains the proofs.

Literature review. The stylized model that we formulate is similar to the one first introduced by Das and Srikant (2000) to model Best-effort type traffic in the data network context. They derived diffusion approximations for this single class system in the so-called Halfin-Whitt heavy traffic regime. In a previous paper, Maglaras and Zeevi (2003*a*) studied a variant of the Das-Srikant model pursuing problems of economic optimization and optimal system design for a single-class system serving only BE-users. The work in Maglaras and Zeevi (2003*a*) covered both profit maximization as well as social welfare objectives adopting an equilibrium formulation that is driven by the treatment in Mendelson and Whang (1990) [see also Basar and Srikant (2002) for a related study of different flavor.] Here, we seek to extend this analysis by considering a canonical two-class system model and in addition consider further design issues, such as admission control and congestion notification mechanisms, and their economic value. In contrast to the single-class case discussed in Maglaras and Zeevi (2003*a*), the analysis of the two class system covered in the current paper hinges on diffusion approximations of the type derived recently in Maglaras and Zeevi (2004). The main result from Maglaras and Zeevi (2004) serves as an auxiliary result in several proofs in the current paper, and is cited in Appendix A for completeness.

The primary motivation to focus on Guaranteed and Best-effort service classes is driven by the communication and information services area (see, e.g., Altman, Orda and Shimkin (2000), Altman and Kushner (1999) Carpenter and Nichols (2002), Gibbens and Kelly (1999) and the references therein). The prism through which we view the system and its performance focuses on the users-level or overall demand-level rather than data flows or packets (similar to the study of Paschalidis and Tsitsiklis (2000)). The notion of service differentiation is of course quite ubiquitous in operations management and service operations. Two specific application areas that are akin to the one studied here include call-centers that process "VIP" and "regular" customers (see, e.g, the recent survey by Gans, Koole and Mandelbaum (2003)), and rental systems that serve customers with reservations as well as "walk-ins" (see, e.g., Savin, Cohen, Gans and Katalan (2002)). In the former, users experience congestion by *waiting* in a queue until agents become available, while in the latter congestion appears in the form of *blocking* when there is no remaining capacity. For a recent discussion of service grades, customer types and scheduling rules in a production system modelled as a multi-class single-server queue, see Van Mieghem (2000).

Our view of the service provider as having complete knowledge of the user (or demand function) characteristics is dubbed "full information" in Van Mieghem (2000). In this setting, as argued in Van Mieghem (2000), the assumption that services are non-substitutable is essentially not restrictive, as the system manager can always select not to serve customers that select the "wrong" class. A similar model to the one we pursue in the current paper, dealing with "incomplete information," substitutable services, and users that have a choice of service level is discussed in Maglaras and Zeevi (2003*b*). Finally, McGill and van Ryzin (1999) provide a recent overview of revenue management that is tangentially related to our work.

A stream of recent research has emphasized the pivotal role played by diffusion limits as a means to analyze large scale service systems. In this context, a particularly useful framework is the many-server heavy-traffic limits pioneered by Halfin and Whitt (1981). The interest in the Halfin-Whitt regime largely stems from its ability to succinctly summarize and elucidate natural statistical economies scale that are present in many large capacity service system. In particular, Whitt (1992) and Garnett, Mandelbaum and Reiman (2002) argue that this regime is a desirable operating point for certain large scale service operations. In Maglaras and Zeevi (2003*a*) the Halfin-Whitt regime is optimal from an economic optimization standpoint in a system that only offers BE-type service. In the current paper the Halfin-Whitt heavy-traffic regime is also seen to be the outcome of economic optimization, viz, fluid-optimal prices induce this type of behavior. For some recent applications and extensions of the Halfin-Whitt results see Whitt (1992), Fleming, Stolyar and Simon (1994), Das and Srikant (2000), Garnett et al. (2002), Puhalskii and Reiman (2000). In the context of call centers, Armony and Maglaras (2004*b*) and Whitt (2004) study the equilibrium behavior of large capacity systems based on the Halfin-Whitt asymptotics, while Whitt (1999) and Armony and Maglaras (2004*a*) study the effect of real-time congestion notification; the latter uses Halfin-Whitt type diffusion limits. Motivated by the skills-based routing problem in call centers, both Atar, Mandelbaum and Reiman (2002) and Harrison and Zeevi (2004) study dynamic scheduling problems in multi-class many-server systems. (For recent surveys of these and other issues related to call center design see Gans et al. (2003) and Whitt (2002).)

2 Model Formulation and Design Objectives

Our stylized system model attempts to capture four important features of the physical system: common and finite processing capacity, lack of resource pooling when the system is under-utilized, differentiated services, and the capability to share processing resources in the Best-effort class.

The system model. The service system is endowed with a finite processing capacity C used to support two non-substitutable services which will also be referred to as classes: "guaranteed-rate" (G) service will be denoted as class 1, and a "best-effort" type service will be denoted as class 2. Hereafter, various quantities will be tagged with subscripts 1 and 2 to denote the two respective classes. Users requesting class i service arrive to the system according to a Poisson process with rate λ_i, and have independent identically distributed (i.i.d.) service requirements that are exponentially distributed with rate μ_i. Note that the two services are linked through the common capacity constraint. The precise details and dynamics of the two service classes are as follows:

(i) *Guaranteed-rate (G) service:* Let $Q_1(t)$ denote the number of G-users in the system at time t and assume, for simplicity, that C is integer valued. Users of this service that are admitted into the system always receive one unit of processing capacity. Since the system has finite capacity, it will not always be possible to deliver this guarantee and thus the service provider will need to exercise some form of admission control. This will be denoted by the non-decreasing process $U = (U(t) : t \geq 0)$, where $U(t)$ counts the cumulative number of such connection requests that have been blocked (i.e., rejected) up to time $t > 0$, with $U(0) = 0$. We will assume that the admission control U is Markovian, that is, the decision on whether to admit a G-user arriving at time t depends only on the number of users of each type currently connected to the system. Note that the guaranteed rate of service offered to these users implies that $Q_1(t) \leq C$ for all times t.

(ii) *Best-effort (BE) service:* BE-users are always admitted into the system and the service provider does not exercise any form of admission control in this service class. When there is sufficient capacity in the system, BE-users receive a nominal allocation of one unit of processing capacity, and otherwise they share available capacity in an egalitarian manner resulting in a degraded processing rate. Specifically, the rate allocated to BE-users at time t is

$$\text{BE service rate} = \begin{cases} 1 & Q_1(t) + Q_2(t) \leq C \\ \frac{C - Q_1(t)}{Q_2(t)} & Q_1(t) + Q_2(t) > C, \end{cases}$$

where $Q_2(t)$ denotes the number of best-effort users in the system at time t. When $Q_1(t) = C$, BE-users temporarily do not receive service but remain connected to the system.

Despite the processor sharing characteristic, it can be verified that the dynamics of the process $(Q_2(t) : t \geq 0)$ are identical to that of an $M/M/C(t)/\infty$ system, where the capacity $C(t) = C - Q_1(t) \geq 0$ is a stochastic process modulated by the number of G-users in the system. The dynamics of $(Q_1(t) : t \geq 0)$ depend on the admission control U that is yet to be specified.

Economic structure and demand model. We assume that the service provider charges a fixed connection fee $p_i > 0$ for each class i user accessing the system. The BE-users perceive the disutility associated with rate degradation through the *excess delay* it induces relative to the nominal sojourn time based on a unit rate allocation.[1] A proxy for this excess delay is inversely proportional to the rate degradation, i.e.,

$$D(t) = \left(\frac{Q_2(t)}{C - Q_1(t)} - 1 \right)^+ = \frac{(Q_1(t) + Q_2(t) - C)^+}{C - Q_1(t)}. \tag{1}$$

We note that in large capacity systems $D(t)/\mu_2$ is an asymptotically accurate estimate of the *actual* excess delay due to a pathwise version of Little's law. To facilitate mathematical analysis, we will take the excess delay to be $D(t) := [(Q_1(t)+Q_2(t)-C)^+]/[(C-Q_1(t))\vee 1]$, where $x \vee y := \max\{x, y\}$. This ensures that the excess delay is finite almost surely. (As will be evident in what follows, this assumption does not restrict the generality of the analysis in any meaningful manner.)

We assume an additive linear delay cost for BE-users, which is $q > 0$ per unit of time of excess delay; the subscript 2 is dropped from D and q since these quantities are only relevant for the BE service class. Thus, the cost of joining the system for G-users is given by the price, p_1, while for BE-users this cost is given by $p_2 + \frac{q}{\mu_2}\mathbb{E}D$, where $\mathbb{E}D$ is the expected steady-state delay (the precise notion of this steady-state is explained below). As a matter of convention, we will denote steady-state quantities with either an '∞' as their time argument or simply by omitting the time

[1] Information/communication service providers often quote a service rate to users, and this is often accompanied by a table that translates rates into waiting times for various job sizes (that the system does not know a priori).

argument altogether when no confusion arises, e.g., $D := D(\infty)$. The arrival rate in each class is then

$$\lambda_1(p_1) \text{ for G-users} \quad \text{ and } \quad \lambda_2(p_2 + (q/\mu_2)\mathbb{E}D) \text{ for BE-users,}$$

where $\lambda_i(\cdot)$ are the respective *demand functions* for each class of service. Note that the rate of G-user connection requests, $\lambda_1(p_1)$, does not depend on the admission control U, but that a fraction of $\lambda_1(p_1)$ will be denied admission in accordance to the control $U(t)$. Note that a fraction of the G-user connection request rate, $\lambda_1(p_1)$, will be denied admission in accordance to the control $U(t)$, but that $\lambda_i(p_1)$ itself does not depend on U. The long-run blocking probability for such users is

$$b(U) := \mathbb{P}(\text{blocking}) = \lim_{t\to\infty} \frac{U(t)}{\lambda_1(p_1)t},$$

which is for now assumed to exist (this is later proved to be the case under a specific admission policy). The demand functions are assumed to be convex, decreasing, continuously differentiable, and such that $\lambda_i(x) \to 0$ as $x \to \infty$, for $i = 1, 2$. The *inverse demand function* will be denoted by $p_i(\lambda)$; i.e., $p_i(\cdot) = \lambda_i^{-1}(\cdot)$. With slight abuse of notation, we will denote the vector of realized arrival rates using the same notation as the demand functions only omitting the argument, i.e., $\lambda = (\lambda_1, \lambda_2)$. Finally, put $\Lambda_i := \max_x \lambda_i(x)$, the maximum demand or *market potential* for each type of service, respectively, which is assumed to be finite. This allows us to normalize the demand functions, writing, for example, the arrival rate into each class as

$$\lambda_i(\cdot) := \Lambda_i\tilde{\lambda}_i(\cdot),$$

where $\tilde{\lambda}_i(\cdot)$ is the normalized demand function taking values in the unit interval.

Equilibrium formulation. As hinted above, we will focus our attention on the *equilibrium steady-state* behavior of the system. To be precise, we say that for some price vector $p = (p_1, p_2)$ and control U the system admits a unique *equilibrium* if there exists a unique steady-state probability distribution for the process $((Q_1(t), Q_2(t)) : t \geq 0)$, such that the expected delay in class 2 when taken w.r.t. to this distribution, $\mathbb{E}D$, induces a time homogenous vector of external arrival rates

$$\lambda_1(p_1) = \Lambda_1\tilde{\lambda}_1(p_1) \quad \text{and} \quad \lambda_2(p_2) = \Lambda_2\tilde{\lambda}_2\left(p_2 + \frac{q}{\mu_2}\mathbb{E}D\right). \tag{2}$$

and these arrival rates together with the steady-state blocking probability defined above are, in turn, consistent with the aforementioned steady-state distribution. For now we will assume that an equilibrium exists and proceed to pose an optimization problem in terms of the pricing and admission control decisions. We will return to this issue in Sections 3 and 4 where we propose a specific admission control policy for which we show that there exists a unique equilibrium. Section 8 will contrast this model with one where state-dependent information is announced to the users.

Design objectives. The economic optimization problem faced by the service provider is to maximize the equilibrium revenue rate generated by the system. This optimum is given by

$$R_* := \sup_{p_1,p_2 \geq 0,\ U} \left\{ p_1\lambda_1(p_1)(1-b(U)) + p_2\lambda_2\left(p_2 + \frac{q}{\mu_2}\mathbb{E}D\right)\right\}, \tag{3}$$

where the optimization is carried out under the equilibrium distribution. Implicit in this expression is the dependence of the congestion effects for G and BE users, namely $b(U)$ and $\mathbb{E}D$, on the admission control U. This formulation assumes that the service provider has full information on the user characteristics and induced demand functions summarized in the five-tuple $(q, \mu_1, \mu_2, \Lambda_1, \Lambda_2)$ and the two normalized demand functions $\tilde{\lambda}_i(\cdot)$, $i = 1, 2$. The design variables in the above optimization problem are the prices levied on each provisioned service and the admission control policy U. Subsequently, in Section 8 we consider the additional decision of selecting the mechanism by which congestion is "fed-back" to the users (i.e., static vs. dynamic information). In what follows, it will be useful to consider a version of the maximization problem stated above where the revenue rates are considered as functions of λ rather than price, p. In particular, put

$$r_i(\lambda_i) := \lambda_i p_i(\lambda) \quad i = 1, 2, \tag{4}$$

where these functions are assumed to be continuously differentiable, strictly concave and increasing in the λ_i's. This formulation, as well as the assumptions accompanying it, are quite standard in the revenue management literature; see, e.g, Gallego and van Ryzin (1994).

Discussion of the modeling assumptions. The model we propose assumes that when the system is under-utilized, spare capacity cannot be redistributed to the users currently in the system; i.e., there is no resource pooling. The reason for this assumption is that most service systems are limited by a maximum processing rate. In the context of communication and information services this is typically due to restricted uploaded and downloaded rates and limited efficiency in executing tasks in parallel. In terms of probabilistic primitives the assumption regarding the Poisson arrival streams is not restrictive, however the exponential distribution of the service times is required for tractability. Note that the expression in (2) implicitly assumes that demand for BE-type service is affected by users assessing their congestion cost based on their average service time, as opposed to the use of their actual (random) service requirement. This follows the modeling framework introduced by Mendelson and Whang (1990), and is reasonable in applications where the user does not know a priori the precise amount of service that he/she will request. In terms of congestion cost, we note that the hierarchical solution approach we propose applies also in the case where delay costs are convex increasing as in Van Mieghem (2000). Finally, our system model assumes that the two service classes are non-substitutable and users cannot select between them upon accessing the system. As pointed out in Van Mieghem (2000), when the service provider has full information on the user characteristics (demand functions) one could allow users a choice of QoS level, then simply penalize users that select the "incorrect" class.

3 Deterministic Analysis

The first step in our analysis is to formulate and solve a deterministic relaxation of the design problem given in (3). To this end, we introduce two new design variables, b and d; the former plays the role of the blocking probability for G-users, and the latter plays the role of the steady-state excess delay suffered by the BE-users. This informal description is meant to indicate the logic behind the deterministic relaxation, and how it is derived from the original optimization problem (3). This deterministic problem is given by

$$\begin{aligned} \max \quad & p_1\lambda_1(p_1)(1-b) + p_2\lambda_2\left(p_2 + (q/\mu_2)d\right) \qquad (5) \\ \text{s.t.} \quad & \frac{\lambda_1(p_1)(1-b)}{\mu_1} + \frac{\lambda_2(p_2+(q/\mu_2)d)}{\mu_2} \leq C \\ & p_1, p_2, d \geq 0, \quad b \in [0,1]. \end{aligned}$$

Note that the objective function is the "same" as the one in (3), and the constraint linking the variables p_1, p_2, b, d is the stability condition that was implicitly satisfied in the stochastic system of the previous section due to blocking of G-users and the regulation of the BE demand via the equilibrium congestion term $\mathbb{E}D$. Treating b, d as optimization variables is of course a relaxation of the original problem, and therefore the value of the optimization problem (5) provides an upper bound on the optimal revenue rate for the stochastic system R_*.

The first observation about the solution to this optimization problem is that in terms of revenue rate maximization it is optimal to never block G-users and never delay BE-users, i.e., $\bar{b} = 0$ and $\bar{d} = 0$, where the overbar notation denotes the solution of the deterministic planning problem. This is proved by contradiction. Suppose that (p_1, p_2, b, d) is optimal with $b > 0$. It is easy to see that the service provider can raise p_1 to p_1' such that $\lambda_1(p_1') = \lambda_1(p_1)(1-b)$ and extract higher revenues while consuming the same capacity per unit time, which contradicts the optimality of (p_1, p_2, b, d). Similarly, one can show that it is never optimal to have BE-users suffer a positive excess delay.

To characterize the optimal prices and associated demand rates for (5) we will assume that the system capacity is scarce in the following sense. Let $p_i^* = \text{argmax}_{p \geq 0}\, p\lambda_i(p)$ denote the *unconstrained* revenue maximizing price for service class i. We will require that

$$\frac{\lambda_1(p_1^*)}{\mu_1} + \frac{\lambda_2(p_2^*)}{\mu_2} \geq C, \qquad (6)$$

i.e., the unconstrained revenue maximizing demand rates consume at least as much capacity as C. This can be motivated by considering a higher-level profit maximization problem that incorporates a convex increasing cost of capacity per unit time, denoted by $H(C)$, in which case it is easy to show that the maximum profit rate $p_1\lambda_1(p_1) + p_2\lambda_2(p_2) - H(C)$ over $\frac{\lambda_1(p_1)}{\mu_1} + \frac{\lambda_2(p_2)}{\mu_2} \leq C$, occurs when the capacity constraint is binding.

Given the discussion above, we can now focus our attention on (5) with $b = d = 0$. Under the assumptions imposed on the primitives, it is readily seen that the resulting two variable optimization problem involves maximizing a (strictly) concave function over a convex set. We also assume that the solution of this deterministic problem is such that it is profitable to offer both service classes so that the problem does not degenerate to one involving only a single class. Analysis of the Lagrangean associated with (5), with b, d set to zero, leads to a precise condition that ensures that the optimum in (5) is achieved at an interior point of the set of feasible rates. This is summarized in the following proposition.

Proposition 1 *Assume that (6) holds and let $\bar{p}_1, \bar{p}_2, \bar{b}, \bar{d}$ denote the maximizer of the deterministic optimization problem (5). Then, $\bar{b} = 0$ and $\bar{d} = 0$, and $\bar{p}_1, \bar{p}_2$ lead to full resource utilization, i.e., $\lambda(\bar{p}_1)/\mu_1 + \lambda_2(\bar{p}_2)/\mu_2 = C$. If in addition, there exist $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1/\mu_1 + \lambda_2/\mu_2 \leq C$ and $r_1'(\lambda_1)\mu_1 = r_2'(\lambda_2)\mu_2$, then then it is optimal to offer both services, i.e., $\lambda(\bar{p}_1), \lambda_2(\bar{p}_2) > 0$.*

The *fluid-optimal demand rates* associated with the solution of (5) are computed through the demand functions, $\bar{\lambda}_i = \lambda_i(\bar{p}_i)$.

Proposed policy. The next step is to articulate a pricing and admission control policy for the original system based on the solution of the deterministic relaxation given above. The pricing policy will be to set the per-access fees for service classes 1 and 2 equal to $\bar{p}_1$ and $\bar{p}_2$, respectively. In terms of admission control decisions, the solution to (5) is not very helpful as it prescribes no blocking and no delay for G and BE users, respectively. Under the pricing structure proposed above, however, it is natural to consider an admission policy that gives "priority" to G-users provided that $\mu_1\bar{p}_1 \geq \mu_2\bar{p}_2$, and to BE-users otherwise. That is, the system gives priority to the class that generates higher revenue per unit of capacity per unit time. Direct analysis of the first order optimality conditions of (5) yield that $\mu_1\bar{p}_1 \geq \mu_2\bar{p}_2$ is equivalent to

$$\bar{\lambda}_1\mu_1 \frac{\partial p_1(\lambda)}{\partial \lambda}\bigg|_{\bar{\lambda}_1} \leq \bar{\lambda}_2\mu_2 \frac{\partial p_2(\lambda)}{\partial \lambda}\bigg|_{\bar{\lambda}_2},$$

where $p_i(\cdot) = \lambda_i^{-1}(\cdot)$. In general, this only provides an implicit condition on the underlying primitives that gives rise to a more "expensive" G-class. This can be further simplified if one assumes a particular form for the demand functions. For example, in the context of linear demand models, where for any price p the demand $\lambda_i(p) = \Lambda_i - \alpha_i p$, this reduces to the condition $\Lambda_1\mu_1/\alpha_1 \geq \Lambda_2\mu_2/\alpha_2$.

In the remainder of the paper we will assume that $\mu_1\bar{p}_1 \geq \mu_2\bar{p}_2$, and thus that G-users receive higher priority. Denoting this policy by $\bar{U}$, we have that $\bar{U}(t)$ only increases (and a G-user is denied admission) at times where upon arrival of a G-user request $Q_1(t) = C$, i.e., the entire capacity is already utilized by high-priority users. Note that in this case $(Q_1(t), t \geq 0)$ has the same dynamics as the number-in-system process in an $M/M/C/C$ system.

For future purposes, it will be convenient to denote the *fluid-optimal revenues*, i.e., the value of the deterministic optimization problem (5), as a function of the fluid-optimal prices,

$$\bar{R}(\bar{p}_1, \bar{p}_2) = p_1\lambda_1(\bar{p}_1) + p_2\lambda_2(\bar{p}_2), \tag{7}$$

and note that by construction this serves as an upper bound for the revenue rate in the underlying stochastic system, i.e., $R_* \leq \bar{R}(\bar{p}_1, \bar{p}_2)$. Finally, the *relative workload contributions* of each service class are defined by

$$\kappa_i = \frac{\bar{\lambda}_i/\mu_i}{\bar{\lambda}_1/\mu_1 + \bar{\lambda}_2/\mu_2} = \frac{\bar{\lambda}_i}{C\mu_i}, \tag{8}$$

and $\kappa_1 + \kappa_2 = 1$. Thus, the κ's represent the fluid-scale fractions of load that emanate from each service class.

4 System Behavior Under the Deterministic Solution

We now study the performance of the system under the policy extracted from the fluid relaxation, taking into account the effect of stochastic variability and congestion.

4.1 Equilibrium analysis under the proposed policy

The first step is to establish that under the policy proposed above there exists a unique equilibrium operating regime. This is addressed in the next two propositions. First, consider a system where the BE-user class is not sensitive to delay, i.e., $q = 0$. (Alternatively, this is a system with no feedback signal.) The next proposition characterizes the stability region for this system, i.e., the set of input rates $\lambda = (\lambda_1, \lambda_2)$ such that the system admits a unique steady-state.

Proposition 2 *For each capacity $C > 0$, and arrival vector $\lambda > 0$ such that $\frac{\lambda_1 \mathbb{P}(Q_1 < C)}{\mu_1} + \frac{\lambda_2}{\mu_2} < C$, the continuous time Markov chain $(Q_1(t), Q_2(t) : t \geq 0)$ admits a unique stationary (steady-state) distribution. Here $\mathbb{P}(Q_1 = C) = 1 - \mathbb{P}(Q_1 < C)$ is the steady-state probability of blocking in an $M/M/C/C$ queue with arrival rate λ_1 and service rate μ_1.*

Note that $(Q_1(t) : t \geq 0)$ admits a unique stationary distribution for any arrival rate λ_1 since the number-in-system is bounded by C. If we assume that under the unique steady-state distribution, the expected delay, $\mathbb{E}D$, is continuous in the BE-class arrival rate, λ_2, then, for the system with feedback ($q > 0$), we have the following result.

Proposition 3 *For each capacity $C > 0$, and price vector $p \geq 0$, there exists a unique steady-state equilibrium.*

The nature of the two-class system and the associated equilibrium formulation make it difficult to pursue a direct analysis of the Markov chain describing the system dynamics. This difficulty is exacerbated when the service rates for the two classes differ, i.e., when $\mu_1 \neq \mu_2$. (This problem has been pointed out in many other studies; see, e.g, Davis (1966) and Williams (1980).) With this in mind, we will first derive approximations to the system equilibrium behavior in the simpler case, where the service rates are identical. This provides a clean illustration of the main structural insights and key results. Subsequently, Section 6 will illustrate how these results extend to the case of non-identical service rates.

4.2 Preliminaries for asymptotic analysis

Our approach will rely on approximations which are accurate in large scale operations, i.e., when the market potential and the system capacity are both large. To derive these approximations, we will let the "scale" of the system grow as follows: for $n = 1, 2, \dots$ and $i = 1, 2$ we set

$$\begin{aligned} C^n &:= n && \text{[capacity grows large]} \\ \Lambda_i^n &:= n\bar{\Lambda}_i && \text{[capacity grows proportionally to the market potential]} \\ \lambda_i^n(\cdot) &:= \Lambda_i^n \tilde{\lambda}_i(\cdot) && \text{[structure of the demand curve is preserved].} \end{aligned} \tag{9}$$

The second and third assumptions imply that the structure of the demand curve is preserved, while its magnitude is scaled up linearly. The proportionality factor, $\bar{\Lambda} = (\bar{\Lambda}_1, \bar{\Lambda}_2)$ are derived from the original system parameters by setting $\bar{\Lambda}_i := \Lambda_i / C$, where Λ_i is the potential demand for ith service class, and the system capacity is C. Note that in the case of linear costs of capacity, $h(C) = h \cdot C$ for some $h > 0$, the assumption that capacity and market potentials grow proportionally would be a *consequence* of the profit maximization objective that incorporates capacity costs; for an illustration of this argument in the single-class context see Maglaras and Zeevi (2003*a*). Finally, we note that under the scalings given in (9) the fluid optimal prices $\bar{p}_i$ and the workload contributions κ_i defined in the previous section are independent of n, while the fluid-optimal demand rates $\bar{\lambda}_i^n$ grow proportionally to n. We use the superscript n to denote quantities that depend on the (growing) system capacity, e.g., ρ^n denotes the system utilization, and the absence of such a superscript will indicate quantities that are independent of n. For two real-valued sequences a^n, b^n we write $a^n = o(b^n)$ if $a^n / b^n \to 0$ as $n \to \infty$. Finally, for any differentiable function $f : \mathbb{R} \to \mathbb{R}$, f' will denote its derivative.

4.3 Main results

System equilibrium characterization. Our first result characterizes the system equilibrium. In particular, it asserts that the fluid optimal-prices induce "high" resource utilization (heavy-traffic) and yet the service quality achieved is high.

Theorem 1 (Equilibrium characterization) *Suppose that $\mu_1 = \mu_2 = \mu$, the conditions of Proposition 1 hold, and assume that demand (λ^n) and capacity (C^n) grow large as in (9) as $n \to \infty$. Consider the sequence of steady-state utilizations ρ^n, delays D^n, and queue lengths Q_1^n, Q_2^n obtained in equilibrium for each n. Then,*

i.) system utilization: $\rho^n = 1 - \frac{\gamma}{\sqrt{n}} + o(1/\sqrt{n})$

ii.) BE-class delay: $\mathbb{E}[D^n] = \frac{d(\gamma)}{\sqrt{n}} + o(1/\sqrt{n})$

iii.) G-class blocking: $\mathbb{P}(Q_1^n = n) = o\left(e^{-cn}\right)$, *for some* $c > 0$

iv.) system congestion: $\mathbb{P}(Q_1^n + Q_2^n \geq n) = \nu(\gamma) + o(1)$

as $n \to \infty$. Furthermore, the parameter γ that characterizes the asymptotic approximations in i) to iv) above can be explicitly computed as the unique solution of the following equation

$$\gamma = -\kappa_2 \frac{q}{\mu} \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)} d(\gamma), \tag{10}$$

where $d(\gamma)$ is given by

$$d(\gamma) = \frac{\phi(\gamma)}{\kappa_2 \gamma (\gamma \Phi(\gamma) + \phi(\gamma))} \ . \tag{11}$$

Here κ_2 is the workload contribution associated with the BE-users for the fluid price vector $\bar{p}$, $\nu(\gamma) := \kappa_2 \gamma d(\gamma)$, and $\Phi(\cdot), \phi(\cdot)$ denote the standard Normal cumulative distribution function and its density, respectively.

Intuition and "proof sketch." The essence of the equilibrium analysis hinges on the delay process D. To this end, the delay experienced by the BE-class (see (1)) satisfies

$$D^n(t) \approx \frac{(Q_1^n(t) + Q_2^n(t) - n)^+}{n - Q_1^n(t)} \ ,$$

where '$\approx$' is used to denote equality up to lower order terms in n. Now, when the mean service requirements are identical in the two classes and neglecting the blocking of G-users, the number-in-system process, $Q_1^n(t) + Q_2^n(t)$, has identical dynamics to the number-in-system process in an $M/M/n$ queue. In particular, $(Q_1^n(t) + Q_2^n(t) - n)^+$ is then simply the queue length at time $t \geq 0$, and the steady-state mean is

$$\mathbb{E}\left(Q_1^n + Q_2^n - n\right)^+ = \frac{\rho^n \mathbb{P}(Q_1^n + Q_2^n \geq n)}{(1 - \rho^n)} \ ,$$

where $Q_i^n := Q_i^n(\infty)$. When $\rho^n = 1 - \gamma/\sqrt{n}$ for some $\gamma > 0$, we have that

$$\mathbb{E}\,(Q_1^n + Q_2^n - n)^+ \approx \sqrt{n}\mathbb{P}(Q_1^n + Q_2^n \geq n)/\gamma \ ,$$

and $\lim_{n\to\infty} \mathbb{P}(Q_1^n + Q_2^n \geq n) = \nu(\gamma) = \phi(\gamma)/(\gamma\Phi(\gamma) + \phi(\gamma))$ (see, e.g., Halfin and Whitt (1981, Proposition 1)). On the other hand, the process Q_1^n is simply the number-in-system in an $M/M/n/n$ queue with $\lambda_1^n = \kappa_1 n\mu$, which follows from (8). (Recall that κ_1, κ_2 are the relative workload contributions from each class of service.) Consequently, $\mathbb{E}Q_1^n \approx \kappa_1 n$, and $n - Q_1^n \approx \kappa_2 n$. Combining these observations, we conclude that $\mathbb{E}D^n \approx d(\gamma)/\sqrt{n}$. Since the blocking probability for the G-class is small, it follows that the utilization is essentially dictated by the above congestion term, in particular, $\rho \approx 1 - \gamma/\sqrt{n}$. Finally, by taking a Taylor expansion of the demand function for the BE-class, $\tilde{\lambda}_2(\cdot)$, one obtains the equilibrium equation (10). While this sketch captures some of the intuition that underlies the actual proof, it is also somewhat misleading. In particular, due to blocking effects a rigorous analysis of the two-class system hinges on diffusion limits.

Implications: performance of the fluid-optimal prices. Theorem 1 suggests that the revenues generated under the fluid-optimal prices ought to be "close" to optimal, due to the "small" degradation attributed to stochastic fluctuations. To turn this into a rigorous statement, let us first introduce the following notation. Let

$$R^n(p_1, p_2, U) := p_1\lambda_1^n(p_1)(1 - b(U)) + p_2\lambda_2^n\left(p_2 + (q/\mu_2)\mathbb{E}D^n\right)$$

denote the revenue rate achieved under any feasible price pair (p_1, p_2) and admission control U in equilibrium, which is assumed to exist, and let

$$R_*^n := \sup\left\{R^n(p_1, p_2, U) : p_1, p_2 \geq 0,\ U\right\}$$

denote the *optimal revenue* for a system with capacity $C^n = n$. Under this notation, $R^n(\bar{p}_1, \bar{p}_2, \bar{U})$ denotes the equilibrium revenue rate under the policy extracted through the fluid relaxation of Section 3. Also, recall that $\bar{R}(\bar{p}_1, \bar{p}_2)$ is the value of the deterministic optimization problem, i.e., the fluid revenue rate generated by the fluid-optimal prices $\bar{p}_1, \bar{p}_2$, and let $\bar{R}^n = n\bar{R}(\bar{p}_1, \bar{p}_2)$ denote the optimal revenue extracted in (5) when $\lambda_i(\cdot)$ is replaced by $\lambda_i^n(\cdot)$ and $C^n = n$.

Theorem 2 (Asymptotic optimality of the deterministic solution) *Under the assumptions of Theorem 1, $\bar{p}_1, \bar{p}_2, \bar{U}$ are asymptotically optimal in the sense that*

$$\frac{R^n(\bar{p}_1, \bar{p}_2, \bar{U})}{R_*^n} \geq 1 - \frac{\alpha}{\sqrt{n}} + o(1/\sqrt{n}), \tag{12}$$

as $n \to \infty$, where $\alpha > 0$ is a function of $\bar{p}_1, \bar{p}_2$. Moreover, $R^n(\bar{p}_1, \bar{p}_2, \bar{U})/\bar{R}^n \to 1$ as $n \to \infty$.

The above result is reminiscent of the one derived by Gallego and van Ryzin (1994) who established "near optimality" of static, fluid-based, pricing rules. A similar result was also derived by

Paschalidis and Tsitsiklis (2000) for a multiclass loss model. While the problem formulation, set up and analysis are quite different, these results are driven by aggregation effects that lead to reduced variability, namely, as the problem scales up, variability only scales as a square-root of the size of the problem. The asymptotic optimality property described above is in the "first order" sense (or fluid scale sense) insofar as it does not yield the best possible constant α. To that end, the effects of the admission control policy are seen to be second order, i.e., are captured in the magnitude of α. In section 7 we will describe a refinement to the fluid-optimal prices that optimizes second order performance given the admission policy derived from the fluid relaxations.

The case where BE service gets priority. As mentioned earlier, one could also consider the situation where the BE-class generates higher revenues per unit of capacity per unit time, and thus receives higher priority from the service provider. In this case, G-users are blocked whenever BE-users suffer rate degradation, and the brunt of congestion is borne by the G-users. The main results we obtained in this section could be derived in this setting as well, and the main structural conclusions should continue to hold (essentially interchanging the two service classes). That is, having the high priority BE-class suffering exponentially small congestion, the G-class being blocked with probability proportional to $\gamma/\sqrt{n}$, and the overall system utilization being again $\rho^n \approx 1 - \gamma/\sqrt{n}$. We will not attempt to rigorously justify these statements, since this would necessitate going well beyond the space limitations of the current paper.

5 Qualitative Insights and Accuracy of the Approximations

The heavy-traffic regime. The operational regime where capacity (C) is large and "matches" demand in a manner that the system probability of congestion is moderate, was first investigated by Halfin and Whitt (1981) in their seminal paper on many-server heavy-traffic limits in the context of the $M/M/n$ queue. As observed in the sequel study by Whitt (1992), large capacity systems that operate in high utilization exhibit *statistical economies of scale*, manifested as stochastic fluctuations which are of order square root the "size" of the system. These economies of scale are the primary reason for the high quality of service that prevails in spite of high utilization. Theorem 1 establishes that the fluid-optimal prices, derived on the basis of a deterministic analysis, lead the system to operate in the so-called Halfin-Whitt regime.

Using Theorem 1 to approximate the performance of a given system. The asymptotic result in Theorem 1 suggests how one can approximate the performance of a system with fixed and finite capacity C. In particular, one first uses the problem primitives (namely, parameters of the demand curve $\tilde{\lambda}(\cdot)$, and the market potential Λ) to solve the deterministic problem in Section 3, arriving at the fluid-optimal prices $\bar{p}_1, \bar{p}_2$. Then, one proceeds to solve (10), deriving the equilibrium parameter γ and computing $d(\gamma)$ given by (11) in Theorem 1. These limiting parameters are then used to approximate key performance measures which are affected by stochastic fluctuations: (i) the utilization is $\rho \approx 1 - \gamma/\sqrt{C}$, and; (ii) the BE-delay is $\mathbb{E}D \approx d(\gamma)/\sqrt{C}$. Moreover, blocking effects in the high priority (G) class are indeed negligible relative to congestion-related effects in the BE-class. The numerical example that follows provides a concrete illustration of the use of Theorem 1 in approximating the behavior of a given system.

Accuracy of the deterministic analysis. The equilibrium operating point characterized in Theorem 1 verifies, in hindsight, the accuracy of the deterministic analysis. In particular, the latter assumes zero delays and blocking effects, deducing that the system operates in 100% utilization, while the former asserts that stochastic effects perturb the fluid operating point by order $1/\sqrt{C}$, as spelled out in statements (i) and (ii) in the previous paragraph. Moreover, Theorem 2 establishes that the revenues generated by the fluid-optimal prices are near optimal, when capacity (C) and market potential (Λ) are both large.

A numerical illustration. Throughout the paper we use the following sample problem as a running example, with the goal of illustrating numerically how the analytical results describe the structural behavior of the system. We assume a linear demand relationship for both services, of the form

$$\lambda_i(p) = \Lambda_i - \alpha_i p,$$

for appropriate parameters Λ_i, α_i for $i = 1, 2$. The α_i's are the price sensitivity parameters of the two demand models. Let $\bar{\lambda}_i$ denote the nominal demand rates for each service class computed through the deterministic revenue maximization problem (5), let $\bar{p}_i$ be the corresponding prices, and let κ_i be the associated relative workload contributions. Pursuing further the analysis presented in Section 3, the first order optimality conditions for (5) are $(\Lambda_i - 2\bar{\lambda}_i)/\alpha_i = \nu/\mu_i$, where ν is the Lagrange multiplier associated with the capacity constraint, which leads to the solution

$$\bar{\lambda}_i = \frac{1}{2}\left(\Lambda_i - \nu\frac{\alpha_i}{\mu_i}\right) \quad \text{where} \quad \nu = \left(\frac{\Lambda_1}{2\mu_1} + \frac{\Lambda_1}{2\mu_1} - C\right)^+ \left(\frac{\alpha_1}{2\mu_1^2} + \frac{\alpha_2}{2\mu_2^2}\right)^{-1}.$$

The fluid prices are then given by $\bar{p}_i = (\Lambda_i - \bar{\lambda}_i)/\alpha_i$.

Figure 1 shows the dependence of the equilibrium congestion term $\mathbb{E}D^n$ and the corresponding traffic intensity ρ^n as we vary the system capacity n. The demand model parameters are chosen so that for $n = 50, 100, \ldots, 450$, $\Lambda_1^n = 1.5 \cdot n$, $\alpha_1^n = n/10$, $\Lambda_2^n = 2 \cdot n$, $\alpha_2^n = n/5$, $\mu_1 = \mu_2 = 1$, $q = 1$. (Under (5), $\kappa_i = .5$ and $\bar{p}_1 = 10$ and $\bar{p}_2 = 7.50$, independent of n.) These results highlight the high accuracy of the proposed asymptotic approximations when compared to the "exact" results based on exhaustive simulation. For the latter, we simulate five sample paths of 2,000,000 events each at different values of $d = \mathbb{E}D^n$, until the sample estimate $\mathbb{E}\hat{D}^n$ is in agreement with the hypothesized parameter d and the system is in equilibrium; henceforth, quantities obtained via simulation will be tagged with a $\hat{}$. Moreover, the structural properties given in Theorem 1 appear to be in force even for systems of moderate capacity. Specifically, the two figures illustrate that the expected delay suffered by the BE-users in equilibrium decays like $d(\gamma)/\sqrt{n}$, and the equilibrium traffic intensity behaves like $1 - \gamma/\sqrt{n}$, where γ is the unique solution of the equilibrium equation (10). For the parameters of this example, $\gamma = .46$ and $d(\gamma) = 2.33$.

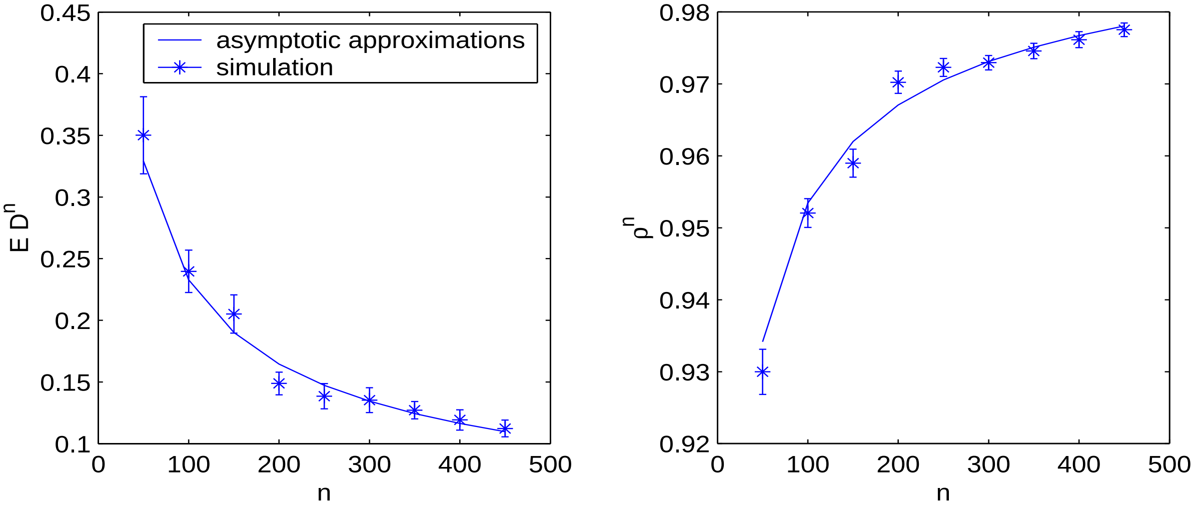

Figure 1: Equilibrium congestion $\mathbb{E}D^n$ and traffic intensity ρ^n under the fluid prices $(\bar{p}_1, \bar{p}_2)$ as a function of the system capacity (n). (Error bars represent pointwise 95% confidence intervals for quantities estimated via simulation.)

The closed-form asymptotic approximations given in Theorem 1 can be used in order to study the sensitivity of system performance to various model parameters such as (q, α_1, α_2). Table 1 reports results obtained via both the asymptotic approximations and exhaustive simulation for a set of representative examples. The parameters used to construct this table are: $n = 100$, $\mu_1 = \mu_2 = 1$, $\Lambda_1 = 150$, and $\Lambda_2 = 200$. As expected, an increase in the delay sensitivity parameter q results in a decrease in the equilibrium delay for the BE users and moreover the arrival rate into the BE service class and the overall traffic intensity also decrease. These effects are "second order" since they depend on the congestion cost which, in turn, behaves like $1/\sqrt{C}$. Similarly, as the price sensitivity parameter for G service increases (this changes the revenue function for this class), the relative workload contributions extracted from the deterministic optimization problem (5) change, and the overall revenue rate decreases; both changes affect the first order behavior of the system by changing the κ_i's that define how nominal capacity is split between the two classes. In all five examples, the revenue rate computed via simulation was very close to the one predicted via the asymptotic approximations, which is given by

$$R^n(\bar{p}_1, \bar{p}_2) \approx \bar{\lambda}_1 \bar{p}_1 + (\bar{\lambda}_2 - \alpha_2 \frac{q}{\mu_2} \mathbb{E}D^n)\bar{p}_2.$$

Moreover, both values were close to $\bar{R}$, obtained from the deterministic optimization problem (5).

(q, α_1, α_2)	κ_1	$(\mathbb{E}\hat{D},\ \hat{\rho})$	$\mathbb{E}D$ (gap)	ρ (gap)	$\hat{R}(\bar{p}_1, \bar{p}_2)$	$R(\bar{p}_1, \bar{p}_2)$ (gap)	$\bar{R}(\bar{p}_1, \bar{p}_2)$
(1,10,20)	.5	(.248, .950)	.234 (5.98%)	.954 (-.42%)	837.80	840.09 (-.27%)	875
(2,10,20)	.5	(.163, .935)	.149 (9.40%)	.940 (-.53%)	826.16	830.23 (-.49%)	875
(4,10,20)	.5	(.102, .919)	.093 (9.68%)	.926 (-.76%)	814.00	819.50 (-.67%)	875
(1,15,20)	.429	(.225, .955)	.219 (2.74%)	.956 (-.10%)	682.16	682.97 (-.12%)	714.29
(1,20,20)	.375	(.228, .954)	.212 (7.55%)	.958 (-.42%)	608.75	611.46 (-.44%)	640.63

Table 1: Sensitivity of the equilibrium behavior (expected delay, system utilization and revenue rate generated) w.r.t. (q, α_1, α_2), in a a system with capacity $C = 100$ operating under the fluid-optimal prices. The expected delay, system utilization and resulting revenue rate $\mathbb{E}D$, ρ and R are computed using the asymptotic expressions, while $\mathbb{E}\hat{D}, \hat{\rho}$, and $\hat{R}$ are simulation-based estimates. $\bar{R}$ denotes the upper bound on the optimal revenues derived from the deterministic analysis, and % relative error are defined as $(\mathbb{E}\hat{D} - \mathbb{E}D)/\mathbb{E}\hat{D} \times 100$ etc.

6 The Case of Non-identical Service Rates

This section describes an approach that allows us to extend our previous results to the case of non-identical service rates (i.e., $\mu_1 \neq \mu_2$). As noted in Section 4, the key element in characterizing the equilibrium is the delay process, D^n, which, in turn, is essentially characterized by the number-in-system process, $Q_1^n + Q_2^n$. The "intuition and proof sketch" that followed Theorem 1 suggests that when the service rates are identical, the number-in-system process has a simple Markovian structure for all n, and its steady-state distribution is simple to characterize. In contrast, when the service rates are not identical, the number-in-system process is not Markovian, and it is no longer simple to characterize its steady-state distribution. While this observation places an obstacle, at least as far as analysis is concerned, the scaling relations of Theorem 1 can be expected to hold true on the basis of diffusion approximations described in Appendix A.

For the purposes of performance analysis, i.e., in order to approximate the system equilibrium behavior, we exploit the following simple observation: given a system with different service rates, we can construct an approximating system that has the same service rates, adjusting arrival rates so as to capture the effect of the difference in the μ_i's. Since this system has equal service rates, it is amenable to the analysis of Section 4. We now provide a skeleton of the approach culminating in an approximation for $\mathbb{E}D^n$.

Consider a system with capacity C, service rates μ_1, μ_2 and arrival rates $\lambda_i = C\kappa_i\mu_i - \gamma_i\sqrt{C}\mu_i$, for some appropriate parameters γ_i such that $\gamma_1 + \gamma_2 > 0$; the latter is equivalent to $\rho < 1$. Note that for any arrival rates λ_1, λ_2 and given values κ_1, κ_2 (extracted from (8)), one can always rewrite the λ_i's in the form given above. For our system $\gamma_1 = 0$ and $\gamma_2 = -\kappa_2 \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)} \frac{q}{\mu_2} d(\gamma) + o(1/\sqrt{C})$; the latter follows from the fact that $\mathbb{E}D \approx d(\gamma)/\sqrt{C}$. Our goal is to approximate $\mathbb{E}D$, and thus characterize the behavior of the underlying system.

The perturbation approximation. Define $\bar{\mu} = (\mu_1 + \mu_2)/2$, and rewrite the service rates as "small" perturbations around that common value $\bar{\mu}$, viz,

$$\mu_i = \bar{\mu}\left(1 - \frac{\zeta_i}{\sqrt{C}}\right), \tag{13}$$

setting $\zeta_i := \sqrt{C}(1 - \mu_i/\bar{\mu})$. Next, rewrite the arrival rates $\lambda_i = C\kappa_i\mu_i - \gamma_i\sqrt{C}\mu_i$ in the form

$$\lambda_i = C\kappa_i\bar{\mu} - \sqrt{C}(\bar{\gamma}_i + \kappa_i\zeta_i)\bar{\mu}, \tag{14}$$

by setting $\bar{\gamma}_i = \gamma_i\mu_i/\bar{\mu}$. Keeping $\zeta_1, \zeta_2, \bar{\gamma}_1, \bar{\gamma}_2$ fixed, define a sequence of systems with

$$C^n = n, \quad \mu_i^n = \bar{\mu}\left(1 - \frac{\zeta_i}{\sqrt{n}}\right) \quad \text{and} \quad \lambda_i^n = n\kappa_i\bar{\mu} - \sqrt{n}(\bar{\gamma}_i + \kappa_i\zeta_i)\bar{\mu}. \tag{15}$$

Note that by setting $n = C$ we recover the parameters of the original system, $(C, \mu_1, \mu_2, \lambda_1, \lambda_2)$. That is, we have embedded the original system in a the sequence of systems defined through (15), the limit of which is tractable since both $\mu_1^n, \mu_2^n \to \bar{\mu}$ as $n \to \infty$, and where the difference between the original values μ_1, μ_2 is captured via the $\bar{\gamma}_i$'s. (See Maglaras and Zeevi (2004, Theorem 2) for details.) Since there is only one value of μ appearing asymptotically, the sum process is now tractable and its steady-state behavior is essentially characterized through the total traffic intensity in the system given by

$$\rho^n \;=\; \sum_i \frac{n\kappa_i\bar{\mu} - \sqrt{n}(\bar{\gamma}_i + \kappa_i\zeta_i)\bar{\mu}}{n\bar{\mu}\left(1 - \frac{\zeta_i}{\sqrt{n}}\right)} \;=\; 1 - \frac{\bar{\gamma}_1 + \bar{\gamma}_2}{\sqrt{n}} + o(1/\sqrt{n})),$$

which can be rewritten in the form $\rho^n = 1 - \bar{\gamma}/\sqrt{n} + o(1/\sqrt{n})$ for

$$\bar{\gamma} = \bar{\gamma}_1 + \bar{\gamma}_2 = \gamma_1\frac{\mu_1}{\bar{\mu}} + \gamma_2\frac{\mu_2}{\bar{\mu}}.$$

In contrast, the original system with service rates μ_1, μ_2 and arrival rates λ_1, λ_2 had total traffic intensity $1 - (\gamma_1 + \gamma_2)/\sqrt{C}$; i.e., the approximating scheme eventually reduces to scaling the γ_i's by $\mu_i/\bar{\mu}$, respectively. Finally, using the results of Section 4 we can compute the congestion cost for the limit system that has same service rates as $\mathbb{E}D = d(\bar{\gamma})$, where $d(\cdot)$ is given in (11). Returning to our original system with $\gamma_1 = 0$ and $\gamma_2 = -\kappa_2 \frac{q}{\mu_2} \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)} d(\gamma) + o(1/\sqrt{C})$, we get the equilibrium equation

$$\bar{\gamma} = -\kappa_2 \frac{q}{\bar{\mu}} \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)} d(\bar{\gamma}). \tag{16}$$

We note that the perturbation approximation pertains to a system that is announcing the true value of the steady-state delay and not the value derived through the perturbation approximation.

That is, the users are responding to the "right" information and not an approximation that is computed through the proposed perturbation approach.

The proposed approximation is quite accurate, as the following numerical study illustrates. Figure 2 studies the equilibrium congestion cost suffered by the BE class as we vary the ratio between μ_2/μ_1. In particular, the congestion cost for the system with $\mu_1 \neq \mu_2$ is computed via simulation and then contrasted against the perturbation approximation which uses $\bar{\mu} = (\mu_1+\mu_2)/2$. The demand model parameters were: $n = 100$, $\Lambda_1^n = 150$, $\alpha_1^n = 10$, $\Lambda_2^n = 200 \cdot \mu_2$, $\alpha_2^n = 20 \cdot \mu_2^2$, $\mu_1 = 1$, $\mu_2 \in \{1, 1.25, 2, 3, 4, 6, 8, 10\}$, $q = 1$ (Under (5), $\kappa_i = .5$ for all parameter choices.). As is evident from this figure (as well as in other test cases studied), the perturbation approximation described above provides an accurate estimate of the congestion cost, and this accuracy degrades, as one would expect, when the ratio of the μ_i's increases.

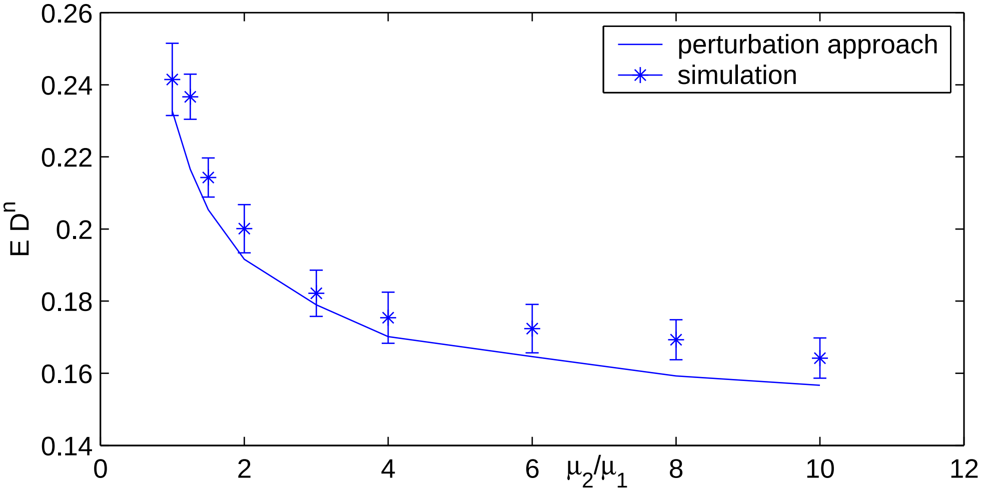

Figure 2: Accuracy of the perturbation approximation in systems with non-identical service rates. The graph depicts the expected delay as a function of μ_2/μ_1. (Error bars represent pointwise 95% confidence intervals for quantities estimated via simulation.)

7 Optimizing Revenues: A Second Order Price Correction

The last three sections have studied the behavior and revenue performance of the fluid prices derived from the solution of the deterministic revenue maximization problem (5). A key insight is that under this pricing policy the system will naturally operate in heavy traffic with a low level of congestion in the BE service class, and the resulting revenue loss due to congestion will be moderate. The latter was captured through the second order term in Theorem 2. This section will apply the asymptotic results derived thus far to approximate the performance of a given system with capacity C that is assumed to be large, and proceed to optimize the second order revenue loss term by appropriately fine tuning the corresponding pricing decisions.

Specifically, for a system with large capacity C, the total load into the system ρ is of the form $1 - \frac{\gamma}{\sqrt{C}}$, the BE congestion $\mathbb{E}D$ is of order $\frac{d(\gamma)}{\sqrt{C}}$, and in equilibrium

$$\lambda_1(\bar{p}_1) = \kappa_1 C\mu_1 \qquad \text{and} \qquad \lambda_2(\bar{p}_2 + \frac{q}{\mu_2}\mathbb{E}D) \approx \kappa_2 C\mu_2 - \sqrt{C}\beta_2\mu_2\frac{q}{\mu_2}d(\gamma),$$

where $\beta_2 = -\frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)}\kappa_2$, and γ is defined via the equilibrium equation $\gamma = \beta_2\frac{q}{\mu_2}d(\gamma)$. This leads to $\sqrt{C}\beta_2\mu_2\frac{q}{\mu_2}d(\gamma)\bar{p}_2$ in lost revenues due to congestion effects, which may be significant.

Refining the fluid-optimal price. Building on the scaling relations given in Theorems 1 and 2, we will consider a pricing rule that incorporates a second order price correction term of the form

$$p_i^* = \bar{p}_i + \frac{\pi_i}{\sqrt{C}} \quad \pi_i \in \mathbb{R}. \tag{17}$$

Such prices do not affect the first order behavior of the system, while introducing a second order correction that affects the equilibrium behavior and the lost revenues due to congestion. Pricing rules of this form have been shown to be asymptotically optimal in Maglaras and Zeevi (2003*a*) for a single-class system offering BE type of service (under the additional assumption that the demand is elastic).

Under the pricing rule (17), and assuming that C is large, we have that

$$\lambda_1(p_1^*) \approx \kappa_1 C\mu_1 - \sqrt{C}\beta_1\mu_1\pi_1 \qquad \text{and} \qquad \lambda_2(p_2^* + \frac{q}{\mu_2}\mathbb{E}D) \approx \kappa_2 C\mu_2 - \sqrt{C}\beta_2\mu_2[\pi_1 + \frac{q}{\mu_2}d(\gamma)],$$

where $\beta_i = -\frac{\tilde{\lambda}_i'(\bar{p}_i)}{\tilde{\lambda}_i(\bar{p}_i)}\kappa_i$, and γ is defined via the equilibrium equation

$$\gamma = \beta_1\pi_1 + \beta_2(\pi_2 + \frac{q}{\mu_2}d(\gamma)). \tag{18}$$

It is again easy to show that for any $\pi_1, \pi_2 \in \mathbb{R}$ this expression has a unique solution $\gamma > 0$ that characterizes the system equilibrium. The system revenues under the pricing rule (17) are

$$\begin{aligned} R(p_1^*, p_2^*) &:= \lambda_1(p_1^*)p_1^* + \lambda_2(p_2^* + q/\mu_2\mathbb{E}D)p_2^* \\ &\approx \kappa_1 C\mu_1\bar{p}_1 + \kappa_2 C\mu_2\bar{p}_2 \\ &\qquad - \sqrt{C}\left[\beta_1\mu_1\bar{p}_1\pi_1 - \kappa_1\mu_1\pi_1 + \beta_2\mu_2\bar{p}_2(\pi_2 + \frac{q}{\mu_2}d(\gamma)) - \kappa_2\mu_2\pi_2\right], \end{aligned} \tag{19}$$

where the '≈' notation implies equality to within lower order terms in C.

Optimizing revenue rates. Given (19) we can formulate a second order optimization problem that determines the price correction factors π_i as follows:

$$\min_{\pi_i \in \mathbb{R}} \quad \left\{ \beta_1\mu_1\bar{p}_1\pi_1 - \kappa_1\mu_1\pi_1 + \beta_2\mu_2\bar{p}_2(\pi_2 + \frac{q}{\mu_2}d(\gamma)) - \kappa_2\mu_2\pi_2 \right\}, \tag{20}$$

subject to the equilibrium condition (18). This problem can be readily solved by searching over the π_i's and using the closed-form expression for $d(\gamma)$ given in (11). For each value of the vector π, the above calculation requires the evaluation of the system equilibrium behavior. If the μ's are the same this reduces to finding the unique solution of (18). If the μ_i's are different, then the equilibrium equation is modified according to the perturbation approximation described earlier, by setting $\bar{\mu} = (\mu_1 + \mu_2)/2$ and replacing μ_2 by $\bar{\mu}$ and β_i by $\bar{\beta}_i = \beta_i(\mu_i/\bar{\mu})$ in (18).

In the context of this second-order analysis, one can also consider a system that offers quality-of-service (QoS) guarantees of the form $\mathbb{E}D \leq \delta$, for some appropriate bound $\delta > 0$. To incorporate these guarantees in an asymptotic sense on needs to add the constraint $d(\gamma) \leq \delta\sqrt{C}$ to the optimization problem posed above in terms of the second order price correction terms π_1, π_2.

Numerical results. We conclude this section with a set of numerical results that illustrate the effect of second order price corrections on system-wide revenues. To isolate the effect of the pricing changes we have kept the service rates (μ_1, μ_2) equal. Given (19), it follows that the magnitude of the revenue improvement due to the second order price corrections is second order, i.e., it grows like the square-root of capacity. Table 2 focuses on the dependence of these refinements, in terms of their effect on pricing decisions and equilibrium revenues, on the demand and delay sensitivity parameters. All reported results were obtained via the proposed asymptotic approximations, since as illustrated in the previous sections these tend to be quite accurate. Specifically, $R(\bar{p}) := R(\bar{p}_1, \bar{p}_2)$ and $R(p^*) := R(p_1^*, p_2^*)$ were computed via (19) for $\pi_1 = \pi_2 = 0$ in the first case and the optimal π_i's obtained from (20) in the second. $\bar{R}(\bar{p}) := \bar{R}(\bar{p}_1, \bar{p}_2)$ is the revenue rate obtained from the deterministic formulation (5). The system parameters were: $C = 100$, $\mu_1 = \mu_2 = 1$, $\Lambda_1 = 150$, and $\Lambda_2 = 200$.

A quick inspection of these results highlights that the absolute magnitude of the revenue improvements is modest. This is not surprising in light of the fact that the system is operating in a regime that is close to heavy traffic and is extracting almost maximum revenues (upper bounded by $\bar{R}(\bar{p}_1, \bar{p}_2)$). However, a more detailed look at the results illustrates that the (%) improvement in terms of the distance from the upper bound $\bar{R}(\bar{p})$ - this is computed as the change in revenues $R(p^*) - R(\bar{p})$ over the sub-optimality gap under the fluid prices $\bar{R}(\bar{p}) - R(\bar{p})$ and is reported in the rightmost column of the table - can be significant.

8 Effects of Congestion Notification

This section considers a system that announces state-dependent congestion information for the BE (or low priority) service class and analyzes the economic implications of this design decision.

(q,α_1,α_2)	$\bar{p}$	p^*	$R(\bar{p})$	$R(p^*)$	$\bar{R}(\bar{p})$	$\frac{R(p^*)-R(\bar{p})}{\bar{R}(\bar{p})-R(\bar{p})}$
(1,2.5,20)	(33.33, 8.33)	(34.11, 7.93)	2455	2460	2500	9.7%
(1,10,20)	(10.00,7.50)	(11.00, 6.96)	840	841	875	1.8%
(1,10,40)	(9.00, 4.00)	(9.68, 3.70)	674	675	700	2.7%
(.25,10,40)	(9.00, 4.00)	(9.24, 3.82)	684	686	700	6.9%
(.1,10,40)	(9.00, 4.00)	(8.16, 4.12)	689	691	700	10.5%
(4,10,40)	(9.00, 4.00)	(9.34, 3.82)	661	662	700	2.6%

Table 2: Sensitivity w.r.t. model parameters (q, α_1, α_2). The table shows the fluid-optimal price $\bar{p}$, the second order corrected price p^*, and the resulting revenues $R(\bar{p})$ and $R(p^*)$. The rightmost column displays the improvement (%) in the sub-optimality gap.

The model. The system announces the state-dependent congestion signal

$$D^d(t) = \frac{(Q_1^d(t) + Q_2^d(t) - C)^+}{C - Q_1^d(t)},$$

which is the excess delay defined in (1). Various quantities that are associated with this system will be tagged with a superscript d, mnemonic for "dynamic," reflecting the real-time nature of this congestion information. The BE-users evaluate the disutility associated with BE service using $D^d(t)$ in place of the steady-state expected congestion cost $\mathbb{E}D$. The state-dependent congestion signal results in a system with state-dependent arrival rate parameters given by

$$\lambda_1^d = \Lambda_1 \tilde{\lambda}_1(p_1) \quad \text{and} \quad \lambda_2^d(t) = \Lambda_2 \tilde{\lambda}_2 \left(p_2 + \frac{q}{\mu_2} D^d(t) \right). \tag{21}$$

Note that under the standing assumption that users do not act strategically in response to the firm's pricing and congestion notification strategy, this model no longer requires an equilibrium analysis.

System behavior under the deterministic solution. Following the scaling assumption given in (9), consider a system with capacity $C^n = n$ and potential demand $\Lambda_i^n = n\bar{\Lambda}_i$. As a starting point we will optimistically assume that the congestion suffered by the BE-class satisfies the scaling relations derived for the system with static information, and scales as

$$D^{n,d}(t) = \frac{D^d(t)}{\sqrt{n}} + o_p(1/\sqrt{n}), \qquad \text{for all } t \geq 0 \tag{22}$$

for some appropriate limit process $D^d(\cdot)$ to be identified later. Here $a_n = o_p(b_n)$ if $a_n/b_n \Rightarrow 0$ as $n \to \infty$. The immediate consequence of this assumption is that under the fluid prices $\bar{p}_1, \bar{p}_2$ the arrival rates in to the two service classes are of the form

$$\lambda_1^{n,d} = \kappa_1 n \mu_1 \qquad \text{and} \qquad \lambda_2^{n,d}(t) = \kappa_2 n \mu_2 - \sqrt{n}\mu_2 \beta_2 \frac{q}{\mu_2} D^d(t) + o_p(\sqrt{n}), \quad \text{for all } t \geq 0, \tag{23}$$

where $\beta_2 = -\kappa_2 \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}(\bar{p}_2)}$. Note that $\lambda_2^{n,d}(t)$ is now a stochastic process. The overall traffic intensity is

$$\rho^n(t) = 1 - \beta_2 \frac{q}{\mu_2} \frac{D^d(t)}{\sqrt{n}} + o_p(1/\sqrt{n}) \quad \text{for all } t \geq 0.$$

In the spirit of the results reported in the previous sections, it is natural to posit that $Q_i^{n,d}(t)$ can be expressed as

$$Q_i^{n,d}(t) = \kappa_i n + \sqrt{n} X_i^{n,d}(t), \tag{24}$$

where the κ_i's were defined in (8), and the process $(X_1^{n,d}(t), X_2^{n,d}(t) : t \geq 0)$ is defined by

$$X_i^{n,d}(t) := \frac{Q_i^{n,d}(t) - \kappa_i n}{\sqrt{n}} \quad \text{for } i = 1, 2. \tag{25}$$

Using the expression in (24), the congestion term can be approximated as follows

$$\begin{aligned} D^{n,d}(t) &= \frac{(Q_1^{n,d}(t) + Q_2^{n,d}(t) - n)^+}{n - Q_1^{n,d}(t)} \\ &= \frac{\sqrt{n}(X_1^{n,d}(t) + X_2^{n,d}(t))^+}{\kappa_2 n - \sqrt{n} X_1^{n,d}(t)} \\ &= \frac{1}{\kappa_2 \sqrt{n}} (X_1^{n,d}(t) + X_2^{n,d}(t))^+ + o_p(1/\sqrt{n}), \end{aligned}$$

which is consistent with (22). To justify the heuristic approach taken above we need to establish that (24) is indeed correct by identifying a well-behaved limit for the process $(X_1^{n,d}(t), X_2^{n,d}(t) : t \geq 0)$. To this end, let $(X^n(t) : t \geq 0)$ and $(X(t) : t \geq 0)$ be $\mathbb{R}^m$-valued continuous time stochastic processes with sample paths in the space of functions having right-continuous paths with left limits. Then, $X^n(\cdot) \Rightarrow X(\cdot)$ denotes weak convergence in this functional space with respect to the Skorohod topology; see, e.g., Billingsley (1968, §3). (Since all limit processes in this paper have continuous sample paths, it suffices to consider the above convergence w.r.t. the uniform metric on compact sets $[0, T]$, with $T < \infty$.) The next proposition justifies the heuristic described above.

Proposition 4 *Assume that demand (λ^n) and capacity (C^n) grow large as in (9) as $n \to \infty$. If $X^{n,d}(0) \Rightarrow \xi$ for some $\xi \in \mathbb{R}^2$, then, under the pricing rule $(\bar{p}_1, \bar{p}_2)$, $X^{n,d}(\cdot) \Rightarrow X^d(\cdot)$ as $n \to \infty$, where X^d is the unique strong solution of the stochastic differential equation:*

$$dX^d(t) = b^d(X^d(t))dt + \Sigma dW(t), \quad X^d(0) = \xi, \tag{26}$$

where $W = (W(t) : t \geq 0)$ is standard Brownian motion in $\mathbb{R}^2$. The infinitesimal drift is given by

$$\begin{aligned} b_1^d(x_1, x_2) &= -\mu_1 x_1 \\ b_2^d(x_1, x_2) &= \begin{cases} -\mu_2 x_2 & x_1 + x_2 \leq 0 \\ -\mu_2 \beta_2 \frac{q}{\mu_2} \frac{x_1 + x_2}{\kappa_2} + \mu_2 x_1 & x_1 + x_2 > 0 \end{cases}, \end{aligned} \tag{27}$$

where $\beta_2 = -\kappa_2 \frac{\tilde{\lambda}_2'(\bar{p}_2))}{\tilde{\lambda}_2(\bar{p}_2)}$ *and* $\Sigma := \mathrm{diag}(\sigma_1, \sigma_2)$*, with* $\sigma_i^2 = 2\mu_i\kappa_i$*. Finally,*

$$\sqrt{n}D^{n,d}(\cdot) \Rightarrow D^d(\cdot) := \frac{(X_1^d(\cdot) + X_2^d(\cdot))^+}{\kappa_2}. \tag{28}$$

That is, $X^{n,d}$ has a well-defined limit, and the approximation of the congestion signal $D^{n,d}(t)$ asserted in (22) is rigorously justified on the basis of (28). The process $X_1^d(\cdot)$ that approximates the fluctuations of the G-users evolves as an Ornstein-Uhlenbeck process that is independent of the congestion information, whereas the drift of the $X_2^d(\cdot)$ process is modulated by the value of the $X_1^d(\cdot)$ process. The congestion suffered by BE users, the demand for BE service, and the associated revenue rate are all functions of the "sum" process $Z^d(\cdot) := X_1^d(\cdot) + X_2^d(\cdot)$. For the case where $\mu_1 = \mu_2 = \mu$, Z^d is a tractable one-dimensional diffusion that solves the stochastic differential equation[2]

$$dZ^d(t) = b_z^d(Z^d(t))dt + \sigma dW(t),$$

where $W = (W(t) : t \geq 0)$ is standard Brownian motion in $\mathbb{R}$, the infinitesimal drift is

$$b_z^d(z) = \begin{cases} -\mu z & z < 0 \\ -\mu\alpha z & z \geq 0, \end{cases} \tag{29}$$

where $\alpha := \beta_2 q/(\kappa_2\mu)$, and the infinitesimal variance is $\sigma^2(z) = 2\mu$. Note that (29) is simply obtained by adding the two drift components in (27).

The above diffusion is comprised of two O-U processes that are "pasted together;" one describes the dynamics when the system has spare capacity, $z < 0$ in diffusion scale, and the other gives the behavior when the system is in the congested state, $z > 0$ in diffusion scale. Using results from Browne and Whitt (1995): (i) when $Z^d < 0$, $Z^d \sim N(0,1)$, where $Z^d := Z^d(\infty)$; (ii) when $Z^d \geq 0$, $Z^d \sim N(0, 1/\alpha)$. Putting the two together we get that $P(Z^d \geq 0) = \frac{\phi(0)}{\phi(0)+\sqrt{\alpha}\phi(0)} = \frac{1}{1+\sqrt{\alpha}}$ and

$$\mathbb{P}(Z^d \leq z | Z^d \leq 0) = 2\Phi(z) \quad z \leq 0 \quad \text{and} \quad \mathbb{P}(Z^d > z | Z^d > 0) = 2\Phi(-z\sqrt{\alpha}) \quad z > 0. \tag{30}$$

Given that $D^d(t) = (Z^d(t))^+/\kappa_2$, a straightforward calculation leads to

$$\mathbb{E}D^d = \sqrt{\frac{2}{\pi}} \frac{1}{(\sqrt{\alpha} + \alpha)\kappa_2}. \tag{31}$$

The value of real-time congestion notification. First, note that the revenue rate extracted at time t depends on the state of the system at that time through the congestion signal $D^{n,d}(t)$. For simplicity, the remainder of this section will restrict attention to the case of identical service rates and prices fixed at $\bar{p} = (\bar{p}_1, \bar{p}_2)$, i.e., $\pi_1 = \pi_2 = 0$, and refer the reader to Sections 6 and 7 for guidelines on possible extension to the general case. With a slight abuse of notation, we will denote

[2] One can obtain a similar process for the case $\mu_1 \neq \mu_2$ akin to the results of Section 6; details are omitted.

the revenue rate at time t by $R(\bar{p}, D^{n,d}(t))$. Then, using the fact that $D^{n,d}(t) = \frac{1}{\sqrt{n}} D^d(t) + o_p(1/\sqrt{n})$, we get that

$$\begin{aligned} R(\bar{p}, D^{n,d}(t)) &:= \lambda_1^{n,d}(t)\bar{p}_1 + \lambda_2^{n,d}(t)\bar{p}_2 \\ &= (\kappa_1 n\mu\bar{p}_1) \;+\; \Big(\kappa_2 n\mu\bar{p}_2 - \sqrt{n}\beta_2\bar{p}_2 q D^d(t)\Big) + o_p(\sqrt{n}). \end{aligned} \tag{32}$$

Finally, using the expected steady-state value $\mathbb{E}D^d$ given in (31) one can approximate the expected revenue rate for that system via a non-rigorous interchange of limits as follows

$$\mathbb{E}R(\bar{p}, D^{n,d}) \approx (\kappa_1 n\mu\bar{p}_1) \;+\; \Big(\kappa_2 n\mu\bar{p}_2 - \sqrt{n}\beta_2\bar{p}_2 q\mathbb{E}D^d\Big). \tag{33}$$

Recall that the revenue rate for the system with static information can be approximated by

$$R(\bar{p}, \mathbb{E}D^n) \approx (\kappa_1 n\mu\bar{p}_1) \;+\; \big(\kappa_2 n\mu\bar{p}_2 - \sqrt{n}\beta_2\bar{p}_2 q\mathbb{E}D\big). \tag{34}$$

Thus,

$$\Delta(\mathbb{E}R^n) := \mathbb{E}R(\bar{p}, D^{n,d}) - R(\bar{p}, \mathbb{E}D^n) \approx \sqrt{n}\beta_2\bar{p}_2 q(\mathbb{E}D - \mathbb{E}D^d). \tag{35}$$

The next theorem establishes that the above difference is strictly positive as n grows large.

Theorem 3 *Suppose that $\mu_1 = \mu_2 = \mu$, and let the conditions of Proposition 1 hold. Then, $\mathbb{E}D^d < \mathbb{E}D$.*

That is, BE-users experience better quality-of-service when real-time congestion information is provided to them. This, in turn, implies that the mean arrival rate into class 2 is larger when real-time congestion information is announced, which leads to the increase in revenues. Using the result of the theorem, we infer that real-time congestion notification results in a gain of order $\sqrt{n}$ in terms of generated revenues.

Numerical results. Figure 3 depicts the increase in revenues and decrease in expected congestion cost that occur in a system with real-time congestion notification. (Note that the simulation of the system with real-time congestion information involves a Markov chain with state-dependent parameters, but does not require a calculation of an equilibrium operating point.) The model parameters were: $n = \{50, 100, \ldots, 450\}$, $\Lambda_1^n = 1.5 \cdot n$, $\alpha_1^n = n/10$, $\Lambda_2^n = 2 \cdot n$, $\alpha_2^n = n/5$, $\mu_1 = \mu_2 = 1$, $q = 1$ (Under (5), $\kappa_i = .5$ and $\bar{p}_1 = 10$ and $\bar{p}_2 = 7.50$, independent of n.). We make three observations about these plots. First, as shown in Theorem 3, real-time congestion information leads to an increase in expected revenues that is proportional to the square root of the capacity, which seems to agree with the results displayed in the figure. Second, the expected delay suffered by BE users in the system with real-time information is indeed smaller. Finally, the variability in the simulation estimates for the change in expected revenues is higher in comparison to other results because here we need to simulate two independent systems.

Acknowledgments: The authors are grateful to the two referees, the associate editor and the area editor for their constructive comments and helpful suggestions.

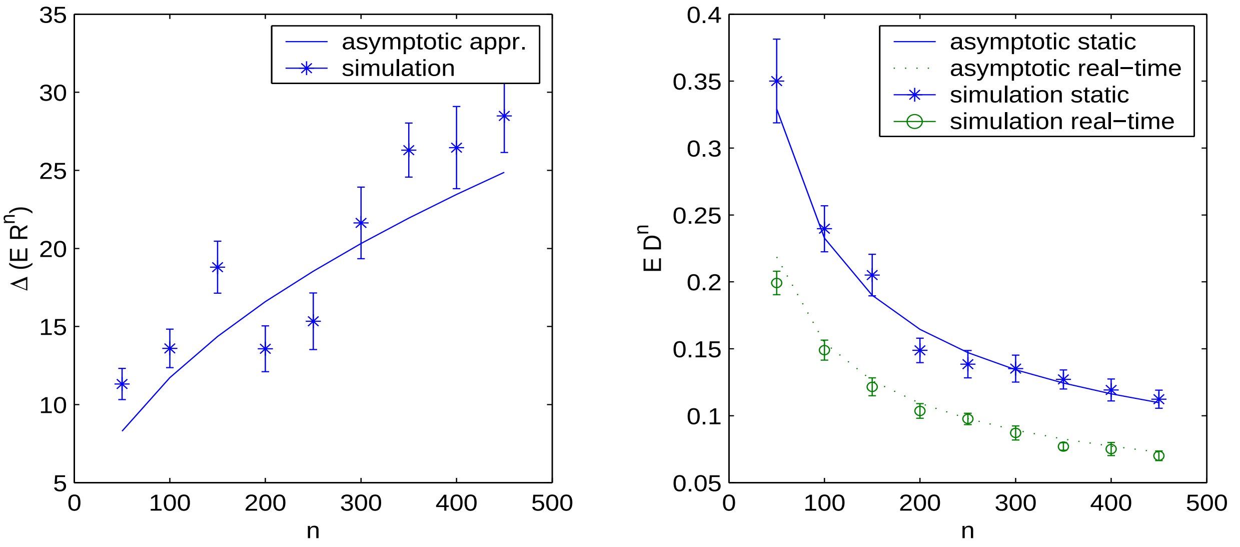

Figure 3: The effect of real-time congestion notification. The left figure shows gains in expected revenues $\Delta(\mathbb{E}R^n)$ as a function of the system capacity (n), and the right figure shows the behavior of expected delay as a function of the system capacity. (Error bars represent pointwise 95% confidence intervals for quantities estimated via simulation.)

A Diffusion Limits: Background and Auxiliary Results

The following theorem, whose proof can be found in Maglaras and Zeevi (2004, Theorem 1 and Corollary 1), characterizes the limiting dynamics in a system with no congestion feedback signal, and assumes that the arrival rates into each class are of the form $\lambda_i^n = \kappa_i\mu_i n - \gamma_i\mu_i\sqrt{n} + o(\sqrt{n})$, for $i = 1, 2$, $n = 1, 2, \ldots$, with γ_i such that $\gamma_1 + \gamma_2 > 0$. The structural implications of this result underlie the proof of Theorem 1, as in that theorem it it shown that equilibrium arrival rates are exactly of the form assumed in the result below. As in the main text, let $X_i^n(\cdot) := n^{-1/2}(Q_i^n(\cdot) - \kappa_i n)$ and $X^n(\cdot) = (X_1^n(\cdot), X_2^n(\cdot))$ and '⇒' denotes weak convergence in the space of functions which are right-continuous with left-limits, with respect to the Skorohod topology; see Billingsley (1968, §3).

Theorem 4 (Maglaras and Zeevi, 2002) *Assume that the arrival rates are of the form $\lambda_i^n = \kappa_i\mu_i n - \gamma_i\mu_i\sqrt{n} + o(\sqrt{n})$, for $i = 1, 2$, $n = 1, 2, \ldots$, with γ_i such that $\gamma_1 + \gamma_2 > 0$. Suppose that $X^n(0) \Rightarrow \xi$ for some $\xi \in \mathbb{R}^2$. Then, $X^n(\cdot) \Rightarrow X(\cdot)$ as $n \to \infty$, where X is a diffusion process. Specifically, X is the unique strong solution of the following stochastic differential equation:*

$$dX(t) = b(X(t))dt + \Sigma dW(t) \quad X(0) = \xi, \tag{36}$$

where $W = (W(t) : t \geq 0)$ is standard Brownian motion in $\mathbb{R}^2$, the infinitesimal drift function $b_i(\cdot)$ for the i'th component is

$$\begin{aligned} b_1(x_1, x_2) &= -\mu_1\gamma_1 - \mu_1 x_1 \\ b_2(x_1, x_2) &= \begin{cases} -\mu_2\gamma_2 - \mu_2 x_2 & x_1 + x_2 \leq 0 \\ -\mu_2\gamma_2 + \mu_2 x_1 & x_1 + x_2 > 0 \end{cases}, \end{aligned} \tag{37}$$

and $\Sigma := \mathrm{diag}(\sigma_1, \sigma_2)$, with $\sigma_i^2 = 2\mu_i\kappa_i$. Moreover, X admits a unique stationary distribution and $X(t) \Rightarrow X(\infty)$ as $t \to \infty$. Finally,

$$\sqrt{n}D^n(\cdot) \Rightarrow \frac{1}{\kappa_2}(X_1(\cdot) + X_2(\cdot))^+.$$

The infinitesimal drift in (37) has an intuitive interpretation: the limit process for the G-users, $X_1(\cdot)$, evolves freely as an Ornstein-Uhlenbeck (O-U) process, while the drift of the limit process for the BE-users, $X_2(\cdot)$, is modulated by the number of excess G-users present in the system. Based on the results given in Theorem 4, and assuming one can justify an interchange of expectation limits on n and t, we anticipate that $\mathbb{E}D^n \approx d/\sqrt{n}$. It turns out that this is sufficient to conclude the structural results (i)-(iv) in Theorem 1. (This interchange argument is rigorously justified for the case of $\mu_1 = \mu_2$ in the proof of Theorem 1, and a similar argument can be employed when the μ's are different.)

B Proofs

Proof of Proposition 2: The proof relies on a relatively straightforward sample path argument imitating the construction in Loynes (1962); details are omitted. ■

Proof of Proposition 3: Using a stochastic ordering argument one can verify that the expected delay $\mathbb{E}D(\lambda_1, \lambda_2)$, considered as an explicit function of the arrival rates, is monotonically increasing in λ_2. In what follows we let d denote $\mathbb{E}D(\lambda_1, \lambda_2)$. Note that

$$\frac{\partial\lambda_1(p_1)}{\partial d} = 0 \quad \text{and} \quad \frac{\partial\lambda_2(p_2 + \frac{q}{\mu_2}d)}{\partial d} < 0.$$

The equilibrium regime can be defined via the solution d^* of the set of equations

$$\lambda_1 = \lambda_1(p_1), \quad \lambda_2 = \lambda_2(p_2 + \frac{q}{\mu_2}d^*), \quad \text{and} \quad d^* = \mathbb{E}D(d^*),$$

where $\mathbb{E}D(d^*)$ denoted the steady state expected waiting time for class 2 service when the arrival rates into classes 1 and 2 are λ_1 and $\lambda_2(p_2 + \frac{q}{\mu_2}d^*)$, respectively. Define the function $h(d) = d - \mathbb{E}D(d)$. First note that $h(0) < 0$, $h(\infty) > 0$, and by assumption $h(\cdot)$ is continuous. Moreover, since $\mathbb{E}D(\lambda_1, \lambda_2)$ is monotonically increasing in λ_2 and λ_2 is monotonically decreasing in d, it follows that $\mathbb{E}D(d)$ is monotonically decreasing in d. This implies that $h(\cdot)$ is monotonically increasing in d, and as a result the equation $h(d) = 0$ has a unique solution, d^* that characterizes the equilibrium regime. (The monotonicity of the functions $f(d) = d$ and $g(d) = \mathbb{E}D(d)$ guarantee that the function $h(d)$ switches sign, and the continuity assumption on ensures the existence of a solution to $h(d) = 0$; i.e., $h(d)$ does not switch sign at a point of discontinuity.) $\blacksquare$

Proof of Theorem 1: We first prove statement (iii) and then statements (i),(ii) and (iv).

Step 1. Proof of (iii). Note that the arrival rate into class 1, namely the G-users, is given by $\lambda_1^n(\bar{p}_1) = \Lambda_1^n \tilde{\lambda}_1(\bar{p}_1) = n\mu\kappa_1$. Since the number of G-users in the system, Q_1^n, follows an $M/M/n/n$ queue, its steady-state is

$$\mathbb{P}(Q_1^n = k) = \frac{a_n^k/k}{\sum_{j=0}^n a_n^j/j} \quad \text{for } k = 0, 1, \ldots, n \tag{38}$$

where $a_n := \lambda_1^n/\mu = \kappa_1 n$, by definition of κ_1. Let Z^n be a r.v. distributed Poisson with mean a_n. Then, multiplying numerator and denominator in (38) by $\exp(-a_n)$ we can express the steady-state of Q_1^n as

$$\mathbb{P}(Q_1^n = k) = \frac{\mathbb{P}(Z^n = k)}{\mathbb{P}(0 \leq Z^n \leq n)} \quad \text{for } k = 0, 1, \ldots, n. \tag{39}$$

The following auxiliary result gives an upper bound on the tail of a Poisson r.v.

Lemma 1 *For any $\epsilon \in (0, 1 - \kappa_1)$ we have that*

$$\mathbb{P}\left(Z^n \geq (\kappa_1 + \epsilon)n\right) \leq e^{-cn} \quad \textit{for } n = 1, 2, \ldots$$

where $c = c(\epsilon) > 0$.

Thus, $\mathbb{P}(0 \leq Z^n \leq n) \to 1$ as $n \to \infty$. To get a bound on the blocking probability, all we need is to bound the probability that Z^n exceeds n. But, Lemma 1 yields exactly the bound asserted in (iii) in the theorem. Finally, we note that the above arguments imply that $Q_1^n/n \Rightarrow \kappa_1$, as $n \to \infty$. To see why this is true, fix $\epsilon > 0$ and note that

$$\begin{aligned} \mathbb{P}(Q_1^n \geq n(\kappa_1 - \epsilon)) &= \frac{\mathbb{P}((Z^n - a_n)/\sqrt{a_n} \geq -\sqrt{n}\epsilon\kappa_1^{-1/2})}{\mathbb{P}(0 \leq Z^n \leq n)} \\ &\to \mathbb{P}(N(0,1) > -\infty) = 1 \end{aligned}$$

as $n \to \infty$, since $(Z^n - a_n)/\sqrt{a_n} \Rightarrow N(0,1)$ by the central limit theorem for a Poisson r.v.

Step 2. In the sequel we will make use of a fictitious system to upper bound various system processes. This is a system without blocking, where a G-user arriving when $Q_1^n(t) = n$ is allowed to join a queue and wait until the first serviced G-user in the system departs. to be served by the first available idle server. We will denote the associated processes by $\tilde{Q}_i^n(\cdot)$, and note that

$\tilde{Q}_i^n(\cdot) \geq Q_i^n(\cdot)$, and that the dynamics of $\tilde{Q}_1^n(\cdot) + \tilde{Q}_2^n(\cdot)$ are that of an $M/M/n$ queue. For this system, if $\rho^n < 1$, then

$$\mathbb{E}(\tilde{Q}_1^n + \tilde{Q}_2^n - n)^+ = \frac{\rho^n \mathbb{P}(\tilde{Q}_1^n + \tilde{Q}_2^n \geq n)}{(1-\rho^n)}. \tag{40}$$

This follows from standard formulas for the steady-state distribution of an $M/M/N$ queue, see, e.g., Halfin and Whitt (1981). Moreover, observe that $\tilde{Q}_1^n(\cdot) + \tilde{Q}_2^n(\cdot) \geq Q_1^n(\cdot) + Q_2^n(\cdot)$ and $n - Q_1^n(\cdot) \geq n - \tilde{Q}_1^n(\cdot)$; i.e., this fictitious system provides a pointwise upper (lower) bound for the dynamics of the number-in-system (available capacity for BE users) in the original system.

Step 3. We now turn our attention to the proof of (i),(ii) and (iv). The BE-users' delay is

$$d^n := \mathbb{E}D^n = \mathbb{E}\left[\frac{(Q_1^n + Q_2^n - n)^+}{(n - Q_1^n) \vee 1}\right] .$$

Suppose that $\liminf_{n\to\infty} d^n > 0$. If d^n does not converge, take a subsequence such that $d^{n_j} \to c$, as $j \to \infty$, where $c > 0$, and for simplicity, let this subsequence also be indexed by n. Then, $\lambda_2^n = \Lambda_2^n \tilde{\lambda}(\bar{p}_2 + (q/\mu)d^n)$ must be such that $\lim_{n\to\infty} \lambda_2^n/n < \kappa_2\mu$. Thus, $\lim_{n\to\infty} \rho^n < 1$, since $\lambda_1^n/n \to \kappa_1\mu$. Using the system defined in Step 2 for $\rho^n < 1$ and using (40) we get that

$$\begin{aligned} \mathbb{E}D^n &\leq \mathbb{E}(\tilde{Q}_1^n + \tilde{Q}_2^n - n)^+ \\ &= \frac{\rho^n \mathbb{P}(\tilde{Q}_1^n + \tilde{Q}_2^n \geq n)}{(1-\rho^n)} \\ &= o(1), \end{aligned}$$

as $n \to \infty$, where the last step follows from (Halfin and Whitt 1981, Proposition 1) that asserts that in an $M/M/n$ system with $\sqrt{n}(1-\rho^n) \to \infty$, $\mathbb{P}(\tilde{Q}_1^n + \tilde{Q}_2^n \geq n) \to 0$. Hence, we have a contradiction and it must be that $d^n \to 0$ as $n \to \infty$.

Step 4. To get the convergence rate of d^n, note that using a Taylor expansion for $\lambda_2(\cdot)$ we have

$$\lambda_2^n = n\kappa_2\mu + \Lambda^n \tilde{\lambda}'(\bar{p}_2)(q/\mu)d^n + o(nd^n),$$

and since $\Lambda^n = n\bar{\Lambda}$, we have that $\rho^n = 1 - cd^n + o(d^n)$ for some $c > 0$, as $n \to \infty$. Suppose that $\sqrt{n}d^n \to d \in (0,\infty)$, or equivalently that $\rho^n = 1 - \gamma/\sqrt{n}$. The next lemma studies the behavior of a system without feedback in this regime. (Its proof is relegated to the end of this appendix.)

Lemma 2 *Consider the two-class system that operates without feedback, and with arrival rates set to be* $\lambda_1^n = \kappa_1\mu n$ *and* $\lambda_2^n = \kappa_2\mu n - \mu\gamma\sqrt{n}$. *Then,*

$$\sqrt{n}\mathbb{E}D^n \to d(\gamma) \qquad \textit{and} \qquad \mathbb{P}(Q_1^n + Q_2^n \geq n) \to \nu(\gamma) \tag{41}$$

where $d(\gamma)$ *was given in (11) and* $\nu(\gamma) := \kappa_2\gamma d(\gamma)$.

That is, for a system where $\sqrt{n}(1-\rho^n) \to \gamma \in (0,\infty)$, $\sqrt{n}d^n \to d(\gamma) \in (0,\infty)$. To establish the equilibrium relation, (10), that holds for a system with feedback operating in the Halfin-Whitt regime, it suffices to consider the second order expansion for λ^n, the total arrival rate into the system given by

$$\lambda^n = n\mu + \sqrt{n}\mu\kappa_2 \frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)}(q/\mu)d(\gamma) + o(\sqrt{n}) \ .$$

Dividing through by $n\mu$ and equating second order terms we obtain the equilibrium condition

$$\gamma = -\kappa_2 \frac{q}{\mu}\frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)}d(\gamma),$$

which establishes (10). From Maglaras and Zeevi (2003a, Proposition 2) we have that this equation has a unique solution, $\gamma > 0$. (The proof of that statement considers the function $h(\gamma) = \gamma + \kappa_2 \frac{q}{\mu}\frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)}d(\gamma)$, and shows that h is continuous and increasing in $(0,\infty)$, $\lim_{\gamma\to 0} h(\gamma) < 0$, and $\lim_{\gamma\to\infty} h(\gamma) > 0$. Hence, $h(\gamma) =$ must have a unique solution.)

To complete the proof it remains to rule out the cases $\sqrt{n}d^n \to 0$ and $\sqrt{n}d^n \to \infty$. To this end, suppose that $\sqrt{n}d^n \to 0$, in which case $\sqrt{n}(1-\rho^n) \to 0$. Observe from (41) and (11) that if we let $\gamma \downarrow 0$, then $d(\gamma) \uparrow \infty$, which contradicts the assumption that $\sqrt{n}d^n \to 0$. Similarly, suppose that $\sqrt{n}d^n \to \infty$, in which case $\sqrt{n}(1-\rho^n) \to \infty$. Again from (41) and (11) we get that as $\gamma \uparrow \infty$, then $d(\gamma) \downarrow 0$, which contradicts the assumption that $\sqrt{n}d^n \to \infty$. Consequently, it must be that $\sqrt{n}d^n \to d(\gamma) \in (0,\infty)$ and $\sqrt{n}(1-\rho^n) \to \gamma \in (0,\infty)$, and γ is defined as the unique solution of equation (10). This establishes assertions (i) and (ii) in the statement of the theorem. Finally, from Lemma 2, $\sqrt{n}(1-\rho^n) \to \gamma$ implies statement (iv) of the theorem. This concludes the proof. ■

Proof of Theorem 2: First, recall that by construction of the deterministic relaxation problem in (5), $R_*^n \le \bar{R}^n$ for all n. By Theorem 1 and Lemma 1, $\mathbb{E}D^n = d/\sqrt{n} + o(1/\sqrt{n})$ and $\mathbb{P}(Q_1^n < n) \le C_1 \exp(-C_2 n)$ for sufficiently large n and constants $C_1, C_2 > 0$. Thus, we can take a Taylor series expansion of the total revenue generated by the fluid-optimal prices

$$\begin{aligned} R^n(\bar{p}_1,\bar{p}_2,\bar{U}) &= \Lambda_1^n \tilde{\lambda}_1(\bar{p}_1)\mathbb{P}(Q_1^n < n)\bar{p}_1 + \Lambda_2^n \tilde{\lambda}_2(\bar{p}_2 + (q/\mu)\mathbb{E}D^n)\bar{p}_2 \\ &= \Lambda_1^n \tilde{\lambda}_1(\bar{p}_1)\bar{p}_1 + \Lambda_2^n \tilde{\lambda}_2(\bar{p}_2)\bar{p}_2 + n\frac{\tilde{\lambda}_2'(\bar{p}_2)}{\tilde{\lambda}_2(\bar{p}_2)}(q/\mu)d/\sqrt{n})\bar{p}_2 + o(1/\sqrt{n}) \\ &= \bar{R}^n(1-\alpha/\sqrt{n}) + o(1/\sqrt{n}) \end{aligned}$$

as $n \to \infty$, using results (ii) and (iii) of Theorem 1. Thus,

$$\frac{R^n(\bar{p}_1,\bar{p}_2,\bar{U})}{\bar{R}^n} = 1 - \frac{\alpha}{\sqrt{n}} + o(1/\sqrt{n})$$

as $n \to \infty$ for some $\alpha > 0$. Since $R_*^n \le \bar{R}^n$ for sufficiently large n, the proof is complete. ■

Proof of Proposition 4: We will first express $D^{n,d}(\cdot)$ in terms of the $X_i^{n,d}(\cdot)$'s, and subsequently obtain the infinitesimal drift for the $Q_i^{n,d}(\cdot)$ and $X_i^{n,d}(\cdot)$ processes, respectively. Then, we will appeal to the proof techniques that underlie Theorem 4 to establish that $X^{n,d}(\cdot) \Rightarrow X^d(\cdot)$.

Using (25) we can rewrite $D^{n,d}(t)$ as

$$D^{n,d}(t) = \frac{(X_1^{n,d}(t) + X_2^{n,d}(t))^+}{\kappa_2\sqrt{n} - X_1^{n,d}(t)/\sqrt{n}}.$$

Suppose that $Q^{n,d}(t) = q^n$ for some $q^n \in S^n = \{(q_1^n, q_2^n) : q_1^n \in \{0, 1, \ldots, n\},\ q_2^n \in \{0, 1, \ldots\}\}$. Also, let $x^n = (q^n - \kappa n)/\sqrt{n}$ such that from (25) $X^{n,d}(t) = x^n$. The congestion signal at time t will be

$$D^{n,d}(q^n) = \frac{(q_1^n + q_2^n - n)^+}{n - q_1^n} = \frac{1}{\sqrt{n}} \frac{(x_1^n + x_2^n)^+}{\kappa_2 - x_1^n/\sqrt{n}}.$$

The arrival rates into the two service classes are given below

$$\lambda_1^{n,d}(q^n) = \kappa_1 n \mu_1 \quad \text{if } q_1^n < n \quad \text{and} \quad \lambda_1^{n,d}(q^n) = 0 \quad \text{otherwise,}$$

and

$$\lambda_2^{n,d}(q^n) = \kappa_2 n \mu_2 - \sqrt{n}\beta_2\mu_2 \frac{q}{\mu_2\kappa_2} \frac{(x_1^n + x_2^n)^+}{1 - x_1^n/\kappa_2\sqrt{n}} + o(\sqrt{n}).$$

With some of abuse of notation we will also refer to $D^{n,d}(q^n)$ and $\lambda_i^{n,d}(q^n)$ by $D^{n,d}(x^n)$ and $\lambda_i^{n,d}(x^n)$, respectively. Under the Markovian dynamics of our system, we have that for any initial state $q^n = (q_1^n, q_2^n) \in S^n$ and $\delta t > 0$, the infinitesimal drift rates for each class are given by

$$\begin{aligned} \mathbb{E}\left[Q_1^n(t + \delta t) - Q_1^n(t) \middle| \, Q^n(t) = q^n\right] &= \left[\lambda_1^n(q^n) - \mu_1 q_1^n\right] \delta t + o(\delta t) \\ \mathbb{E}\left[Q_2^n(t + \delta t) - Q_2^n(t) \middle| \, Q^n(t) = q^n\right] &= \left[\lambda_2^n(q^n) - \mu_2\left((n - q_1^n) \wedge q_2^n\right)\right] \delta t + o(\delta t) \quad , \end{aligned} \tag{42}$$

as $\delta t \downarrow 0$. Similarly, the infinitesimal variance for each class is

$$\begin{aligned} \mathbb{E}\left[(Q_1^n(t + \delta t) - Q_1^n(t))^2 \mid Q^n(t) = q^n\right] &= \left[\lambda_1^n(q^n) + \mu_1 q_1^n\right] \delta t + o(\delta t) \\ \mathbb{E}\left[(Q_2^n(t + \delta t) - Q_1^n(t))^2 \mid Q^n(t) = q^n\right] &= \left[\lambda_2^n(q^n) + \mu_2\left((n - q_1^n) \wedge q_2^n\right)\right] \delta t + o(\delta t) \ . \end{aligned} \tag{43}$$

Finally,

$$\mathbb{E}\left[(Q_1^n(t + \delta t) - Q_1^n(t))\,(Q_2^n(t + \delta t) - Q_2^n(t)) \ \mid Q^n(t) = q^n\right] = o(\delta t) \quad \text{for all } n = 1, 2, \ldots. \tag{44}$$

Using (42)-(44) we can derive the infinitesimal rates for the $X^{n,d}$ process. Specifically,

$$\begin{aligned} \frac{1}{\delta t}\mathbb{E}\left[X_1^n(t + \delta t) - X_1^n(t) \mid X^n(t) = x^n)\right] &= -\mu_1 x_1^n + o(1/\sqrt{n}) \\ \frac{1}{\delta t}\mathbb{E}\left[X_2^n(t + \delta t) - X_2^n(t) \mid X^n(t) = x^n)\right] &= -\mu_2\beta_2 \frac{q}{\mu_2} \frac{x_1^n + x_2^n}{\kappa_2} - \mu_2 x_2^n + \mu_2(x_1^n + x_2^n)^+ + o(1/\sqrt{n}) \end{aligned}$$

for small δt and large n. (Note the similarity between these expressions and the limiting infinitesimal drift given in the statement of the Proposition.) Similar expressions can be obtained for the infinitesimal variance. Using standard weak convergence arguments for Markov processes, as in

Maglaras and Zeevi (2004), we can now establish that $X^{n,d}(\cdot) \Rightarrow X^d(\cdot)$. To complete the proof of the proposition note that

$$D^{n,d}(t) = \frac{(X_1^d(t) + X_2^d(t))^+}{\kappa_2\sqrt{n}} + o(1/\sqrt{n}),$$

and apply the continuous-mapping theorem. ■

Proof of Theorem 3: The main challenge is to compare the explicit performance characterization for the system with real-time information given in (22) with the implicit equilibrium characterization given for the system with static information in (10) and (11).

We start by giving a skeleton of the proof. Step 1: We will study a fictitious system where instead of $D^{n,d}(t)$, the system manager announces the state independent BE congestion estimate $\frac{\xi}{\kappa_2\sqrt{n}}$. In terms of limiting behavior, this replaces the state-dependent drift term $-\mu\alpha z^+$ that appears in (29), with the constant $-\mu\beta_2\frac{q}{\mu}\frac{\xi}{\kappa_2}$. This new system is governed by the behavior given in Section 4 with $\gamma(\xi) = \beta_2\frac{q}{\mu}\frac{\xi}{\kappa_2}$. Denote by $\tilde{Z}(\xi)$ the steady state random variable associated with the "sum" process in this system. Step 2: Set $\xi = \mathbb{E}(Z^d)^+$. We demonstrate that $\mathbb{E}(\tilde{Z}(\xi))^+ > \mathbb{E}(Z^d)^+$, which implies that $\mathbb{E}\tilde{D}(\xi) := \mathbb{E}(\tilde{Z}(\xi))^+/\kappa_2 > \mathbb{E}D^d$. Step 3: Now consider the function $h(\xi) = \xi - \mathbb{E}(\tilde{Z}(\xi))^+$ It is easy to verify that $h(\cdot)$ is continuous, increasing, and that $h(0) < 0$ and $h(\infty) > 0$. This implies that the equation $h(\xi) = 0$ has a unique solution ξ^*, which defines the equilibrium of the system analyzed in Section 4; i.e., the equilibrium expected congestion cost for BE users is $\mathbb{E}D^* = \frac{\xi^*}{\kappa_2}$. Given (ii) above we get that for $\xi = \mathbb{E}(Z^d)^+$, $h(\mathbb{E}(Z^d)^+) < 0$, which by the monotonicity property of $h(\cdot)$ implies that $\mathbb{E}(Z^d)^+ < \mathbb{E}(Z)^+$, and, in turn, that $\mathbb{E}D^d < \mathbb{E}D$.

Step 1: Consider the fictitious system that announces to arriving BE-users the congestion estimate $\xi/\kappa_2\sqrt{n}$, for the particular choice $\xi = \mathbb{E}(Z^d)^+$. From (31) we get that

$$\xi = \mathbb{E}(Z^d)^+ = \sqrt{\frac{2}{\pi}}\frac{1}{\alpha + \sqrt{\alpha}},$$

and note that $\mathbb{E}D = \xi/\kappa_2$. Using the analysis of Section 4, we have that the corresponding limit system is one with parameter $\tilde{\gamma} = \alpha\xi = \sqrt{\frac{2}{\pi}}\frac{\alpha}{\alpha+\sqrt{\alpha}}$. For such a system,

$$\mathbb{E}(\tilde{Z}(\xi))^+ = \frac{\phi(\tilde{\gamma})}{\tilde{\gamma}(\tilde{\gamma}\Phi(\tilde{\gamma}) + \phi(\tilde{\gamma}))}.$$

Step 2: The goal is to show that $\mathbb{E}(\tilde{Z}(\xi))^+ < \xi$, when $\xi = \mathbb{E}(Z^d)^+$. For any fixed $\alpha > 0$, define

$$g(\alpha) := \frac{\mathbb{E}(\tilde{Z}(\xi))^+}{\xi} = \delta(\tilde{\gamma})\frac{\pi}{2}(1 + \sqrt{\alpha})^2 \quad \text{where} \quad \xi = \mathbb{E}(Z^d)^+,$$

and $\delta(\tilde{\gamma}) = \frac{\phi(\tilde{\gamma})}{\tilde{\gamma}\Phi(\tilde{\gamma})+\phi(\tilde{\gamma})}$. We wish to show that for all $\alpha > 0$, $g(\alpha) > 1$. Note that $g(\alpha)$ is continuous in α for all $\alpha > 0$. To establish that $g(\alpha) > 1$, it suffices to show that $\lim_{\alpha\downarrow 0} g(\alpha) > 1$, and that $g(\alpha)$ is monotonically increasing in α.

Note that as $\alpha \to 0$, $\tilde{\gamma} \to 0$, $\delta(\tilde{\gamma}) \to 1$ and $\lim_{\alpha\to 0} g(\alpha) = \frac{\pi}{2} > 1$. Also, as $\alpha \to \infty$, $\tilde{\gamma} \to \sqrt{\frac{2}{\pi}}$, $\delta(\tilde{\gamma}) \to \delta(\sqrt{\frac{2}{\pi}}) \in (0,1)$ and $\lim_{\alpha\to 0} g(\alpha) = \infty$. To complete the proof that $g(\alpha) > 1$ for all $\alpha > 0$, it suffices to show that $g'(\alpha) \geq 0$. It will be convenient to express $\sqrt{\alpha}$ as a function of $\tilde{\gamma}$, through $\sqrt{\alpha} = \frac{\tilde{\gamma}}{\sqrt{\frac{2}{\pi}-\tilde{\gamma}}}$, and rewrite all expressions in terms of $\tilde{\gamma}$. Specifically, with some abuse of notation we will analyze the function

$$g(\tilde{\gamma}) = \delta(\tilde{\gamma})\frac{\pi}{2}\frac{1}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})^2}.$$

Since $\frac{\partial\tilde{\gamma}}{\partial\alpha} > 0$, it suffices to show that $g(\tilde{\gamma})$ is increasing in $\tilde{\gamma}$. To that end we have that

$$g'(\tilde{\gamma}) = \frac{1}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})^2}\left[\delta'(\tilde{\gamma}) + \frac{2\delta(\tilde{\gamma})}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})}\right], \quad \text{where} \quad \delta'(\tilde{\gamma}) = -\delta(\tilde{\gamma})\left(\tilde{\gamma} + \frac{\Phi(\tilde{\gamma})}{\tilde{\gamma}\Phi(\tilde{\gamma}) + \phi(\tilde{\gamma})}\right).$$

Grouping terms we get that

$$g'(\tilde{\gamma}) = \frac{\delta(\tilde{\gamma})}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})^2}\underbrace{\left[\frac{2}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})} - \tilde{\gamma} - \frac{\Phi(\tilde{\gamma})}{\tilde{\gamma}\Phi(\tilde{\gamma}) + \phi(\tilde{\gamma})}\right]}_{:=f(\tilde{\gamma})}.$$

To conclude that $g'(\tilde{\gamma}) \geq 0$ it suffices to show that $f(\tilde{\gamma}) \geq 0$ for all $\tilde{\gamma} \in [0, \sqrt{\frac{2}{\pi}})$. Note that $f(0) = 0$ and that $\lim_{\tilde{\gamma}\to\sqrt{\frac{2}{\pi}}} f(\tilde{\gamma}) = \infty$. Finally, straightforward calculations gives

$$f'(\tilde{\gamma}) = \frac{2}{(\sqrt{\frac{2}{\pi}} - \tilde{\gamma})^2} - 1 - \delta(\tilde{\gamma}) + \frac{\Phi^2(\tilde{\gamma})}{(\tilde{\gamma}\Phi(\tilde{\gamma}) + \phi(\tilde{\gamma}))^2} \geq 0.$$

This implies that $f(\tilde{\gamma}) \geq 0$, and thus $g'(\tilde{\gamma}) \geq 0$, for all $\tilde{\gamma} \in [0, \sqrt{\frac{2}{\pi}})$. It follows that $g(\alpha) \geq \frac{\pi}{2}$ for all $\alpha > 0$, which completes the proof of step 2. ■

Proof of Lemma 2: First, note that by Lemma 1, the arrival rate into the system due to class 1 customers that are admitted is $\lambda_1^n = n\kappa_1\mu + o(\sqrt{n})$ (since the blocking effects are lower order than $1/\sqrt{n}$). Setting $\mu_1 = \mu_2 = \mu$, we have by Theorem 4 that $\sqrt{n}D^n(\cdot) \Rightarrow (X_1(\cdot) + X_2(\cdot))^+$, where $X(\cdot) = (X_1(\cdot), X_2(\cdot))$ is the 2-dimensional diffusion process identified in Theorem 4. Moreover, $(X(t) : t \geq 0)$ admits a unique stationary distribution. Let $X_i^n := (Q_i^n - \kappa_1 n)/\sqrt{n}$ for $i = 1, 2$, where $X^n = (X_1^n, X_2^n)$ has the stationary distribution in the nth system in the sequence. (The existence and uniqueness of this distribution is established in Proposition 2.) We will next establish an "interchange argument" which concludes that $\sqrt{n}D^n \Rightarrow \kappa_2^{-1}(X_1 + X_2)^+$, where $D^n := D^n(\infty)$ and $X_i^n := X_i^n(\infty)$. We then prove that $\{\sqrt{n}D^n\}$ is uniformly integrable, from which it follows that $\sqrt{n}\mathbb{E}D^n \to \kappa_2^{-1}\mathbb{E}(X_1 + X_2)^+$. The latter is then seen to be equal to $d(\gamma)$.

Step 1. We first prove that $X^n \Rightarrow X$, where X is equal in distribution to $X(\infty)$, the stationary marginal of the limiting diffusion $(X(t) : t \geq 0)$. From the proof of Theorem 1 in Maglaras and

Zeevi (2004) and Lemma 11.2.2 in Strook and Varadhan (1979) we have that the sequence of generators corresponding to the Markov processes $X^n(\cdot)$ converges uniformly on compact sets to the generator of $X(\cdot)$. Thus, appealing to Theorem 4.9.10 in Ethier and Kurtz (1986), we have that any weak limit of the sequence of stationary distributions corresponding to $(X^n(t) : t \geq 0)$ must be a stationary distribution of $(X(t) : t \geq 0)$. But since the limit process has a unique stationary distribution, all weak limit points must correspond to this distribution. Thus, all that is left is to establish that $\{X_n\}$ is tight, and therefore must have a subsequence that converges weakly (see, e.g., section 13 in chapter 3 of Billingsley (1968)).

Step 2. The Poisson limit theory that was used in Step 1 of the proof of Theorem 1 establishes that $\{X_1^n\}$ is tight. Now, for $\{X_2^n\}$, observe that $X_2^n = (X_1^n + X_2^n) - X_1^n$, thus, it suffices to show that $\{X_1^n + X_2^n\}$ is tight. To prove tightness, consider the following two systems. Let $\tilde{Q}_i^n(t)$ denote the number-in-system of class i users, in a system which is identical to the original one, with the exception that G-users wait in queue when there is no capacity available to serve them. Then, the proof of Theorem 1 establishes that $Q_i^n(t) \leq \tilde{Q}_i^n(t)$, for $i = 1, 2$, all $n \geq 1$ and all $t \geq 0$, almost surely. Let $\hat{Q}_i^n(t)$ denote the number-in-system of class i users, in a system which is identical to the original one, only here BE-users are blocked when the total number-in-system from both classes exceeds capacity, i.e., when $\hat{Q}_1^n + \hat{Q}_2^n \geq n$. For this system we have $\hat{Q}_i^n(t) \leq Q_i^n(t)$, for $i = 1, 2$, all $n \geq 1$ and all $t \geq 0$, almost surely. Consequently, we have that

$$\frac{(\hat{Q}_1^n + \hat{Q}_2^n - n)}{\sqrt{n}} \leq \frac{(Q_1^n + Q_2^n - n)}{\sqrt{n}} \leq \frac{(\tilde{Q}_1^n + \tilde{Q}_2^n - n)}{\sqrt{n}} \; .$$

Since $(\tilde{Q}_1^n(\cdot) + \tilde{Q}_2^n(\cdot))$ has the dynamics of the number-in-system in an $M/M/n$ queue, it follows from the results of Halfin and Whitt (1981, Lemma 1) that $\mathbb{E}|(\tilde{Q}_1^n + \tilde{Q}_2^n - n)/\sqrt{n}|^4$ is bounded uniformly in n when the arrival rate is such that $\sqrt{n}(1 - \rho^n) \to \gamma > 0$. Thus, the upper bound is uniformly integrable, and hence tight. Now, $(\hat{Q}_1^n(\cdot) + \hat{Q}_2^n(\cdot))$ has the dynamics of the number-in-system in an $M/M/n/n$ queue with arrival rate such that $\sqrt{n}(1 - \rho^n) \to \gamma > 0$. Then, the same argument used for X_1^n applies here as well, and we conclude that $\{(\hat{Q}_1^n + \hat{Q}_2^n - n)/\sqrt{n}\}$ is tight. This establishes that $\{X_1^n + X_2^n\}$ is tight, and thus $\{X^n\}$ is tight as well. Finally, we conclude that $X^n \Rightarrow X$, where $X := X(\infty)$, and the specification of $(X(t) : t \geq 0)$ is given in Theorem 4. Thus, by the continuous mapping theorem we have that $\sqrt{n}D^n \Rightarrow \kappa_2^{-1}(X_1 + X_2)^+$.

Step 3. To prove that $\{\sqrt{n}D^n\}$ is uniformly integrable, it suffices to show that $\sup_n \mathbb{E}|\sqrt{n}D^n|^2 < \infty$. To this end, note that

$$\begin{aligned} \mathbb{E}|\sqrt{n}D^n|^2 &= \mathbb{E}\left[\left(\frac{(Q_1^n + Q_2^n - n)^+}{\sqrt{n}}\right)^2 \frac{n^2}{[(Q_1^n - n) \vee 1]^2}\right] \\ &\leq \left(\mathbb{E}\left[\frac{(\tilde{Q}_1^n + \tilde{Q}_2^n - n)^+}{\sqrt{n}}\right]^4\right)^{1/4} \left(\mathbb{E}\left[\frac{n^4}{[(Q_1^n - n) \vee 1]^4}\right]\right)^{1/4} \end{aligned}$$

which follows from the Cauchy-Schwartz inequality and the bounding system described in Step 2. As noted above, $\mathbb{E}|(\tilde{Q}_1^n + \tilde{Q}_2^n - n)/\sqrt{n}|^4$ is bounded uniformly in n when the arrival rate is such that $\sqrt{n}(1-\rho^n) \to \gamma > 0$. We now turn to the second term on the right-hand-side. Fix $\epsilon > 0$ such that $\kappa_1 + \epsilon < 1$ (this is feasible since $\kappa_1 < 1$). Then,

$$\begin{aligned}
\mathbb{E}\left[\frac{n^4}{[(Q_1^n - n) \vee 1]^4}\right] &= \sum_{j=0}^{n} \frac{n^4}{[(j-n) \vee 1]^4} \mathbb{P}(Q_1^n = j) \\
&= \sum_{j=0}^{\lfloor(\kappa_1+\epsilon)n\rfloor} \frac{n^4}{[(j-n) \vee 1]^4} \mathbb{P}(Q_1^n = j) + \sum_{j=\lfloor(\kappa_1+\epsilon)n\rfloor+1}^{n} \frac{n^4}{[(j-n) \vee 1]^4} \mathbb{P}(Q_1^n = j) \\
&\leq C_1 \mathbb{P}(Q_1^n \leq \lfloor(\kappa_1+\epsilon)n\rfloor) + C_2 n^4 \mathbb{P}(Q_1^n > (\kappa_1+\epsilon)n)
\end{aligned}$$

where the last step follows from the fact that the terms $\{n/(n-j)^4\}$ are bounded by a constant in the first summation on the right-hand-side, and bounded by n^4 in the second summation on the right-hand-side. Now, by Lemma 1 in the proof of Theorem 1, we have that

$$\mathbb{P}(Q_1^n > (\kappa_1+\epsilon)n) \leq C_1 \exp(-C_2 n)$$

where the constants depend on ϵ. Thus,

$$\sup_n \left\{ n^4 \mathbb{P}(Q_1^n > (\kappa_1+\epsilon)n) \right\} < \infty \ .$$

Since $\{\sqrt{n}D^n\}$ is uniformly integrable, we have that $\sqrt{n}\mathbb{E}D^n \to \kappa_2^{-1}\mathbb{E}(X_1+X_2)^+$. But when $\mu_1 = \mu_2 = \mu$, $Z = X_1 + X_2$ has the simple stationary distribution identified in Theorem 1 of Halfin and Whitt (1981); see also Maglaras and Zeevi (2004). Specifically,

$$\begin{aligned}
\mathbb{P}(Z \leq z | Z \leq 0) &= \Phi(\gamma+z)/\Phi(\gamma) \quad z \leq 0 \\
\mathbb{P}(Z > z | Z > 0) &= \exp(-z\gamma) \qquad\quad z > 0.
\end{aligned}$$

This gives the expression for $d(\gamma)$ given in (11). Finally, $\mathbb{P}(Q_1^n + Q_2^n \geq n) = \mathbb{P}(X_1^n + X_2^n > 0)$ and the latter converges to $\mathbb{P}(Z > 0)$ which by the above is easily seen to be equal to $\kappa_2\gamma d(\gamma) =: \nu(\gamma)$. This concludes the proof. ■

References

Altman, E. and Kushner, H. (1999), 'Admission control for combined guaranteed performance and best effort communications systems under heavy traffic', *SIAM J. Control and Opt.* **37**, 1780–1807.

Altman, E., Orda, A. and Shimkin, N. (2000), 'Bandwidth allocation for guaranteed versus best effort service categories', *Queueing Systems* **36**, 89–105.

Armony, M. and Maglaras, C. (2004*a*), 'Contact centers with a call-back option and real-time delay information', *Oper. Res.* . To appear.

Armony, M. and Maglaras, C. (2004*b*), 'On customer contact centers with a call-back option: customer decisions, routing rules and system design', *Oper. Res.* . To appear.

Atar, R., Mandelbaum, A. and Reiman, M. (2002), 'Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy traffic'. Working paper, Technion, Israel.

Basar, T. and Srikant, R. (2002), 'Revenue-maximizing pricing and capacity expansion in a many-users regime', *In Proc. IEEE Infocom, New York, NY* .

Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley and Sons, Inc.

Browne, S. and Whitt, W. (1995), Piecewise-linear diffusion processes, *in* J. H. Dshalalow, ed., 'Advances in Queueing: Theory, Methods, and Open Problems', CRC Press, Inc., pp. 463–480.

Carpenter, B. E. and Nichols, K. (2002), 'Differentiated services in the internet', *Proc. IEEE* **90**, 1479–1494.

Das, A. and Srikant, R. (2000), 'Diffusion approximations for a single node accessed by congestion controlled sources', *IEEE Trans. Aut. Control* **45**, 1783–1799.

Davis, R. H. (1966), 'Waiting time distribution of a multi-server priority queueing system', *Oper. Res.* **14**, 133–136.

Ethier, S. N. and Kurtz, T. G. (1986), *Markov Processes: Characterization and Convergence*, Wiley, New York.

Fleming, P., Stolyar, A. and Simon, B. (1994), Heavy traffic limit for a mobile phone system loss model, *in* 'Proc. of 2^{nd} Int. Conf. on Telecomm. Syst. Mod. and Analysis, Nashville, TN'.

Gallego, G. and van Ryzin, G. (1994), 'Optimal dynamic pricing of inventories with stochastic demand over finite horizons', *Manag. Sci.* **40**, 999–1020.

Gans, N., Koole, G. and Mandelbaum, A. (2003), 'Telephone call centers: Tutorial, review, and research prospects', *Manufacturing & Service Operations Management* **5**, 79–141.

Garnett, O., Mandelbaum, A. and Reiman, M. (2002), 'Designing a call center with impatient customers', *Manufacturing & Service Operations Management* **4**, 208–227.

Gibbens, R. and Kelly, F. (1999), 'Resource pricing and the evolution of congestion control', *Automatica* **35**, 1969–1985.

Halfin, S. and Whitt, W. (1981), 'Heavy-traffic limits for queues with many exponential servers', *Oper. Res.* **29**, 567–588.

Harrison, J. and Zeevi, A. (2004), 'Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime', *Oper. Res.* . To appear.

Loynes, R. M. (1962), 'The stability of a queue with non-independent inter-arrival times and service times', *Proc. Cambridge Philos. Soc.* **58**, 497–520.

Maglaras, C. and Zeevi, A. (2003*a*), 'Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations', *Manag. Sci.* **49**, 1018–1038.

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

Maglaras, C. and Zeevi, A. (2004), 'Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels', *Math. Oper. Res.* . To appear.

McGill, J. and van Ryzin, G. (1999), 'Revenue management: reserach overview and prospects', *Transportation Science* **33**, 233–256.

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

Paschalidis, I. and Tsitsiklis, J. N. (2000), 'Congestion-dependent pricing of network services', *IEEE/ACM Trans. on Networking* **8**, 171–184.

Puhalskii, A. and Reiman, M. (2000), 'The multiclass GI/PH/N queue in the Halfin-Whitt regime', *Adv. Appl. Prob.* **32**, 564–595.

Savin, S., Cohen, M., Gans, N. and Katalan, Z. (2002), 'Capacity management in rental businesses with heterogeneous customer bases'. Working paper, Columbia University.

Strook, D. W. and Varadhan, S. R. S. (1979), *Multidimensional Diffusion Processes*, Springer-Verlag, New York.

Van Mieghem, J. (2000), 'Price and service discrimination in queueing systems: incentive compatability of G$c\mu$ scheduling', *Manag. Sci.* **46**, 1249–1267.

Whitt, W. (1992), 'Understanding the efficiency of multi-server service systems', *Manag. Sci.* **28**, 708–723.

Whitt, W. (1999), 'Improving service by informing customers about anticipated delays', *Manag. Sci.* **45**, 192–207.

Whitt, W. (2002), 'Stochastic models for the design and management of customer contact centers: Some research directions'. Working paper, Columbia University.

Whitt, W. (2004), 'How multiserver queues scale with growing congestion-dependent demand', *Oper. Res.* . To appear.

Williams, T. M. (1980), 'Nonpreemptive multi-server priority queues', *J. Opl. Res. Soc.* **31**, 1105–1107.