Diffusion Approximations for a Markovian Multi-Class Service System with "Guaranteed" and "Best-Effort" Service Levels

Constantinos Maglaras
Columbia University

Assaf Zeevi*
Columbia University

First version December 2002, last revision July 2003

To appear in *Mathematics of Operations Research*

Abstract

This paper considers a Markovian model of a service system motivated by communication and information services. The system has finite processing capacity and offers multiple grades of service. The highest priority users receive a "guaranteed" processing rate, while lower priority users *share* residual capacity according to their priority level and therefore may experience service degradation; hence the term "best effort." This paper focuses on performance analysis for this class of systems. We consider the Halfin-Whitt heavy-traffic regime where the arrival rate and system processing capacity both grow large in a way that the traffic intensity approaches one. We first derive a multi-dimensional diffusion approximation for the system dynamics, and subsequently obtain a more tractable diffusion limit based on an intuitive "perturbation approach." This method enables us to compute various closed form approximations to steady-state as well as transient congestion-related performance measures. Numerical examples illustrate the accuracy of these approximations.

Short Title: Diffusion approximations for "guaranteed" and "best-effort" services

Keywords: diffusion approximations, service systems, static priorities, shared resources, differentiated services, many server limits, Halfin-Whitt regime, parameter perturbations

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

1 Introduction

In recent years there has been an explosive growth in the usage of the Internet and web-based applications, including internet telephony, streaming audio, e-mail and information retrieval. 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 quality-of-service (QoS) requirements and willingness to pay. The canonical example of service differentiation in this context consists of *guaranteed-rate* and *best-effort* service grades. The former "guarantees" a rate to each such user, while the latter typically corresponds to users *sharing* the residual capacity not allocated to users with guaranteed-rate connections.

This paper introduces a simple and tractable model for service systems that offer differentiated services of this type, and derives a set of diffusion approximations for studying the system dynamics. Specifically, we consider a service provider that operates a finite set of processing resources and offers guaranteed-rate (G) and best-effort (BE) types of service. Once a guaranteed-rate user is accepted into the system, he/she will always receive their contracted nominal amount of capacity. (While this formulation can be viewed as an abstraction of the probabilistic service level guarantees that are used in practice, the asymptotic analysis pursued in the sequel is not "sensitive" to this distinction.) There are $m-1$ priority levels of best-effort service. Best-effort users of a given priority level are allocated capacity as follows: if their total rate requirement is less than the available capacity that is not used up by higher priority guaranteed and best-effort users they will each receive their nominal allocation; otherwise, they will *share* the remaining capacity in an egalitarian manner, and consequently experience service rate degradation. (When there is no remaining capacity in the system, the allocated rate is zero.) This capacity allocation mechanism can be viewed as a static priority policy that distinguishes between a high priority (G) and multiple low-priority (BE) classes of service. We assume that connection requests arrive according to independent Poisson processes, and the various user types may have different processing requirements, in the form of exponential service times with different rates.

The system dynamics can be described informally as follows. The number of guaranteed-rate users in the system is a "free" process that evolves according to the state of an $M/M/C/C$ system, where C is the system capacity. The number of best-effort users of a given priority level in the system, on the other hand, evolves as an $M/M/C(t)$ system with infinite waiting room; the available capacity at time t, $C(t)$, is stochastically modulated by the number of guaranteed and higher priority best-effort users present in the system. Natural measures of performance for this system are the blocking probability for guaranteed-rate users, and the congestion-related cost due to service rate degradation experienced by best-effort users. Despite its simple structure, exact analysis of this multi-class system is not straightforward and relies either on simulation or on numerical methods that offer little insight as to its structural behavior. In contrast, the goal of this paper is to develop approximations that support the derivation of closed form performance measures and provide some qualitative structural insights on the behavior of such systems. [Subsequent work builds on these approximations to address problems of economic optimization and optimal system design; see Maglaras and Zeevi (2003*b*).]

Our analysis is based on the asymptotic regime introduced by Halfin and Whitt (1981). This regime is defined by letting capacity grow large and concurrently letting the system utilization approach 100% at an appropriate rate. Specifically, Halfin and Whitt (1981) considered this asymptotic regime in the context of an $M/M/C$ queue as C grows large. By setting the traffic intensity parameter $\rho = 1 - \frac{\gamma}{\sqrt{C}}$ for some $\gamma > 0$, Halfin and Whitt (1981, Proposition 1) observed that the probability that an arriving customer experiences queueing delays is strictly less than 1. Thus, high utilization can go "hand-in-hand" with moderate levels of congestion. In this regime, Halfin and Whitt (1981, Theorem 1) established that the number of users in the system can be expressed roughly as $C + \sqrt{C}X(t)$ for C sufficiently large, where $X(t)$ is a simple one dimensional diffusion process. (This also implies that queueing delays are, asymptotically, of order $1/\sqrt{C}$.) The diffusion process $X(t)$ can then be used to obtain simple approximations for the system behavior, and congestion-related performance measures. Finally, the statistical economies of scale identified above suggest that the Halfin-Whitt regime may be a desirable operating point for large scale service systems.

The goal of this paper is to develop tractable approximations for a multi-class service system with guaranteed and best-effort service levels by studying its asymptotic behavior in the Halfin-Whitt regime. Broadly speaking, letting the system capacity and the arrival rates into each service class grow large such that the overall traffic intensity approaches one according to the Halfin-Whitt rate, we derive two different diffusion approximations for the system dynamics. The two approximations, briefly explained below, admit direct interpretations in terms of the original system model, and ultimately lead to a tractable characterization of its performance. Numerical experiments show that both approximations are very accurate when compared with the simulated behavior of the underlying Markovian system.

The main contributions of this paper are the following.

i.) *Heavy-traffic limits.* We first derive a multi-dimensional diffusion approximation for the system dynamics (Theorem 1), and use this result to characterize the congestion level experienced by the best-effort users of the various priority levels (Corollary 1). The limiting diffusion is shown to admit a unique stationary distribution under a natural stability condition (Proposition 2).

ii.) *An infinitesimal parameter perturbation approach.* The multi-dimensional diffusion limit identified in Theorem 1 is unfortunately not tractable from a performance analysis standpoint. However, it turns out that the analysis is greatly simplified if the service rates of the various user classes are identical. To exploit this observation, we consider the service rates as appropriate "small" perturbations around a common value (where the term "small" reflects the fact that these perturbations are asymptotically negligible). Specifically, we express the service rates (μ_i, $i = 1, \ldots, m$) as follows

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

where $\zeta_i \in \mathbb{R}$ are appropriate constants, C is the system capacity, and $\bar{\mu}$ is an appropriately selected common value, e.g., $\bar{\mu} = \max_i \mu_i$ or $\bar{\mu} = (\sum_i \mu_i)/m$.

iii.) *Tractable diffusion approximations.* We first derive diffusion limits in the Halfin-Whitt regime using the aforementioned infinitesimal perturbations (Theorem 2). Here we let C grow large, the traffic intensity approach one, and scale the service rates according to the relationship given above keeping $\bar{\mu}$ and ζ constant. The perturbation vector ζ captures the effects of the difference between the service rates, by appropriately modifying the drift term in the multi-dimensional diffusion limit. Numerical results illustrate the accuracy of this "perturbed diffusion" limit by comparing the resulting steady-state distribution to the original diffusion approximation identified in Theorem 1, and to the discrete stochastic system model (Figure 1).

iv.) *Congestion-related performance approximations.* Using the infinitesimal parameter perturbations, we show that the one-dimensional process that summarizes system congestion converges to a simple one-dimensional diffusion (Corollary 2). This limit process supports closed form computation of steady-state performance measures as it admits a simple steady-state distribution. This distribution is seen to be an accurate approximation of the actual system congestion in steady-state (Figures 2 and 3). Transient performance calculations identify the time scales at which the system reaches high levels of congestion (Proposition 3), as well as "recovery times" from these congested states.

v.) *An alternative perturbation approach.* Rather than applying a parameter perturbation on the pre-limit processes, one can use a similar idea directly to the diffusion limit derived in Theorem 1. This is explored in Section 7 via a change-of-measure argument that relies on Girsanov's transformation to derive transient system characteristics (Proposition 4).

The remainder of the paper is structured as follows. Sections 1 - 3 are introductory. Specifically, this section concludes with a short literature review, Section 2 formulates the stochastic system model, and Section 3 introduces the Halfin-Whitt limiting regime. Focusing on this regime, Section 4 derives a multi-dimensional diffusion limit and studies some of its basic characteristics. Section 5 contains the main contributions of the paper: it develops an alternative diffusion limit based on the perturbation argument highlighted above, and develops closed form characterizations of steady-state and transient performance measures associated with the system congestion. Section 6 discusses the actual use of these approximations and contains numerical results that illustrate their accuracy. Section 7 explores an alternative perturbation approach using a change-of-measure idea based on Girsanov's transformation of probability measures. Section 8 contains some brief concluding remarks.

The stylized model that we posit with capacity constraints and shared resources was first introduced by Das and Srikant (2000) to model best-effort type traffic. They derive diffusion approximations for this single class system in the Halfin-Whitt heavy traffic regime. Our work is strongly influenced by theirs, and seeks to extend the analysis to a multi-class setting [for related work of a different flavor see Bean, Gibbens and Zachary (1995)]. 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 system serving a single class of best-effort users [see also Basar and Srikant (2002) for a related study of different flavor.] The primary motivation to focus on guaranteed and best-effort service classes is driven by the communication and information services area [see, e.g., Carpenter and Nichols (2002), Gibbens and Kelly (1999), Odlyzko (1999) and references therein]. Similar systems arise in other contexts as well, for example, call-centers that process "VIP" and "regular" customers, and rental systems that serve customers with reservations as well as "walk-ins." In the former, users experience congestion by waiting in a queue until agents becomes available, while in the latter congestion appears in the form of *blocking* when there is no remaining capacity. Applications and extensions of the Halfin-Whitt results can be found in Whitt (1992), Fleming, Stolyar and Simon (1994), Garnett, Mandelbaum and Reiman (2002), Puhalskii and Reiman (2000), Borst, Mandelbaum and Reiman (2003), Armony and Maglaras (2003*b*), Armony and Maglaras (2003*a*), Whitt (2003), Atar, Mandelbaum and Reiman (2002) and Harrison and Zeevi (2004).

The interest in the Halfin-Whitt regime largely stems from its ability to succinctly capture the natural statistical economies scale that are present in many large capacity service system. In particular, Whitt (1992) and Garnett et al. (2002) argue that this regime is in some sense a desirable nominal operating point for such large scale service operations. In Maglaras and Zeevi (2003*a*) it is further shown, under an elastic demand assumption, that the Halfin-Whitt regime is optimal from an economic optimization standpoint.

Multi-class service systems in the Halfin-Whitt regime have been studied both in the context of performance analysis by Mandelbaum, Massey and Reiman (1998) and Puhalskii and Reiman (2000). Dynamic control formulations are pursued by Harrison and Zeevi (2004) and Atar et al. (2002). Our work is more closely related to the former set of references. Specifically, Mandelbaum et al. (1998) derive limiting diffusion models for service networks with time varying parameters including multi-class nodes under preemptive priority rules; the limiting regime considered in Mandelbaum et al. (1998) is related yet not identical to the one studied by Halfin and Whitt. Puhalskii and Reiman (2000) derive diffusion approximations for a multi-class queue with phase type service time distributions and non-preemptive priorities operating in the Halfin-Whitt heavy-traffic regime. These papers differ from ours in several regards. First, we start with a different formulation of the original system model. Second, our main goal is to develop *tractable* performance approximations for the system in question, dealing explicitly with the complication that arises from different service rates via the aforementioned perturbation approach. Finally, in terms of proof techniques, this paper makes use of "classical" weak convergence arguments, whereas the proofs in Mandelbaum et al. (1998) and Puhalskii and Reiman (2000) make use of strong approximations and martingale methods, respectively, due to the different objective, modelling framework and probabilistic assumptions invoked in those papers.

2 Model Formulation

Our system model attempts to capture four important features of the physical system: finite capacity, absence of resource pooling when the system is under-utilized, differentiated service levels, and the capability to share processing resources among best-effort users. These features will be made explicit in the sequel. Because realistic models of a processor sharing discipline can be quite complicated to analyze, we purse a simpler and more tractable stylized formulation. Specifically, we consider a system with m *service grades* (or *classes*), with class 1 denoting the guaranteed-rate or high-priority class, and classes $2, \ldots, m$ being the different best-effort classes that are labelled according to their priority level, i.e., class i has higher priority than all classes $j > i$, $i, j = 2, \ldots, m$. In the sequel, various quantities will be tagged with subscripts $1, \ldots, m$ to denote their association with a respective class of service.

The guaranteed-rate users are assumed to arrive according to a Poisson process with rate λ_1. Requests for guaranteed-rate service engage *one unit of capacity* each for i.i.d. exponentially distributed amounts of time with rate μ_1, provided that the total number of guaranteed users currently connected is less than the system capacity C; otherwise, they are denied service. (Without loss of generality, we take C to be an integer corresponding to the number of *nominal processing resources.*) The number of guaranteed-rate (G) users in the system at time t will be denoted by $Q_1(t)$. The process $Q_1 = (Q_1(t) : t \geq 0)$ has the same dynamics as the number-in-the-system process in an $M/M/C/C$ system, i.e., an $M/M/C$ queue with no waiting room.

The best-effort users of class i ($i = 2, \ldots, m$) are assumed to arrive according to a Poisson process with rate λ_i, independent of the arrival process of the guaranteed users and of other best-effort (BE) users. Requests for best-effort service are always admitted into the system. When there is sufficient capacity in the system not used up by the guaranteed-rate and higher priority best-effort users (that correspond to classes $j < i$), class i BE-users are allocated a nominal processing rate corresponding to one unit of capacity, and when capacity is not sufficient to support the nominal rate allocation, the BE-users share the available processing capacity in an egalitarian manner. The latter leads to congestion in the form of degradation of the processing rate allocated to them. Specifically,

$$\text{BE class } i \text{ service rate at time } t \;=\; \frac{(C - \sum_{j<i} Q_j(t))^+}{Q_i(t)} \wedge 1. \tag{1}$$

where $Q_i(t)$ denotes the number of class i best-effort users in the system at time t. Here and in the sequel we set $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$, and let $x \wedge y := \min(x, y)$ and $x \vee y := \max(x, y)$. The service rate degradation will lead to an *excess delay* encountered by BE-users that is proportional to

$$D_i(t) = \left(\frac{Q_i(t)}{C - \sum_{j<i} Q_j(t)} - 1 \right)^+ .$$

That is, this quantity is proportional to the delay encountered by a BE-user in excess of the processing time under nominal allocation. Here "nominal" refers to the case where the user is allocated one full unit of capacity, i.e., he/she suffers no congestion due to sharing, and the associated processing time for such a user is an i.i.d. exponentially distributed random variable with rate μ_i.

Despite the processor sharing characteristic of the BE-class, the dynamics of the process $Q_i = (Q_i(t) : t \geq 0)$ are essentially that of an $M/M/C_i(t)/\infty$ system, where the capacity $C_i(t) := (C - \sum_{j<i} Q_j(t))^+$ is a stochastic process modulated by the number of higher priority users in the system; this can be verified by a quick inspection of the birth-death rates of the associated Markov Chain. (When $C_i(t) = 0$, we assume that BE-users temporarily do not receive service but remain connected to the system.) This *stochastic capacity* interpretation of the BE dynamics is interesting in its own right, and manifests itself in other multi-class service systems such as call centers and rental (loss) systems.

Our focus will be on performance analysis of the Continuous Time Markov Chain (CTMC) $Q = (Q_1, \ldots, Q_m)$. As it turns out, the key performance driver in this G/BE multi-class service system is the congestion suffered by the low priority classes. With this in mind our emphasis will be on highlighting the natural scaling relations that prevail in such systems and characterizing the congestion processes $(D_i(t) : t \geq 0)$.

3 The Halfin-Whitt Asymptotic Regime

Let us consider now a sequence of models, each having the structure described in section 2, indexed by $n = 1, 2, \ldots$, attaching a superscript n to the notation established previously in order to indicate the dependence of a parameter or process on the capacity of the system $C^n = n$. (Accordingly, the absence of such a superscript shows that the quantity in question is independent of n.)

The analysis that we pursue considers a particular heavy-traffic regime where capacity grows large and utilization approaches one. The theoretical foundations of this heavy-traffic regime were laid out in the seminal paper of Halfin and Whitt (1981). They considered an $M/M/n$ queue with infinite buffer and a single class of arriving customers requiring service at rate μ. Letting the arrival rate λ^n increase with n so that $\rho^n = \lambda^n/(n\mu) \to 1$, they characterized the limiting behavior of a sequence of normalized state processes $X^n(t) = (Q^n(t) - n)/\sqrt{n}$ and set $X^n = (X^n(t); t \geq 0)$, where Q^n here is the number of customers in the system. In what follows, '$\Rightarrow$' denotes weak convergence in the space $D[0,\infty)$, or the associated product space $D^m[0,\infty)$, endowed with the usual Skorohod topology [see, e.g., Billingsley (1999)]. The following result is due to Halfin and Whitt (1981, Theorem 2) [the application of this result to a system with shared resources is discussed in Maglaras and Zeevi (2003*a*)].

Proposition 1 [Halfin and Whitt (1981)] *If $\sqrt{n}(1-\rho^n) \to \gamma > 0$, and $X^n(0) \Rightarrow \xi \in \mathbb{R}$, then $X^n \Rightarrow X$, where the limit X is a diffusion process with infinitesimal drift function*

$$b(x) = \begin{cases} -\mu x - \mu\gamma & x \leq 0 \\ -\mu\gamma & x > 0 \end{cases},$$

and infinitesimal variance (or diffusion coefficient) given by $\sigma^2 = 2\mu$.

The limiting diffusion X in Proposition 1 is essentially obtained by "pasting together" an Ornstein-Uhlenbeck (O-U) process and a negative-drift Brownian motion; for further details on diffusions with piecewise linear drift coefficients the reader is referred to Browne and Whitt (1995). The fact that the properly centered and scaled occupancy process X^n has a weak limit as stated in Proposition 1 has many important consequences, and gives rise to several insights that have been pointed out in Section 1 and will be discussed in more detail in the sequel as well.

In the following section we seek to extend Proposition 1 to the multi-class system described in Section 2, i.e., to derive diffusion approximations for the system with guaranteed and best-effort users. We note that Halfin and Whitt extended their result to allow for a renewal arrival process. While such an extension is also possible for our model, for brevity it will not be pursued in this paper. To this end, recall that we denote the guaranteed users as class 1, and the best effort users as classes $2,\ldots,m$. It will be convenient for future purposes to define some additional notation. First, let

$$\kappa_i = \lim_{n\to\infty} \frac{(\lambda_i^n/\mu_i)}{\sum_{j=1}^m (\lambda_j^n/\mu_j)} \qquad \text{for } i = 1,\ldots,m. \tag{2}$$

The sum appearing in the denominator on the right side of (2) represents the total workload input rate to the system (that is, average time units of work arriving per time unit), thus κ_i is the fraction of that input attributed to class i. Following Harrison and Zeevi (2004) we hereafter refer to $\kappa = (\kappa_1,\ldots,\kappa_m)$ as the vector of *relative workload contributions.* By definition $\kappa_1 + \cdots + \kappa_m = 1$, and we assume that $\kappa_1,\ldots,\kappa_m > 0$ (otherwise some of the service classes could be omitted from the analysis). We define as usual the system's *traffic intensity parameter*

$$\rho^n = \frac{1}{n}\sum_{j=1}^m \frac{\lambda_j^n}{\mu_j} \quad . \tag{3}$$

For simplicity, we shall vary only the arrival rate parameters into each class as the capacity n increases. Following the seminal work of Halfin and Whitt (1981), we want to do this in such a way that $\sqrt{n}(1-\rho^n) \to \gamma$ as $n \to \infty$, where γ is a positive real number representing the system's "excess capacity," in a suitable asymptotic sense. Given that both the average service rates μ_i and the asymptotic relative workload contributions κ_i are fixed and given, our "heavy-traffic" assumption posits that the arrival rates into each service class are given by

$$\lambda_i^n = \kappa_i \mu_i n - \gamma_i \mu_i \sqrt{n} \tag{4}$$

for $i = 1, \ldots, m$, $n = 1, 2, \ldots$, and appropriate constants γ_i. (That is, the system's processing rate matches demand up to "small," $\sqrt{n}$, perturbations.) Using the above together with (3) we have that

$$\rho^n = 1 - \gamma/\sqrt{n} \tag{5}$$

for all $n = 1, 2, \ldots$, where $\gamma := \gamma_1 + \ldots + \gamma_m$. (In the sequel we will impose that $\gamma > 0$ to ensure stability.)

4 Diffusion Approximations

This section develops an asymptotic approximation for the dynamics of the multi-class system introduced in Section 2, appealing to the many-server heavy-traffic limits discussed in the previous section. What we prove in the sequel (Section 4.1) is an analogue to the Halfin-Whitt result (Proposition 1) for our m-class system. This result will consist of a diffusion approximation to the underlying CTMC that describes the dynamics of the service system with shared resources and guaranteed and best-effort service grades. Specifically, we prove that a properly normalized version of the CTMC converges weakly to a certain multi-dimensional diffusion process. This process consists of $m-1$ O-U processes (that describe the dynamics of the guaranteed and $m-2$ highest priority best-effort classes in diffusion scale), and another process whose drift depends on the state of the aforementioned O-U processes (this process describes the dynamics of the lowest-priority best-effort class, in diffusion scale). Subsequently, in Section 4.2, we specify a parameter condition that ensures that this approximating diffusion admits a stationary distribution, which is crucial for a steady-state analysis.

4.1 Heavy traffic limits

Formulation and main result. Prior to stating our main result which considers heavy-traffic limits for the Markovian model introduced in Section 2, we first provide a formal (non-rigorous) derivation for a system with 2 classes, namely a guaranteed-rate service level and a single BE one. This two-class case reveals most of the underlying intuition.

Suppose the system is in state q at some point in time, i.e., $Q(t) = q$, where $q = (q_1, q_2)$. Then, q_1 users of class 1 are "in service," each user is allocated a single unit of capacity, and q_2 users of class 2 are "in service," and each such user receives $((n - q_1)/q_2) \wedge 1$ units of capacity (this allocation may be equal to zero, in which case the users temporarily experience no service). With the Markovian dynamics described in Section 2, the probability of a class 1 service completion in the next t time units is $q_1\mu_1 t + o(t)$ for small t, where $f(t) = o(t)$ if $f(t)/t \to 0$ as $t \to 0$. The corresponding probability of class 2 service completion is $((n - q_1) \wedge q_2)\mu_2 t + o(t)$. Similarly, the probability of a new class i connection arriving is $\lambda_i t + o(t)$ for small t. To this end, if a user of class 1 arrives in state $q_1 = n$, then he/she are denied service (i.e., $\lambda_1 = 0$ in that state). These transition intensities completely specify the probabilistic structure of the continuous time Markov chain (CTMC) that describes the dynamics of the system, and lives on the state space

$$S^n = \{(q_1^n, q_2^n) : q_1^n \in \{0, 1, \ldots, n\},\ q_2^n \in \{0, 1, \ldots\}\} \ . \tag{6}$$

Given the transition intensities described above, we have that for any initial state $q^n = (q_1^n, q_2^n) \in S^n$ and $t > 0$, the infinitesimal *drift rates* for each class are given by

$$\begin{aligned} \mathbb{E}\left[Q_1^n(t) - Q_1^n(0) \middle|\ Q^n(0) = q^n\right] &= \left[\lambda_1^n \mathbb{I}_{\{q_1^n < n\}} - \mu_1 q_1^n\right] t + o(t) \\ \mathbb{E}\left[Q_2^n(t) - Q_2^n(0) \middle|\ Q^n(0) = q^n\right] &= \left[\lambda_2^n - \mu_2\left((n - q_1^n) \wedge q_2^n\right)\right] t + o(t) \quad , \end{aligned} \tag{7}$$

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

$$\begin{aligned} \mathbb{E}\left[(Q_1^n(t) - Q_1^n(0))^2 \mid Q^n(0) = q^n\right] &= \left[\lambda_1^n \mathbb{I}_{\{q_1^n < n\}} + \mu_1 q_1^n\right] t + o(t) \\ \mathbb{E}\left[(Q_2^n(t) - Q_1^n(0))^2 \mid Q^n(0) = q^n\right] &= \left[\lambda_2^n + \mu_2\left((n - q_1^n) \wedge q_2^n\right)\right] t + o(t) \ . \end{aligned} \tag{8}$$

Finally,

$$\mathbb{E}\left[\left(Q_1^n(t) - Q_1^n(0)\right)\left(Q_2^n(t) - Q_2^n(0)\right) \ \mid Q^n(0) = q^n\right] = o(t) \quad \text{for all } n = 1, 2, \ldots. \tag{9}$$

Now, define the normalized state processes

$$X_i^n(t) := \frac{Q_i^n(t) - \kappa_i n}{\sqrt{n}} \quad \text{for } i = 1, \ldots, m \tag{10}$$

and set $X_i^n = (X_i^n(t) : t \geq 0)$ and $X^n = (X_1^n, \ldots, X_m^n)$. The infinitesimal rates described above in (7)-(9) together with the scaling assumption embodied in (4) suggest that the normalized state process X^n should converge to a limiting diffusion process. For example, for the process X_1 the drift rates above and the heavy-traffic assumption suggest that for $x_1 \in \mathbb{R}$ and n sufficiently large,

$$\begin{aligned} \mathbb{E}\left[X_1^n(t) - X_1^n(0) \mid X_1^n(0) = x_1\right] &= \frac{(\kappa_1 n\mu_1 - \gamma_1\sqrt{n}\mu_1)t - \mu_1(\kappa_1 n + x_1\sqrt{n})t}{\sqrt{n}} + o(t/\sqrt{n}) \\ &= -\mu_1(\gamma_1 + x_1)t + o(t/\sqrt{n}) \end{aligned}$$

for sufficiently small t. (Note that by construction $\kappa_1 n + \sqrt{n}x_1 < n$, which will guarantee that at least for small time intervals the boundary $q_1^n = n$ is inaccessible.) This, in turn, suggests that the limiting process for X_1^n, representing the dynamics of the guaranteed-rate users in diffusion scale, will be given by an O-U process, while the dynamics of the best-effort class, given by the limit of X_2^n, will have a piece-wise linear drift which depends on the state of the aforementioned O-U process. The following result provides a rigorous justification for this suggested asymptotic behavior for the m class case.

Consider a sequence of systems with capacity $C^n = n$ and arrival rates $\lambda_i^n = n\kappa_i\mu_i - \gamma_i\sqrt{n}\mu_i$ for some constants $\kappa_i > 0$ and $\gamma_i \in \mathbb{R}$ for $i = 1, \ldots, m$, such that $\sum_{i=1}^m \kappa_i = 1$.

Theorem 1 *Suppose that for some $\xi \in \mathbb{R}^m$, $Q_i^n(0) = \lfloor n\kappa_i + \sqrt{n}\xi_i \rfloor$ for $i = 1, \ldots, m$. Then, $X^n \Rightarrow X$ in $D^m[0,\infty)$ as $n \to \infty$, where X is a diffusion process. Specifically, X is the unique strong solution of the following stochastic differential equation:*

$$dX(t) = b(X(t))dt + \Sigma dW(t), \quad X(0) = \xi \quad , \tag{11}$$

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

$$\begin{aligned} b_i(x) &= -\mu_i\gamma_i - \mu_i x_i & & i = 1, \ldots, m-1 \\ b_m(x) &= \begin{cases} -\mu_m\gamma_m - \mu_m x_m & \sum_{i=1}^m x_i \leq 0 \\ -\mu_m\gamma_m + \mu_m \sum_{i=1}^{m-1} x_i & \sum_{i=1}^m x_i > 0 \end{cases} & & , \end{aligned} \tag{12}$$

and $\Sigma := \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$, with $\sigma_i^2 = 2\mu_i\kappa_i$.

Discussion and implications. To recapitulate, the heavy-traffic limits for the normalized number of users connected to the system in the $m-1$ high-priority classes are given by one dimensional O-U processes, while the respective limit for the lowest-priority BE-users is given by a more complicated diffusion whose drift function depends on all state variables. In particular, its drift is "modulated" by the aforementioned O-U processes. For a system with capacity $C^n = n$, Theorem 1 suggests the approximation

$$Q_i^n(t) \approx \kappa_i C^n + \sqrt{C^n} X_i(t) \tag{13}$$

for $i = 1, \ldots, m$, where the rigorous meaning of this approximation is captured in the above limit theorem, namely, $(Q_i^n(\cdot) - \kappa_i n)/\sqrt{n} \Rightarrow X_i(\cdot)$, as $n \to \infty$. These observations give rise to several important insights. First, the excess number of users in the system is small relative to the capacity. Specifically,

$$[Q_1^n(t) + \cdots + Q_m^n(t)] - C^n = \mathcal{O}_p(\sqrt{C^n}),$$

where a sequence of random variables $\{Y^n\}$ is $\mathcal{O}_p(a^n)$ if Y^n/a^n is bounded in probability. Second, the blocking probability experienced by the G-users is of the form $\mathbb{P}(Q_1^n = n) \approx e^{-cn}$ for some appropriate $c > 0$, which is asymptotically negligible. Third, the congestion manifested as excess delay experienced by the best-effort users in classes $i = 2, \ldots, m-1$ is negligible $\sqrt{n}D_i^n(t) \Rightarrow 0$, while for the lowest-priority class

$$D_m^n(t) = \mathcal{O}_p\left(\frac{1}{\sqrt{C^n}}\right).$$

To see why this is true, observe that

$$\begin{aligned}
D_m^n(t) &= \frac{(\sum_{i=1}^m Q_i^n(t) - n)^+}{n - \sum_{j<m} Q_j^n(t)} \\
&\approx \frac{\sqrt{n}\,(\sum_{i=1}^m X_i(t))^+}{\kappa_m n - \sqrt{n}\sum_{j<m} X_j(t)} \\
&= \frac{1}{\kappa_m\sqrt{n}}\left(\sum_{i=1}^m X_i(t)\right)^+\left(1 - \frac{\sum_{j<m} X_j(t)}{\kappa_m\sqrt{n}}\right)^{-1} \\
&\approx \frac{1}{\kappa_m\sqrt{n}}\left(\sum_{i=1}^m X_i(t)\right)^+\left(1 + \frac{\sum_{j<m} X_j(t)}{\kappa_m\sqrt{n}}\right) \\
&\approx \frac{1}{\kappa_m\sqrt{n}}\left(\sum_{i=1}^m X_i(t)\right)^+, \qquad (14)
\end{aligned}$$

where "$X_n \approx Y_n$" means $\sqrt{n}(X_n - Y_n)$ converges to zero in probability, as n goes to infinity. This derivation is made rigorous in the following corollary to Theorem 1. Let $D_i^n = (D_i^n(t) : t \geq 0)$ for $i = 2, \ldots, m$.

Corollary 1 *Under the conditions of Theorem 1, $\sqrt{n}D_i^n(t) \Rightarrow 0$ for $i = 2, \ldots, m-1$, and*

$$\sqrt{n}D_m^n \Rightarrow \frac{1}{\kappa_m}\left(\sum_{i=1}^m X_i\right)^+.$$

Thus, the key in analyzing the congestion in the system is the behavior of the "sum process" $X_1 + \cdots + X_m$, which represents the stochastic fluctuations of the total number of users connected to the system around the nominal level when viewed in diffusion scale.

Remark 1 (Alternative models of G/BE systems.) There are other alternative multi-class service models that one could have considered. One such alternative would allow for $m-1$ "guaranteed-rate" service classes and one best-effort class corresponding to the lowest priority class, where higher priority guaranteed-rate connections could "eject" lower priority guaranteed-rate connections if at some instance in time the entire capacity was used up by the guaranteed-rate classes $1, \ldots, m-1$. Yet another alternative would be to consider a system with m classes operating under a priority rule that allows users to queue if there is no available capacity upon their arrival. In the former, congestion arises in the form of blocking or ejections of guaranteed-rate users, and excess delay experienced by the lowest-priority best-effort users due to sharing of processing capacity (that leads to a degradation of their allotted service rate). In the latter, congestion is manifested as queueing delays. The results of Theorem 1 show that only the lowest priority class suffers congestion, and the same observation would hold true if one were to analyze the two alternative models described above. In particular, the probability of blocking or "ejections" of guaranteed-rate users is asymptotically negligible in the first model, and the queueing delays suffered by users of the $m-1$ high priority classes is negligible in the second. While the dynamics of these two models and those of the system model considered in this paper differ for each finite n, they all have the same asymptotic behavior.

4.2 Steady-state properties

For the approximations proposed in the previous section to be useful, we need to characterize the transient, and more importantly, steady-state behavior of the limiting multi-dimensional diffusion. While the first $m-1$ components of the process X are simple O-U processes and clearly admit a steady-state, the last component has more complicated structure and thus it is not clear a-priori under what conditions a steady-state will exist for X. The next proposition resolves this issue.

Proposition 2 *Let X be the unique strong solution to (11) with drift and infinitesimal variance specified in Theorem 1. Then, X is a positive recurrent diffusion, in particular, it admits a unique stationary distribution π if and only if $\gamma := \sum_{i=1}^{m} \gamma_i > 0$. Moreover, $X(t) \Rightarrow X(\infty)$ as $t \to \infty$, where $X(\infty)$ has distribution π.*

The typical approach to characterizing the stationary distribution π is either via the basic adjoint relation that asserts $\mathbb{E}_\pi[(\mathcal{A}f)(X(0))] = 0$ for all functions f that are twice continuously differentiable with compact support. Here $\mathcal{A}$ is the generator of X given in (32) in the appendix. This characterization of the stationary distribution is sometimes referred to as the *basic adjoint relation*; for further details see, e.g, Ethier and Kurtz (1986, §4) or Karatzas and Shreve (1991). An alternative approach which is less implicit characterizes the stationary distribution as the solution to $\mathcal{A}^*\pi = 0$, where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$ [see, e.g., Karatzas and Shreve (1991, p. 282)]. Unfortunately, the drift $b(\cdot)$ in our two-dimensional diffusion process is not sufficiently smooth to make use of the adjoint characterization of the stationary distribution. Moreover, the partial differential equation whose solution is the sought stationary distribution is typically intractable and does not lead to closed form solutions.

5 Tractable Diffusion Approximations via Infinitesimal Perturbation Asymptotics

The asymptotic regime described in Section 3 yields diffusion approximations for the original stochastic system. These asymptotics can be used to approximate the dynamics of a finite capacity system using (13) and, in particular, approximate the behavior of the excess delay D^n using (14). Unfortunately, the diffusion identified in Theorem 1 does not support closed form calculations of steady-state and transient quantities. This section proposes an approximation scheme that gives rise to slightly different diffusion limits which are more tractable than those established in Theorem 1.

5.1 Infinitesimal perturbation asymptotics

Basic formulation. The key observation is the following. If the service rates for the various grades of service would be equal, then the number of users in the system, given by $Q_1^n + \ldots + Q_m^n$, would have almost identical dynamics to the number-in-system in an $M/M/n$ queue. Thus, the essence of our proposed approach is to view the more complicated case with different service rates as an *infinitesimal perturbation* around the more tractable single-class case. The term "infinitesimal" refers to the fact that the suggested parameter perturbations are of second order, as will be explained shortly, and thus only affect the constant drift terms in the limiting diffusion.

To formalize this approach, we consider the following sequence of service rates

$$\mu_i^n = \bar{\mu}\left(1 - \frac{\zeta_i}{\sqrt{n}}\right), \tag{15}$$

for appropriate $\zeta_i \in \mathbb{R}$, where $\bar{\mu}$ is interpreted as the "nominal" rate of service, e.g., $\bar{\mu} = \max_i \mu_i$ or $\bar{\mu} = \frac{1}{m}\sum_i \mu_i$. The size of the perturbation coefficients ζ reflect the extent to which the actual service rates differ, relative to the square root of the system's capacity. Next, rewrite the arrival rates $\lambda_i = n\kappa_i\mu_i - \gamma_i\sqrt{n}\mu_i$ in the form

$$\lambda_i = n\kappa_i\bar{\mu} - \sqrt{n}(\bar{\gamma}_i + \kappa_i\zeta_i)\bar{\mu}, \tag{16}$$

by setting

$$\bar{\gamma}_i = \gamma_i\frac{\mu_i}{\bar{\mu}} \quad i = 1, \ldots, m. \tag{17}$$

In the next section we explain in more detail how one selects the parameters $\zeta, \bar{\gamma}$ starting from a system of original interest described by (C, μ, λ). That is, we have embedded the original system into the sequence of systems defined through (15), the limit of which is tractable since both $\mu_1^n, \ldots, \mu_m^n \to \bar{\mu}$ as $n \to \infty$, and where the difference of the original values $\mu_1, \ldots, \mu_m$ is captured via the $\bar{\gamma}_i$'s. (The reader will note that the relative workload contributions are calculated in "fluid scale" and are therefore unaffected by the infinitesimal perturbations in the service rates.) From (15)-(16) we obtain the traffic intensity

$$\begin{aligned}\rho^n &= \sum_i \frac{n\kappa_i\bar{\mu} - \sqrt{n}(\bar{\gamma}_i + \kappa_i\zeta_i)\bar{\mu}}{n\bar{\mu}\left(1 - \frac{\zeta_i}{\sqrt{n}}\right)} \\ &= 1 - \frac{\bar{\gamma}_1 + \cdots + \bar{\gamma}_m}{\sqrt{n}} + o(1/\sqrt{n})), \end{aligned} \qquad (18)$$

which can be rewritten in the form $\rho^n = 1 - \bar{\gamma}/\sqrt{n} + o(1/\sqrt{n})$ for $\bar{\gamma} = \bar{\gamma}_1 + \cdots + \bar{\gamma}_m = \sum_i \gamma_i \frac{\mu_i}{\bar{\mu}}$. In contrast, the original system with service rates μ_i and arrival rates λ_i had total traffic intensity $1-(\gamma_1+\cdots+\gamma_m)/\sqrt{n}$; i.e., the approximating scheme eventually reduces to scaling the γ_i's by $\mu_i/\bar{\mu}$, respectively.

Main results. Keeping $\zeta, \bar{\gamma}$ fixed, define a sequence of systems where $C^n = n$, μ_i^n and λ_i^n scale according to (15) and (16), respectively. In what follows we suppress the dependence of various quantities on the perturbation vector ζ to keep the notation more transparent, with the understanding that all processes superscripted by n are constructed by implicitly assuming the asymptotic (15) above. Specifically, let $Q_i^n(t)$, for $i = 1, \ldots, m$, denote the class i number of users in the system for the "perturbed" system and let $S^n(t) := Q_1^n(t) + \cdots + Q_m^n(t)$ denote the sum process. As before, set $Y_i^n(t) := (Q_i^n(t) - \kappa_i n)/\sqrt{n}$, for $i = 1, \ldots, m$, and let $Y_i^n = (Y_i^n(t) : t \geq 0)$. Our next result establishes that the normalized "perturbed" process $Y^n = (Y_1^n, \ldots, Y_m^n)$ converges weakly to a diffusion process that is closely related to the one given in Theorem 1. This limiting diffusion leads to a very simple approximation to the sum process $S^n = (S^n(t) : t \geq 0)$.

Theorem 2 *Fix $\zeta \in \mathbb{R}^m$, and let the service rates scale as in (15) and the arrival rates scale as in (16). Suppose that for some $\xi \in \mathbb{R}^m$, $Q_i^n(0) = \lfloor n\kappa_i + \sqrt{n}\xi_i \rfloor$ for $i = 1, \ldots, m$, and that (18) holds. Then, $Y^n \Rightarrow Y$ in $D^m[0, \infty)$ as $n \to \infty$, where Y is a diffusion process. Specifically, Y is the unique strong solution of the following stochastic differential equation:*

$$dY(t) = b(Y(t))dt + \Sigma dW(t), \quad Y(0) = \xi \quad ,$$

where $W = (W(t) : t \geq 0)$ is standard Brownian motion in $\mathbb{R}^m$. The infinitesimal drift function $b_i(\cdot)$ for the ith component is

$$\begin{aligned} b_i(y) &= -\bar{\mu}\bar{\gamma}_i - \bar{\mu}y_i & & i = 1, \ldots, m-1 \\ b_m(y) &= \begin{cases} -\bar{\mu}\bar{\gamma}_m - \bar{\mu}y_m & \sum_{i=1}^m y_i \leq 0 \\ -\bar{\mu}\bar{\gamma}_m + \bar{\mu}\sum_{i=1}^{m-1} y_i & \sum_{i=1}^m y_i > 0 \quad , \end{cases} \end{aligned} \qquad (20)$$

and $\Sigma := \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$, with $\sigma_i^2 = 2\bar{\mu}\kappa_i$, when $\bar{\mu}$ and $\bar{\gamma} = (\bar{\gamma}_1, \ldots, \bar{\gamma}_m)$ are defined in (15) and (17), respectively.

In terms of the normalized sum process, $Z^n(t) = (S^n(t) - n)/\sqrt{n}$, a direct application of the continuous mapping theorem yields

Corollary 2 *Under the assumptions of Theorem 2, $Z^n \Rightarrow Z$ in $D[0,\infty)$, where $Z = Y_1 + \cdots + Y_m$ is a diffusion process, with drift*

$$b(z) = \begin{cases} -\bar{\mu}(\bar{\gamma} + z) & z < 0 \\ -\bar{\mu}\bar{\gamma} & z \geq 0, \end{cases} \tag{21}$$

and diffusion coefficient $\sigma^2(z) = 2\bar{\mu}$, where $\bar{\gamma} = \sum_{i=1}^m \bar{\gamma}_i$.

The above diffusion limit has identical structure to the Halfin-Whitt diffusion given in Proposition 1. The goal of the next section is to derive approximations for system performance on the basis of this limiting diffusion.

5.2 Steady-state and transient performance analysis

Steady-state approximations. We start by repeating the observation that the probability of blocking guaranteed-rate users vanishes exponentially fast and the diffusion approximation does not involve any boundary-related terms. Assuming that $\gamma > 0$, and as a result $\bar{\gamma} > 0$ as well, we can appeal to Proposition 2 to conclude that Y is positive recurrent and admits a unique stationary distribution π. The key feature of Y is that the limiting diffusion $Z = Y_1 + \cdots + Y_m$ now admits a simple stationary distribution given by

$$\begin{aligned} \mathbb{P}(Z(\infty) > z | Z(\infty) > 0) &= e^{-z\bar{\gamma}} \quad z > 0 \\ \mathbb{P}(Z(\infty) \leq z | Z(\infty) \leq 0) &= \Phi(\bar{\gamma} + z)/\Phi(\bar{\gamma}) \quad z \leq 0, \end{aligned} \tag{22}$$

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function. For further details see, e.g, Halfin and Whitt (1981) and Browne and Whitt (1995) for a general discussion of diffusions with piece-wise linear drift functions. With this formulation, the probability that a best-effort user will find the system in "shared mode" is such that $\mathbb{P}(\text{congestion}) = \mathbb{P}(S^n(\infty) > n) \to \nu \in (0,1)$, where

$$\nu = [1 + \sqrt{2\pi}\bar{\gamma}\Phi(\bar{\gamma})e^{\bar{\gamma}^2/2}]^{-1} , \tag{23}$$

for further details see Halfin and Whitt (1981). Note that the latter statement implicitly assumes an interchange argument, namely, that one can interchange the diffusion limit, obtained by letting $n \to \infty$, and the limits w.r.t. to the time variable that give rise to the steady-state version by letting $t \to \infty$. While this interchange is usually difficult to establish, one can in fact show that it is valid for the G/BE system under consideration; see Maglaras and Zeevi (2003*b*, Lemma 2). The above discussion together with Corollary 1 suggest the following very simple approximation for the "excess delay" experienced by the best-effort users, viz,

$$\mathbb{E}D_m^n(\infty) \approx \frac{\mathbb{E}[Z(\infty)]^+}{\sqrt{n}} = \frac{[1 + \sqrt{2\pi}\bar{\gamma}\Phi(\bar{\gamma})e^{\bar{\gamma}^2/2}]^{-1}}{\sqrt{n}\bar{\gamma}} .$$

Transient approximations. We start our analysis by characterizing the first passage time of the system to a "congested state." (A particular case of interest is time elapsed until the number of users in the system exceeds capacity, which, in turn, corresponds to the point where best-effort users start receiving a degraded service rate.) Let $T_b = \inf\{t \geq 0 : Z(t) \geq b\}$, where $b \geq 0$, and let $\mathbb{E}_z[\cdot] := \mathbb{E}[\cdot|Z(0) = z]$. We then have the following result.

Proposition 3 *Fix $b \geq 0$ and $z \leq b$. Then, for $Z(t) = Y_1(t) + \cdots + X_m(t)$,*

$$\mathbb{E}_z T_b = \begin{cases} \left(\frac{\sqrt{2\pi}}{\bar{\mu}\bar{\gamma}} e^{\bar{\gamma}^2/2}\Phi(\bar{\gamma}) + \frac{1}{\bar{\mu}\bar{\gamma}^2}\right)\left(e^{\bar{\gamma}b} - e^{\bar{\gamma}z}\right) - \frac{(b-z)}{\bar{\mu}\bar{\gamma}} & z > 0 \\ \\ \frac{\sqrt{2\pi}}{\bar{\mu}} \int_z^0 e^{(\bar{\gamma}+x)^2/2}\Phi(x+\bar{\gamma})dx + \left(\frac{\sqrt{2\pi}}{\bar{\mu}\bar{\gamma}} e^{\bar{\gamma}^2/2}\Phi(\bar{\gamma}) + \frac{1}{\bar{\mu}\bar{\gamma}^2}\right) e^{\bar{\gamma}b} & \\ -\frac{\sqrt{2\pi}}{\bar{\mu}\bar{\gamma}} e^{\bar{\gamma}^2/2}\Phi(\bar{\gamma}) - \frac{(b\bar{\gamma}+1)}{\bar{\mu}\bar{\gamma}^2} & z \leq 0 \quad , \end{cases} \tag{24}$$

While the exact distribution of T_b is not tractable, we can use the above computation of its mean to produce an approximation that is valid for high congestion levels. In particular, we have that for large values of b,

$$T_b \approx C_1 e^{\bar{\gamma}b} V \tag{25}$$

where V is a random variable that follows an exponential distribution with mean 1, and C_1 is a constant independent of b that can be explicitly identified on the basis of (24). The rigorous meaning of "$\approx$" is captured by the limit theorem $C_1^{-1}e^{-\bar{\gamma}b}T_b \Rightarrow V$ as $b \to \infty$. To this end, the key observation is that Z is an ergodic one-dimensional diffusion process that has the origin, say, as a classical regeneration point (which follows by the strong Markov property.) Thus, the aforementioned limit theorem follows from standard limit theory for ergodic regenerative processes [see, e.g., Keilson (1966)].

Remark 2 This result suggests the following heuristic approximation for the original system with capacity $C^n = n$. Suppose that the system has initially q_i^n class i users connected to it, for $i = 1, \ldots, m$. Suppose the "congested" state of interest is $b^n > \sum_i q_i^n$. Put $T^n(b^n) = \inf\{t \geq 0 : \sum_i Q_i^n(t) \geq b^n\}$. Then, the suggested approximation to the first passage time in the system with capacity C^n is

$$\mathbb{E}_q T^n(b^n) \approx \mathbb{E}_{z^n} T_{\beta^n}$$

where $z^n = (\sum_i q_i^n - n)/\sqrt{n}$, $\beta^n = (b^n - n)/\sqrt{n}$ and T_b is the first passage time for Z defined above. Moreover, using (25) we can approximate the distribution of this passage time as follows

$$T^n(b^n) \approx C_1 \exp\left(\bar{\gamma}(b^n - n)/\sqrt{n}\right) V$$

which we anticipate to be valid for large values of capacity n and congestion levels b^n.

The above discussion has focused on the time scales that lead to high levels of congestion. In particular, the time it takes to reach a level b is roughly exponential in b, indicating that extreme congestion is not observed in relatively "short" to "moderate" time scales. A question of separate interest concerns the so-called *recovery time*, i.e., the time it takes the system to return to an uncongested state from a congested one. To this end, observe that the dynamics of the process Z are such that when $Z > 0$, the process has behavior identical to that of a negative drift Brownian motion. It therefore follows from standard arguments that

$$\mathbb{E}_z T_b = \frac{z-b}{\mu}$$

for $z \geq b \geq 0$. In particular, $\mathbb{E}_z T_0 = z/\mu$, and the recovery time is linear in the level of congestion z.

6 Numerical Study and Quality of the Proposed Approximations

This section reports on a set of numerical results that compare the steady state distributions of (a) the original G/BE system, (b) the multi-dimensional diffusion approximation given in Section 4, and (c) the approximation based on the perturbation approach of Section 5. These results illustrate the accuracy of the proposed diffusion approximations, as well as its dependence on the different model parameters such as the system capacity, the traffic intensity, and the difference between the service rate requirements for the various classes. For simplicity, the running example in this section restricts attention to the two-class system with a single guaranteed and best-effort class.

Determination of the appropriate diffusion model parameters. First, we outline how starting from the parameters of the original system one can compute the appropriate parameters $(\kappa_i, \gamma_i, \zeta_i)$ of the two diffusion models. As a running example consider a system with: $C = 100$, $\mu_1 = 1$, $\mu_2 = 2$, $\lambda_1 = 47.5$, and $\lambda_2 = 95$ (here $\rho = .95$).

Approximation based on two-dimensional diffusion of Section 4. The parameters κ_i and γ_i are computed through (2) and (4), respectively. In our example,

$$\kappa_i = \frac{\lambda_i/\mu_i}{\lambda_1/\mu_1 + \lambda_2/\mu_2} \quad \Rightarrow \quad \kappa_1 = 47.5/(47.5 + 95/2) = .5 \quad \text{and} \quad \kappa_2 = 1 - \kappa_1 = .5,$$

and

$$\gamma_i = \frac{\kappa_i \mu_i C - \lambda_i}{\mu_i \sqrt{C}} \quad \Rightarrow \quad \gamma_1 = \frac{.5 \cdot 1 \cdot 100 - 47.5}{\sqrt{100}} = .25 \quad \text{and} \quad \gamma_2 = \frac{.5 \cdot 2 \cdot 100 - 95}{2 \cdot \sqrt{100}} = .25.$$

Approximation based on the perturbation approach of Section 5. Set $\bar{\mu} = \max(\mu_1, \mu_2)$ and compute the ζ_i's using (15) as follows

$$\zeta_i = \sqrt{C}\left(1 - \frac{\mu_i}{\bar{\mu}}\right) \quad \Rightarrow \quad \zeta_1 = \sqrt{100}(1 - 1/2) = 5 \quad \text{and} \quad \zeta_2 = 0.$$

As in the previous approximation, the relative workload contributions are given by $\kappa_1 = \kappa_2 = .5$. The $\bar{\gamma}_i$'s are computed through (17) to give

$$\bar{\gamma}_1 = \gamma_1\mu_1/\bar{\mu} = .125 \quad \text{and} \quad \bar{\gamma}_2 = \gamma_2\mu_2/\bar{\mu} = .250.$$

Finally, $\bar{\gamma} = \sum_i \bar{\gamma}_i = 0.375$.

Diffusion based approximations. The results we present in this section contrast the steady-state distribution of the discrete system, obtained via simulation of the CTMC, with the proposed diffusion approximations. The distributions associated with the multi-dimensional diffusion derived in Section 4 were obtained by simulating the diffusion process. This is done by first discretizing over time, and then simulating the discrete time model that has appropriate dynamics and is driven by a Gaussian noise term. (In the sequel, whenever we say that a diffusion was simulated we will be referring to this procedure.) Note that the diffusion approximation for the number of guaranteed users in the system is an O-U process whose steady-state distribution has a simple Gaussian structure. Specifically, under this approximation $X_1(\infty) \sim N(-\gamma_1, \kappa_1)$ which suggests that

$$Q_1 \sim N(\kappa_1 C - \sqrt{C}\gamma_1, \kappa_1 C).$$

For the approximations obtained via the perturbation approach, the distributions of the guaranteed-rate users and of the total number of users in the system are known in closed form, whereas the corresponding distribution for the best-effort users was obtained by simulating the perturbed diffusions. Specifically, the distribution for the limiting number of guaranteed users is given by $Y_1(\infty) \sim N(-(\gamma_1' - \kappa_1\zeta_1), \kappa_1)$, which suggests the following approximation for the number of guaranteed-rate users:

$$Q_1' \sim N(\kappa_1 C - \sqrt{C}(\gamma_1' - \kappa_1\zeta_1), \kappa_1 C),$$

where the $'$ is used to denote that this approximation was based on the perturbation approach. The total number-of-users in the system is approximated through the distribution described subsequent to Theorem 2. In particular, this distribution is obtained by "pasting together" a truncated Gaussian and an exponential distribution. The Gaussian portion describes the distribution when the number of users falls short of the system capacity, and the exponential portion when the number of users exceeds capacity. Following (22) and (23) we have that the approximate distribution for the total number of users in the system is given by: $\mathbb{P}(Q_1' + Q_2' > C) = \nu$, $\mathbb{P}(Q_1' + Q_2' > C + z | Q_1' + Q_2' > C) = e^{-z\bar{\gamma}/\sqrt{C}}$, $z > 0$, and $\mathbb{P}(Q_1' + Q_2' \leq C + z | Q_1' + Q_2' \leq C) = \Phi(\bar{\gamma} + z/\sqrt{C})/\Phi(\bar{\gamma})$, $z \leq 0$, where $\Phi(\cdot)$ denotes the standard Gaussian c.d.f., and $\nu = [1 + \sqrt{2\pi}\bar{\gamma}\Phi(\bar{\gamma})e^{\bar{\gamma}^2/2}]^{-1}$. [For further details, see e.g., Halfin and Whitt (1981) and Browne and Whitt (1995).]

Numerical results and discussion. All figures presented in the sequel have been transformed (by appropriately re-scaling and translating from the diffusion scale) so as to correspond with the original scale of the underlying Markovian system.

1. General overview of proposed approximations: structural properties and accuracy.

Figure 1 compares the proposed approximations with the behavior of the actual system for the parameters of the example given above. Apart from the striking accuracy of the approximations, we make two observations: (i) the distribution for the number of guaranteed-rate users in the system is very close to normal, as predicted by both diffusion approximations; (ii) the distribution of the total number of users in the system is well approximated by a mixture of a Gaussian and an exponential random variable. The latter agrees with the closed form characterization obtained via the perturbation approach in Section 5, and with the first order term of an "asymptotic expansion" derived based on Girsanov's transformation which is discussed in the next section.

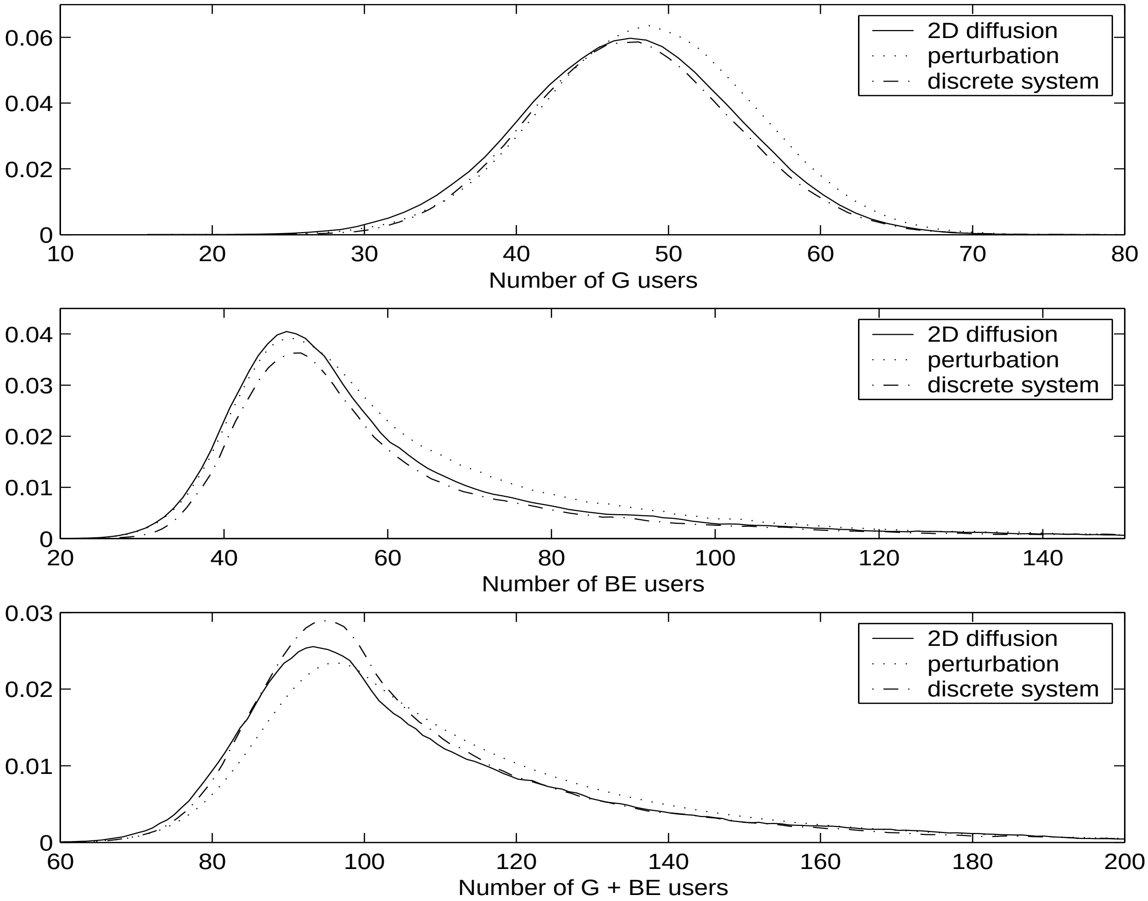

Figure 1: **Comparison of the steady-state densities for (a) discrete system, (b) 2-D diffusion, and (c) perturbed diffusion.** The parameters are $C = 100$, $\mu_1 = 1$, $\mu_2 = 2$, $\lambda_1 = 47.5$, $\lambda_2 = 95$.

2. Accuracy of the proposed approximations as a function of different model parameters.

In the sequel we only report results for the sum process. This effectively summarizes the congestion experienced by the best-effort users, which, in turn, is the key performance measure for the system.

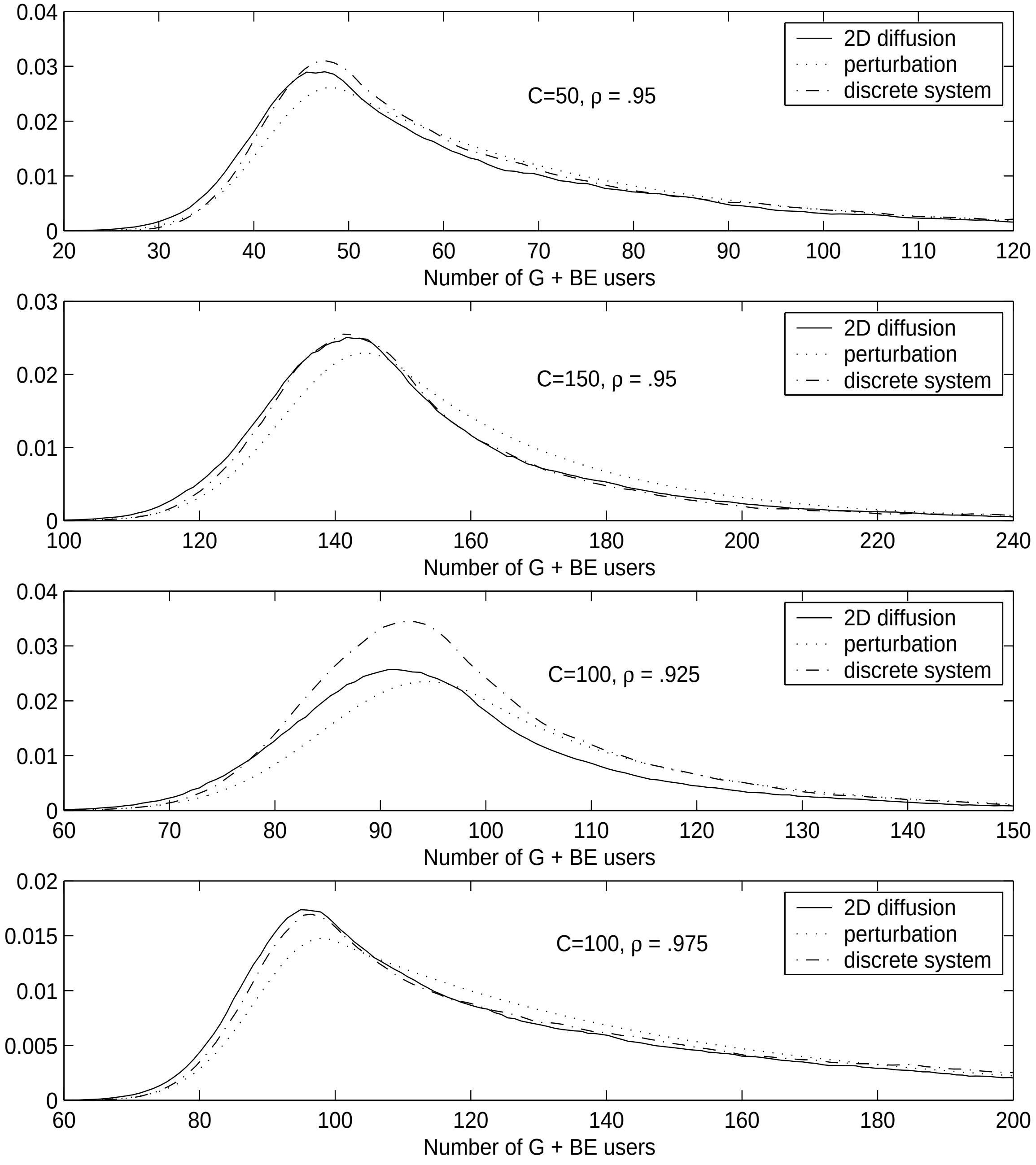

Figure 2: **Comparison of the steady-state densities at different levels of capacity and traffic intensity.** In all systems, $\mu_1 = 1$, $\mu_2 = 2$. Top panel: $C = 50$, $\lambda_1^{50} = 23.75$ and $\lambda_2^{50} = 47.5$. Second panel: $C = 150$, $\lambda_1^{150} = 71.25$ and $\lambda_2^{150} = 142.5$. In the third and fourth panels $C = 100$. Third panel: $\lambda_1 = 46.25$, $\lambda_2 = 92.5$ and $\rho = .925$. Fourth panel: $\lambda_1 = 48.75$, $\lambda_2 = 97.5$ and $\rho = .975$.

a. Dependence on the system capacity C.

The top two panels of Figure 2 show how the various approximations behave as we vary the system capacity to $C = 50$ and $C = 150$ while keeping the traffic intensity constant. They illustrate that the accuracy of the approximations is very good even for smaller levels of capacity. (Indeed, similar results were obtained for even smaller values for C.) It is interesting to note that the quality of the approximations seems to degrade when the capacity is higher ($C = 150$). This is easy to explain by expressing the traffic intensity in the form $\rho \approx 1 - \frac{\gamma}{\sqrt{C}}$ following the Halfin-Whitt asymptotics. Since $\rho = .95$ in both systems, the system with capacity $C = 50$ operates "closer" to heavy traffic as measured by the magnitude of the parameter γ, relative to the one with $C = 150$. Hence, the degradation of the quality of the approximations in the latter case.

b. Dependence on the aggregate traffic intensity.

Pursuing the last remark a little further, the third and fourth plots in Figure 2 shows results for the original system with $C = 100$ when the traffic intensity was set equal to .925 and .975, respectively. Consistent with the explanation given above, we see that the quality of the approximation improves as the system operates "closer" to the heavy traffic regime.

c. Dependence on the difference between the μ_i's.

This comparison is particularly important in the context of the perturbation approach. Figure 3 contains the associated results. The main observations from this figure are: (i) the accuracy of the approximation based on the perturbation approach depends on the ratio μ_1/μ_2 and not on the absolute difference $|\mu_1 - \mu_2|$; (ii) the quality of the approximation is reasonable for ratios μ_1/μ_2 in the range $(1/5, 5)$; and, (iii) the quality of the approximation degrades differently when μ_1/μ_2 grows large or μ_1/μ_2 gets small. The first two points are evident from the form of the limiting diffusions and the fact that the perturbation approach captures the effect of the difference in service rates by replacing γ_i by $\bar{\gamma}_i = \gamma_i\mu_i/\bar{\mu}$. Numerically, this is illustrated by the second and fourth panels of Figure 3 that depict systems with $C = 100$ and $\mu_1 = 2, \mu_2 = 1$ and $\mu_1 = 20, \mu_2 = 10$, respectively. The third observation can also be explained by close inspection of the limiting diffusions. The key point is that when $\mu_2 \ll \mu_1$, then the drift term contributed by the guaranteed-rate users dominates the approximation. Consequently, the resulting density for the sum process approaches the Gaussian density of the guaranteed users. This is illustrated by the densities of the Markovian system and the two-dimensional diffusion of Section 4; the perturbation approach is not very accurate in this parameter regime. In contrast, when $\mu_2 \gg \mu_1$, the opposite effect becomes true and makes the exponential tail of the distribution more pronounced.

The persistence of the Gaussian-Exponential form of the simulated distribution for the sum process, suggests that this may indeed be "close" to the form of the actual distribution corresponding to the multi-dimensional process obtained in Section 4. Explicit characterization and computation of the joint distribution for the multi-dimensional limit process remains an open problem.

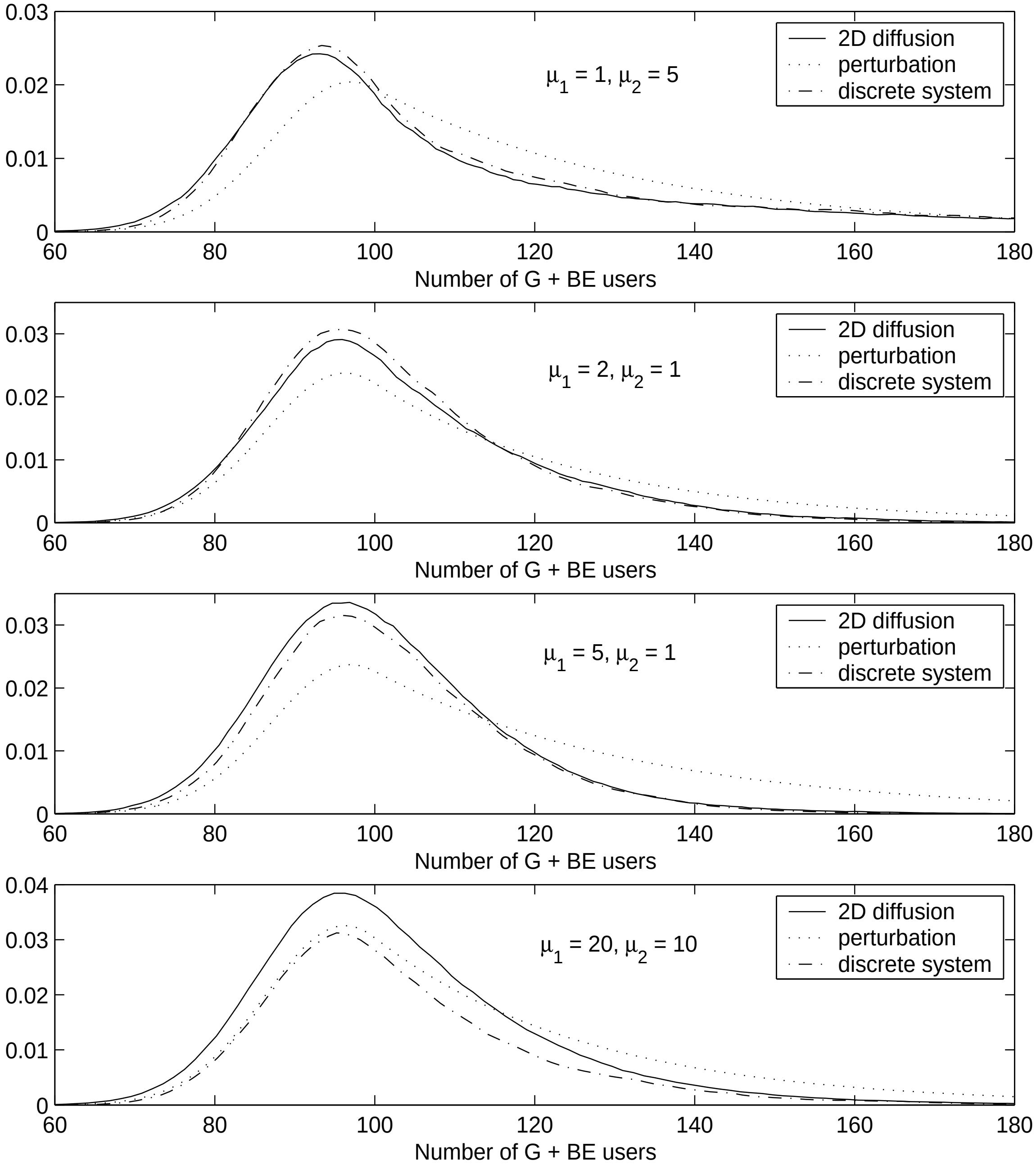

Figure 3: **Comparison of the steady-state densities for different ratios μ_1/μ_2.** In all systems: $C = 100$, $\lambda_1 = 47.5 \cdot \mu_1$, $\lambda_2 = 47.5 \cdot \mu_2$ ($\rho = .95$). Top panel: $\mu_1 = 1, \mu_2 = 5$; Second panel: $\mu_1 = 2, \mu_2 = 1$; Third panel: $\mu_1 = 5, \mu_2 = 1$; Fourth panel: $\mu_1 = 20, \mu_2 = 10$.

7 An Alternative Approach to Deriving Performance Analysis Approximations

This section explores an alternative approach to the perturbation asymptotics described in the previous two sections. In particular, here we apply a similar idea to the "post-limit" processes, i.e., the diffusion process identified in Theorem 1. Using the Girsanov transformation, we can formally identify a version of the "perturbed diffusion" described in Section 5 as the result of a change-of-measure applied to the original diffusion limits given in Theorem 1. For simplicity, the following discussion restricts attention to the two-class system that has a single guaranteed and best-effort class.

To illustrate the approximation we develop in this section, let us focus on transient expectations of the form $\mathbb{E}_x f(X(t))$, where $\mathbb{E}_x[\cdot] = \mathbb{E}[\cdot|X(0) = x]$. The latter are typically computed by solving the Kolmogorov backwards partial differential equation (PDE). Put $u(t, x) = \mathbb{E}_x f(X(t))$. Then, u is the solution to

$$\frac{\partial u(t,x)}{\partial t} = b(x) \cdot \nabla u(t,x) + \frac{1}{2}\sigma_1^2 \frac{\partial^2 u(t,x)}{\partial x_1^2} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 u(t,x)}{\partial x_2^2},$$

subject to the initial condition $u(0, x) = f(x)$, where ∇u is the gradient of u relative to the spatial variable x. Unfortunately, this PDE appears to be intractable.

Given the discussion in the previous sections a major role in the analysis of the system is played by the sum process $Z = X_1 + X_2$, thus we focus on functions f of the sum of the coordinate processes. [For example, the value of the function $f(x) = (x_1 + x_2)^+$ is directly related to the overall congestion level encountered by best-effort users when the system is in state x.] Again, the key observation is that if $\mu_1 = \mu_2 = \mu$, then $Z = X_1 + X_2$ is a diffusion process whose structure is identical to the Halfin-Whitt (H-W) diffusion identified in Proposition 1. This suggests that if the difference between the two service rates is small, we can view the dynamics of the sum process Z as a "small perturbation" of the Halfin-Whitt diffusion. In particular, it may be possible to expand the original performance measure in a Taylor series where the "first order term" is the performance measure derived on the basis of a diffusion process with drift

$$b(z) = \begin{cases} -\mu(\gamma_1 + \gamma_2) - \mu z & z \le 0 \\ -\mu(\gamma_1 + \gamma_2) & z > 0 \end{cases}, \qquad (26)$$

and infinitesimal variance $\sigma^2(z) = 2\mu$.

This approach can be made mathematically rigorous by appealing to Girsanov's transformation of probability measures; see, e.g., Karatzas and Shreve (1991, §5.3.B). In particular, the approach we pursue here is inspired by the recent Girsanov-based approach to analyzing non-stationary queues proposed by Glynn (2001); see also (Ward and Glynn 2003) that use a similar approach to derive transient approximations for a reflected O-U process. For the purpose of the next result, we assume for simplicity that $\mu_1 > \mu_2$.

Proposition 4 *Suppose that under the probability measure Q, $W = (W_1, W_2) = ((W_1(t), W_2(t)) : t \ge 0)$ is a standard Brownian motion in $\mathbb{R}^2$ and $Y = (Y_1, Y_2)$ is the solution to*

$$\begin{aligned} dY_1(t) &= -(\mu_1\gamma_1 + \mu_2 Y_1(t))dt + \sigma_1 dW_1(t) \\ dY_2(t) &= -\left(\mu_2\gamma_2 + \mu_2 \left(Y_1(t) + Y_2(t)\right)^- + \mu_2 Y_1(t)\right) dt + \sigma_2 dW_2(t) \end{aligned}$$

subject to $Y(0) = y \in \mathbb{R}^2$. Put $\alpha = \mu_1 - \mu_2$, and

$$M(t;\alpha) = \exp\left(-\frac{\alpha}{\sigma_1} \int_0^t Y_1(s) dW_1(s) - \frac{\alpha^2}{2\sigma_1^2} \int_0^t Y_1^2(s) ds \right) \ .$$

Then, $M = (M(t;\alpha) : t \geq 0)$ is a martingale adapted to the filtration generated by the Brownian motion W. Moreover, for each $t \geq 0$, $dP = M(t;\alpha) dQ$ is a probability measure and under P the process $Y = (Y(t) : t \geq 0)$ is identical in law to the solution X of the stochastic differential equation given in Theorem 1.

Suppose that X is the two-dimensional solution of the SDE in Theorem 1. Then, under P, $\mathbb{E}_x^P f(Y(t)) = \mathbb{E}_x f(X(t))$. Moreover, according to Proposition 4

$$\mathbb{E}_x^P f(Y(t)) = \mathbb{E}^Q \left[f(Y(t)) M(t;\alpha) \right]$$

provided f is non-negative, where $\mathbb{E}^Q[\cdot]$ is the expectation operator associated with the probability measure Q of Proposition 4. If f grows at most exponentially, we can show that a derivative according to α can be interchanged with the expectation operator, so that

$$\mathbb{E}_x f(X(t)) = \mathbb{E}_x^Q f(Y(t)) + \mathbb{E}_x^Q \left[f(Y(t)) M^{(1)}(t;0) \right] + O(\alpha^2)$$

as $\alpha \downarrow 0$, where

$$\begin{aligned} M^{(1)}(t;0) &= \left. \frac{\partial M(t;\alpha)}{\partial \alpha} \right|_{\alpha=0} \\ &= -\frac{1}{\sigma_1} \int_0^t Y_1(s) dW_1(s) \\ &= -\frac{1}{2\sigma_1^2} Y_1^2(s) + \frac{1}{2\sigma_1^2} Y_1^2(0) - \frac{\gamma_1 \mu_1}{\sigma_1^2} \int_0^t Y_1(s) ds - \frac{\mu_2}{\sigma_1^2} \int_0^t Y_1^2(s) ds + \sigma_1 t/2 \ . \end{aligned}$$

As mentioned previously, a quantity of particular interest is $\mathbb{E}(X_1(t) + X_2(t))^+$. To this end, taking $f(x) = (x_1 + x_2)^+$ we obtain

$$\begin{aligned} \mathbb{E}_x (X_1(t) + X_2(t))^+ &= \mathbb{E}_x^Q (Y_1(t) + Y_2(t))^+ \\ &\quad + \alpha \left(\frac{y_1^2}{2\sigma_1^2} + \sigma_1 \frac{t}{2} \right) \mathbb{E}_x^Q (Y_1(t) + Y_2(t))^+ - \frac{\alpha}{2\sigma_1^2} \mathbb{E}_x^Q \left[(Y_1(t) + Y_2(t))^+ Y_1(t)^2 \right] \\ &\quad - \alpha \frac{\gamma_1 \mu_1}{\sigma_1^2} \int_0^t \mathbb{E}_x^Q \left[(Y_1(t) + Y_2(t))^+ Y_1(s) \right] ds \\ &\quad - \alpha \frac{\mu_2}{\sigma_1^2} \int_0^t \mathbb{E}_x^Q \left[(Y_1(t) + Y_2(t))^+ Y_1^2(s) \right] ds + o(\alpha) \end{aligned}$$

as $\alpha \downarrow 0$.

While the above analysis does not lead to a completely tractable computation, it does illuminate a key characteristic of the problem that has been exploited via the change-of-measure approach, namely, a substantial simplification is obtained if $\mu_1 = \mu_2$. In particular, the process $Z = Y_1 + Y_2$ is then a diffusion under the transformed probability measure Q, whose drift is given in (26) and with constant infinitesimal variance equal to $\sqrt{2\mu_2}$. The expansion also identifies a "correction factor" that accounts for the fact that these rates are not identical. Thus, the results bear some structural resemblance to those derived in Section 5, where there the approximation was arrived at by applying an infinitesimal parameter perturbation in the "pre-limit" process, i.e., in the original CTMC. The latter approach seems to lead to a much more tractable analysis, and is thus preferred to the "post-limit" parameter perturbation approach illustrated above.

8 Concluding Remarks

Motivated by emerging technologies in the areas of information and communication services, this paper has proposed and analyzed a mathematical model of a system that offers guaranteed and best effort types of service. Practical systems in these application domains typically comprise of two classes, one for guaranteed and one for best-effort service, which as it turns out capture most of the benefits of service differentiation through prioritized service grades. This was illustrated by the analysis of its multi-class analogue that essentially "reduces" to the two-class system in the sense of having diffusion limits that "do not distinguish" between the top $m - 1$ priority classes. Several interesting directions of future work arise. One is to use these performance approximations for the purpose of designing such service systems. The problem would consist of choosing the system's capacity, the menu of service grades to be offered, and the optimal prices that maximize the system's profitability. A related question is to study the effect of congestion notification that the users receive to the system dynamics. These issues are addressed in a follow-up paper (Maglaras and Zeevi 2003*b*). A second interesting question is that of control of the service rate allocated to the different users in order to maximize profits. The perturbation approach developed in this paper seems a promising starting point for this problem as well. Finally, an interesting open problem is to characterize the steady-state distribution of the limiting multi-dimensional diffusion.

Acknowledgements: The authors are grateful to the two referees and the area editor for their constructive comments regarding the paper.

A Proofs

Proof of Theorem 1: Let us first verify that the SDE given in the theorem indeed admits a (unique) strong solution. To this end, note that the drift function $b(\cdot)$ is continuous and $\|b(x)\| \leq C(1 + \|x\|)$ for some finite constant C, where $\|\cdot\|$ denotes the usual Euclidean norm. Now, the drift function $b_i(\cdot)$ is linear and therefore Lipschitz continuous for all $i = 1, \dots, m - 1$. As for the drift function

$b_m(\cdot)$, fix $x, y \in \mathbb{R}^m$ and note that if $\sum_i x_i > 0$ and $\sum_i y_i > 0$, or if $\sum_i x_i \leq 0$ and $\sum_i y_i \leq 0$, then clearly $|b_m(x) - b_m(y)| \leq C|x-y|$. Now, if $\sum_i x_i \leq 0$ and $\sum_i y_i > 0$ then

$$b_m(x) - b_m(y) = -\mu_m x_m - \mu_m \sum_{i<m} y_i$$

and since $x_m \leq -\sum_{i<m} x_i$ we have that $b_m(x) - b_m(y) \geq \mu_m \sum_{i<m}(x_i - y_i)$. Similarly since $\sum_{i<m} y_i > -y_m$ we have that $b_m(x) - b_m(y) < \mu_m(y_m - x_m)$. Thus, it follows that

$$|b_m(x) - b_m(y)| \leq C\|x-y\|$$

for some finite constant C. A similar argument applies when $\sum_i x_i > 0$ and $\sum_i y_i \leq 0$. Thus, by Karatzas and Shreve (1991, Theorem 5.2.9) the SDE given in the theorem admits a unique strong solution.

The proof of weak convergence builds on general convergence results for Markov chains; our main reference for the latter is Strook and Varadhan (1979, §11.2). The typical approach reduces the task of establishing weak convergence to convergence of the infinitesimal mean and infinitesimal variance of the normalized process X^n, and in addition verification that the "jump size" becomes negligible (ensuring continuity of the sample paths of the limit process). The sufficiency of this set of conditions is known as Reboledo's theorem, see, e.g., Theorem 7.4.1 in Ethier and Kurtz (1986). Fix $n \geq 1$, $x \in \mathbb{R}^m$ and $\varepsilon > 0$. Let

$$\begin{aligned}
b_i^n(x) &= n\mathbb{E}\left[X_i^n(1/n) - X_i^n(0) \mid X^n(0) = x\right] \quad \text{for } i = 1, \ldots, m \\
\Sigma_{ij}^n(x) &= n\mathbb{E}\left\{\left[X_i^n(1/n) - X_i^n(0)\right]\left[X_j^n(1/n) - X_j^n(0)\right] \mid X^n(0) = x\right\} \quad \text{for } i, j = 1, \ldots, m \\
\Delta^n(x) &= n\mathbb{P}\left(\|X^n(1/n) - X^n(0)\| > \epsilon \mid X^n(0) = x\right) ,
\end{aligned}$$

where $X_i^n(\cdot) = (Q_i^n(\cdot) - \kappa_i n)/\sqrt{n}$. We will now prove that

$$\sup_{\|x\| \leq R} |b_i^n(x) - b(x)| \to 0 \quad \text{for } i = 1, \ldots, m \tag{27}$$

$$\sup_{\|x\| \leq R} |\Sigma_{ij}^n(x) - \Sigma_{ij}| \to 0 \quad \text{for } i, j = 1, \ldots, m \tag{28}$$

$$\sup_{\|x\| \leq R} \Delta^n(x) \to 0 \tag{29}$$

as $n \to \infty$ for each fixed $R > 0$. Statements (27)-(29) are equivalent to conditions (2.4)-(2.6) in Strook and Varadhan (1979, p.268). To prove (27), note that upon substituting $X_i^n = (Q_i^n - \kappa_i n)/\sqrt{n}$ into $b_i^n(x)$ we have that

$$b_i^n(x) = n^{-1/2}\mathbb{E}\left[Q_i^n(1/n) - Q_i^n(0) \mid Q^n(0) = \kappa_i n + x\sqrt{n}\right] .$$

Using the transition rates given in (7) adapted for to the m-class setting we have that

$$\begin{aligned}
b_1^n(x) &= n^{-1/2}\left[\lambda_1^n - \mu_1(\kappa_1 n + x_1\sqrt{n})\right] + o(1) \\
&= n^{-1/2}\left(\mu_1\kappa_1 n - \gamma_1\mu_1\sqrt{n} - \mu_1(\kappa_1 n + x_1\sqrt{n})\right) + o(1) \\
&= -\gamma_1\mu_1 - \mu_1 x_1 + o(1)
\end{aligned}$$

and for $i = 2, \ldots, m$

$$\begin{aligned} b_i^n(x) &= n^{-1/2}\left[\lambda_i^n - \mu_i\left([n - \sum_{j<i}\kappa_j n - \sum_{j<i} x_j\sqrt{n}] \wedge [\kappa_i n + x_i\sqrt{n}]\right)\right] + o(1) \\ &= n^{-1/2}\left[\mu_i\kappa_i n - \gamma_i\mu_i\sqrt{n} - \sqrt{n}\mu_i\left([n\sum_{j>i}\kappa_j - \sqrt{n}\sum_{j\le i} x_j] \wedge 0\right) - \mu_i\kappa_i n + \mu_i x_i\sqrt{n}\right] + o(1) \\ &= -\mu_i\gamma_i - \mu_i x_i \mathbb{I}_{\{\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j\le i} x_j \le 0\}} - \mu_i(\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j<i} x_j)\mathbb{I}_{\{\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j\le i} x_j > 0\}} + o(1), \end{aligned}$$

where the $o(1)$ denotes a term that converges to zero as $n \to \infty$, uniformly in x. Note that the last term of $b_i^n(x)$ becomes asymptotically negligible for all $i < m$. Here we have used the fact that for any fixed x and initial state of the Markov chain, $Q^n(0) = \kappa n + x\sqrt{n}$, the probability that $Q_1^n(1/n) = n$ is identically zero for sufficiently large n, thus the boundary is inaccessible (i.e., "blocking" of the guaranteed-rate users does not occur). The above derivations verify (27) and the uniform convergence of b^n to the drift function given by (12) in Theorem 1. For the infinitesimal variance, observe that

$$\begin{aligned} \Sigma_{11}^n(x) &= n^{-1}\left[\lambda_1^n + \mu_1(\kappa_1 n + x_1\sqrt{n})\right] + o(1) \\ &= 2\kappa_1\mu_1 + o(1) \end{aligned} \tag{30}$$

using the infinitesimal transition rate given in (8). A similar calculation yields that $\Sigma_{ii}^n(x) = 2\kappa_i\mu_i + o(1)$ for all $i = 2, \ldots, m$ and $\Sigma_{ij}^n = o(1)$ for all $i \neq j$. Thus, we have verified (28). To establish (29) it suffices to note that the Markov chain Q^n is essentially a birth-death process, thus, the jump size is bounded. In particular, it is straightforward to show that

$$\mathbb{E}\left[|X_i^n(1/n) - X_i^n(0)|^4 \mid X^n(0) = x\right] = \frac{C}{n^2} + o(1/n^2),$$

and using Markov's inequality we have that

$$\Delta^n(x) \le \frac{C}{n}$$

for some finite positive constant C.

Finally, with the above condition verified, we note that the continuity of the limiting drift $b(\cdot)$ together with the fact that the martingale problem for (b, Σ) is well posed [the latter follows straightforwardly from Theorem 10.2.2 in Strook and Varadhan (1979)], proves the asserted convergence $X^n \Rightarrow X$ following Theorem 11.2.3 in Strook and Varadhan (1979) [see also Corollary 7.4.2 in Ethier and Kurtz (1986)]. This concludes the proof. ■

Proof of Corollary 1: Recall that

$$D_i^n(t) := \left(\frac{\sum_{j\le i} Q_j^n(t) - n}{n - \sum_{j<i} Q_j^n(t)}\right)^+ .$$

Thus, using the definition of the normalized state process $X_i^n = (Q_i^n - \kappa_i n)/\sqrt{n}$, and recalling that the fluid scale workload contributions satisfy $\kappa_1 + \cdots + \kappa_m = 1$ we have that for $i = 2, \ldots, m-1$

$$\begin{aligned}\sqrt{n} D_i^n(t) &= n \left(\frac{\sum_{j \le i} X_j^n(t) - \sqrt{n} \sum_{j>i} \kappa_j}{\sum_{j \ge i} \kappa_j n - \sqrt{n} \sum_{j<i} X_j^n(t)} \right)^+ \\ &= \left(\frac{\sum_{j \le i} X_j^n(t) - \sqrt{n} \sum_{j>i} \kappa_j}{\sum_{j \ge i} \kappa_j - n^{-1/2} \sum_{j<i} X_j^n(t)} \right)^+ .\end{aligned}$$

For the lowest-priority best-effort class

$$D_m^n(t) := \left(\frac{\sum_j Q_j^n(t) - n}{n - \sum_{j<m} Q_j^n(t)} \right)^+ ,$$

which can be rewritten as

$$\begin{aligned}\sqrt{n} D_m^n(t) &= n \left(\frac{\sum_j X_j^n(t)}{\kappa_m n - \sqrt{n} \sum_{j<m} X_j^n(t)} \right)^+ \\ &= \left(\frac{\sum_j X_j^n(t)}{\kappa_m - n^{-1/2} \sum_{j<m} X_j^n(t)} \right)^+ .\end{aligned}$$

Given the convergence $X^n \Rightarrow X$ in Theorem 1, an application of the continuous mapping theorem concludes the proof. ■

Proof of Proposition 2: We will establish positive recurrence by means of an appropriately chosen Lyapunov function. (By positive recurrence we mean that the expected hitting time of any compact set is finite, and for this it suffices to focus on the time it takes the process to return to a compact set, when it starts outside of it.) To elucidate the logic supporting the selection of our Lyapunov function, we first consider in detail the case where $m = 2$. Subsequently, we will construct the Lyapunov function that is used to prove stability in the general case where $m > 2$, omitting the tedious verification calculations as these simply replicate the two-class case. Finally, we will prove the necessity of the stability condition $\gamma > 0$.

Step 1. Let $m = 2$, assume that $\gamma > 0$ and put

$$f(x) = C_1(1/2)x_1^2 + C_2 x_1 + C_3\sqrt{1 + x_2^2} + C_4, \tag{31}$$

for $x := (x_1, x_2) \in \mathbb{R}^2$, and certain constants $C_1, C_3 > 0$, and $C_2 \in \mathbb{R}$ to be specified in the sequel. The remaining constant C_4 is chosen so that $C_2^2/(2C_1) < C_4$. (This last condition guarantees that $(C_1/2)x_1^2 + C_2 x_1 + C_4 > 0$.) The function f is then non-negative and satisfies $f(x) \to \infty$ as $\|x\| \to \infty$. Since f is twice continuously differentiable, we can apply to it the differential operator $\mathcal{A}$

[the generator of the diffusion (11) given in Theorem 1]. We then have

$$\begin{aligned}(\mathcal{A}f)(x) &:= b(x)\cdot\nabla f(x) + \frac{1}{2}\left(\sigma_1^2\frac{\partial^2 f(x)}{\partial x_1^2} + \sigma_2^2\frac{\partial^2 f(x)}{\partial x_2^2}\right)\\ &= -C_2\mu_1\gamma_1 - (C_1\mu_1\gamma_1 + C_2\mu_1)x_1 - C_1\mu_1 x_1^2 - C_3\mu_2\gamma_2\frac{x_2}{\sqrt{1+x_2^2}}\\ &\quad -C_3\mu_2\frac{x_2^2}{\sqrt{1+x_2^2}}\mathbb{I}\{x_1+x_2\le 0\} + C_3\mu_2 x_1\frac{x_2}{\sqrt{1+x_2^2}}\mathbb{I}\{x_1+x_2>0\}\\ &\quad +(1/2)C_1\sigma_1^2 + (1/2)C_3\sigma_2^2\frac{1}{(1+x_2^2)^{3/2}}\ ,\end{aligned}\tag{32}$$

where $\mathbb{I}\{B\}$ is the indicator of the set B. Observe that since $\sigma_i^2/2 = 2\kappa_i\mu_i/2$, then $\sigma_i^2/2 \le \mu_i$.

Step 2. Our goal is to show that there exists an $\varepsilon > 0$ and $R > 0$ such that

$$(\mathcal{A}f)(x) \le -\varepsilon \quad \text{for all } x \text{ such that } \|x\| > R. \tag{33}$$

The above is a typical Foster-Lyapunov condition [for more details see, e.g., Meyn and Tweedie (1994)]. A close inspection of (32) reveals that when $x_1 + x_2 \le 0$, we have that $(\mathcal{A}f)(x) \le -K_1x_1^2 + K_2x_1 + K_3 - K_4|x_2|$ for suitable constants $K_i > 0$, for all x such that $\|x\|$ is sufficiently large. Thus, we can verify condition (33). Consider now the case where $x_1 + x_2 > 0$.

Case (i): $x_2 < 0$, and $x_1 + x_2 > 0$. Then, clearly if x is such that $\|x\| > R$, say, we must have that $x_1 > \sqrt{R/2}$. A close inspection of (32) reveals that $(\mathcal{A}f)(x) \le -K_1x_1^2 + K_2x_1 + K_3$, for suitable constants $K_i > 0$, for all x such that $\|x\|$ is sufficiently large. Condition (33) can then be satisfied for an appropriate choice of $\varepsilon, R > 0$.

Case (ii): $x_2 > 0$ and $x_1 \in \mathbb{R}$ such that $x_1+x_2 > 0$. First, we observe that if x_1 is large in magnitude, then the analysis of case (i) above follows, and condition (33) is met. Thus, the interesting case is where $x_2 > 0$ is "large," and x_1 is "small." To this end, we can write

$$\begin{aligned}(\mathcal{A}f)(x) &\le -C_2\mu_1\gamma_1 - (C_1\mu_1\gamma_1 + C_2\mu_1)x_1 - C_1\mu_1x_1^2 - C_3\mu_2\gamma_2 g(x_2)\\ &\quad +C_3\mu_2x_1g(x_2) + 2C_1\mu_1,\end{aligned}\tag{34}$$

where $g(x_2) = x_2/(1+x_2^2)^{1/2}$. Here we have used the fact that for x_2 sufficiently large we can set $(1/2)C_3\sigma_2^2(1+x_2^2)^{-3/2} \le C_1\mu_1$. Now, note that $g(x_2) \to 1$ as $x_2 \to \infty$. Then, let us re-write the expression on the right-hand-side of (34) as follows,

$$\psi(x_1,x_2) = Ax_1^2 + B(x_2)x_1 + C(x_2)$$

with

$$\begin{aligned}A &= -C_1\mu_1\\ B(x_2) &= -(C_1\mu_1\gamma_1 + C_2\mu_1 - C_3\mu_2 g(x_2))\\ C(x_2) &= -C_2\mu_1\gamma_1 - C_3\mu_2\gamma_2 g(x_2) + 2C_1\mu_1\ .\end{aligned}$$

We will now show that $\psi(x_1, x_2) < 0$ for all x_2 sufficiently large, and for all $x_1 \in \mathbb{R}$. To establish this, fix an arbitrary $x_1 \in \mathbb{R}$, and let us examine

$$\psi(x_1, \infty) := \lim_{x_2 \to \infty} \psi(x_1, x_2) = Ax_1^2 + B(\infty)x_1 + C(\infty).$$

We now fix the constants C_i in the definition of the Lyapunov function f as follows:

$$\begin{aligned} C_3 &= (c + \gamma_1^2)/((\gamma_1 + \gamma_2)\mu_2) \\ C_2 &= (C_3\mu_2 - \gamma_1)/\mu_1 \\ C_1 &= 1/\mu_1 \ , \end{aligned} \tag{35}$$

where $c > 0$ is specified shortly. With this choice, note that C_3 is indeed strictly positive, as required, since $\gamma_1 + \gamma_2 > 0$ by assumption. Setting, for example, $c = 3$ in (35), it follows by straightforward algebra that $A = -1$, $B(\infty) = 0$, $C(\infty) = -1$, so $(B(\infty))^2 - 4AC(\infty) = -4$. Thus, it must be that $\psi(x_1, \infty) < 0$ for all $x_1 \in \mathbb{R}$. By continuity, it is also clear that for sufficiently large x_2 we have that $(B(x_2))^2 - 4AC(x_2) \leq -2$, say. Thus, for all x_2 such that $x_2 > R_2$, for an appropriately chosen constant R_2, we have that $\psi(\cdot, x_2) < 0$. Consequently, since $(\mathcal{A}f)(x) \leq \psi(x)$, we can find a pair (ε, R) such that condition (33) is satisfied.

Step 3. Let $\varepsilon, R > 0$ be such that (33) is satisfied (the previous step ensures the existence of such constants). Let $K = \{x : \|x\| \leq R\}$. Fix an initial state $X(0) = x \in \mathbb{R}^2$ and let $X = (X(t) : t \geq 0)$ be the diffusion process that is the strong solution to the stochastic differential equation (11) whose existence and uniqueness are established in Theorem 1. Using Itô's formula we then have

$$f(X(t)) = f(X(0)) + \int_0^t (\mathcal{A}f)(X(s))ds + \int_0^t \nabla f(X(s)) \cdot \Sigma dB(t) \ .$$

Now, since the drift of X, $b(\cdot)$ given in (33), satisfies $\|b(x)\| \leq C(1 + \|x\|)$, then for any fixed initial state $X(0) = x \in \mathbb{R}^2$ we have that $\mathbb{E}_x \int_0^t \|X(s)\|^2 ds < \infty$ [cf. Karatzas and Shreve (1991, Theorem 5.2.9)], where $\mathbb{E}_x[\cdot] := \mathbb{E}[\cdot | X(0) = x]$. (The process is non-explosive and in $L2$.) Consequently, since $\|\nabla f(x)\| \leq C\|x\|$ for some $C < \infty$, the stochastic integral in the right-hand-side above is an L^2-martingale. Now, for $x \in K^c$ we have $(\mathcal{A}f)(x) \leq -\varepsilon$. Put $T_K = \inf\{t \geq 0 : X(t) \in K\}$ and fix $t \geq 1$. Since $t \wedge T_K$ is bounded, we can apply Doob's optional stopping theorem [see, e.g., Karatzas and Shreve (1991, Theorem 1.3.22)] and thus have

$$\mathbb{E}_x f(X(t \wedge T_K)) = f(x) + \mathbb{E}_x \int_0^{t \wedge T_K} (\mathcal{A}f)(X(s))ds \ .$$

Since, $f \geq 0$ and $x \in K^c$ and $(\mathcal{A}f)(\cdot) \leq -\varepsilon$ on K^c, we have that

$$\mathbb{E}_x[t \wedge T_K] \leq \frac{f(x)}{\varepsilon}$$

and using the Monotone Convergence Theorem we have that $\mathbb{E}_x T_K \leq f(x)/\varepsilon$ for all $x \in K^c$ which establishes positive recurrence. Finally, we can use Ethier and Kurtz (1986, Lemma 9.7 and Theorem

9.9) to establish that the mean occupation measure is tight, and thus that a stationary distribution π exists. Now, since X has constant diffusion matrix coefficients, and this matrix is full rank, the resulting diffusion is non-degenerate and hence the stationary distribution π is unique. Moreover, $P_x(\cdot, t) := \mathbb{P}_x(X(t) \in \cdot) \to \mathbb{P}_\pi(\cdot)$ as $t \to \infty$, thus X is ergodic [see Has'minskii (1979, p.129-131)]. The intuition here is that the full rank of the diffusion matrix $\Sigma\Sigma^\top$ guarantees that the process "visits the whole state space," and cannot get "trapped" in any one region. Thus, we have that $X(t) \Rightarrow X(\infty)$ where $X(\infty)$ is distributed according to π.

Step 4. To extend the above arguments to the case where $m > 2$, we construct a Lyapunov function which is the obvious multi-dimensional analogue of (31). Specifically, let $\gamma := \sum_{i=1}^m \gamma_i$ and put

$$f(x) = \sum_{i<m} (1/2) C_{1,i} x_i^2 + \sum_{i<m} C_{2,i} x_i + C_3 \sqrt{1 + x_m^2} + C_4$$

where C_4 is set so that f is non-negative, and the constants $C_{1,i}, C_{2,i}$ for $i = 1, \ldots, m-1$, and C_3, C_4, are chosen based on the logic explained in Step 2. In particular, analogously to the calculation carried out in that step, we set

$$\begin{aligned} C_3 &= \frac{c + \sum_{i<m} \gamma_i^2}{\gamma \mu_m} \\ C_{2,i} &= (C_3 \mu_m - \gamma_i)/\mu_i \\ C_{1,i} &= 1/\mu_i, \end{aligned}$$

where $c > 0$ is a constant that ensures that $(\mathcal{A}f)(x)$ is negative for x outside some compact set. (Recall, in the case of $m = 2$, setting $c = 3$ in (35) was seen to be one particular choice that guarantees that this "negative drift" condition holds.) Straightforward algebra, essentially repeating the derivations in Steps 1 and 2, yields that f is a valid Lyapunov function. Step 4 can then be repeated verbatim.

Step 5. To prove necessity of the stability condition, suppose that $\gamma \leq 0$ and X is positive recurrent. Consider the case where $m = 2$ and $\mu_1 = \mu_2$. It follows that $Z = X_1 + X_2$ is the Halfin-Whitt diffusion, see Section 3 Proposition 1, with drift $b(z) = -\gamma\mu - \mu z \mathbb{I}\{z \leq 0\}$. But then Z is clearly either null-recurrent or transient, since its dynamics when $Z \geq 0$ are given by a driftless or positive drift Brownian motion, when $\gamma = 0$ or $\gamma < 0$, respectively. Thus, it cannot be that X is positive recurrent, in contradiction. This concludes the proof. ■

Proof of Theorem 2: The proof follows exactly the same steps in the proof of Theorem 1. Repeating the derivation of the infinitesimal drift rates for the m component processes, as in Theorem 1, only now using $\mu_i^n = \bar{\mu}(1 - \zeta_i/\sqrt{n})$ for $i = 1, \ldots, m$, we get that for $i = 1$

$$\begin{aligned} b_1^n(y) &= n^{-1/2} \left[\lambda_1^n - \mu_1^n (\kappa_1 n + y_1 \sqrt{n}) \right] + o(1) \\ &= n^{-1/2} \left(\bar{\mu} \kappa_1 n - \bar{\gamma}_1 \bar{\mu} \sqrt{n} - \bar{\mu} \kappa_1 n - \bar{\mu} y_1 \sqrt{n} \right) + o(1) \\ &= -\bar{\mu} \bar{\gamma}_1 - \bar{\mu} y_1 + o(1) \end{aligned}$$

and similarly, for $i = 2, \ldots, m$

$$\begin{aligned} b_i^n(y) &= n^{-1/2}\left[\lambda_i^n - \mu_i^n\left([n - \sum_{j<i}\kappa_j n - \sum_{j<i} y_j\sqrt{n}] \wedge [\kappa_i n + y_i\sqrt{n}]\right)\right] + o(1) \\ &= n^{-1/2}\left[\bar{\mu}\kappa_i n - \bar{\gamma}_i\bar{\mu}\sqrt{n} - \sqrt{n}\bar{\mu}\left([n\sum_{j>i}\kappa_j - \sqrt{n}\sum_{j\le i} y_j] \wedge 0\right) - \bar{\mu}\kappa_i n + m\bar{u}y_i\sqrt{n}\right] + o(1) \\ &= -\bar{\gamma}\bar{\mu} - \bar{\mu}y_i\mathbb{I}_{\{\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j\le i} y_j \le 0\}} - \bar{\mu}(\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j<i} y_j)\mathbb{I}_{\{\sqrt{n}\sum_{j>i}\kappa_j - \sum_{j\le i} y_j > 0\}} + o(1), \end{aligned}$$

where again the last term in b_i^n is asymptotically negligible is $i < m$. It is easily seen that the infinitesimal variance is not affected by the parameter perturbation, and therefore the results given in (30) hold in this case as well. Thus, proceeding as in the proof of Theorem 1, we have the required convergence of the infinitesimal drift and diffusion functions on compact sets to the limits given in (20). The rest of the calculations follow exactly those in the proof of Theorem 1, we therefore omit the details. ■

Proof of Corollary 2: The proof is a straightforward application of the continuous mapping theorem; the details are omitted. ■

Proof of Proposition 3: Let $Z = (Z(t) : t \ge 0)$ be given by the solution to

$$\begin{aligned} dZ(t) &= -\left(\mu\gamma - \mu(Z(t))^-\right)dt + \sigma dW(t) \\ Z(0) &= z, \end{aligned}$$

where $(x)^- = -\min(x, 0)$, $\sigma^2 = 2\mu$ and W is standard Brownian motion in $\mathbb{R}^2$. The proof follows in 3 steps.

Step 1. Fix $b > 0$, $z \le b$ and put $u(z)$ equal to the expression given in (24) for $\mathbb{E}_z T_b$. Straightforward algebra yields that u satisfies

$$\begin{aligned} &(\mathcal{A}u)(z) = -1 \\ &u(b) = 0\ , \end{aligned}$$

where

$$\mathcal{A} := -\left(\mu\gamma - \mu(z)^-\right)\frac{d}{dz} + \frac{\sigma^2}{2}\frac{d^2}{dz^2}\ .$$

To verify that u is twice continuously differentiable for all z, we need only verify that this holds at the origin. (Clearly the “two pieces” given in (24) are twice continuously differentiable for all $z \neq 0$.) Straightforward calculations then establish that $u(0^+) = u(0^-)$, $u'(0^+) = u'(0^-)$ and $u''(0^+) = u''(0^-)$, where u', u'' denote, respectively, the first and second derivative of u. Here $u(0^+) = \lim_{x\downarrow 0} u(x)$ and $u(0^+) = \lim_{x\uparrow 0} u(x)$ with the equivalent definitions holding for the respective derivatives. Thus, we have that u is twice continuously differentiable everywhere.

Step 2. The key observation is that u' as well as u'' are bounded for $z \le b$. To see this, observe that for $z \in [0, b]$ this is clear since u along with its derivatives are continuous. For $z \le 0$ we can differentiate (24) to get

$$\begin{aligned} u'(z) &= -\frac{\sqrt{2\pi}}{\mu} \exp((\gamma + z)^2/2)\Phi(z + \gamma) \\ u''(z) &= -\frac{\sqrt{2\pi}}{\mu}(\gamma + z) \exp((\gamma + z)^2/2)\Phi(z + \gamma) - \frac{1}{\mu}. \end{aligned}$$

Now, using estimates on the tail of the Normal distribution we have that $(1-\Phi(x)) \sim x^{-1}(2\pi)^{-1} \exp(-x^2/2)$ as $x \to \infty$ [see, e.g., Feller (1968, p. 175)] which proves the assertion. Applying Itô's differential rule to $u(Z(t))$ we have that

$$u(Z(t)) = u(Z(0)) + \int_0^t (\mathcal{A}u)(Z(s))ds + \int_0^t \sigma u'(Z(s))dW(s) \ .$$

Since u' is bounded, the last integral on the right hand side is a martingale adapted to the Brownian motion W. Thus, the Optional Stopping Theorem gives that

$$\mathbb{E}_z u(Z(t \wedge T_b)) = u(z) - \mathbb{E}_z[t \wedge T_b] \tag{36}$$

using the fact that $(\mathcal{A}u) = -1$. Because u is non-negative, we have that $\mathbb{E}_z[t \wedge T_b] \le u(z)$, and using the Monotone Convergence Theorem we get that $\mathbb{E}_z[T_b] \le u(z)$ thus $T_b < \infty$, almost surely for all $z < b$.

Step 3. To finish the proof, we need to apply another limit/expectation interchange argument for $\mathbb{E}_z u(Z(t \wedge T_b))$. To this end, let

$$M(t \wedge T_b) = \int_0^{t \wedge T_b} \sigma u'(Z(s))dW(s)$$

then, since u' is bounded for all $z \le b$, $M = (M(t) : t \ge 0)$ is a martingale adapted to the filtration of W. Moreover, $M^2(t \wedge T_b) - \langle M \rangle(t \wedge T_b)$ is a martingale, where the latter term is the quadratic variation process of M given by $\int_0^t \sigma^2(u'(Z(s)))^2 ds$. Therefore, using the Optional Sampling Theorem for the bounded stopping time $t \wedge T_b$ we have that

$$\begin{aligned} \mathbb{E}_z M^2(t \wedge T_b) &= \mathbb{E}_z \langle M \rangle (t \wedge T_b) \\ &\le \sigma^2 \sup_{x \le b}\{(u'(x))^2\} \mathbb{E}_z T_b \ . \end{aligned}$$

Since we have already established that $\mathbb{E}_z T_b < \infty$, and the right hand side is independent of t, we have that $\{M(t \wedge T_b)\}$ is a uniformly integrable family relative to $t \ge 0$. Consequently, since $u(Z(t \wedge T_b)) \le u(z) + M(t \wedge T_b)$, we also have that $\{u(Z(t \wedge T_b))\}$ is uniformly integrable and therefore upon taking $t \to \infty$ we obtain from (36) that $\mathbb{E}_y u(Z(T_b)) = u(z) - \mathbb{E}_z(T_b)$ and since $u(Z(T_b)) = u(b) = 0$, the proof is complete. ■

Proof of Proposition 4: We first prove that M is a martingale adapted to the filtration generated by the Brownian motion W. Let $T_n = \inf\{t \geq 0 : |Y_1(t)| \geq n\}$. Since $|Y(t \wedge T_n)| \leq n$ for all $t \geq 0$, it follows that $M(t \wedge T_n; \alpha)$ is integrable. By construction, $M(t \wedge T_n; \alpha)$ is a martingale adapted to the filtration of W_1 [see, e.g., Karatzas and Shreve (1991, p. 191)]. Thus, setting $\mathcal{F}_s = \sigma(W(u) : u \in [0, s])$ to be the filtration generated by the two-dimensional Brownian motion for, we have that

$$\mathbb{E}[M(t \wedge T_n; \alpha)|\mathcal{F}_s] = M(s \wedge T_n; \alpha)$$

almost surely, for all $t > s \geq 0$. To establish that M is a martingale, we seek a bound on $M(t \wedge T_n; \alpha)$ that is uniform in n and integrable, so that we may apply an interchange of limit/conditional expectation. To this end, Itô's lemma gives us

$$Y_1^2(t) = Y_1^2(0) - 2\int_0^t (\mu_1\gamma_1 + \mu_2 Y_1(s))Y_1(s)ds + \sigma_1^2/2t + 2\sigma_1 \int_0^t Y_1(s)dW_1(s)$$

thus,

$$\begin{aligned} M(t;\alpha) &= \exp\left(-\frac{\alpha}{2\sigma_1^2}Y_1^2(t) + \frac{\alpha}{2\sigma_1^2}Y_1^2(0) - \frac{\alpha\gamma_1\mu_1}{\sigma_1^2}\int_0^t Y_1(s)ds \right.\\ &\qquad \left. -\frac{2\alpha\mu_2 + \alpha^2}{2\sigma_1^2}\int_0^t Y_1^2(s)ds + \alpha t/2\right) . \end{aligned}$$

Now, using the fact that $\alpha > 0$, and plugging in $Y_1(0) = y_1$ we have that

$$M(t;\alpha) \leq \underbrace{\exp(C(1 + y_1^2 + t))\exp\left(-\frac{\alpha\gamma_1\mu_1}{\sigma_1^2}\int_0^t Y_1(s)ds\right)}_{R(t;\alpha)} .$$

Since Y_1 is an O-U process, it is Gaussian and therefore $\int_0^t Y_1(s)ds$ follows a Normal distribution with mean $m(t)$ and variance $v^2(t)$, where $m(t) = C_1 t$, and $v^2(t) = C_2 t$ for some finite constants C_1, C_2. Thus, $\mathbb{E}\exp\{\theta\int_0^t Y_1(s)ds\} = \exp(\theta m(t) + \theta^2/2v^2(t))$ which is finite for all t and θ. It then follows straightforwardly that $M(t \wedge T_n; \alpha) \leq R(t; \alpha)$ almost surely, for all $t \geq 0$ and uniformly in n, where $\mathbb{E}_{y_1} R(t;\alpha) \leq \exp(C(1 + y_1^2 + t)$. Thus, an application of the Dominated Convergence Theorem for conditional expectations yields that

$$\lim_{n\to\infty} \mathbb{E}[M(t \wedge T_n; \alpha)|\mathcal{F}_s] = M(s;\alpha)$$

which establishes that M is a martingale. Finally, the rest of the assertions in the statement of the theorem are a standard consequence of Girsanov's theorem, cf. Karatzas and Shreve (1991, Proposition 5.3.6). This concludes the proof. ■

References

Armony, M. and Maglaras, C. (2003*a*), 'Contact centers with a call-back option and real-time delay information', *Oper. Res.* . To appear.

Armony, M. and Maglaras, C. (2003*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* .

Bean, N. G., Gibbens, R. J. and Zachary, S. (1995), 'Asymptotic analysis of single resource loss systems in heavy traffic with applications to integrated networks', *Adv. Appl. Prob.* **27**, 273–292.

Billingsley, P. (1999), *Convergence of Probability Measures*, 2nd ed., Wiley, New York.

Borst, S., Mandelbaum, A. and Reiman, M. (2003), 'Dimensioning large call centers', *Oper. Res.* . To appear.

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**(9), 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.

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

Feller, W. (1968), *An Introduction to Probability Theory and its Applications, Vol. I*, 3rd ed., 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'.

Garnett, O., Mandelbaum, A. and Reiman, M. (2002), 'Designing a call center with impatient customers', *Manufacturing & Service Operations Management* **4**(3), 208–227.

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

Glynn, P. W. (2001), 'Performance analysis for nonstationary queues'. 11th Informs Applied Probability Conference, New York, 2001.

Halfin, S. and Whitt, W. (1981), 'Heavy-traffic limits for queues with many exponential servers', *Oper. Res.* **29**(3), 567–588.

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

Has'minskii, R. Z. (1979), *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff.

Karatzas, I. and Shreve, S. (1991), *Brownian Motion and Stochastic Ccalculus*, 2nd ed., Springer-Verlag, New York.

Keilson, J. (1966), 'A limit theorem for passage times in ergodic regenerative processes', *Ann. Math. Stat.* **37**, 866–870.

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

Maglaras, C. and Zeevi, A. (2003*b*), 'Pricing and design of differentiated services: Approximate analysis and structural insights'. Working paper, Columbia University.

Mandelbaum, A., Massey, W. and Reiman, M. (1998), 'Strong approximations for Markovian service networks', *Queueing Systems* **30**, 149–201.

Meyn, S. P. and Tweedie, R. I. (1994), *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

Odlyzko, A. (1999), 'Paris metro pricing for the internet', *In Proc. ACM Conf. on Elec. Comm.* pp. 140–147.

Puhalskii, A. and Reiman, M. (2000), 'The multiclass GI/PH/N queue in the Halfin-Whitt regime', *Adv. Appl. Prob.* **32**(2), 564–595.

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

Ward, A. and Glynn, P. W. (2003), 'Properties of the reflected Ornstein-Uhlenbeck process', *Queueing Systems* . To appear.

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

Whitt, W. (2003), 'How multiserver queues scale with growing congestion-dependent demand', *Oper. Res.* . To appear.