OPERATIONS RESEARCH
Vol. 52, No. 2, March–April 2004, pp. 271–292
ISSN 0030-364X | EISSN 1526-5463 | 04 | 5202 | 0271

DOI 10.1287/opre.1030.0088
© 2004 INFORMS

On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules, and System Design

Mor Armony
Stern School of Business, New York University, 44 West 4th Street, Suite 8-66, New York, New York 10012, marmony@stern.nyu.edu

Constantinos Maglaras
Columbia Business School, 409 Uris Hall, 3022 Broadway, New York, New York 10027, c.maglaras@columbia.edu

Organizations worldwide use contact centers as an important channel of communication and transaction with their customers. This paper describes a contact center with two channels, one for real-time telephone service, and another for a postponed call-back service offered with a guarantee on the maximum delay until a reply is received. Customers are sensitive to both real-time and call-back delay and their behavior is captured through a probabilistic choice model. The dynamics of the system are modeled as an $M/M/N$ multiclass system. We rigorously justify that as the number of agents increases, the system's load approaches its maximum processing capacity. Based on this observation, we perform an asymptotic analysis in the many-server, heavy traffic regime to find an asymptotically optimal routing rule, characterize the unique equilibrium regime of the system, approximate the system performance, and finally, propose a staffing rule that picks the minimum number of agents that satisfies a set of operational constraints on the performance of the system.

Subject classifications: service networks; call centers; heavy traffic; service level guarantees; choice models; Nash equilibrium; Halfin-Whitt regime.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received September 2001; revision received May 2002; accepted May 2003.

1. Introduction

Organizations worldwide use contact centers as an important channel of communication and transaction with their customers. The most prevalent form of communication is the telephone, but, with the proliferation of the Internet, other channels such as e-mail or online real-time support are becoming widespread. Corporate contact centers range in size from a few agents in one office to several hundred or even thousands that are located in geographically dispersed locations spread out around the world. Their socioeconomic importance in today's business landscape cannot be overstated: "There are approximately 7,000,000 agents now working in 70,000 call centers in the United States, with an annual growth rate of up to 20% in agent positions," "70% of all customer interaction occurs in the call center," "personnel (staffing) costs account for over 65% of the running costs of a call center," etc.; see Call center statistics (2001) for many interesting statistics. From a modeling point of view, contact centers can be viewed as large systems, operating in a stochastic environment, at very high agent utilization rates. Due to their inherent complexity, common practice is to either use simplified formulas and simulation to predict system performance, or to employ some form of approximate analysis that leads to good, closed-form characterizations of their behavior. Our work falls into the latter category.

This paper studies a contact center with many identical agents (or servers) and two service modes: (1) traditional (real-time) telephone service, and (2) postponed (call-back) service. To make the call-back option attractive to callers, it is coupled with a quality-of-service guarantee on the maximum delay before receiving a reply. Arriving customers are informed (or know from prior experience) of the expected waiting time for real-time service. Based on this information and the delay guarantee for the call-back option, they decide whether to join the queue, leave their number to be called back later, or balk. In turn, the actual waiting time depends on the customer's decisions. Hence, we are interested in an equilibrium operating mode in which the customers' reactions to the announced information results in a pair of arrival rates into each class that induce a steady-state regime that is consistent with the information announced. The control decision faced by the system manager is the routing of the jobs; namely, upon service completion, should the agent take an online call or call a customer back. We study the system performance under a general customer choice mechanism and an appropriately designed routing rule.

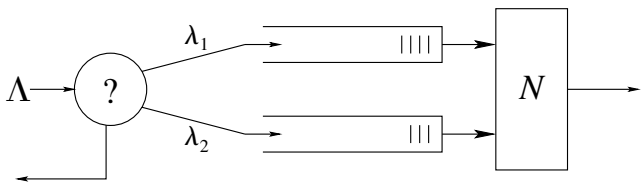
There are four goals of this paper. First, at the lowest level, we address the operational issue of routing service requests originating in the two channels in a way that maximizes the quality of service experienced by the real-time

customers subject to the delay constraint on the postponed service. Second, at the system level, we characterize the equilibrium operating regime, if it exists, analyze performance measures of interest, and gain insights on their dependence on system and choice model parameters. Third, from an economic viewpoint, we design the most cost-effective system that achieves a desired level of quality of service with the minimum number of agents. Finally, from a managerial perspective, we demonstrate the performance improvements realized by introducing the call-back option.

The call center under investigation is modeled in this paper as a two-class $M/M/N$ system. We refer to the real-time customers as class 1 customers, whereas customers who choose to be called back are referred to as class 2. The two types of service are assumed to have equal mean processing times. The arrival rates of the two classes are determined through the customers' decisions. These decisions are made according to a probabilistic choice model that captures the trade-off between the value of receiving service and the cost of waiting associated with the two options. The interplay between the arrival rates and the steady-state expected waiting time may be described as a game in strategic form. In this context, a Nash equilibrium describes the situation in which the steady-state expected waiting time resulting from certain arrival rates, will induce the same arrival rates if used as the waiting time announced to customers. A schematic of this system is shown in Figure 1.

To motivate the integration of a call-back option into a call center we illustrate its effect on performance using an example summarized in Table 1. The table compares via simulation a system without the call-back option with one that offers this option with a delay guarantee of 10 minutes. The latter employs a threshold routing rule outlined later on. Note that the steady-state expected waiting time in equilibrium is reduced by almost a factor of three, the probability of waiting more than 20 seconds (a typical measure in call centers) is reduced by a factor of six, while more traffic is being served by the system. Intuitively, this is happening because the customer group is segmented into two classes, one in which workload can be stored or delayed for future processing. This flexibility improves system performance. Finally, the total arrival rate into both systems is very close to the total service capacity of 50 calls/min., which numerically demonstrates that rational customer choice behavior "brings" the system into heavy traffic.

Figure 1. Schematic of a contact center.



Note. Class 1 corresponds to telephone service, and class 2 is for the call-back option. Λ is the aggregate arrival rate, λ_i is the arrival rate into class i.

Table 1. Performance comparisons.

	No Call-Back	Call-Back with Delay $\leqslant$ 10 Minutes
E(waiting time of class 1) (sec.)	14.0	5.4
Std(waiting time of class 1) (sec.)	21.04	11.68
Total actual arrival rate (arrivals/min.)	47.38	48.23
Fraction of customers who choose call-back	—	18%
P(waiting in class 1 $>$ 20 sec.)	0.25	0.04
P(balking)	0.033	0.016
P("late" response to class 2 customers)	—	0.013
E(waiting time for "late" call-backs) (min.)	—	10.68
Std(waiting time for "late" call-backs) (min.)	—	0.55

Note. $N = 50$ servers, overall potential arrival rate = 49 requests/min., 1 min. mean service time, 2 min. average patience time for telephone service, 20 min. average patience time for a call-back.

Our analysis throughout this paper is based on the many-server, heavy traffic asymptotic regime of Halfin and Whitt (1981). This limiting regime (i.e., the system approaching heavy traffic as the number of servers grows large) can be rigorously justified under very general assumptions on the customer choice behavior (§4). This asymptotic mode of analysis allows us to devise an asymptotically optimal routing policy, analytically characterize the unique equilibrium operating regime along with the system's performance, and propose an analytic expression for the appropriate staffing level for this system. In more detail, the limiting regime is used in the following three areas:

1. *Delay Specifications and Optimal Routing.* Due to the randomness of the service system, the delay constraint for class 2 customers cannot always be satisfied. In the limiting regime, however, the system manager can indeed guarantee that the waiting time encountered by class 2 customers never exceeds its upper bound. Analysis of the associated control problem yields a simple characterization of the optimal policy that minimizes the limiting expected waiting time experienced by class 1 customers among all policies that asymptotically guarantee the delay bound for class 2; these are the so-called *asymptotically compliant* policies (see §3 or Plambeck et al. 2001). This is a threshold rule that gives priority to class 1 as long as the class 2 queue length is below an appropriately chosen threshold that is easy to specify; beyond this threshold class 2 gets priority.

2. *Equilibrium Analysis.* It is common to use diffusion approximations for performance analysis of queueing systems. In this paper, we extend this idea and use the limiting regime to study the system's equilibrium behavior. Specifically, we establish that the limit system has a unique, stable equilibrium, which is characterized as the solution of a nonlinear equation. The limiting equilibrium is used

to approximate the equilibrium in the original system. This use of the limit system is novel and seems to provide structural insights and accurate numerical results. In contrast, direct analysis of the Markov chain model is very hard, and one has to rely on extensive simulation to estimate the system's equilibrium behavior—this is only feasible for small systems.

3. *System Design.* We focus on the staffing problem of choosing the minimum number of servers for this two-class system that guarantees a certain set of performance specifications; typical examples involve bounds on the expected waiting time for class 1 customers, bounds on the probability that the waiting time exceeds some threshold, etc. The asymptotic analysis provides a simple characterization of the number of servers needed to meet these specifications, and validates the familiar "square-root" staffing rule that has been advocated in the literature in the context of simpler service systems; see, for example, Kolesar and Green (1998) and Borst et al. (2004). This rule suggests that the number of servers should be of the form $N^* = R + x^*\sqrt{R}$, where R is the total load into the system, and x^* is an appropriate multiple that is directly computable from the specifications and the choice model parameters.

Methodologically, the novelty in this paper is in addressing these three points (control, equilibrium analysis, design) within this limiting regime, and appropriately translating its solutions into simple rules that may be used in the original system.

The remainder of this paper is structured as follows. We conclude this section with a literature survey. Section 2 describes the model. Section 3 provides an asymptotic analysis for the system as the number of agents grows to infinity and the traffic intensity grows to one; we call this the "rationalized" regime, which is typical of large call centers. Section 3 also gives basic results about the limiting system and derives the optimal scheduling policy. Section 4 analyzes the equilibrium behavior of the system, and justifies that the "rationalized" regime of §3 is indeed the natural equilibrium regime for the system under rational customer choice behavior. Section 5 exploits these asymptotic results to approximate system performance and proposes near-optimal staffing rules. Section 6 gives concluding remarks.

The literature on call centers is quite extensive. It starts with a plethora of results on the structure and performance of the $M/M/N$ system that can be found in most textbooks on stochastic models. A significant fraction of recent work has been on developing good staffing rules for large call centers (see Kolesar and Green 1998 and references therein) with particular emphasis on nonstationary arrival streams and the development of practical solutions that can be implementable in a true system with eight-hour shifts, breaks, etc. Typically, these papers assume a single class of customers, do not explicitly model customer choice behavior, and use direct analysis of the Markov chain to derive their results. While for the single-class model such an approach is feasible; it does not scale to multiclass systems, or, more so, to networks, and it often gives numerical rather than analytic characterizations of quantities of interest. Also within the nonstationary arrivals framework, Whitt (1999) demonstrates how postponing the service of less urgent jobs to off-peak hours can smooth out the load of the system.

Two-class $M/M/N$ systems related to ours have been studied in the following two papers. Brandt and Brandt (1999) analyze the Markov chain associated with a two-class system with impatient customers, fixed arrival rates, no service level guarantees, and a static priority policy that prioritizes the real-time channel (class 1). They characterize the steady-state distribution for the number of high priority customers in the system via a set of integral equations, and give a similar approximation for the low priority class. Gans and Zhou (2003) study a two-class system, where one class has a fixed arrival rate and is subject to a probabilistic service level guarantee, and the other class is an infinitely backlogged queue that awaits to be processed. They develop a routing policy that gives priority to the class with the service level guarantee and only serves the other class when the number of idle servers is above a certain threshold. This policy maximizes the throughput of the infinite backlogged class subject to the quality-of-service constraint of the incoming call channel. This model is probably the closest to ours. The main difference in assumptions lies in the fact that in their model it is assumed that arrival rates are exogenous (with one infinitely backlogged queue); in contrast, we assume endogenous arrival rates, which are the result of system equilibrium.

A paper that has motivated a lot of recent work on asymptotic methods for the analysis of multiserver systems is due to Halfin and Whitt (1981). In their paper, they perform an asymptotic analysis for $M/M/N$ systems in the form of a simple diffusion process. It also provides useful insights about the scaling phenomena in these systems. This work has been extended in several ways: Jennings et al. (1996) seem to be the first to have used the Halfin-Whitt regime in the context of call centers; they use this asymptotic regime to characterize staffing levels with time-varying demand. Fleming et al. (1994) and Garnett et al. (2002) have added the notion of abandonment (that is, customers renege after having waited in the queue for some time). Puhalskii and Reiman (2000) have analyzed multiclass systems with renewal arrival and phase-type processing time distributions with and without priorities. A detailed asymptotic analysis of dimensioning rules for single-class call centers has been done by Borst et al. (2004). Finally, in a recent paper, Whitt (2003) studies single-class multiserver systems in which arrival rates depend on system performance, including the scenario that leads to the limiting regime proposed by Halfin and Whitt (1981) that is also used here.

Customer behavior has been analyzed by Hassin and Haviv (1995), Mandelbaum and Shimkin (2000), and

recently by Zohar et al. (2002) in the context of modeling rational abandonment. In Hassin and Haviv (1995) rational abandonments are considered in an $M/M/1$ queue, and it is shown that the only rational abandonments are the trivial ones, that is, those that occur upon a customer's arrival, or once the service is no longer needed by him or her. Mandelbaum and Shimkin (2000) extend this result to a call center (modeled as an $M/M/N$ queue), and propose a realistic modeling framework under which nontrivial abandonments occur in equilibrium. Zohar et al. (2002) analyze a model of rational abandonments in which customers' patience depends on specific performance measures such as the expected waiting time, rather than the whole waiting time distribution. It also addresses the learning process of the customers and how the system evolves into equilibrium.

Our work combines rational decision making with an asymptotic analysis in the Halfin-Whitt regime. Similar to Mandelbaum and Shimkin (2000), customer behavior is explained through a probabilistic choice model and the equilibrium regime is analyzed. The control problem solved in this paper is related to that of Brandt and Brandt (1999) and Gans and Zhou (2003). In contrast to these papers, our asymptotic mode of analysis yields simple but accurate closed-form characterizations of the system equilibrium and performance, that are then used in the economic analysis of the system. In common with Puhalskii and Reiman (2000) we analyze a two-class system, under considerably different policies, however, and establish a state space collapse result (see Proposition 3.1). The issue of abandonment is not addressed in this paper. Another paper by the authors (Armony and Maglaras 2004) studies a related problem of a contact center with a call-back option in which customers are informed of their state-dependent anticipated delay.

2. The Model and the Routing Problem

The service system has N identical servers and provides two types of service: (1) *real-time* service, where users join a FIFO queue (queue 1, here); and (2) *postponed* (call-back) service, where users leave a message and the system calls them back within D_2 time units (this is queue 2). We assume that once a class 2 customer leaves a service request he/she is available until he/she gets called back. Clearly, the upper bound on the waiting time for class 2 service requests is not meaningful in a conventional sense, because the waiting times are unbounded random variables. However, as we will show later on, this constraint can be guaranteed in an appropriate asymptotic regime that characterizes the equilibrium behavior of the system as the number of servers grows large. Both classes have identical processing requirements, and service times are independent, exponential random variables with mean m (and rate $\mu = 1/m$).[1] The system parameters N and D_2 are assumed to be fixed; §5.2 will address the service provider's problem of system design to optimize a profitability criterion.

Customer Behavior. Customers arrive according to a Poisson process with rate Λ. Upon arrival they have three *choices*: (1) join queue 1 and wait to be processed, (2) leave a message for postponed service in queue 2, or (3) balk and do not join the system. We denote by λ_1, λ_2, λ_0 the rates at which customers join class 1, class 2, or balk, respectively. Clearly, $\Lambda = \lambda_1 + \lambda_2 + \lambda_0$.

Given λ_1, λ_2, let W_i denote the steady-state waiting time for class i jobs (this does not include their service time), and let $\mathbf{E}W_i$ denote the respective expected values. Arriving customers are informed of (a) the steady state expected waiting time in class 1, $\mathbf{E}W_1$, and (b) the delay D_2 within which they will receive a call-back should they select option (2).[2] Based on their knowledge of $(\mathbf{E}W_1, D_2)$, they decide whether to join the system and what type of service to request. That is, customers use long-run average information to assess their utility for real-time service (class 1), and the guaranteed upper bound on anticipated delay for postponed (call-back) service for the latter.

The key trade-off callers are faced with choosing between the real-time service and the postponed one is analogous to the trade-off between "best effort" and "guaranteed" type of service. In particular, it is to be expected that some customers may choose to get a call-back to free their time (as well as their phone line) to attend to other matters. This may be true even if $\mathbf{E}W_1 \ll D_2$. The call-back option may be even more attractive if customers are reliably being called back within the promised deadline. Finally, the call-back option can also be perceived as an e-mail option in contact centers that offer both types of services (phone and e-mail); in such cases, some customers may naturally prefer the e-mail service, even though the corresponding response time is significantly longer.

A Mathematical Model of Choice Behavior. We assume that there is a continuum of customer types, indexed by τ, that are differentiated by their preferences; hereafter, a superscript τ will denote dependence on user type. User preferences are determined as follows:

(a) The utility for real-time service (i.e., choose (1) and join queue 1) is $u_1^\tau(\mathbf{E}W_1)$.

(b) The utility for leaving a request for a call-back within D_2 time units is $u_2^\tau(D_2)$.

Both $u_1^\tau(\cdot)$, $u_2^\tau(\cdot)$ are continuously differentiable with respect to τ, W_1, D_2 are nonincreasing in W_1, D_2, respectively, and $u_i^\tau(\infty) < 0$; i.e., the utility has negative values if $\mathbf{E}W_1$ or D_2 are sufficiently large because in such cases the cost of waiting exceeds the value obtained by receiving service, making such choices undesirable. For $i = 1, 2$, $u_i^\tau(0)$ represents the utility for receiving immediate service (no wait), which gets depreciated as the customer has to wait either online $(i = 1)$ or offline to be called back $(i = 2)$. Without loss of generality, we assume that the utility of not joining is zero; that is, $u_0 = 0$. Customers choose the type of service that maximizes their own utility according to

$$\max\{0, u_1^\tau, u_2^\tau\}; \tag{1}$$

that is, a type τ customer will join queue 1 if $u_1^\tau \geqslant u_2^\tau$ and $u_1^\tau \geqslant 0$, leave a service request in queue 2 if $u_2^\tau > u_1^\tau$ and $u_2^\tau \geqslant 0$, and balk if $u_1^\tau, u_2^\tau < 0$. In principle, the utilities may depend on the entire distributions of W_1 and W_2, however, this does not appear to be very realistic (due to bounded rationality arguments). We will make the simplifying assumption that $u_1^\tau(\cdot)$ and $u_2^\tau(\cdot)$ are only functions of $\mathbf{E}W_1$ and D_2, respectively, and that $\mathbf{P}^\tau(u_1(0,\tau) > u_2(0,\tau)) > 0$ and that $\mathbf{P}^\tau(u_2(0,\tau) > u_1(0,\tau)) > 0$. The interpretation of the latter condition is that the two service modes are not perfect substitutes of each other, and this is reflected in their respective utilities.

Finally, the customer type is a random variable. Let P^τ be the probability distribution over the set of customer types, which for simplicity is assumed to be the positive real line. We require that the type distribution has a continuous density function and that for all finite $x \geqslant 0$, $\mathbf{P}^\tau(u_i^\tau(x) \geqslant 0) > 0$. The type of each customer is chosen according to P^τ and is independent of all other customers' types. Given this setup,

$$\lambda_1(\mathbf{E}W_1, D_2) = \Lambda \mathbf{P}^\tau(u_1^\tau(\mathbf{E}W_1) \geqslant u_2^\tau(D_2) \text{ and } u_1^\tau(\mathbf{E}W_1) \geqslant 0), \tag{2}$$

$$\lambda_2(\mathbf{E}W_1, D_2) = \Lambda \mathbf{P}^\tau(u_2^\tau(D_2) > u_1^\tau(\mathbf{E}W_1) \text{ and } u_2^\tau(D_2) \geqslant 0), \tag{3}$$

and $\lambda_0(\mathbf{E}W_1, D_2) = \Lambda - \lambda_1(\mathbf{E}W_1, D_2) - \lambda_2(\mathbf{E}W_1, D_2)$. For $i = 1, 2$ we assume that $\lambda_i(\cdot,\cdot)$ is continuously differentiable with respect to both arguments, and that $\lambda_i(0,0) > 0$. In addition, we assume that the total aggregate arrival rate into the system, $\lambda_a(\mathbf{E}W_1, D_2) = \lambda_1(\mathbf{E}W_1, D_2) + \lambda_2(\mathbf{E}W_1, D_2)$ is strictly decreasing in both arguments. Finally, we define $\Lambda_{\text{eff}} := \lambda_1(0,0) + \lambda_2(0,0)$ to be the maximal overall (effective) arrival rate into the system which is achieved when both $\mathbf{E}W_1 = 0$ and $D_2 = 0$.

Example. The Multinomial Logit Model. A simple example, which is used later on for illustrative purposes, is one with linear waiting costs and the specific structure of the *Multinomial Logit Model* (*MNL*) (see Anderson et al. 1996, §2.6). Neither of these two assumptions is necessary for our analysis, but will be used in numerical examples. In this case, for appropriate constants r_i and c_i, we have

$$u_1 = r_1 - c_1 \mathbf{E}W_1 \quad \text{and} \quad u_2 = r_2 - c_2 D_2,$$

and the total utility for each choice is given by $u_i^\tau = u_i + \tau_i$, where τ_i are IID double exponential (Gumbel) distributed with parameter ν; that is, τ_i is the random part of the utility that differentiates among customer types. Following earlier remarks, it is natural (but not necessary) to consider parameters such that $r_1 \geqslant r_2$ and $c_1 > c_2$. For the MNL model (2) and (3) simplify to

$$\lambda_i = \Lambda \frac{e^{u_i/\nu}}{\sum_{j=0,1,2} e^{u_j/\nu}}. \tag{4}$$

Although we do not necessarily advocate the use of the MNL model for call-center applications, we use it here as an example due to its simplicity.

The Equilibrium Model. Putting it all together, the model we analyze is a two-class $M/M/N$ system. The arrivals into classes 1 and 2 are Poisson with rates $\lambda_1(\mathbf{E}W_1, D_2)$ and $\lambda_2(\mathbf{E}W_1, D_2)$. The service rate for both classes is μ. The service manager has control with respect to routing decisions (i.e., whether to process class 1 or class 2 customers) at each server. The interplay between arrival rates and expected waiting times, and the dependence of both on the routing rule employed, may be viewed as a game in strategic form. In particular, in equilibrium, the expected waiting time for class 1 when the arrival rates are $\lambda_i(\mathbf{E}W_1, D_2)$, will be $\mathbf{E}W_1$.[3]

Stability. We will assume that D_2 is sufficiently large such that $\lambda_2(\infty, D_2) = \Lambda P^\tau(u_2^\tau(D_2) \geqslant 0) < N\mu$. In this case, the system has enough capacity to at least cope with the arrivals into queue 2 (by giving them preemptive priority) and is therefore stabilizable. Of course, this condition does not guarantee that the delay constraint for these customers is met, and in practice $\lambda_2(\infty, D_2)$ should be considerably smaller than $N\mu$ for the waiting time encountered by class 2 customers to be close to or smaller than D_2.

The Queue Length Threshold Control Policy. The system manager (SM) has discretion as to the routing of jobs to the various servers. In the model specified above a routing rule takes the form of an allocation rule $\{(T_1(t), T_2(t)),\ t \geqslant 0\}$, where $T_i(t)$ is the cumulative time allocated into processing class i jobs in $[0, t)$. The policy should also be nonanticipating; roughly speaking, this says that decisions made at time t only use information that becomes available in $[0, t]$. It is reasonable to expect (although not required) that a policy will be nonidling in the sense that servers may idle only when both queues are empty.

Treating the service network as a profit center, the SM should choose a policy that maximizes the aggregate arrival rate into the system, which is positively correlated with the profit (or value) it generates. Given that the aggregate arrival rate is decreasing in $\mathbf{E}W_1$, a naive problem formulation would be to choose a nonpreemptive routing policy to

$$\begin{array}{ll} \text{minimize} & \mathbf{E}W_1 \\ \text{subject to} & W_2 \leqslant D_2. \end{array} \tag{5}$$

This is, of course, meaningless, because the waiting time for class 2 customers is a random variable for which the upper bound constraint cannot be guaranteed. One way to pose a well-defined control problem is to replace the upper bound constraint by a probabilistic one of the form $\mathbf{P}(W_2 > D_2) \leqslant \epsilon$ for some small $\epsilon > 0$. Instead, we take a different approach. Our analysis in §3 focuses on an asymptotic regime that corresponds to systems with many servers operating close to the heavy traffic regime; that is, where the aggregate arrival rate is close to the system's processing capacity $N\mu$. In this regime, the problem formulation in (5) becomes meaningful, and the constraint $W_2 \leqslant D_2$ can

be guaranteed almost surely. This is related to the idea of asymptotic compliance that ensures that as N grows large, class 2 tardiness becomes negligible; details are given in §3.

Let $Q_i(t)$ be the number of class i jobs in queue (but not in service) at time t. The solution of (5) in this asymptotic regime suggests the use of the following nonpreemptive, head-of-line policy:

THRESHOLD RULE. *If* $Q_2(t) \geqslant \lambda_2 D_2$, *give priority to class* 2, *otherwise give priority to class* 1.

The intuition behind this policy is as follows. Note that the average number of class 2 customers that arrive into queue 2 in D_2 time units is $\lambda_2 D_2$. Our threshold routing rule attempts to keep the class 2 queue length less than or equal to $\lambda_2 D_2$. This, in turn, implies that on average queue 2 will comprise of customers that have arrived in the past D_2 time units, and hence the delay constraint will tend to be satisfied. This connection between the queue length and delay experienced by jobs in the queue is based on an observation made in Maglaras and Van Mieghem (2004). In §3 we prove that in an appropriate asymptotic regime (where the number of servers grows large) this policy always satisfies the delay constraints and is optimal for the problem in (5).

Many other policies could also be considered, including ones that make routing decisions using the age information of the jobs in the system. However, given the asymptotic optimality and simplicity of threshold policies, other alternatives will not be pursued here.

We conclude our model description with two remarks. First, the total arrival rate Λ is assumed to be fixed. This is not realistic, because in most service systems, such as call centers, there is a very pronounced variation depending on time-of-day, day-of-week, promotional offers, etc. If the nonstationarity is slowly varying relative to the system dynamics, then such systems have been typically analyzed using a pointwise stationary approximation, where the performance at time t is approximated by the steady-state performance of the stationary system with constant arrival rates given by $\lambda(t)$; see Green and Kolesar (1991) and Jennings et al. (1996). Also, the approach taken by Whitt (1999) of differentiating between calls according to their level of urgency, and postponing the less pressing calls to off-peak hours to alleviate the load during busier times, may be adapted to our framework. Here, class 2 calls will naturally be considered the less urgent ones. This extension will not be pursued here.

Second, the primitive data for this model are the total arrival rate Λ, the service rate μ, the distribution of types P^τ, and the utility functions u_i^τ. It is relatively simple to estimate Λ and μ. An interesting problem that will not be broached here is the estimation of customer preferences using observed data. For the general model described above, this estimation procedure is very difficult. In practice, one should use a family of parameterized choice models that are validated through experimental evidence and for which efficient estimation procedures can be constructed. For example, the parameters of the MNL choice model are simple to estimate (see Talluri and van Ryzin 2004 and the references therein). On the positive side, as the number of agents increases and the traffic intensity approaches one, the actual waiting times encountered by both classes of customers will decrease to zero, and thus the customer choice behavior can be approximated (through a Taylor expansion that is derived in §4) by a linear model of the form $\lambda_1(\mathbf{E}W_1) = \lambda_1(0) - \kappa \mathbf{E}W_1$, whose parameters can easily be estimated using real data.

3. Asymptotic Analysis for Systems with Many Servers and Exogenously Given Demand

Despite the fairly simple model described above, explicit analysis of the Markov chain is very involved, and the equilibrium regime can only be computed via exhaustive simulation. Instead, focusing on large systems (i.e., systems with many agents), which are typical in the context of modern contact centers, we pursue an "appropriate" asymptotic analysis that is tractable and becomes accurate as the number of agents grows large. This section motivates and develops an approximating model for the original system with a large number of servers for the simple case where the arrival rates are exogenously given (i.e., do not depend on $\mathbf{E}W_1$, D_2). We also highlight the natural scaling relations that prevail in such systems and provide the basic building block for the analysis of the system's equilibrium behavior, undertaken in the next section.

Operating Regimes. To motivate the subsequent analysis we start by identifying the "physical" modes of operation for the system. The quantity that we focus on is the probability that a randomly selected customer arriving to the system will have to wait before getting served. Following the taxonomy given in Garnett et al. (2002), we consider three modes of operation.

- **Cost-driven regime:** The system is undercapacitated and customers almost always wait, $\mathbf{P}(\text{wait} > 0) \approx 1$.
- **"Rationalized" regime:** The system's capacity is *balanced* and customers may have to wait but not always, $\mathbf{P}(\text{wait} > 0) \approx \alpha \in (0, 1)$. We also refer to this as the *Halfin-Whitt regime*.
- **Quality-driven regime:** The system is overcapacitated and customers almost never wait, $\mathbf{P}(\text{wait} > 0) \approx 0$.

The cost-driven regime underemphasizes congestion effects, the quality-driven regime focuses on service quality, while the "rationalized" regime achieves a balance between operating costs and quality of service. As advocated in Garnett et al. (2002) this seems to be the natural operating regime to consider. Borst et al. (2004) have shown in the context of a single-class model that the economically optimal capacity level, trading off congestion and staffing costs, puts the system in this regime. Finally, recent work by Maglaras and Zeevi (2003) has shown that for a related

model with pricing decisions this is the revenue maximizing and socially optimal regime.

The remainder of this section analyzes the system of interest in the rationalized regime, and under the simplifying assumption that the arrival rates into each class are exogenously given, i.e., do not depend on $\mathbf{E}W_1$, D_2. This will build the required background to address the equilibrium analysis in §4, where we also prove that the natural system equilibrium places the system in the rationalized regime.

The Halfin-Whitt Regime. Consider a system with N servers. Let $Q_i^N(t)$ denote the number of jobs in queue i at time t, and let $Z_i^N(t)$ be the total number of class i jobs present in the system (i.e., in queue or in service) at time t; superscript N will be attached to all relevant quantities to denote their dependence on the size of the system. The arrival rates are λ_1^N and λ_2^N. Define the aggregate arrival rate by $\lambda_a^N = \lambda_1^N + \lambda_2^N$. It is easy to see that the evolution of the total number of customers in the system, given by $Z_1^N(t) + Z_2^N(t)$, behaves precisely like that of an $M/M/N$ system with arrival rate λ_a^N. This is independent of the specific details of the routing rule, provided that it is nonidling. Hereafter, the notation Z_i^N and Q_i^N without time argument will denote the steady-state random variable. Also, the notation "$\Rightarrow$" is used to denote weak convergence in $D[0,\infty)$ (see, e.g., Billingsley 1968, §§14 and 15), or convergence in distribution, and a tilde denotes that the relevant quantity is associated with the limiting system. Halfin and Whitt established the following results (see Halfin and Whitt 1981, Proposition 1, Theorems 2 and 3).

THEOREM 3.1. *As* $N \to \infty$,

$$\lim_{N\to\infty} \mathbf{P}(wait > 0) = \lim_{N\to\infty} \mathbf{P}(Z_1^N + Z_2^N > N) = \alpha \in (0,1), \qquad (6)$$

if and only if

$$\rho^N := \frac{\lambda_a^N}{N\mu} = 1 - \frac{\beta}{\sqrt{N}} + o\left(\frac{1}{\sqrt{N}}\right), \quad \beta > 0 \qquad (7)$$

(i.e., $\sqrt{N}(1-\rho^N) \to \beta$ *as* $N \to \infty$*) where* $\alpha = [1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}]^{-1}$ *and* $\Phi(\cdot)$ *is the standard normal cumulative distribution function. Assume that* (6) *or* (7) *hold and define*

$$X^N(t) = \frac{(Z_1^N(t) + Z_2^N(t)) - N}{\sqrt{N}}.$$

Then, if $X^N(0) \Rightarrow \widetilde{X}(0)$, $X^N(\cdot) \Rightarrow \widetilde{X}(\cdot)$, *where* $\widetilde{X}(\cdot)$ *is a one-dimensional diffusion process with infinitesimal drift* $m(x)$ *given by*

$$m(x) = \begin{cases} -\mu\beta, & x \geqslant 0, \\ -\mu(x+\beta), & x < 0, \end{cases}$$

and constant infinitesimal variance 2μ. *For* $\beta > 0$, *the steady-state distribution of* $\widetilde{X}(\cdot)$ *is given by* $\mathbf{P}(\widetilde{X} > 0) = \alpha$, $\mathbf{P}(\widetilde{X} > x \mid \widetilde{X} > 0) = e^{-x\beta}$, $x > 0$, *and* $\mathbf{P}(\widetilde{X} \leqslant x \mid \widetilde{X} \leqslant 0) = \Phi(\beta + x)/\Phi(\beta)$, $x \leqslant 0$.

The interpretation of the process $X^N(\cdot)$ is as follows: when $X^N(t) > 0$, it is equal to the scaled total number of jobs in both queues, whereas when $X^N(t) < 0$, it is equal to the scaled number of idle servers in the system. This result highlights the natural scaling relations that prevail in this many-server asymptotic regime. Namely, for systems with balanced capacity (that is, when the system is neither systematically underutilized nor overutilized) the natural scale that emerges is of order $\sqrt{N}$. Specifically, the total number of customers in the system is approximately $Z^N = N + \sqrt{N}X$, which implies that both queue lengths or the total number of idle servers in the system are of order $\sqrt{N}$. In addition, it implies that the waiting times encountered in the system are of order $1/\sqrt{N}$; N servers take $\mathcal{O}(1/\sqrt{N})$ to clear a backlog of $\mathcal{O}(\sqrt{N})$ jobs. This has an important design implication on the choice of the upper bound for class 2 delay D_2^N; in particular, it is plausible to assume that it scales according to

$$D_2^N = \frac{\widetilde{D}_2}{\sqrt{N}} \qquad (8)$$

for some appropriate value of $\widetilde{D}_2 > 0$. This choice makes the delay guarantee to be of the right order of magnitude relative to the actual waiting times encountered in the system.

Finally, these scaling relations highlight the fact that large systems that operate close to heavy traffic can still offer a high-quality service; the waiting times go to zero even as the traffic intensity grows to one. This provides some explanation about the "optimality" of the rationalized regime; see Borst et al. (2004) and Maglaras and Zeevi (2003) for details of related results in various settings.

Analysis of the Two-Class System in the Halfin-Whitt Regime. Next, we study the asymptotic class-level behavior of the two-class system, as it approaches the Halfin-Whitt regime. In particular, consider a sequence of systems indexed by the number of servers N with aggregate arrival rate λ_a^N of the form

$$\lambda_a^N = N\mu - \beta\sqrt{N}\mu. \qquad (9)$$

In addition, we assume that the arrival rates for each customer class is exogenously given, according to $\lambda_1^N = \eta\lambda_a^N$ and $\lambda_2^N = (1-\eta)\lambda_a^N$ for some $\eta \in (0,1)$. For example, consider the system simulated in Table 1 that has 50 servers, $\mu = 1$ (per min.), $D_2 = 10$ min., and arrival rates $\lambda_1 = 39.5$ and $\lambda_2 = 8.7$ (in requests/min.). This makes $\lambda_a = 48.2$, $\beta := (50\mu - \lambda_a)/\sqrt{50}\mu = 0.25$, $\eta := \lambda_1/\lambda_a = 0.82$, and $\widetilde{D}_2 := D_2\sqrt{N} = 10\sqrt{50} = 70.7$.

The following results describe the asymptotic behavior of the two-class system as $N \to \infty$, D_2^N and λ_a^N scale according to (8) and (9) respectively, and β and η stay constant. (This implies that $\sqrt{N}(1-\rho^N) \to \beta$, as required in (7).) The limiting service capacity is $N\mu/N = \mu$ and the limiting arrival rates are $\tilde{\lambda}_1 := \lim_{N\to\infty} \lambda_1^N/N = \eta\mu$

and $\tilde{\lambda}_2 := \lim_{N\to\infty} \lambda_2^N/N = (1-\eta)\mu$. Recall that routing decisions are made according to the queue-length threshold policy defined earlier: if $Q_2^N < \theta^N$, the high priority class is 1; otherwise, it is class 2. Given the scalings outlined above, one can see that the appropriate size of the threshold θ^N is also of order $\sqrt{N}$. Specifically, we consider threshold parameters of the form

$$\theta^N = \tilde{\theta}\sqrt{N} \tag{10}$$

for some $\tilde{\theta} > 0$. It is worth noting that the threshold does not vanish in the limit system. Indeed, $\tilde{\theta}$ is integral in the description of the limiting policy that specifies when the system manager should switch priorities. Its value will be selected later on to ensure that in the limit system all class 2 customers commence service within $\widetilde{D}_2$ time units. Our use of the term "threshold policy" is different from that appearing in, for example, Bell and Williams (2001), where thresholds are synonymous for safety stocks (see also Kelly and Laws 1993, Harrison 1996, Maglaras 2000, Teh and Ward 2002) that are asymptotically negligible and are used to prevent the system from incurring undesirable idleness when some of the queue lengths get depleted.

Irrespective of the choice of θ^N, this policy is nonidling, and thus satisfies the assumptions and properties of the Halfin-Whitt result (Theorem 3.1 above). Our first proposition focuses on the behavior of the queue lengths for class 1 and class 2 customers.

PROPOSITION 3.1 (STATE SPACE COLLAPSE). *Assume that*

$$\left(\frac{Q_1^N(0)}{\sqrt{N}}, \frac{Q_2^N(0)}{\sqrt{N}}\right) \to ((\widetilde{X}(0)-\tilde{\theta})^+, \widetilde{X}(0)^+ \wedge \tilde{\theta})$$

in probability. Then, for every $t \geqslant 0$, *as* $N \to \infty$,

$$X_1^N(t) := \frac{Q_1^N(t)}{\sqrt{N}} \;\Rightarrow\; (\widetilde{X}(t)-\tilde{\theta})^+ \quad \textit{and}$$
$$X_2^N(t) := \frac{Q_2^N(t)}{\sqrt{N}} \;\Rightarrow\; \widetilde{X}(t)^+ \wedge \tilde{\theta}.$$

(All proofs are given in the Appendix.) The limit queue-length processes will be denoted by $\widetilde{X}_1(\cdot)$ and $\widetilde{X}_2(\cdot)$. The first observation is that the class-level queue lengths can be expressed solely in terms of the limit of the one-dimensional total queue-length process, which is well defined (see Theorem 3.1); the name "state space collapse" reflects this dimension reduction. Intuitively, this result hinges on the observation that asymptotically the class 2 queue length cannot exceed $\tilde{\theta}$. Indeed, when $Q_2^N(t) = \theta^N$, class 2 gets higher priority, while class 2 jobs are arriving at a rate λ_2^N and servers are becoming available at a rate of $N\mu$ (because all of them are busy). In the limit, servers become available much faster than the rate at which class 2 customers arrive into the system, and hence the limiting class 2 queue length always stays below the threshold $\tilde{\theta}$. It follows that if $\widetilde{X}(t)^+ > \tilde{\theta}$, the remaining jobs $(\widetilde{X}(t)-\tilde{\theta})^+$ must be held in queue 1. If $\widetilde{X}(t)^+ < \tilde{\theta}$, then $\widetilde{X}_2(t) < \tilde{\theta}$, and by an analogous argument one could show that $\widetilde{X}_1(t) = 0$ and $\widetilde{X}_2(t) = \widetilde{X}(t)^+$.

Let $W_i^N(t)$ denote the virtual waiting time for class i jobs at time t (this is the time a virtual class i customer would have to wait if he/she arrived at time t). Given that queue lengths are of order $\mathcal{O}(\sqrt{N})$ and there are N servers processing customers, the waiting times are of order $\mathcal{O}(1/\sqrt{N})$ and in the limit they appear instantaneous. Hence, $\widetilde{X}^+(\cdot)$ and $\widetilde{X}_i(\cdot)$ stay constant over this infinitesimal time—this is the so-called *snapshot principle*, first derived by Reiman (1984)—and similarly to Little's law the following holds:

PROPOSITION 3.2. *For every* $t \geqslant 0$, *as* $N \to \infty$,

$$\begin{aligned}
&\sqrt{N}W_1^N(t) \;\Rightarrow\; \widetilde{W}_1(t) = \frac{(\widetilde{X}(t)-\tilde{\theta})^+}{\tilde{\lambda}_1} \quad \textit{and}\\
&\sqrt{N}W_2^N(t) \;\Rightarrow\; \widetilde{W}_2(t) = \frac{\widetilde{X}(t)^+ \wedge \tilde{\theta}}{\tilde{\lambda}_2}.
\end{aligned} \tag{11}$$

A direct consequence of this proposition is that in steady state,

$$\begin{aligned}
\mathbf{E}\widetilde{W}_1 &= \frac{1}{\tilde{\lambda}_1}\mathbf{E}(\widetilde{X}-\tilde{\theta})^+ = \frac{1}{\tilde{\lambda}_1}\alpha\int_{\tilde{\theta}}^{\infty}(x-\tilde{\theta})\beta e^{-\beta x}\,dx\\
&= \frac{1}{\tilde{\lambda}_1}\frac{\alpha}{\beta}e^{-\beta\tilde{\theta}}.
\end{aligned} \tag{12}$$

Because we are interested in approximating the steady-state behavior of the original N-server system using the steady-state behavior of the limiting diffusion, we need to establish a limiting relationship between the two. While typically weak convergence results of the underlying processes (as in Propositions 3.1 and 3.2) need not imply convergence of associated steady-state quantities, similarly to Halfin and Whitt (1981), such a result can be established for our model; see Whitt (1974) for a discussion of this issue.

PROPOSITION 3.3. *For* $i = 1, 2$, *let* W_i^N *denote the class* i *steady-state waiting time associated with the* N*-server system, and let* $\widetilde{W}_i$ *be the steady state of class* i *limiting waiting time process,* $\widetilde{W}_i(\cdot)$. *Then,*

$$\sqrt{N}W_i^N \;\Rightarrow\; \widetilde{W}_i \qquad \textit{as } N\to\infty \textit{ for } i = 1, 2, \tag{13}$$

and

$$\sqrt{N}\mathbf{E}W_i^N \to \mathbf{E}\widetilde{W}_i \qquad \textit{as } N\to\infty \textit{ for } i = 1, 2. \tag{14}$$

The scaling behavior of W_2^N is consistent with that of D_2^N in (8). In the limiting regime where $N \to \infty$, the requirement of delay compliance is strengthened to one where *all* class 2 requests receive service within $\widetilde{D}_2$ time units; i.e., $\widetilde{W}_2(t) \leqslant \widetilde{D}_2$ for all $t \geqslant 0$. The following definition is adapted from Plambeck et al. (2001).

DEFINITION 3.1. A policy π is said to be asymptotically compliant if

$$\left[\sqrt{N}W_2^{N,\pi}(t) - \widetilde{D}_2\right]^+ \Rightarrow 0,$$

where the superscript π denotes the dependence of $W_2^N(\cdot)$ on the policy.

We now turn to the problem of choosing the right threshold $\tilde{\theta}$ that will guarantee the delay constraint in the limiting system; that is, we look for a value of $\tilde{\theta}$ that will make the threshold policy asymptotically compliant. Let $A_i^N(t)$ be the number of customers that have arrived into queue i in $[0, t]$. Using an observation made in Maglaras and Van Mieghem (2004) we get

$$W_2^N(t) \leqslant D_2^N \quad \forall t \;\Leftrightarrow\; Q_2^N(t) \leqslant A_2^N(t) - A_2^N(t - D_2^N) \quad \forall t;$$

i.e., no class 2 customer has been waiting for more than D_2^N if all the customers currently in queue 2 arrived within the last D_2^N time units. Scaling both sides appropriately and taking the limit as $N \to \infty$, we obtain

$$\widetilde{W}_2(t) \leqslant \widetilde{D}_2 \quad \forall t$$
$$\Leftrightarrow\; \widetilde{X}_2(t) \leqslant \lim_{N\to\infty} \frac{A_2^N(t) - A_2^N(t - D_2^N)}{\sqrt{N}} = \tilde{\lambda}_2 \widetilde{D}_2 \quad \forall t.$$

Given that $\widetilde{X}_2(t) \leqslant \tilde{\theta}$ for all $t \geqslant 0$, and that this upper bound is achieved in the limiting system with probability one, we need to set

$$\tilde{\theta} = \tilde{\lambda}_2 \widetilde{D}_2. \tag{15}$$

Note that this is consistent with the scaling relations in place, because $\theta^N = \lambda_2^N D_2^N \approx \sqrt{N}\tilde{\lambda}_2\widetilde{D}_2$.

PROPOSITION 3.4. *Consider any nonidling, nonpreemptive, asymptotically compliant policy π, and assume that the limit queue-length and waiting time processes, $\widetilde{X}_i^\pi(\cdot)$, $\widetilde{W}_i^\pi(\cdot)$, exist. Let π^* denote the threshold policy with $\tilde{\theta} = \tilde{\lambda}_2\widetilde{D}_2$. Then we have*

$$\widetilde{W}_1^{\pi^*}(t) \leqslant \widetilde{W}_1^{\pi}(t) \quad \forall t \geqslant 0 \;\; \textit{w.p. } 1.$$

That is, the threshold rule defined through (15) is pointwise optimal among all nonidling policies that converge to some well-defined limit, and satisfy the delay constraint $\widetilde{W}_2(\cdot) \leqslant \widetilde{D}_2$. Clearly, the threshold policy also minimizes $\mathbf{E}\widetilde{W}_1$ subject to $\widetilde{W}_2(t) \leqslant \widetilde{D}_2$ for all t. A few comments are in place regarding the restriction of Proposition 3.4 to nonidling policies. In the N-server system, intentional idling may be desirable—or even optimal—to reserve capacity for the call-back customers that need to satisfy a quality-of-service constraint (e.g., see the policy derived by Gans and Zhou 2003). In contrast, in the limiting system, this quality-of-service constraint can be satisfied without incurring any idleness, which motivated the restriction to nonidling policies. The optimality of the threshold policy in the limiting model could be established in a richer setting that allows for intentional idling. This would make the analysis significantly more complex without generating additional practical insights, and will not be pursued here.

4. Equilibrium Analysis in the Many-Server Regime

This section studies the original system of interest, where customer choices depend on system performance. This will involve an equilibrium analysis. While a direct analysis of the system equilibrium using the associated Markov chain is theoretically possible, the expressions one gets are so complicated that it is hard to proceed either analytically or numerically.

Our approach uses the asymptotic results described above to analyze the system's equilibrium behavior. Given a system with N servers, the first step is to derive the appropriate limiting model. Assuming for now the validity of the rationalized regime, this amounts to finding the right parameters for the limiting system, much like the calculation of β, η done earlier. The second step is to study this limit system and show that it has a unique and stable equilibrium operating point that is simple to characterize and to evaluate numerically. Finally, to justify this analysis, we prove that the rationalized regime emerges as the natural equilibrium point in large contact centers. In §5 we use this characterization of equilibrium behavior to get analytic performance approximations for the N-server system, and to propose simple rules for dimensioning such systems.

We use the notation $\approx$ to denote an equality to within a quantity that is $o(\sqrt{N})$. Let $\mathbf{E}W_1^N$ and $\mathbf{E}\widetilde{W}_1$ (or in shorthand notation w_1^N and $\tilde{w}_1$) denote the steady-state expected waiting time for class 1 customers in equilibrium in the N-server and the limiting system, respectively. Proposition 3.3 implies that under the rationalized regime and for large N, $w_1^N = \tilde{w}_1/\sqrt{N} + o(1/\sqrt{N})$. Using this and (8) we first approximate the asymptotic arrival rates into each service class by invoking a Taylor expansion as follows:

$$\begin{aligned}
\lambda_1^N(w_1^N, D_2^N) &= \Lambda^N \mathbf{P}^\tau(u_1^\tau(w_1^N) \geqslant u_2^\tau(D_2^N)^+) \\
&= \Lambda^N \mathbf{P}^\tau\left(u_1^\tau\left(0 + \frac{\tilde{w}_1}{\sqrt{N}} + o\left(\frac{1}{\sqrt{N}}\right)\right)\right. \\
&\qquad \left. \geqslant u_2^\tau\left(\frac{\widetilde{D}_2}{\sqrt{N}}\right)^+\right) \qquad (16) \\
&\approx \Lambda^N \left[\mathbf{P}^\tau(u_1^\tau(0) \geqslant u_2^\tau(0)^+)\right. \\
&\qquad + \frac{\tilde{w}_1}{\sqrt{N}} \left.\frac{\partial \mathbf{P}^\tau(u_1^\tau(w) \geqslant u_2^\tau(0)^+)}{\partial w}\right|_{w=0} \\
&\qquad \left. + \frac{\widetilde{D}_2}{\sqrt{N}} \left.\frac{\partial \mathbf{P}^\tau(u_1^\tau(0) \geqslant u_2^\tau(d)^+)}{\partial d}\right|_{d=0}\right] \\
&:= \Lambda^N\left(\gamma_1 + \kappa_1 \frac{\tilde{w}_1}{\sqrt{N}} + \zeta_1 \frac{\widetilde{D}_2}{\sqrt{N}}\right), \qquad (17)
\end{aligned}$$

for the obvious choice of γ_1, κ_1, and ζ_1.[4] Similarly,

$$\begin{aligned}
&\lambda_2^N(w_1^N, D_2^N) \\
&\quad = \Lambda^N \mathbf{P}^\tau(u_2^\tau(D_2^N) \geqslant u_1^\tau(w_1^N)^+)
\end{aligned}$$

$$= \Lambda^N \mathbf{P}^\tau\left(u_2^\tau\left(\frac{\widetilde{D}_2}{\sqrt{N}}\right) \geqslant u_1^\tau\left(0+\frac{\tilde{w}_1}{\sqrt{N}}+o\left(\frac{1}{\sqrt{N}}\right)\right)^+\right)$$

$$\approx \Lambda^N\Bigg[\mathbf{P}^\tau(u_2^\tau(0) \geqslant u_1^\tau(0)^+) + \frac{\tilde{w}_1}{\sqrt{N}}\frac{\partial \mathbf{P}^\tau(u_2^\tau(0) \geqslant u_1^\tau(w)^+)}{\partial w}\Bigg|_{w=0} + \frac{\widetilde{D}_2}{\sqrt{N}}\frac{\partial \mathbf{P}^\tau(u_2^\tau(d) \geqslant u_1^\tau(0)^+)}{\partial d}\Bigg|_{d=0}\Bigg]$$

$$:= \Lambda^N\left(\gamma_2+\kappa_2\frac{\tilde{w}_1}{\sqrt{N}}+\zeta_2\frac{\widetilde{D}_2}{\sqrt{N}}\right). \tag{18}$$

Note that $\kappa_1, \zeta_2 \leqslant 0$, $0 \leqslant \kappa_2 \leqslant -\kappa_1$, and $0 \leqslant \zeta_1 \leqslant -\zeta_2$. The effective arrival rate into the system (that is achieved as w_1^N, $D_2^N \to 0$) is given by $\Lambda_{\text{eff}}^N = \Lambda^N(\gamma_1+\gamma_2)$. For any Λ^N, we can rewrite Λ_{eff}^N in the form $N\mu - \delta\sqrt{N}\mu$ for the appropriate choice of δ. Then, the appropriate value for the parameter β is derived by

$$\lambda_1^N(w_1^N, D_2^N) + \lambda_2^N(w_1^N, D_2^N)$$

$$\approx \Lambda^N(\gamma_1+\gamma_2) + \Lambda^N(\kappa_1+\kappa_2)\frac{\tilde{w}_1}{\sqrt{N}} + \Lambda^N(\zeta_1+\zeta_2)\frac{\widetilde{D}_2}{\sqrt{N}}$$

$$= (N\mu - \delta\sqrt{N}\mu)\left(1+\frac{\kappa_1+\kappa_2}{\gamma_1+\gamma_2}\frac{\tilde{w}_1}{\sqrt{N}}+\frac{\zeta_1+\zeta_2}{\gamma_1+\gamma_2}\frac{\widetilde{D}_2}{\sqrt{N}}\right)$$

$$\approx N\mu - \sqrt{N}\mu\left(\delta - \frac{\kappa_1+\kappa_2}{\gamma_1+\gamma_2}\tilde{w}_1 - \frac{\zeta_1+\zeta_2}{\gamma_1+\gamma_2}\widetilde{D}_2\right). \tag{19}$$

Let $\gamma = \gamma_1+\gamma_2$, $\kappa = \kappa_1+\kappa_2$, $\zeta = \zeta_1+\zeta_2$, and note that κ, $\zeta \leqslant 0$. In the notation of §3,

$$\lambda_a^N(w_1^N, D_2^N) \approx N\mu - \beta(\tilde{w}_1, \widetilde{D}_2)\sqrt{N}\mu,$$

where

$$\beta(\tilde{w}_1, \widetilde{D}_2) = \delta - \frac{\kappa}{\gamma}\tilde{w}_1 - \frac{\zeta}{\gamma}\widetilde{D}_2. \tag{20}$$

The limiting arrival rates into the system are given by

$$\frac{1}{N}\lambda_1^N(w_1^N, D_2^N) \to \frac{\gamma_1}{\gamma_1+\gamma_2}\mu := \tilde{\lambda}_1(0,0) \quad \text{and}$$
$$\frac{1}{N}\lambda_2^N(w_1^N, D_2^N) \to \frac{\gamma_2}{\gamma_1+\gamma_2}\mu := \tilde{\lambda}_2(0,0). \tag{21}$$

Hence, the proposed approximation is to replace the N-server system with the diffusion model derived in the previous section with limiting arrival rates given in (21) and $\beta(\tilde{w}_1, \widetilde{D}_2)$ given in (20). For a steady-state distribution and an equilibrium operating regime to exist, $\beta(\tilde{w}_1, \widetilde{D}_2)$ must be positive when $\tilde{w}_1$ is the expected waiting time for class 1 service in equilibrium. In this case, (12) implies that in equilibrium, the expected waiting time for class 1 service is

$$\tilde{w}_1 = \frac{1}{\tilde{\lambda}_1(0,0)}\frac{\alpha(\beta(\tilde{w}_1, \widetilde{D}_2))}{\beta(\tilde{w}_1, \widetilde{D}_2)}e^{-\beta(\tilde{w}_1, \widetilde{D}_2)\tilde{\theta}}, \tag{22}$$

where $\tilde{\theta} = \tilde{\lambda}_2(0,0)\widetilde{D}_2$.

Proposition 4.1. *The limit system specified through Theorem* 3.1, *Propositions* 3.1 *and* 3.2, *with limiting arrival rates specified in* (21), D_2^N *satisfying* (8), *and* β *given in* (20), *has a unique, stable equilibrium point given by the unique solution,* $\tilde{w}_1$, *of* (22) *subject to* $\beta(\tilde{w}_1, \widetilde{D}_2) > 0$.

The equilibrium $\tilde{w}_1$ is characterized implicitly through Equation (22). The equilibrium is said to be *stable* in the sense that the limiting system starting from any initial condition (i.e., any arrival rate vector) will eventually adapt and reach this unique point. We have not been able to find an explicit solution for $\tilde{w}_1$ even for simple choice models. It is easy to solve (22) numerically, however (by searching over $\tilde{w}_1$), and compute the equilibrium point as a function of the class 2 delay bound $\widetilde{D}_2 = D_2^N\sqrt{N}$, the number of servers N, the service rate μ, and the parameters of the choice model.[5] For example, under the MNL model the relevant parameters are the κ_is, ζ_is, and γ_is given by

$$\gamma_1 = \frac{e^{r_1/\nu}}{1+e^{r_1/\nu}+e^{r_2/\nu}}, \quad \kappa_1 = -\frac{c_1}{\nu}\gamma_1(1-\gamma_1), \quad \zeta_1 = \gamma_1\gamma_2\frac{c_2}{\nu},$$
$$\gamma_2 = \frac{e^{r_2/\nu}}{1+e^{r_1/\nu}+e^{r_2/\nu}}, \quad \kappa_2 = \frac{c_1}{\nu}\gamma_1\gamma_2, \quad \zeta_2 = -\frac{c_2}{\nu}\gamma_2(1-\gamma_2),$$

and

$$\tilde{\lambda}_1(0,0) = \mu\frac{e^{r_1/\nu}}{e^{r_1/\nu}+e^{r_2/\nu}}, \qquad \tilde{\lambda}_2(0,0) = \mu - \tilde{\lambda}_1(0,0).$$

Justification of the Rationalized Regime Systems with Many Servers. Thus far we have assumed that the system operates in the rationalized regime. Next, we establish that if customers make decisions according to a choice model that satisfies the assumptions outlined in §2, then the equilibrium arrival rates will indeed satisfy the defining assumption of the rationalized regime. We prove this result under two assumptions:

Assumption 1. *Balanced capacity*: *We assume that the number of servers is selected in a way that "almost matches" the total potential demand for the system. Specifically,*

$$\Lambda_{\text{eff}}^N = \lambda_1^N(0,0)+\lambda_2^N(0,0) = N\mu - \delta\sqrt{N}\mu \quad \text{for some } \delta \in \mathbf{R}.$$

Assumption 2. *Uniqueness of equilibrium*: *We assume that for any* N, *there exists a unique equilibrium point characterized by the steady-state expected waiting time in queue* 1, *denoted by* w_1^N.

Assumption 1 is not very restrictive. One way to explain what it means it to consider again an example that one may want to analyze. Specifically, for the system with $N = 50$ and $\Lambda_{\text{eff}} = 49$ that was analyzed in Table 1, one would proceed by rewriting Λ_{eff} in the form $50 - \delta\sqrt{50} = 49 \Rightarrow \delta = 0.14$. In trying to approximate the 50-server system via the asymptotic analysis of the last two sections, one would scale Λ_{eff} as a function of N in a way that keeps δ always

equal to 0.14. This assumes that the system manager does not intentionally under- or overcapacitate the system; this staffing guideline has been shown to be economically optimal in various settings in Borst et al. (2004) and Maglaras and Zeevi (2003). It remains to show that under the customer choice model described in §2, the waiting times manifest themselves in a way that the resulting equilibrium point satisfies the assumptions of the "rationalized" regime.

As an alternative to Assumption 1, one could potentially write Λ_{eff}^N in the form $\Lambda_{\text{eff}}^N = N\mu(1+\delta)$, and scale demand according to this relation. This would keep the system either intentionally under- or overutilized. This approach has been pursued by Whitt (2003) in recent work, where (among other results) he derives the limiting equilibrium for a single-class $M/M/N$ queue under the assumption that $\delta > 0$. His results imply that the system operates in the cost-driven regime and the limiting equilibrium analysis becomes substantially simpler. Given the economic optimality of the "rationalized" regime, we advocate that this regime may be more suitable for call centers operating in steady state. On the other hand, the regime studied by Whitt is suitable for system analysis when there is a sudden increase in the total demand that is not followed by a corresponding surge in staffing, which leads to a systematically undercapacitated system.

Assumption 2 seems plausible. All numerical results suggest that there always exists a unique equilibrium point. This is certainly correct asymptotically (Proposition 4.1). We have been unable, however, to prove this fact in a straightforward manner, mainly because there are no closed-form expressions of the two-class system behavior for a finite number of servers.

To prove that the rationalized regime is the appropriate operating regime for the system in equilibrium, one needs to establish (according to Theorem 3.1) that under Assumptions 1 and 2, either condition (6) or (7) is satisfied. We prove this result in two steps. The first step (Proposition 4.2) shows that the equilibrium traffic intensity goes to one (i.e., the system goes to heavy traffic). The second (Proposition 4.3) establishes that the traffic intensity satisfies (7).

PROPOSITION 4.2. *Suppose that Assumptions* 1 *and* 2 *hold, D_2^N scales according to* (8)*, and that for all N sufficiently large, the system is stabilizable (i.e., $\lambda_1^N(\infty,0) + \lambda_2^N(\infty,0) < N\mu$). Then, if w_1^N is the steady-state expected waiting time for class* 1 *in equilibrium,*

$$\lim_{N\to\infty} \rho^N(w_1^N, D_2^N) = \lim_{N\to\infty} \frac{\lambda_1^N(w_1^N, D_2^N) + \lambda_2^N(w_1^N, D_2^N)}{N\mu} = 1. \tag{23}$$

The next proposition shows that the rate at which the traffic intensity approaches one is the one required by (7). In addition to Assumptions 1 and 2 we require that in the limit the system does not become degenerate in the sense that almost all customers join only queue 1. This roughly says that the system retains its multiclass nature even in the asymptotic regime. The associated technical condition is that $\gamma_i > 0$, $i = 1, 2$; recall the definitions of γ_i through (17) and (18).

PROPOSITION 4.3. *Under the assumptions of Proposition* 4.2 *and the additional condition that the constants γ_1, γ_2 defined through* (17)*,* (18) *are strictly positive,*

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho^N(w_1^N, D_2^N)) = \beta \quad \textit{for some } \beta > 0. \tag{24}$$

That is, the equilibrium point of the system converges to a limiting model that satisfies the defining assumption of the "rationalized" regime. This validates the model approximations proposed so far, as well as the ones that are derived in the next section.

5. Approximations and System Design

The previous sections proposed a model approximation for the N-server system. Here, the steady-state distribution of this asymptotic approximation is used to derive estimates for quantities of interest in the original system. Their accuracy is checked numerically. Finally, the system design problem of choosing the minimum number of servers needed to satisfy a set of performance specifications is addressed.

5.1. Performance Approximations

The steady-state performance of the N-server system can be approximated using the limiting diffusion process derived in §§3 and 4. The approximations are summarized in Table 2. For illustrative purposes, the derivation for two of these entries are included next. A numerical example is used to compare these approximations with results obtained via exhaustive simulation.

For the purposes of this section one should think of N and D_2^N as fixed parameters. Let $\beta^* = \beta(\tilde{w}_1, \widetilde{D}_2)$ and $\alpha = [1 + \sqrt{2\pi}\beta^*\Phi(\beta^*)e^{\beta^{*2}/2}]^{-1}$, where $\widetilde{D}_2 = \sqrt{N}D_2^N$, $\tilde{w}_1$ is the unique solution of (22), and $\beta(\tilde{w}_1, \widetilde{D}_2)$ is given

Table 2. Approximations of performance measures of the N-server system.

$\mathbf{P}(W_1^N > y)$ for $y \geqslant 0$	$\alpha e^{-(1-\rho^N)(\lambda_2^N D_2^N + \lambda_1^N y)}$
$\mathbf{E}W_1^N$	$\frac{1}{\lambda_1^N}\frac{\alpha}{1-\rho^N}e^{-(1-\rho^N)\lambda_2^N D_2^N}$
$\text{Var}[W_1^N]$	$\frac{2\mathbf{E}W_1^N}{\lambda_1^N(1-\rho^N)} - (\mathbf{E}W_1^N)^2$
$\mathbf{E}W_2^N$	$\frac{1}{\lambda_2^N}\frac{\alpha}{1-\rho^N}\left(1 - e^{-(1-\rho^N)\lambda_2^N D_2^N}[1 + (1-\rho^N)\lambda_2^N D_2^N]\right)$
$\text{Var}[W_2^N]$	$\frac{2\mathbf{E}W_2^N(1+\beta^*\theta)}{\lambda_2^N(1-\rho^N)} - (\mathbf{E}W_2^N)^2$

in (20). Also, let $\tilde{\lambda}_i = \tilde{\lambda}_i(0,0)$, $i = 1, 2$; see (21).[6] We first approximate the steady-state expected value of class 1 waiting time. This can be done by observing that due to Proposition 3.3, $W_1^N \approx \widetilde{W}_1/\sqrt{N}$. Hence,

$$\mathbf{E}W_1^N \approx \frac{\tilde{w}_1}{\sqrt{N}} = \frac{1}{\tilde{\lambda}_1\sqrt{N}}\frac{\alpha}{\beta^*}e^{-\beta^*\tilde{\theta}}.$$

Recall that $\beta^* \approx \sqrt{N}(1-\rho^N)$, $\tilde{\theta} = \theta^N/\sqrt{N} = \lambda_2^N D_2^N/\sqrt{N}$, and $\tilde{\lambda}_1 \approx \lambda_1^N/N$. Then,

$$\mathbf{E}W_1^N \approx \frac{1}{\lambda_1^N}\frac{\alpha}{1-\rho^N}e^{-(1-\rho^N)\lambda_2^N D_2^N}, \tag{25}$$

where the equilibrium arrival rates are given by

$$\lambda_i^N = \lambda_i^N(\mathbf{E}W_1^N, D_2^N) \approx \lambda_i^N\left(\frac{\tilde{w}_1}{\sqrt{N}}, D_2^N\right), \quad i = 1, 2.$$

Similarly, to calculate $\mathbf{P}(W_1^N > y)$, observe that the distribution of W_1^N may be approximated by that of $\widetilde{W}_1/\sqrt{N}$ (again, due to Proposition 3.3). For $\tilde{\theta} = \tilde{\lambda}_2\widetilde{D}_2$ and for any $y > 0$,

$$\begin{aligned}\mathbf{P}(W_1^N > y) &\approx \mathbf{P}(\widetilde{W}_1 > y\sqrt{N}) \\ &= \mathbf{P}(\widetilde{X} > \tilde{\theta} + \tilde{\lambda}_1 y\sqrt{N}) \\ &= \alpha e^{-\beta^*(\tilde{\theta}+\tilde{\lambda}_1\sqrt{N}y)} \\ &\approx \alpha e^{-(1-\rho^N)(\lambda_2^N D_2^N + \lambda_1^N y)}.\end{aligned}$$

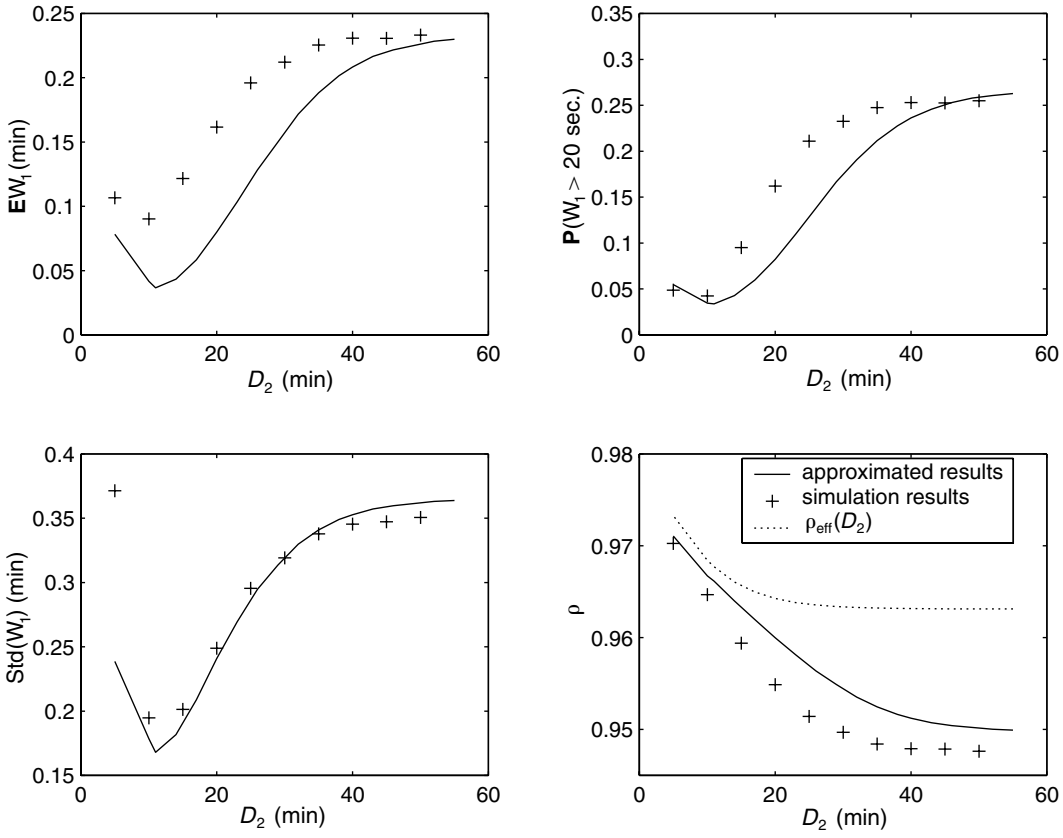
In particular, the probability that a class 1 customer will have to wait is given by

$$\mathbf{P}(W_1^N > 0) \approx \alpha e^{-(1-\rho^N)\lambda_2^N D_2^N}.$$

The remaining entries of Table 2 are derived along the same lines. It is worth noting that the diffusion analysis does not provide a meaningful approximation of the probability of a call-back violating its delay specification (that is, $\mathbf{P}(W_2^N > x)$ for some $x \geqslant D_2^N$). The reason is that in the diffusion model the waiting time distribution for class 2 calls is bounded above by $\widetilde{D}_2$, and thus the probability of violating this upper bound is always zero. Such estimates can be obtained through a more detailed analysis—potentially using large deviations arguments—that studies how the queue-length and waiting time processes converge to their limits.

We conclude this section with the numerical example of Figure 2 that studies a 50-server system and compares the equilibrium performance computed through extensive simulation with the one approximated using the asymptotic analysis. To find the equilibrium regime using simulation we proceeded as follows: (1) hypothesize a value for the steady-state expected waiting time for class 1, $\mathbf{E}W_1^N$, compute the corresponding arrival rates using the choice model, and simulate to estimate the actual steady-state expected

Figure 2. Comparison of equilibrium behavior computed via (a) simulation and (b) asymptotic approximations for $N = 50$, $\mu = 1$, $\rho_{\text{eff}} = 0.98$, $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, $\nu = 0.3$.



waiting time for this set of arrival rates; and (2) repeat until the hypothesized value for $\mathbf{E}W_1^N$ agrees with the one estimated via simulation.[7]

It is important to note that the numerical results depicted in Figure 2 depend strongly on the choice parameters of the MNL model. On the other hand, the overall nature of these results, and in particular the list of observations given below, are consistent with all the examples we examined using different sets of the MNL choice parameters. Also, the accuracy of the heavy traffic approximation improves as N grows. This is important from a practical viewpoint, because contact centers can have several hundred or thousands of servers, and cannot be analyzed efficiently using simulation. In such cases the approximate analysis is particularly useful. The key observations are:

(a) The system operates close to heavy traffic; this verifies Propositions 4.2 and 4.3.

(b) The introduction of the call-back option improves overall performance. For example, in the system analyzed in Figure 2, when $D_2^N = 10$ min., $\mathbf{E}W_1^N$ is reduced by 60% in comparison with a single-class system (no call-back), while the traffic intensity is increased by 1.5%. Similar performance improvements are observed in the variance of the waiting time for class 1, and in the probability that W_1^N exceeds some acceptable upper bound (typically 20 sec.); these results were reported in Table 1. The simulation results also show that in equilibrium and for all choices of D_2^N, the proportion of late call-backs (i.e., $\mathbf{P}(W_2^N > D_2^N)$) is between 0.03 and 0.01, and on average late calls violate their respective upper bounds by 3%–8%. To improve the probability of meeting the delay specification for class 2 customers, the SM should add a safety margin and switch priorities at $\lambda_2^N D_2^N - \epsilon$ for some $\epsilon > 0$ that can be selected experimentally.

(c) In equilibrium, $\mathbf{E}W_1^N$ is a decreasing function of D_2^N for small values of D_2^N, and it approaches a constant value as D_2^N grows large that corresponds to a system without the call-back option. This dependence was consistent among all examples that we studied, and it can be justified through an analysis of the limiting model. This suggests that we can always choose an optimal D_2^N that lies between these two extreme cases.

5.2. Choosing Staffing Levels

An important application of the analytical results of §§3 and 4 is the design of contact centers with multiple channels of communication that offer quality-of-service (delay) guarantees. We focus on the staffing problem that involves choosing the minimum number of servers to satisfy a set of performance specifications that are typically encountered in contact centers, such as

- The expected waiting time for real time calls $\leqslant$10 seconds,
- 80% of all calls are answered within 20 seconds,
- $\leqslant$1% balking probability.

The mathematical formulation is as follows:

$$\min\{N: \mathbf{E}W_1^N \leqslant w_e, \mathbf{P}(W_1^N \geqslant y) \leqslant \epsilon_1, \mathbf{P}(\text{balking}) \leqslant \epsilon_b\}, \tag{26}$$

with typical values for these specifications given above.

With no analytic characterization of performance, the design problem in (26) must be addressed via simulation. Following our previous comments on this approach, optimizing over the number of servers in this way is quite involved: one has to find the equilibrium regime for different values of N and then optimize, which can be quite expensive for large systems. In contrast, we again proceed by using the analytic performance approximations given above. This is computationally simple, accurate, and provides useful insights about the appropriate structure of the optimal solution.

The previous section has analyzed the asymptotic system behavior when the total arrival rate into the system is of the form $\lambda_a^N = N\mu - \beta\sqrt{N}\mu$. Denote by R_a the offered load into the system $R_a = \lambda_a^N/\mu$; this is a unitless quantity, which is often expressed in Erlangs. We can rewrite the number of servers in terms of R_a as follows:

$$N = R_a + \beta\sqrt{R_a} + o(\sqrt{R_a}).$$

The analysis in §§3 and 4 suggests that the appropriate solution to the staffing problem should also take a form of this nature. This agrees with the "square-root" laws that have been proposed for single-class systems by Kolesar and Green (1998) and Garnett et al. (2002), and rigorously justified by Borst et al. (2004).

For the purposes of this section, let $D_2^N = d_2$ denote the delay bound on the call-back option, and let $\Lambda_{\text{eff}}(d_2) = \lambda_1^N(0, d_2) + \lambda_2^N(0, d_2)$ be the effective arrival rate into the system given this delay bound. Both are assumed to be fixed and superscripts are dropped from $\Lambda_{\text{eff}}, d_2$ to indicate that these quantities are exogenously given and do not scale with N.[8] The goal is to find the minimum staffing level of the form $N = R + x\sqrt{R}$ that satisfies (26), where $R = \Lambda_{\text{eff}}(d_2)/\mu$ and x is the design parameter.

Bounds on the Expected Waiting Time for Class 1 of the Form $\mathbf{E}W_1^N \leqslant w_e$. Because $\sqrt{N}W_1^N \approx \widetilde{W}_1$, this constraint can be approximated by $\mathbf{E}\widetilde{W}_1 \leqslant \sqrt{N}w_e$. Following (22) this is equivalent to

$$\frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta)}{\beta}e^{-\beta\tilde{\theta}} \leqslant \sqrt{N}w_e,$$

where the notation $\alpha(\beta)$ underlines the explicit dependence of α on β, and $\tilde{\theta} = \tilde{\lambda}_2 d_2\sqrt{N}$. In the sequel, we rewrite this as a constraint on β, and subsequently on x. First, note that $\tilde{\lambda}_1$ is independent of $\mathbf{E}\widetilde{W}_1$, and it is constant in the above expression. Because $N = R + x\sqrt{R}$, when R is large, $\sqrt{N} \approx \sqrt{R}$. So, the bound $\sqrt{N}w_e$ can be replaced by $\tilde{w}_e := \sqrt{R}w_e$, and $\tilde{\theta}$ can be approximated by $\tilde{\lambda}_2 d_2\sqrt{R}$.

Hence, the constraint on $\mathbf{E}\widetilde{W}_1$ can be approximated by a constraint on the equilibrium parameter β as follows:

$$\frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta)}{\beta}e^{-\beta\tilde{\theta}} \leqslant \tilde{w}_e \Leftrightarrow \beta \geqslant \beta_{\tilde{w}_e},$$

$$\text{where} \quad \beta_{\tilde{w}_e} = \inf\left\{\beta > 0\colon \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta)}{\beta}e^{-\beta\tilde{\theta}} \leqslant \tilde{w}_e\right\}.$$

The value of $\beta_{\tilde{w}_e}$ can easily be computed numerically. Second, as in (20), β can be rewritten in the form $\beta = \delta_{d_2} - (\kappa/\gamma)\mathbf{E}\widetilde{W}_1$, where δ_{d_2} is the appropriate constant for which $\Lambda_{\text{eff}}(d_2) := \lambda_1(0, d_2) + \lambda_2(0, d_2) = N\mu - \delta_{d_2}\sqrt{N}\mu$; i.e., δ_{d_2} represents the slack capacity of the system. Indeed, using the fact that $R = \Lambda_{\text{eff}}/\mu$, the above expression can be rewritten as $R = N - \delta_{d_2}\sqrt{N}$, which implies that $N = R + \delta_{d_2}\sqrt{R} + o(\sqrt{R})$. Hence, δ_{d_2} is equal to the design parameter x. Using the fact that $\mathbf{E}\widetilde{W}_1 \leqslant \tilde{w}_e$, the constraint on β is implied by the following constraint on δ_{d_2}:

$$\delta_{d_2} \geqslant \beta_{\tilde{w}_e} + \frac{\kappa}{\gamma}\tilde{w}_e,$$

which is simply a lower bound on the design parameter x; i.e., $x \geqslant \beta_{\tilde{w}_e} + (\kappa/\gamma)\tilde{w}_e$.

Probabilistic Constraints on Waiting Time W_1^N of the Form $\mathbf{P}(W_1^N \geqslant y) \leqslant \epsilon_1$. Typical parameters in call centers are $y = 20$ seconds and $\epsilon_1 = 0.2$. This constraint can be approximated by

$$\mathbf{P}(W_1^N \geqslant y) \approx \mathbf{P}(\widetilde{W}_1 \geqslant \sqrt{R}y) = \alpha(\beta)e^{-\beta(\tilde{\theta}+\tilde{\lambda}_1\sqrt{R}y)} \leqslant \epsilon_1 \;\Rightarrow\; \beta \geqslant \beta_{(y,\epsilon_1)},$$

where $\beta_{(y,\epsilon_1)} = \inf\{\beta > 0\colon \alpha(\beta)e^{-\beta(\tilde{\theta}+\tilde{\lambda}_1\sqrt{R}y)} \leqslant \epsilon_1\}$. This implies an upper bound on $\mathbf{E}\widetilde{W}_1$:

$$\beta \geqslant \beta_{(y,\epsilon_1)} \;\Rightarrow\; \mathbf{E}\widetilde{W}_1 \leqslant \tilde{w}_{(y,\epsilon_1)} := \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta_{(y,\epsilon_1)})}{\beta_{(y,\epsilon_1)}}e^{-\beta_{(y,\epsilon_1)}\tilde{\theta}}.$$

Using the same reasoning as above, we have that $\beta \geqslant \beta_{(y,\epsilon_1)}$ and $\mathbf{E}\widetilde{W}_1 \leqslant \tilde{w}_{(y,\epsilon_1)}$ is implied by the following condition on δ_{d_2}:

$$\delta_{d_2} \approx x \geqslant \beta_{(y,\epsilon_1)} + \frac{\kappa}{\gamma}\tilde{w}_{(y,\epsilon_1)}.$$

Bounds on the Balking Rate of the Form $\mathbf{P}(\text{balking}) \leqslant \epsilon_b$. Using (19),

$$\mathbf{P}(\text{balking}) = 1 - \frac{\lambda_1^N(\mathbf{E}W_1^N, d_2) + \lambda_2^N(\mathbf{E}W_1^N, d_2)}{\Lambda_{\text{eff}}(d_2)} \approx -\frac{\kappa}{\gamma}\mathbf{E}W_1^N.$$

Hence, the constraint $\mathbf{P}(\text{balking}) \leqslant \epsilon_b$ is equivalent to an upper bound constraint on $\mathbf{E}W_1^N$,

$$\mathbf{E}W_1^N \leqslant -\frac{\gamma}{\kappa}\epsilon_b := w_b$$

(recall that $\gamma/\kappa < 0$). This is of the same form of the constraint $\mathbf{E}W_1^N \leqslant w_e$ studied above, and results in the following condition:

$$\delta_{d_2} \approx x \geqslant \beta_{\tilde{w}_b} + \frac{\kappa}{\gamma}\tilde{w}_b \;\Leftrightarrow\; \delta_{d_2} \approx x \geqslant \beta_{\tilde{w}_b} - \epsilon_b\sqrt{R},$$

where $\tilde{w}_b := w_b\sqrt{R}$.

Hence, the staffing problem of (26) is solved by $N = R + x^*\sqrt{R}$, where

$$x^* = \max\left(\beta_{\tilde{w}_e} + \frac{\kappa}{\gamma}\tilde{w}_e, \beta_{(y,\epsilon_1)} + \frac{\kappa}{\gamma}\tilde{w}_{(y,\epsilon_1)}, \beta_{\tilde{w}_b} - \epsilon_b\sqrt{R}\right), \tag{27}$$

and all of the quantities involved in this expression can be evaluated using one-dimensional parameter searches and Equation (22).

An alternative formulation of the staffing problem would be in terms of a profitability criterion of the form $\pi(N) = (\lambda_1^N + \lambda_2^N)p - \lambda_{\text{balk}} \cdot q - cN$, where $\$p$ is the profit per served customer, $\$q$ is the penalty (lost goodwill, etc.) incurred per customer that decides not to join because he/she found the system too congested, $\lambda_{\text{balk}} = \Lambda_{\text{eff}}(d_2) - (\lambda_1^N + \lambda_2^N)$ is the fraction of customers that choose not to join the system, and $\$c$ is the operating cost per unit time per server (see Borst et al. 2004 for a detailed analysis in this direction).

We conclude with a numerical example. Suppose that customers make decisions according to the MNL model with parameters $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.02$, $\nu = 0.3$. The server speed is $\mu = 1$ service request/min., the load is $R = 100$ service requests/min., and the delay bound for class 2 service is $d_2 = 90$ min. The specifications are

$$\mathbf{E}W_1^N \leqslant 10 \text{ sec.}, \quad \mathbf{P}(W_1^N > 20 \text{ sec.}) \leqslant 0.2, \quad \text{and} \quad \mathbf{P}(\text{balking}) \leqslant 0.01.$$

Using the asymptotic expressions of the previous subsection we can plot these specifications as a function of the number of servers N. The results are shown in Figure 3. The required number of servers is $N = 102$. We call this the "analytic" solution. In contrast, using (27), we have that in the limit system $\tilde{d}_2 \approx 90\sqrt{R} = 900$,

$$\tilde{\lambda}_1(0, 90) = \mu\frac{e^{r_1/\nu}}{e^{r_1/\nu} + e^{(r_2 - c_2\cdot 90)/\nu}},$$

$$\tilde{\lambda}_2(0, 90) = \mu - \tilde{\lambda}_1(0, 90),$$

and $\tilde{\theta} = \tilde{d}_2\tilde{\lambda}_2(0, 90)$. Then,

$$\beta_{\tilde{w}_e} + \frac{\kappa}{\gamma}\tilde{w}_e = 0.1558, \qquad \beta_{(y,\epsilon_1)} + \frac{\kappa}{\gamma}\tilde{w}_{(y,\epsilon_1)} = 0.1305,$$

$$\beta_{\tilde{w}_b} - \epsilon_b\sqrt{R} = 0.1440,$$

which implies that $x^* = 0.1558$ and $N^* = \lceil R + 0.1558\sqrt{R}\rceil = 102$ servers. Throughout the examples we

examined, we found that the staffing rule of (27) is always within one server of the "analytic" result obtained by plotting the specifications for all N (using the asymptotic approximations) and picking the minimum number of servers that leads to an acceptable design. It is also close to the results obtained via simulation; this was only checked for smaller systems because it is computationally tedious. Overall, the rule proposed in (27) is intuitive and simple to compute given specific parameters for the choice model.

6. Concluding Remarks

This paper analyzed a contact center that offers two service modes: real-time telephone service and postponed (call-back) service with a guarantee on the maximum delay until a reply is received. Customers choose which channel to use based on a probabilistic choice model. An asymptotic analysis was used to develop a near-optimal scheduling rule, analytic approximations for the system's equilibrium behavior, performance measures, and the appropriate staffing criteria. This mode of analysis is accurate for systems with many servers that operate close to heavy traffic—this is the canonical operating regime for such systems.

The key findings of this paper are: (1) service systems can improve their performance substantially by offering multiple channels of service (such as the call-back option), even when these are accompanied by performance guarantees, (2) the scheduling policy needed to guarantee the delay specification for the call-back service is a simple threshold rule, (3) the system equilibrium and performance measures are easy to characterize and compute via a tractable asymptotic analysis, and (4) "square-root" staffing rules are still near optimal. The main methodological contribution is to illustrate how one can formulate an appropriate, tractable asymptotic model that captures in a nontrivial way the equilibrium behavior of the system.

Several interesting areas of future research arise. First, one could consider the problem where the system manager announces state-dependent information about the anticipated delay in the real-time queue. This model is analyzed in another paper by the authors (Armony and Maglaras 2004). The main obstacles there are to come up with the right state-dependent estimate for the waiting time for class 1 customers, and to analyze the multiclass system with state-dependent arrival rates into both classes. In contrast with the current paper, this does not involve an equilibrium analysis. The results of the current paper provide some useful background in addressing the first issue raised above.

Other extensions would be to allow for customers selecting the call-back option to be able to schedule a time window where the call-back will be placed (much like in a reservation system). Another one is to allow for nonstationary arrivals. Finally, one needs to consider how these results change in a multiclass setting with specialized and cross-trained agents.

Appendix. Proofs

Proof of Proposition 3.1

We start with an outline of the proof.

1. *Sufficient Conditions for State Space Collapse.* First, we establish that to prove the statement of the proposition it suffices to show that for an appropriate sequence $\{b^N\}$ such that $b^N \to 0$ as $N \to \infty$,

$$\begin{aligned}&\mathbf{P}\left(\sup_{0\leqslant t\leqslant T}\left|X_1^N(t+b^N)-(X^N(t)-\tilde{\theta})^+\right|>\epsilon\right)\to 0,\\ &\mathbf{P}\left(\sup_{0\leqslant t\leqslant T}\left|X_2^N(t+b^N)-X^N(t)^+\wedge\tilde{\theta}\right|>\epsilon\right)\to 0.\end{aligned}\tag{28}$$

We focus on how to establish the second part of (28). The other statement follows similarly. Let $X_2^N(s;x,y)$ be the scaled class 2 queue length at time s starting from $X^N(0)=x$ and $X_2^N(0)=y$; this descriptor assumes that service and interarrival times start afresh at time 0. As in Puhalskii and Reiman (2000, Lemma 3),

$$\begin{aligned}&\mathbf{P}\left(\sup_{0\leqslant t\leqslant T}|X_2^N(t+b^N)-X^N(t)^+\wedge\tilde{\theta}|>\epsilon\right)\\ &\quad\leqslant\mathbf{P}\left(\sup_{t\leqslant T}X^N(t)>C\right)\\ &\qquad+\mathbf{P}\left(\sup_{x\leqslant C,\,0\leqslant y\leqslant x^+}|X_2^N(b^N;x,y)-x^+\wedge\tilde{\theta}|>\epsilon\right),\end{aligned}\tag{29}$$

where C is an arbitrary upper bound for $X^N(t)$. From Puhalskii and Reiman (2000, Equation (3.39)),

$$\lim_{C\to\infty}\limsup_{N\to\infty}\mathbf{P}\left(\sup_{t\leqslant T}X^N(t)>C\right)=0.$$

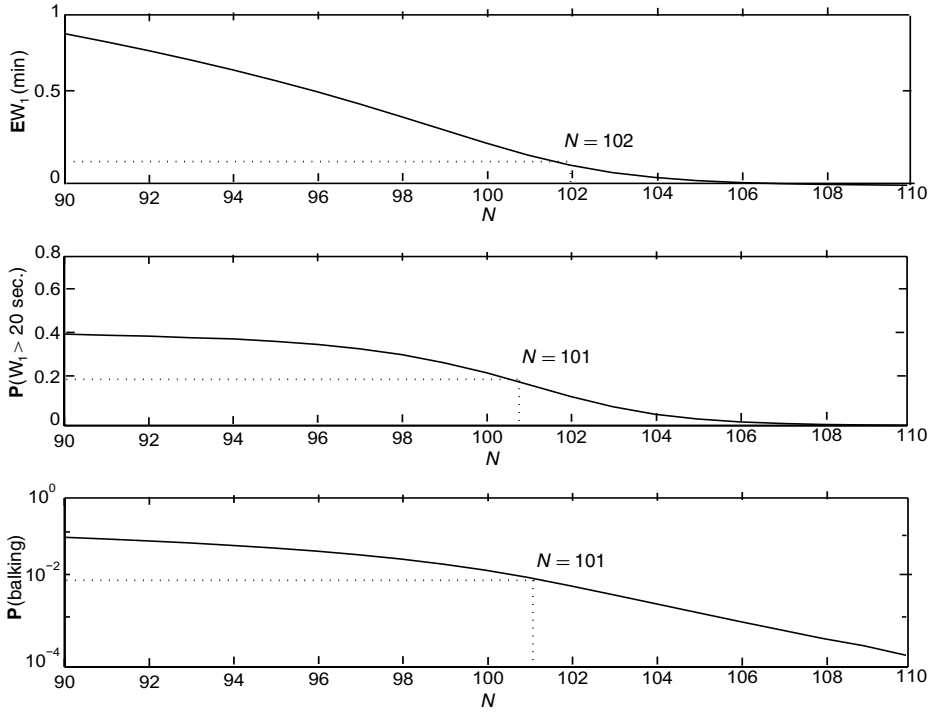
It remains to show that the second term in (29) also goes to zero as $N\to\infty$. This is done by analyzing a set of fluid scaled processes.

2. *Convergence of the Fluid Scale Process.* This part mimics the framework proposed by Bramson (1998), but does not make explicit use of his results. Specifically, we will analyze the limits of the scaled queue-length processes defined according to

$$\bar{Q}^N(\cdot;x,y):=\frac{Q^N(\cdot/\sqrt{N};x,y)}{\sqrt{N}}.$$

This process has a deterministic "fluid" limit that describes the state evolution over time periods of $\mathcal{O}(1/\sqrt{N})$ over which the queues can change by $\mathcal{O}(\sqrt{N})$; that is, N servers working for $\mathcal{O}(1/\sqrt{N})$ time. We will first derive this fluid limit, and then show that starting from any initial condition, the fluid limit processes converge in finite time to the appropriate limits $(x-\tilde{\theta})^+$, $x^+\wedge\tilde{\theta}$, where x is the initial condition for the total queue-length process. Moreover, if $x<C$ for any $C>0$, then the time it takes to reach this

Figure 3. Staffing example.



Note. $R = 100$ calls/min., $\mu = 1$/min., and $d_2 = 90$ min. Choice model parameters are $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.02$, $\nu = 0.3$. The design specifications are $\mathbf{E}W_1 \leqslant 10$ sec., $\mathbf{P}(W_1 > 20 \text{ sec.}) \leqslant 0.2$, and $\mathbf{P}(\text{balking}) \leqslant 0.01$. The optimal number of servers is 102.

target is less than or equal to $s^* = (C \vee \tilde{\theta})/(\tilde{\lambda}_1 \wedge \tilde{\lambda}_2)$. By the definition of the fluid scaled process, this implies that

$$\left| X_2^N\left(\frac{s^*}{\sqrt{N}}; x, y\right) - x^+ \wedge \tilde{\theta} \right| \to 0 \quad \text{w.p. 1.} \tag{30}$$

Set $b^N = s^*/\sqrt{N}$ and note that $b^N \to 0$ as $N \to \infty$.

3. *Uniform Convergence of the Fluid Scale Processes.* To complete the proof of (29), we show that the convergence in (30) is uniform in x, y.

We next give the details of the proof.

1. *Sufficient Conditions for State Space Collapse.* The sufficiency of (28) follows from known results on weak convergence theory. Specifically, by Proposition 7 in Glynn (1990), convergence in probability in the sup norm implies weak convergence; that is, (28) implies that $(X_1^N(t + b^N), X_2^N(t + b^N)) \Rightarrow ((\widetilde{X}(t) - \tilde{\theta})^+, \widetilde{X}(t)^+ \wedge \tilde{\theta})$. Because $t + b^N \to t$, an application of a random time change argument (see Glynn 1990, Proposition 5) establishes the statement of the proposition.

The next step is to show that the second term in (29) converges to 0 as $N \to \infty$. This is done in two steps. First, we establish (30), and second, show that this is uniform in x, y.

2. *Convergence of the Fluid Scale Process.* To prove (30), consider the sequence of initial conditions $X^N(0) = x$ and $X_2^N(0) = y$, $x \leqslant C$, $0 \leqslant y \leqslant x^+$. Define fluid scaled processes under our threshold policy as follows:

$$(\overline{Q}^N(\cdot; x, y), \overline{X}^N(\cdot; x, y), \overline{T}^N(\cdot; x, y))$$

$$:= \left(\frac{Q^N(\cdot/\sqrt{N}; x, y)}{\sqrt{N}}, X^N\left(\frac{\cdot}{\sqrt{N}}; x, y\right), \frac{T^N(\cdot/\sqrt{N}; x, y)}{\sqrt{N}} \right).$$

The notation $T_i^N(t)$ refers to the total time devoted by all servers to class i jobs up to time t; that is, $T_i^N(0) = 0$, $0 \leqslant \dot{T}_i^N(t) \leqslant N$, and $\dot{T}_1^N(t) + \dot{T}_2^N(t) \leqslant N$. Here, $Q^N(\cdot; x, y)$, $X^N(\cdot; x, y)$, and $T^N(\cdot; x, y)$ denote the respective processes given that $X^N(0) = x$ and $X_2^N(0) = y$.

We will show that the fluid limit processes starting from an arbitrary initial condition (x, y) reach the desired target positions $((x - \tilde{\theta})^+, x^+ \wedge \tilde{\theta})$ in finite time.

First, we analyze the limit of $\overline{X}^N$. For the N-server system, recall that $A_i^N(t)$ denotes the number of class i arrivals up to time t, and let $S(t)$ be the number of service completions when the one server has allocated t time units processing jobs, and let $E_i^N(t)$ be the number of class i service completions up to time t. For any vector process Y, $|Y| = \sum_i Y_i$. Clearly, $(|Z|^N, |T|^N)$ describe an $M/M/N$ queue for which $|Z|^N(t) = z^N + |A|^N(t) - S(|T|^N(t))$, which implies that

$$\overline{X}^N(s; x, y) = x + \frac{|A|^N(s/\sqrt{N})}{\sqrt{N}} - \frac{S(|T|^N(s/\sqrt{N}; x, y))}{\sqrt{N}}.$$

Note that $|T|^N$ is uniformly Lipschitz with constant N, and thus $|\overline{T}|^N(\cdot; x, y)$ is Lipschitz with unit constant and the family $\{|\overline{T}|^N\}$ is relatively compact. Hence, there exists a converging subsequence $\{N_j\}$ for which $|\overline{T}|^{N_j}(\cdot; x, y) \to |\overline{T}|(\cdot; x, y)$, where $|\overline{T}|$ is some limit allocation process.

Using the functional strong law of large numbers, (9), and the key renewal theorem, we have that

$$\frac{|A|^N(s/\sqrt{N})}{\sqrt{N}} \to \mu s \quad \text{and}$$

$$\frac{S(|T|^N(s/\sqrt{N}; x, y))}{\sqrt{N}} \to \mu|\overline{T}|(s; x, y),$$

where the convergence is almost surely (a.s.) uniform on compact sets in s (u.o.c.). Passing to the fluid limit,

$$\overline{X}^N(s; x, y) \to \overline{X}(s; x, y) = x + \mu(s - |\overline{T}|(s; x, y)) \quad \text{a.s. u.o.c.}$$

Also note that

$$|\overline{T}|^N(s; x, y) = \int_0^{s/\sqrt{N}} \frac{|Z|^N(\tau; x, y) \wedge N}{\sqrt{N}}\, d\tau = s - \frac{1}{\sqrt{N}} \int_0^s \overline{X}^N(\tau; x, y)^-\, d\tau \to s,$$

where $x^- = -\min(0, x)$. Substituting into the differential equation derived above we get that $\forall s \geqslant 0$, $\overline{X}(s; x, y) = x$, and in particular, the limit of the total queue-length process satisfies $\overline{X}(s; x, y)^+ = x^+$.

Let $E^N(t)$ be the departure process from the system. The queue lengths are given by

$$Q_1^N(t) = Q_1^N(0) + A_1^N(t) - \int_0^t \mathbf{1}\big(Q_2^N(s^-) < \theta^N \text{ and } Q_1^N(s^-) \geqslant 1\big)\, dE^N(s),$$

$$Q_2^N(t) = Q_2^N(0) + A_2^N(t) - \int_0^t \mathbf{1}\big(Q_2^N(s^-) \geqslant \theta^N \text{ or } \big\{Q_1^N(s^-) = 0 \text{ and } Q_2^N(s^-) \geqslant 1\big\}\big)\, dE^N(s).$$

To take the fluid limits of the queue lengths, note that the class level cumulative allocations are uniformly Lipschitz with constant N, and thus $\overline{T}_i^N(\cdot; x, y)$ are Lipschitz with unit constant and the family $\{\overline{T}^N\}$ is relatively compact. Hence, for every subsequence $\{N_j\}$ for which $\overline{T}_i^{N_j}(\cdot; x, y) \to \overline{T}_i(\cdot; x, y)$, $\overline{T}_i$ is some limit allocation process, and the convergence is with probability one (w.p. 1) and is u.o.c. Thus, for such a subsequence $\{N_j\}$, for almost all sample paths $(\overline{Q}_1^{N_j}(\cdot; x, y), \overline{Q}_2^{N_j}(\cdot; x, y)) \to (\overline{Q}_1(\cdot; x, y), \overline{Q}_2(\cdot; x, y))$. If $x \leqslant 0$, then $\overline{Q}_1(s; x, y) + \overline{Q}_2(s; x, y) = X^+(s) = 0$ for all $s \geqslant 0$. If $x > 0$, then from the analysis of the $(\overline{X}^N, |\overline{T}|^N)$ processes we have that $E^{N_j}(s) \to \mu|\overline{T}|(s) = \mu s$, and using Dai and Williams (1995, Lemma 2.4) we get that

$$\overline{Q}_1(s; x, y) = (x - y)^+ + \tilde{\lambda}_1 s - \mu \int_0^s \mathbf{1}\big(\overline{Q}_2(s; x, y) < \tilde{\theta} \text{ and } \overline{Q}_1(s; x, y) > 0\big)\, ds,$$

$$\overline{Q}_2(s; x, y) = y + \tilde{\lambda}_2 s - \mu \int_0^s \mathbf{1}\big(\overline{Q}_2(s; x, y) \geqslant \tilde{\theta} \text{ or } \big\{\overline{Q}_1(s; x, y) = 0 \text{ and } \overline{Q}_2(s; x, y) > 0\big\}\big)\, ds.$$

It now follows that for all $x \leqslant C$ and $0 \leqslant y \leqslant x^+$, $\overline{Q}_1(s; x, y) = (x - \tilde{\theta})^+$, and $\overline{Q}_2(s; x, y) = x^+ \wedge \tilde{\theta}$, for all

$$s \geqslant \max\left(\frac{\overline{Q}_1(0; x, y)}{\tilde{\lambda}_2}, \frac{(\tilde{\theta} - \overline{Q}_2(0; x, y))^+}{\tilde{\lambda}_2}, \frac{\overline{Q}_2(0; x, y)}{\tilde{\lambda}_1}\right) \leqslant \frac{C^+ \vee \tilde{\theta}}{\tilde{\lambda}_1 \wedge \tilde{\lambda}_2} := s^*.$$

That is,

$$\big|\overline{Q}_2^{N_j}(s^*; x, y) - x^+ \wedge \tilde{\theta}\big| \to 0 \tag{31}$$

with probability one. Thus, we have shown that (31) holds for every subsequence $\{N_j\}$ such that $\lim_{j\to\infty} \overline{Q}_2^{N_j}(s^*; x, y)$ exists. It is left to show that (31) holds when the subsequence $\{N_j\}$ is replaced by the whole sequence $\{N\}$. It suffices to show that $\liminf_{N\to\infty} \overline{Q}_2^N(\cdot; x, y) = \limsup_{N\to\infty} \overline{Q}_2^N(\cdot; x, y)$, w.p. 1. By contradiction, and without loss of generality, suppose that with positive probability, $\limsup_{N\to\infty} \overline{Q}_2^N(\cdot; x, y) > x^+ \wedge \tilde{\theta}$. In particular, this implies that there exists a subsequence $\{N_l\}$, such that $\lim_{l\to\infty} \overline{Q}_2^{N_l}(s^*; x, y) \in (x^+ \wedge \tilde{\theta}, \infty]$, with positive probability. By the relative compactness of $\overline{T}_2^{N_l}$, there exists a further subsequence $\{N_{l_j}\}$ such that $\overline{T}_2^{N_{l_j}}(\cdot; x, y) \to \overline{T}_2(\cdot; x, y)$. Hence, (31) holds for the subsequence $\{N_{l_j}\}$, which is a contradiction. That is, (31) implies (30).

3. *Uniform Convergence of the Fluid Scale Processes.* Set $b^N = s^*/\sqrt{N}$ and note that $b^N \to 0$ as $N \to \infty$. It remains to show that (30) is uniform in (x, y); that is, it is true with the supremum in front of the expression on the left-hand side of (30). This will prove that the second term in (29) goes to 0 as $N \to \infty$, and completes the proof.

We follow Dai (1995, Lemma 4.1). Suppose that the convergence is not uniform in x, y. Then, there exists an $\epsilon > 0$ and a sequence $\{(x^{N_l}, y^{N_l})\}$ with $x^{N_l} \leqslant C$ and $y^{N_l} \leqslant x^{N_l+}$ for all N_l, and

$$\left|X_2^{N_l}\left(\frac{s^*}{\sqrt{N_l}}; x^{N_l}, y^{N_l}\right) - x^{N_l+} \wedge \tilde{\theta}\right| \geqslant \epsilon \quad \forall\ N_l.$$

Because (x^{N_l}, y^{N_l}) is bounded, it has a converging subsequence with limit (x_0, y_0) with $x_0 \leqslant C$ and $y_0 \leqslant x_0^+$. Without loss of generality, assume that the original sequence converges to (x_0, y_0). By an argument analogous to the one above, it follows that there exists N_1 large enough such that for all $N_l > N_1$,

$$\left|X_2^{N_l}\left(\frac{s^*}{\sqrt{N_l}}; x^{N_l}, y^{N_l}\right) - x_0{}^+ \wedge \tilde{\theta}\right| < \frac{\epsilon}{2} \quad \text{w.p. 1},$$

and similarly, there exists N_2 large enough such that for all $N_l > N_2$,

$$\left|x^{N_l} - x_0\right| < \frac{\epsilon}{2}.$$

This leads to a contradiction. Hence, (30) is true uniformly in x, y. This establishes (28) (through (29)) and completes the proof. □

Proof of Proposition 3.2

The result follows from Puhalskii's (1994) invariance principle. We prove it by a direct application of Lemma A.2 of Puhalskii and Reiman (2000) for $i = 1, 2$. In the notation of Puhalskii and Reiman, K_i^N is equal to Q_i^N, the limits of process $L_i^N(t)$ follow from the FCLT for the arrival processes, and $D_i^N(t)/N \to \tilde{\lambda}_i$ in probability. The remainder of their lemma goes through unchanged. □

Proof of Proposition 3.3

First, we show that the steady-state queue lengths converge to the right limit. That is, we show that for $i = 1, 2$, $Q_i^N/\sqrt{N} \Rightarrow \widetilde{X}_i$, where Q_i^N corresponds to class i steady-state queue length, and $\widetilde{X}_i$ is the steady state of the limiting process $\widetilde{X}_i(\cdot)$. This is established via a slight modification of Proposition 3.1. Specifically, assume that the steady-state distributions of the class level queue lengths for the N-server system exist. Start the system at time $t = 0$ with an initial queue length vector drawn from the steady-state joint distribution. The results of Proposition 3.1 show that for all $t > 0$ (not $t \geqslant 0$ as in Proposition 3.1 because the initial condition is arbitrary),

$$\begin{aligned} &\frac{Q_1^N(t)}{\sqrt{N}} \Rightarrow (\widetilde{X}(t) - \tilde{\theta})^+ = \widetilde{X}(t)_1, \\ &\frac{Q_2^N(t)}{\sqrt{N}} \Rightarrow \widetilde{X}(t)^+ \wedge \tilde{\theta} = \widetilde{X}(t)_2, \end{aligned} \tag{32}$$

where $\widetilde{X}(\cdot)$ is the diffusion process obtained as the weak limit of the process $(Z_1^N(\cdot) + Z_2^N(\cdot) - N)/\sqrt{N}$, initialized by its steady-state distribution. Theorem 1 of Halfin and Whitt (1981) then implies that $\widetilde{X}(t)$ is distributed according to the steady-state distribution of the diffusion process characterized in Theorem 3.1. Hence, the steady-state queue-length distributions converge to the corresponding steady state of their diffusion limits.

Next, observe that by Little's Law, in steady state

$$\sqrt{N}\mathbf{E}W_i^N = \mathbf{E}\left[\frac{\sqrt{N}Q_i^N}{\lambda_i^N}\right].$$

Hence, to prove (14), it suffices to show that in steady state,

$$\mathbf{E}\left[\frac{\sqrt{N}Q_i^N}{\lambda_i^N}\right] \Rightarrow \mathbf{E}\widetilde{W}_i = \mathbf{E}\left[\frac{\widetilde{X}_i}{\tilde{\lambda}_i}\right].$$

Because (32) has already been established, it remains to show that the family of random variables $\{Q_i^N/\sqrt{N}\}$ is uniformly integrable (U.I.). This follows from the fact that the class level queue lengths are bounded above by the total queue length of the $M/M/N$ queue given by $\{(Q_1^N + Q_2^N)/\sqrt{N}\}$, which is itself U.I. (see Halfin and Whitt 1981, Corollary 1 and Lemma 1).

Finally, (13) follows from (32) and Proposition 3.2. □

Proof of Proposition 3.4

For any nonidling policy π, the total queue length of the two-class system is equal to that of an $M/M/N$ queue, and thus $X^{N,\pi} \Rightarrow \widetilde{X}$. Let $\widetilde{X}_i^\pi$ be the associated limit queue-length processes that are assumed to exist. Lemma A.2 of Puhalskii and Reiman (2000) is still valid and, as in Proposition 3.2, $\widetilde{W}_i^\pi(t) = \widetilde{X}_i^\pi(t)/\tilde{\lambda}_i$ for $i = 1, 2$. Because π is asymptotically compliant (i.e., $[\sqrt{N}W_2^{N,\pi}(t) - \widetilde{D}_2]^+ \Rightarrow 0$), it follows that

$$\widetilde{W}_2^\pi(t) \leqslant \widetilde{D}_2 \quad \forall t \geqslant 0 \;\Leftrightarrow\; \widetilde{X}_2^\pi(t) \leqslant \tilde{\lambda}_2\widetilde{D}_2 \quad \forall t \geqslant 0.$$

Note that π^* is asymptotically compliant. From Proposition 3.2,

$$\sqrt{N}W_2^{N,\pi^*}(t) - \widetilde{D}_2 \;\Rightarrow\; \frac{\widetilde{X}^{\pi^*}(t)^+ \wedge \tilde{\lambda}_2\widetilde{D}_2}{\tilde{\lambda}_2} - \widetilde{D}_2 \leqslant 0 \quad \forall\, t \geqslant 0.$$

The limiting control problem is the following:

$$\begin{aligned} \min \quad & \mathbf{E}\widetilde{W}_1 \\ \text{subject to} \quad & \widetilde{W}_2(t) \leqslant \widetilde{D}_2 \quad \forall t \geqslant 0, \\ & \widetilde{W}_i(t) = \frac{\widetilde{X}_i(t)}{\tilde{\lambda}_i}, \\ & \widetilde{X}_1(t) + \widetilde{X}_2(t) = \widetilde{X}(t)^+, \\ & \qquad \text{where } \widetilde{X} \text{ was defined in Theorem 3.1.} \end{aligned}$$

It is easy to show that the proposed threshold policy, denoted by π^*, is pointwise optimal. For every sample path ω, let $\widetilde{X}^+(t, \omega)$ denote the total queue-length trajectory. This is the same for all nonidling policies. Note that for all ω, all $t \geqslant 0$, and for all π such that $\widetilde{X}_2^\pi(t) \leqslant \tilde{\lambda}_2\widetilde{D}_2$,

$$\begin{aligned} \widetilde{X}_1^{\pi^*}(t, \omega) &= \arg\min\{X_1 \colon X_1 + X_2 = \widetilde{X}^+(t, \omega),\ X_2 \leqslant \tilde{\lambda}_2\widetilde{D}_2\} \\ &\leqslant \widetilde{X}_1^\pi(t, \omega). \end{aligned} \tag{33}$$

Because $\widetilde{W}_1^\pi = \widetilde{X}_1^\pi/\tilde{\lambda}_1$, this completes the proof. Given that (33) holds w.p. 1 for all $t \geqslant 0$, it follows that the threshold policy π^* also minimizes (the weaker objective) $\mathbf{E}\widetilde{W}_1$ subject to $\widetilde{W}_2(t) \leqslant \widetilde{D}_2$, for all t. □

Proof of Proposition 4.1

We need to prove that (22) has a unique and stable solution. We start with existence and uniqueness. We abbreviate $\tilde{\lambda}_i(0,0)$ by $\tilde{\lambda}_i$, $i=1,2$, and $\beta(\tilde{w}_1, \widetilde{D}_2)$ by $\beta(w)$. Let

$$h(w) = w - \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta(w))}{\beta(w)}e^{-\beta(w)\tilde{\theta}}.$$

It suffices to show that the equation $h(w)=0$ has a unique root, denoted by $\tilde{w}_1$.

Note that

$$\beta(w) > 0 \;\Leftrightarrow\; w > \underline{w} := \left(\delta\frac{\gamma}{\kappa} - \widetilde{D}_2\frac{\zeta}{\kappa}\right)^+.$$

We show that the equation $h(w)=0$ has a unique solution in $(\underline{w}, \infty)$.

From (20) and the expression for α given in Theorem 3.1, it follows that h is continuous in w in $(\underline{w}, \infty)$. Hence, we can choose $\epsilon > 0$ sufficiently small such that $h(\underline{w}+\epsilon) < 0$. Direct calculation verifies that

$$\frac{\partial h(w)}{\partial w} = 1 - \frac{1}{\tilde{\lambda}_1}\frac{1}{\beta(w)}e^{-\beta(w)\tilde{\theta}}\frac{\partial\beta}{\partial w}\left[\frac{\partial\alpha}{\partial\beta} - \frac{\alpha}{\beta(w)} - \alpha\tilde{\theta}\right] \geqslant 1,$$

where the last inequality follows from the fact that $\partial\alpha/\partial\beta < 0$ and $\partial\beta/\partial w = -\kappa/\gamma > 0$. It follows that there exists a unique solution to Equation (22), denoted by w_1^*, that satisfies $w_1^* > \underline{w}$. As a side comment, note that in addition, if $\underline{w} = 0$,

$$\tilde{w}_1 \leqslant \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta(0))}{\beta(0)}e^{-\beta(0)\tilde{\theta}},$$

and a similar bound can be obtained if $\underline{w} > 0$ by careful selection of ϵ above.

To prove stability, we perturb the system from its equilibrium $\tilde{w}_1$ and show that the expected waiting time will eventually return to $\tilde{w}_1$. We assume that at time t, the steady-state expected waiting time for class 1 service is $\tilde{w}_1(t)$. The corresponding arrival rates are given by $\tilde{\lambda}(t)$, while $\beta(t)$ will denote the appropriately scaled distance from heavy traffic. In the limit system, changes in $\tilde{w}_1(t)$ only affect $\beta(t)$ and not $\tilde{\lambda}(t)$. This follows from (21). As a result, fluctuations in the arrival rates are only captured in $\beta(t)$.

Let

$$\mathbf{E}\widetilde{W}_1(t) = \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta(t))}{\beta(t)}e^{-\beta(t)\tilde{\theta}},$$

and define $x(t) = \mathbf{E}\widetilde{W}_1(t) - \tilde{w}_1$,

$$\frac{dx(t)}{dt} = \frac{d\mathbf{E}\widetilde{W}_1(t)}{dt} = \frac{d\mathbf{E}\widetilde{W}_1(t)}{d\beta(t)}\frac{d\beta(t)}{dt}.$$

It is easy to see that $d\mathbf{E}\widetilde{W}_1(t)/d\beta(t) < 0$. Let $\beta(t^-)$ denote the β value at time t^- that induced the expected waiting time $\mathbf{E}\widetilde{W}_1(t)$. If $\mathbf{E}\widetilde{W}_1(t) < \tilde{w}_1$, then

$$\beta(t^-) > \beta(w) \quad\text{and}\quad \beta(t^+) = \delta - \widetilde{D}_2\frac{\zeta}{\gamma} - \frac{\kappa}{\gamma}\mathbf{E}\widetilde{W}_1(t) < \beta(w).$$

This implies that

$$\frac{d\beta(t)}{dt}\begin{cases} >0, & \mathbf{E}\widetilde{W}_1(t) > \tilde{w}_1, \\ <0, & \mathbf{E}\widetilde{W}_1(t) < \tilde{w}_1, \\ =0, & \mathbf{E}\widetilde{W}_1(t) = \tilde{w}_1, \end{cases} \quad\text{and}$$

$$\frac{dx(t)}{dt}\begin{cases} >0, & x(t)<0 \Rightarrow \mathbf{E}\widetilde{W}_1(t) < \tilde{w}_1, \\ <0, & x(t)>0 \Rightarrow \mathbf{E}\widetilde{W}_1(t) > \tilde{w}_1, \\ =0, & x(t)=0 \Rightarrow \mathbf{E}\widetilde{W}_1(t) = \tilde{w}_1, \end{cases}$$

which proves that $x(t)\to 0$, and equivalently, $\mathbf{E}\widetilde{W}_1(t)\to\tilde{w}_1$ as $t\to\infty$. □

Proof of Proposition 4.2

Let $\rho^N = \rho^N(w_1^N, D_2^N)$. First, we prove that $\liminf_{N\to\infty}\rho^N \geqslant 1$. By contradiction, suppose that there is a subsequence $\{N_j\}$ of $\{N\}$ such that $\lim_{j\to\infty}\rho^{N_j} = 1-\phi < 1$. We will show that this implies that $\lim_{j\to\infty} w_1^{N_j} = 0$. Let $Q^N(t)$ be the total queue length (jobs in the system but not in service) of system N at time t. Let Q^N be the steady-state total system queue length, and let Z^N be the steady-state total number of jobs in the system (i.e., in queue and in service). Let $\alpha^N := \mathbf{P}(Z^N \geqslant N)$. Then, one can show that $\mathbf{E}Q^N = \mu\rho^N(1-\rho^N)^{-1}\alpha^N$ (this follows from Halfin and Whitt 1981, Lemma 1) and the fact that $\mathbf{E}Q^N = \sum_{k=N}^{\infty} k(\mathbf{P}(Z^N=k) - N\alpha^N)$. The following lemma shows that $\lim_{j\to\infty}\mathbf{E}Q^{N_j} = 0$.

Lemma A.1 *Suppose that* $\lim_{j\to\infty}\rho^{N_j} = 1-\phi < 1$. *Then,* $\lim_{j\to\infty}\alpha^{N_j} = 0$.

Proof. As in Halfin and Whitt (1981, Proposition 1), one can show that $\alpha^N = [1+\gamma^N/\xi^N]^{-1}$, where

$$\gamma^N = \sum_{k=0}^{N-1}\frac{1}{k!}(N\rho^N)^k e^{-N\rho^N} \quad\text{and}\quad \xi^N = \frac{(N\rho^N)^N e^{-N\rho^N}}{N!(1-\rho^N)}.$$

Now, γ^N can be thought of as $\mathbf{P}(S^N \leqslant N-1)$, where S^N is a Poisson random variable with parameter $N\rho^N$, and thus,

$$\begin{aligned}\gamma^N &= \mathbf{P}(S^N \leqslant N-1) \\ &= \mathbf{P}(Y^N := (N\rho^N)^{-1/2}[s^N - N\rho^N] \leqslant \nu^N),\end{aligned}$$

where

$$\nu^N = (1-\rho^N)N^{1/2}\rho^N - 1/2 - (N\rho^N)^{-1/2}.$$

Because Y^N converges weakly to a standard normal random variable (see Halfin and Whitt 1981, p. 574),

and $\lim_{j\to\infty}\nu^{N_j}=\infty$, we have $\lim_{j\to\infty}\gamma^{N_j}=1$. Now, using Stirling's formula, we obtain

$$\xi^N \approx \frac{\exp(N[1-\rho^N+\log\rho^N])}{\sqrt{2\pi N}(1-\rho^N)},$$

which implies that $\lim_{j\to\infty}\xi^{N_j}=0$ (note that $1-\rho^{N_j}+\log\rho^{N_j}<0$ for all j such that $\rho^{N_j}<1$). Therefore, $\lim_{j\to\infty}\alpha^{N_j}=0$. □

To complete the proof of the first part of the proposition, note that $0\leqslant Q_1^N(t)\leqslant Q^N(t)$ for all t and N. Therefore, $\lim_{j\to\infty}\mathbf{E}Q_1^{N_j}=0$. From Little's law it follows that $\lim_{j\to\infty}w_1^{N_j}=0$. This implies that for every $\epsilon>0$, there exists $j(\epsilon)$ such that for all $j>j(\epsilon)$, $w_1^{N_j}<\epsilon$. On the other hand, Assumption 1 implies that for all $\epsilon>0$, there exists $j'(\epsilon)$ such that $\rho_{\text{eff}}^{N_j}\geqslant 1-\epsilon$ for all $j>j'(\epsilon)$. The continuity of the demand function for all N then leads to a contradiction. The second part of the proof is to show that $\limsup_{N\to\infty}\rho^N\leqslant 1$. This is true, because if for any N, $\rho^N>1$, it would imply that $w_1^N=\infty$, which is a contradiction because the system is stabilizable. □

Proof of Proposition 4.3

Let w_1^N be the steady-state expected class 1 waiting time in equilibrium, and denote by $\lambda_i^N:=\lambda_i^N(w_1^N,D_2^N)$, $i=1,2$, and $\rho^N:=\rho^N(w_1^N,D_2^N)$. We start by observing that the arrival rates into each class may be written as $\lambda_i^N=f_i(w_1^N,D_2^N)\Lambda_{\text{eff}}^N$, where $f_i(0,0)>0$, and $f:=f_1+f_2$ is strictly decreasing and $f(0,0)=1$. Hence, from Assumption 1 the traffic intensity into the system can be expressed as

$$\rho^N=f(w_1^N,D_2^N)\left(1-\frac{\delta}{\sqrt{N}}\right). \tag{34}$$

From Proposition 4.2, $\lim_{N\to\infty}\rho^N=1$. This implies that $\lim_{N\to\infty}w_1^N=0$, which in turn implies that $\lim_{N\to\infty}(\lambda_i^N/N)=f_i(0,0)\mu$, where λ_i^N is the arrival rate into class i in equilibrium.

Let Q_i^N be the steady-state queue length of class i ($i=1,2$) in equilibrium, and let Z^N be the steady-state total number of jobs in the system in equilibrium. Following Theorem 3.1, it is sufficient (and necessary) to show that

$$\lim_{N\to\infty}\mathbf{P}(Z^N\geqslant N)=\alpha \tag{35}$$

for some $\alpha\in(0,1)$. We will first rule out the cases $\alpha=0$ and $\alpha=1$. Then, we will show that this limit is unique by considering two converging subsequences $\{N_j\}_{j=1}^\infty$ and $\{N_k\}_{k=1}^\infty$, such that $\lim_{j\to\infty}\mathbf{P}(Z^{N_j}\geqslant N_j)=\alpha_1$ and $\lim_{k\to\infty}\mathbf{P}(Z^{N_k}\geqslant N_k)=\alpha_2$, and proving that $\alpha_1=\alpha_2$.

Note that from (34) we get that

$$\liminf_{N\to\infty}\sqrt{N}(1-\rho^N)\geqslant\delta. \tag{36}$$

Consider a subsequence $\{N_j\}_{j=1}^\infty$ for which $\lim_{j\to\infty}\mathbf{P}(Z^{N_j}\geqslant N_j)$ exists and is equal to some $\alpha\in[0,1]$. From Halfin and Whitt (1981, Proposition 1) it follows that if $\alpha=0$, then $\lim_{j\to\infty}\sqrt{N_j}(1-\rho^{N_j})=\infty$. Using standard formulas for the single-class $M/M/N$ system we get that the total queue length satisfies

$$\mathbf{E}[Q_1^N+Q_2^N]=\frac{\mathbf{P}(Z^N\geqslant N)\rho^N N\mu}{N(1-\rho^N)};$$

hence, $\alpha=0$ implies that

$$\lim_{j\to\infty}\frac{\mathbf{E}[Q_1^{N_j}+Q_2^{N_j}]}{\sqrt{N_j}}=0.$$

Specifically,

$$\lim_{j\to\infty}\frac{\mathbf{E}[Q_1^{N_j}]}{\sqrt{N_j}}=0,$$

which together with Little's law implies that

$$\lim_{j\to\infty}\sqrt{N_j}w_1^{N_j}=\lim_{j\to\infty}\frac{\mathbf{E}[Q_1^{N_j}]/\sqrt{N_j}}{\lambda_1^{N_j}/N_j}=0.$$

Now, from the Taylor expansion of (19) applied to the subsequence $\{N_j\}_{j=1}^\infty$ with $\tilde{w}=0$, we get that

$$\lim_{j\to\infty}\sqrt{N_j}(1-\rho^{N_j})=\delta-\frac{\zeta}{\gamma}<\infty.$$

This is a contradiction.

To rule out the possibility that $\alpha=1$, note that if indeed $\alpha=1$, then $\lim_{j\to\infty}\sqrt{N_j}(1-\rho^{N_j})=0$. That is, $\beta=0$, which implies that $\sqrt{N_j}w_1^{N_j}\to\infty$. At the same time, as has been observed in the beginning of this proof, $\lim_{j\to\infty}w_1^{N_j}=0$. Rederiving (19) in this case where $w_1^{N_j}=0+o(1)$ yields $\sqrt{N_j}(1-\rho^{N_j})\to\infty$, which is a contradiction.

Now, suppose that $\alpha_1,\alpha_2\in(0,1)$ are limits of two different subsequences, and $\alpha_1\neq\alpha_2$. Theorem 3.1 applied to the two subsequences implies that $\lim_{j\to\infty}\sqrt{N_j}(1-\rho^{N_j})=\beta_1$ and $\lim_{k\to\infty}\sqrt{N_k}(1-\rho^{N_k})=\beta_2$ for some $\beta_1,\beta_2\in(0,\infty)$ with $\beta_1\neq\beta_2$. Now, consider two subsequences of systems with arrival rates $\lambda_i^{N_j}$ and $\lambda_i^{N_k}$, respectively. Proposition 3.2 applied to these two subsequences implies that (11) holds when N is replaced by N_j or N_k, respectively.

Next, we use the equilibrium analysis associated with Proposition 4.1 with N replaced by N_j or N_k. This implies that (22) will apply with two different values of α and β, which is a contradiction to the uniqueness of the solution of (22). □

Endnotes

1. The assumption that the service times for the two classes are exponentially distributed with equal means are modeling idealizations that make the analysis of the limiting diffusions tractable. Some results that allow for phase-type distributions can be found in Puhalskii and Reiman (2000). A new approach that leads to a tractable analysis for systems with different service rates has been recently developed in Maglaras and Zeevi (2004).

2. Alternatively, one can assume that customers have formed an accurate estimate of $\mathbf{E}W_1$ from prior experience.

3. To be precise, we say that the system admits a unique *equilibrium* if there exists a unique steady-state probability distribution of the underlying continuous time Markov chain, such that the expected waiting time for class 1 users when taken w.r.t. this distribution, $\mathbf{E}W_1$, induces time homogenous arrival rates λ_i through (2) and (3) that, in turn, are consistent with the aforementioned steady-state distribution.

4. That is, for large N, customer choice behavior can be approximated by a linear demand model.

5. Note, however, that these results do not show that the N-server system itself has a unique and stable equilibrium, as this would have to involve explicit analysis of the $M/M/N$ model. Numerical evidence in §5 confirms that the actual behavior of the N-server queue is close to what the asymptotic analysis predicts, making the existence and uniqueness of an equilibrium regime plausible.

6. In computing the various quantities of interest for a system with finite N and a given D_2^N, we only need to take a Taylor expansion with respect to w_1^N (and not D_2^N). For example, the equivalent of (16) would be $\lambda_1^N(w_1^N, D_2^N) \approx \Lambda^N \mathbf{P}^\tau(u_1^\tau(0+\tilde{w}_1/\sqrt{N}+o(1/\sqrt{N})) \geqslant u_2^\tau(D_2^N)^+)$.

7. Another example where simulation was used for performance analysis of a call center can be found in Saltzman and Mehrotra (2001), where they study a system with two classes of jobs, with static priorities.

8. In principle, d_2 is also a design parameter, but its selection is often influenced by what is acceptable and/or expected by the customers for the specific service that is being offered.

Acknowledgments

The authors thank two referees, Ward Whitt, the associate editor, as well as Marty Reiman, Bill Massey, Avi Mandelbaum, and Assaf Zeevi for many helpful comments on this work. In particular, the comments made by Ward Whitt led to the proof of Proposition 4.3, which was earlier posed as an assumption.

References

Anderson, S. P., A. de Palma, J.-F. Thisee. 1996. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* Forthcoming.

Bell, S. L., R. J. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* **11** 608–649.

Billingsley, P. 1968. *Convergence of Probability Measures*. John Wiley and Sons, New York.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.

Brandt, A., M. Brandt. 1999. On a two-queue priority system with impatience and its applications to a call center. *Methodology Comput. Appl. Probab.* **1** 191–210.

Call center statistics. 2001. www.callcenternews.com/resources/statistics.shtml.

Dai, J. G. 1995. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

Dai, J. G., R. J. Williams. 1995. Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory Probab. Its Appl.* **50** 3–53.

Fleming, P., A. Stolyar, B. Simon. 1994. Heavy traffic limit for a mobile phone system loss model. *Proc. 2nd Internat. Conf. Telecomm. Syst. Mod. Anal.* Nashville, TN.

Gans, N., Y.-P. Zhou. 2003. A call-routing problem with service-level constraints. *Oper. Res.* **51**(2) 255–271.

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

Glynn, P. W. 1990. Diffusion approximations. D. Heyman, M. Sobel, eds. *Stochastic Models*. *Handbooks in OR & MS*, Vol. 2. North-Holland, Amsterdam, The Netherlands, 145–198.

Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37**(1) 84–97.

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

Harrison, J. M. 1996. The BIGSTEP approach to flow management in stochastic processing networks. F. Kelly, S. Zachary, I. Ziedins, eds. *Stochastic Networks: Theory and Applications*. Oxford University Press, Oxford, U.K., 57–90.

Hassin, R., M. Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett.* **17** 41–45.

Jennings, O., A. Mandelbaum, W. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42**(10) 1383–1394.

Kelly, F. P., C. N. Laws. 1993. Dynamic routing in open queueing models: Brownian models, cut constraints and resource pooling. *Queueing Systems* **13** 47–86.

Kolesar, P. J., L. V. Green. 1998. Insights on service system design from a normal approximation to Erlang's delay formula. *Prod. Oper. Management* **7** 282–293.

Maglaras, C. 2000. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Ann. Appl. Probab.* **10**(3) 897–929.

Maglaras, C., J. Van Mieghem. 2004. Admission and sequencing control under delay constraints with applications to GPS and GLQ. *Eur. J. Oper. Res.* Forthcoming.

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

Maglaras, C., A. Zeevi. 2004. Diffusion approximations for a Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.* Forthcoming.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. Leadtime constraints in stochastic processing networks under heavy traffic conditions. *Queueing Systems* **39** 23–54.

Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19**(2) 946–954.

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

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Math. Oper. Res.* **9** 441–458.

Saltzman, R. M., V. Mehrotra. 2001. A call center uses simulation to drive strategic change. *Interfaces* **31**(3) 87–101.

Talluri, K., G. van Ryzin. 2004. Revenue management under a general discrete-choice model of demand. *Management Sci.* Forthcoming.

Teh, Y.-C., A. Ward. 2002. Critical thresholds for dynamic routing in queueing networks. *Queueing Systems* **42** 297–316.

Whitt, W. 1974. Heavy traffic limits for queues: A survey. A. B. Clarke, ed. *Mathematical Methods in Queueing Theory*. *Lecture Notes in Econom. and Math. Systems*, Vol. 98. Springer-Verlag, New York, 307–350.

Whitt, W. 1999. Using different response-time requirements to smooth time-varying demand for service. *Oper. Res. Lett.* **24** 1–10.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51**(4) 531–542.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in telequeues: Theory and empirical support. *Management Sci.* **48** 566–583.