OPERATIONS RESEARCH
Vol. 54, No. 3, May–June 2006, pp. 419–435
ISSN 0030-364X | EISSN 1526-5463 | 06 | 5403 | 0419

DOI 10.1287/opre.1060.0285
© 2006 INFORMS

Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, a-bassamboo@northwestern.edu

J. Michael Harrison
Graduate School of Business, Stanford University, Stanford, California 94305, harrison_michael@gsb.stanford.edu

Assaf Zeevi
Graduate School of Business, Columbia University, New York, New York 10027, assaf@gsb.columbia.edu

This paper analyzes a call center model with m customer classes and r agent pools. The model is one with doubly stochastic arrivals, which means that the m-vector λ of instantaneous arrival rates is allowed to vary both temporally and stochastically. Two levels of call center management are considered: staffing the r pools of agents, and dynamically routing calls to agents. The system manager's objective is to minimize the sum of personnel costs and abandonment penalties. We consider a limiting parameter regime that is natural for call centers and relatively easy to analyze, but apparently novel in the literature of applied probability. For that parameter regime, we prove an asymptotic lower bound on expected total cost, which uses a strikingly simple distillation of the original system data. We then propose a method for staffing and routing based on linear programming (LP), and show that it achieves the asymptotic lower bound on expected total cost; in that sense the proposed method is *asymptotically optimal*.

Subject classifications: stochastic model applications; call centers; queueing; dynamic routing; fluid limits; doubly stochastic; asymptotic analysis; discrete review; performance bounds; abandonments; staffing.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received June 2004; revised March 2005; accepted May 2005.

1. Introduction

This paper is concerned with two central problems in the management of a telephone call center. The first is a *static design problem* that determines staffing levels according to which agents will later be assigned to work schedules. The second is a *dynamic control problem* whose solution determines the real-time assignment of incoming calls to agents. While these two goals are clearly interrelated, their complexity has led most researchers to treat them separately, in a hierarchical manner. The method we propose in this paper simultaneously addresses both problems.

We consider a call center model with m customer classes and r agent pools. As usual in operations research studies, we view a call center as a queueing system, frequently referring to callers as "customers" and to call center agents as "servers." Each of the pools consists of interchangeable servers whose common skills dictate the possible customer classes that these agents can serve, and the speed at which such service is delivered. There can be more than one pool that serves a particular customer class, and conversely, there can be more than one customer class that is served by a particular agent pool.

Customers of the various classes arrive randomly over time, and those who cannot be served immediately wait in a (possibly virtual) infinite-capacity buffer. Two important assumptions are made in this regard to capture recognized "real-world" phenomena. First, we assume that customers of any given class will abandon their calls if forced to wait too long before commencement of service (see Gans et al. 2003 for further discussion). Second, we allow the arrival rates for the various customer classes (expressed in units such as calls per minute) to be both temporally and stochastically variable, i.e., the m-vector of instantaneous arrival rates is itself a stochastic process. As Gans et al. (2003) acknowledge in §4.4 of their survey paper, such a view is realistic, although most published papers on both call center staffing and dynamic routing treat average arrival rates as known and constant over the relevant planning period.

We assume that there are two types of costs: the direct and indirect variable costs associated with agents staffing the various pools, which we call "personnel costs"; and abandonment costs that capture the penalty associated with "lost business." The objective of the system manager is to minimize the sum of these two operating costs in selecting a staffing level for each pool and then a routing rule by which calls will be assigned to servers. (A precise description of this call center model and details of the various probabilistic assumptions are deferred to §2.)

For any given staffing decision, the dynamic routing problem faced by the system manager is the following. First, whenever a customer arrives and there exist one or more idle servers who can handle that customer's class, the system manager must choose between routing the customer immediately to one of them versus having the customer wait for later disposition. If the customer is to be routed immediately, there may be a further choice regarding the server pool to which it will be routed. Second, each time a server completes the processing of a customer and there exist waiting customers of one or more classes that the server can handle, the system manager must choose between routing one of those customers to the server immediately versus idling the server in anticipation of future arrivals. These resource allocation decisions are conditioned on system status information at the time of the choice, including the number of customers waiting in the various buffers and the number of idle servers in the various pools.

In the context of a multiclass/multipool call center, the problem in the previous paragraph is often referred to as *skills-based routing* (see Gans et al. 2003, §5.1 for further discussion). This dynamic routing problem is quite difficult to address by means of exact analysis, even under simplifying Markovian assumptions. In fact, even in the case where average arrival rates are constant and known, Gans et al. (2003, §5.1) describe the dynamic routing problem as extremely challenging, with most work to date done on specific problem instances, using various approximations, and often resulting only in implicit characterization of routing rules. In light of this, it is not surprising that staffing decisions and routing objectives are most often treated in a hierarchical manner as essentially separate problems.

The approach we propose in this paper does not attempt to disentangle design (staffing) and control (routing) decisions. In particular, it jointly optimizes over both objectives in a manner that gives rise to a simple staffing algorithm and an explicit characterization of dynamic routing policies. The implementation of this method is straightforward, and it will be shown to be optimal in a precise mathematical sense. To substantiate that last statement in a setting where demand rates may vary both temporally and stochastically, we propose a novel asymptotic regime and an approximation method that gives rise to several important insights.

Throughout the remainder of this paper, when we speak of the system manager's *dynamic control problem*, that is understood to mean the skills-based routing problem described above. This more abstract terminology makes for economy of expression, and also promotes a symmetric view of the system manager's problem, which can be viewed either as one of routing customers or as one of allocating servers. The main contributions of our paper can then be summarized as follows.

(1) We propose a new *asymptotic parameter regime* for studying call centers. In this regime service rates and abandonment rates are accelerated in a linear manner, while the arrival rates grow super-linearly. The key feature of this *two-scale* parameter regime is that the limiting system "equilibrates instantly" and the dynamic control problem becomes tractable (see Proposition 1).

(2) We develop an asymptotic lower bound on achievable expected cost, referred to hereafter as an *asymptotic performance bound*, that uses a strikingly simple distillation of the original system data (see Theorem 1).

(3) We establish *asymptotic optimality* of a simple staffing and dynamic control policy based on linear programming (LP). That is, we prove that our LP-based method achieves the asymptotic performance bound referred to above when the arrival rate process is directly observable (see Theorem 2).

(4) In the case where the arrival rate process is not observable, we describe a policy that estimates arrival rates "on the fly" and uses these values as "plug in" estimates in the previous dynamic control policy. For a suitable class of estimators (see Proposition 3), this approach is shown to be asymptotically optimal (see Theorem 3). Based on these ideas, we develop a *discrete-review nonpreemptive policy* that is more suitable for implementation purposes, and prove it is asymptotically optimal (see Theorem 4).

Numerical examples will be advanced to validate the accuracy of the approximations discussed above.

Existing Analytical Approaches and Related Work. As indicated by Gans et al. (2003, §5.1), both staffing and dynamic routing problems in multiclass/multipool call centers are essentially outside the reach of exact analytical methods (for an exception, see, e.g., Gans and Zhou 2003). Thus, most research on these problems has focused on various forms of approximations, a particularly prominent role being played by two asymptotic regimes.

The first is the so-called "conventional" heavy-traffic regime. Here, the number of servers is held fixed while service and arrival rates are accelerated linearly in such a way that system utilization approaches one. In this manner, and under appropriate regularity conditions, one can derive so-called heavy-traffic limit theorems which provide rigorous approximations to the original system dynamics (cf. Whitt 2001). Harrison and Lopez (1999) is essentially the first study in which a dynamic control problem is explicitly solved in the multiclass/multipool setting using conventional heavy-traffic limit theory (see also Gans and van Ryzin 1997, Harrison 1998, Bell and Williams 2001, and Mandelbaum and Stolyar 2004).

The second regime considered in the literature to date is the so-called many-server heavy-traffic regime, which was first made rigorous by Halfin and Whitt (1981). In this regime, the arrival rate and the number of servers are increased in a fixed proportion to each other while the system utilization approaches one. There is general accord that this regime is more appropriate for describing the dynamics of a call center than the conventional heavy-traffic regime; see, e.g., Whitt (1992), Garnett et al. (2002), and the recent survey by Gans et al. (2003). In terms of dynamic control, Harrison and Zeevi (2004) and Atar et al. (2004) are the first to analyze a multiclass single-pool system in the Halfin-Whitt regime and to characterize the optimal control policy. Unfortunately, this requires one to solve a nonlinear partial differential equation whose dimension is equal to the number of customer classes, and is therefore not a practical means of deriving implementable control policies. Armony (2005) analyzes staffing and routing decisions in a single-class/multipool system operating in the Halfin-Whitt regime (see also Armony and Maglaras 2004). Finally, Armony et al. (2005) studies staffing and server allocation decisions in a multiclass/single-pool system under the assumption that service rates for all classes are identical.

The two strands of research summarized above assume that the arrival rates do not exhibit any stochastic or temporal variation. However, statistical evidence suggests that demand patterns observed in real call centers exhibit such properties (see Brown et al. 2005, §1.1). When arrival rates are allowed to vary with time, simplified system dynamics in the form of fluid-limit differential equations can often be derived, yet are difficult to solve (see, e.g., Mandelbaum et al. 1998 and Whitt 2006, as well as the references therein). Jennings et al. (1996) analyze a particular case of a staffing problem in a single-class single-pool setting with time-varying demand. Some implications of demand uncertainty are discussed in Chen and Henderson (2001). A recent paper by Wallace and Whitt (2005) investigates with the aid of simulation several ideas for staffing and routing decisions in a multiclass/multipool system.

The asymptotic regime described in this paper is closely related to the concept of pointwise stationary approximations, which was first described in the context of a simple Markovian queueing model with nonstationary arrivals by Green and Kolesar (1991) and subsequently made rigorous by Whitt (1991) (for further refinements, see Massey and Whitt 1998). The asymptotic regime that is used in these papers involves uniform acceleration of transition rates in the underlying Markov chain, i.e., accelerating arrival rates and service rates by the same factor.

The point of departure for our current work is the recent paper by Harrison and Zeevi (2005), which describes a staffing method for a call center with multiple customer classes and multiple agent pools under arrival rates that vary temporally and stochastically. That method reduces the staffing problem, whose objective is to minimize the sum of personnel costs and expected abandonment costs, to a static stochastic program which takes the form of an LP with recourse. Numerical experiments in Harrison and Zeevi (2005) indicate that this optimization problem results in "near optimal" staffing vectors. Moreover, it is informally argued that the minimum value of the objective function yields a lower bound on system performance. This paper is largely concerned with a rigorous derivation of this bound, and an articulation of staffing and control policies that achieve it.

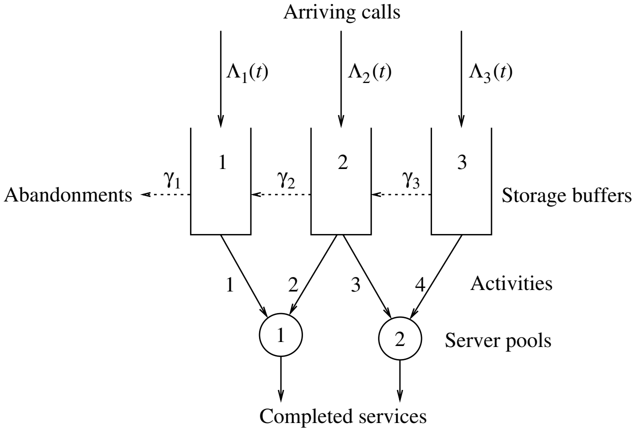
The remainder of this paper is organized as follows. Section 2 provides a precise description of the call center model and economic objective. Section 3 describes the asymptotic parameter regime used in later analysis. Section 4 gives the main results, and §5 presents "picture proofs" of these results by means of simulation experiments. Section 6 concludes with some remarks and directions for future research. Proofs of the main results are given in Appendix A. Appendix B, which contains proofs of auxiliary results, is available as an online companion at http://or.pubs.informs.org/Pages/collect.html.

2. Problem Formulation

In our general call center model, there are m customer classes and r server pools. Server pool k consists of b_k interchangeable servers ($k = 1, \ldots, r$), and servers in a given pool may be *cross-trained* to handle customers of several different classes. By the same token, there may be several pools that are able to handle a given customer class. Customers of the various classes arrive randomly over time according to a doubly stochastic Poisson process with *instantaneous arrival rates* given by $\Lambda_1(t), \ldots, \Lambda_m(t)$; a more precise definition will be given later. Those customers who cannot be served immediately wait in a (possibly virtual) infinite-capacity buffer that is dedicated to their specific class. An example with $m = 3$ customer classes and $r = 2$ server pools is shown schematically in Figure 1.

To describe server capabilities, we shall use the notion of processing "activities," following Harrison and Lopez (1999). There are a total of n processing activities available to the system manager in our call center model, each of which corresponds to agents from one particular pool serving customers of one particular class (activities are denoted by solid arrows leading from buffers to server pools in Figure 1). For each activity $j = 1, \ldots, n$, we denote by $i(j)$ the customer class being served, by $k(j)$ the server pool involved, and by μ_j the associated mean service rate (that is, the reciprocal of the mean of the associated service time distribution). The actual service times are taken to be exponentially distributed random variables with the above rates, these being independent of one another and also of the arrival processes. Note that we allow the service time

Figure 1. A call center with three customer classes, two agent pools, and four activities.



distribution of a customer to depend on both the customer's class and on the pool to which the server belongs.

An important assumption of our model is that customers of any given class will abandon their calls if forced to wait too long for the commencement of service; abandoned calls are represented by the horizontal dotted arrows emanating from the storage buffers in Figure 1. Specifically, there is associated with each class i customer an exponentially distributed "impatience" random variable τ that has mean $1/\gamma_i$, independent of the impatience random variables characterizing other customers, and of service times and arrival processes. The customer will abandon the call when his or her waiting time in queue (exclusive of service time) reaches a total of τ time units. This assumption is quite standard in call center modelling (cf. Garnett et al. 2002, Harrison and Zeevi 2004, and Gans et al. 2003).

As stated in the introduction, our problem formulation and analysis emphasize an operating environment in which the instantaneous arrival rates are random and time varying, consistent with the observations made in Brown et al. (2005, §1.1). In addition, service times and the impatience random variables associated with individual customers are exponentially distributed, are independent of one another, and are independent of the arrival processes. To spell out this structure more precisely, we take as given a complete probability space $(\Omega, \mathscr{H}, \mathbb{P})$ on which are defined m continuous, nonnegative, integrable arrival rate processes, $\Lambda_i = (\Lambda_i(t)\colon 0 \leqslant t \leqslant T)$, satisfying $\mathbb{E}[\int_0^T \Lambda_i(s)\,ds] < \infty$ for $i = 1, \ldots, m$, plus $3m$ mutually independent Poisson processes, each with a unit intensity parameter, which are denoted $N_i^{(l)} = (N_i^{(l)}(t)\colon 0 \leqslant t < \infty)$ for $i = 1, \ldots, m$ and $l = 1, 2, 3$. The Poisson processes $N_i^{(l)}$ are further taken to be independent of the arrival rate processes Λ_i. We use the processes $N_i^{(1)}$ to construct arrivals in our model, defining

$$F_i(t) := N_i^{(1)}\left(\int_0^t \Lambda_i(s)\,ds\right) \quad \text{for } i = 1, \ldots, m \text{ and } 0 \leqslant t \leqslant T. \tag{1}$$

This is a standard construction of a doubly stochastic Poisson process (cf. Bremaud 1981); we interpret $F_i(t)$ as the cumulative number of class i arrivals up to time t. The unit-rate Poisson processes $N_i^{(2)}$ and $N_i^{(3)}$ will be used to construct service completions and abandonments, respectively, under a given dynamic control policy, via relationships analogous to (1).

For future purpose, it will be useful to introduce the following matrices. Let R and A be an $m \times n$ matrix and an $r \times n$ matrix, respectively, defined as follows: for each $j = 1, \ldots, n$ set $R_{ij} = \mu_j$ if $i = i(j)$ and $R_{ij} = 0$ otherwise, and set $A_{kj} = 1$ if $k = k(j)$ and $A_{kj} = 0$ otherwise. Thus, one interprets R as an *input-output matrix*, precisely as in Harrison and Lopez (1999): its (i, j)th element specifies the average rate at which activity j removes class i customers from the system. Also, A is a *capacity consumption matrix* as in Harrison and Lopez (1999): its (k, j)th element is 1 if activity j draws on the capacity of server pool k and is zero otherwise. We define an $m \times n$ matrix B by setting $B_{ij} = 1$ if $i(j) = i$ and $B_{ij} = 0$ otherwise; elements of this matrix show which server pools conduct which activities. Finally, let $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_m)$ denote the *abandonment rate matrix*.

Control Formulation and Objective. The system manager confronts a two-stage decision problem. First, the system manager chooses a *staffing vector* $b = (b_1, \ldots, b_r)$ in $\mathbb{R}_+^r$, whose kth component is the number of servers to be employed during the specified planning period for server pool k; by assumption, this decision cannot be revised as actual demand is observed during the period.

Second, the system manager chooses a *dynamic control policy* that determines how the calls of various customer classes are routed to server pools. The mathematical approach that we shall adopt in formulating the dynamic control problem may appear both clumsy and erroneous at first glance. The apparent error is that certain physically important constraints are deleted in our formulation, or to put it another way, our definition of an admissible control is overly generous. The seemingly clumsy aspect of our formulation is that we speak in terms of *control processes*, as opposed to specifying the controls as functions of observed states. However, the approach we adopt is an efficient one mathematically, given the specific objectives of this paper, and we shall discuss the "correctness" of our formulation after the formal mathematical definitions have been laid out.

A *dynamic control* is defined as a stochastic process $X = (X(t)\colon 0 \leqslant t \leqslant T)$ taking values in $\mathbb{R}_+^n$, whose sample paths are right continuous with left limits and Lebesgue integrable. Writing $X(t) = (X_1(t), \ldots, X_n(t))$, we interpret $X_j(t)$ as the number of servers engaged in activity j at time t. A dynamic control X is said to be *admissible* with respect to a staffing vector b if there exist processes Z and Q, both having time domain $[0, T]$, both taking values in $\mathbb{R}_+^m$, and both necessarily unique (see below), that jointly satisfy conditions (2)–(4) below for all $t \in [0, T]$. As an aid to intuition, it is useful to have the following interpretations from the outset: $Z_i(t)$ represents the number of class i customers in the system at time t (we call Z the *headcount process*, and Z_i is its ith component); $Q_i(t)$ represents the number of class i customers in the buffer that are waiting for service at time t (we call Q the *queue-length process*, and Q_i is its ith component). The essential relationships among these processes are the following:

$$AX(t) \leqslant b, \tag{2}$$

$$Q(t) = Z(t) - BX(t) \geqslant 0, \tag{3}$$

$$Z_i(t) = F_i(t) - N_i^{(2)}\left(\int_0^t (RX)_i(s)\,ds\right) - N_i^{(3)}\left(\int_0^t \gamma_i Q_i(s)\,ds\right) \geqslant 0 \quad \text{for all } i = 1, \ldots, m. \tag{4}$$

The second term on the right-hand side of (4) is interpreted as the cumulative number of class i service completions up to time t, while the third term represents cumulative class i abandonments; according to (4), the instantaneous departure rate for class i customers due to abandonments is $\gamma_i Q_i$, and the instantaneous departure rate for class i due to service completions is $\sum \mu_j X_j$, where the sum is taken over activities j that serve class i. This is consistent with the verbal model description provided earlier.

Our first constraint (2) simply requires that the number of servers in various pools that are engaged in some activity (as opposed to idle) at time t cannot exceed the total number of servers in each pool. In the second constraint (3), $BX(t)$ is a vector whose components represent the number of servers allocated to various customer classes at time t. The constraint therefore prohibits allocating to a given class a number of servers which exceeds the headcount in that class. The final admissibility condition (4) is the system dynamics equation.

Given a dynamic control X, the headcount process Z and the queue-length process Q can be viewed as the unique solution of (3) and (4): one simply constructs the paths of Z and Q from jump to jump in accordance with those relationships, starting from time zero. Because the primitive processes $N_i^{(l)}$ are independent Poisson processes, the probability of simultaneous jumps (for example, a service completion and an abandonment occurring simultaneously) is zero, and hence there almost surely exists *at most* one pair (Z, Q) satisfying (3) and (4).

Of course, the usual way to describe a dynamic control policy is in state feedback form. Having done so, one could then define the associated stochastic processes Z and Q as the solution of a system of stochastic equations, and *finally* our process X could be defined by applying the state-feedback rule to the trajectory of (Z, Q). By taking X as the primitive specification of a control policy, we are able to eliminate a whole level of mathematical description in developing our theory of asymptotic optimality.

Next, we describe the economic objective of the system manager. Let $p = (p_1, \ldots, p_m)$ be the *penalty cost vector*, where p_i is the cost associated with abandonment of a class i customer, and let $c = (c_1, \ldots, c_r)$ be the *personnel cost vector*, where c_k is the cost of employing a server in pool k for the entire planning horizon $[0, T]$. The objective of the system manager is to choose a staffing vector b and an admissible dynamic control X that jointly minimize the sum of personnel costs and expected abandonment penalties for the various customer classes, which is given by

$$c \cdot b + \mathbb{E}\left[\sum_{i=1}^{m} p_i N_i^{(3)}\left(\int_0^T \gamma_i Q_i(s)\,ds\right)\right], \tag{5}$$

where $c \cdot b$ represents the inner product between the vectors c and b.

Discussion. A reader may reasonably object that our problem formulation suffers from the following errors of omission. First, we allow noninteger values for the staffing levels b_k and the server allocation $X_j(t)$. This is obviously unrealistic, although the key relationships (2)–(4) make mathematical sense even without the integrality restriction. (The stochastic process Z is automatically integer valued, but Q may take noninteger values if X does.) Second, we implicitly allow the system manager to interrupt services at will, without any associated penalty. Finally, our definition of an admissible control does not rule out clairvoyance on the part of the system manager. A realistic formulation would require that the control X be nonanticipating in an appropriate sense, but we do not do so for the following reason.

The asymptotic lower bound derived later in this paper applies to any family of staffing vectors and admissible controls, regardless of whether they have the defects enumerated above. We will eventually construct a family of LP-based policies (that is, staffing vectors and dynamic controls) that are integer valued, nonpreemptive, and suitably nonanticipating, and show that these policies achieve the asymptotic performance bound. Thus, in the limiting parameter regime that we consider, the errors of omission enumerated above do not allow the system manager to significantly reduce cost. The fact that we grant the system manager excessive power in our formulation simply strengthens our results.

3. An Asymptotic Parameter Regime

A Parametric Family of System Models. Let f: $\mathbb{R}_+ \to \mathbb{R}_+$ be a super-linear function, meaning that $x^{-1}f(x) \to \infty$ as $x \to \infty$. Let us define a sequence of system models indexed by $\kappa \in \mathbb{N}$. In the κth system, the arrival process is doubly stochastic with rate $\Lambda^\kappa(\cdot) = f(\kappa)\Lambda(\cdot)$, the input-output matrix is $R^\kappa = \kappa R$, and the abandonment matrix is $\Gamma^\kappa = \kappa\Gamma$. Thus, the arrival rates into all classes scale up super-linearly, while all service and abandonment rates scale up linearly. Because the servers work κ times faster, we also scale up personnel costs by a factor of κ, meaning that the personnel cost vector for the κth system is $c^\kappa = \kappa c$. One can express this assumption verbally by saying that the effective cost of capacity (that is, the expected cost of processing any given set of customers using any given set of activities) remains constant as κ varies. Because the arrivals are scaled up by a super-linear function $f(\cdot)$ while the service rates are only scaled up linearly, the number of servers required for nominal operation should increase without bound. Thus, this parameter regime is characterized by large arrival rates, a large number of servers, short service requirements, and impatient customers.

For each system in the sequence indexed by κ, the system manager must choose a staffing vector b^κ and a dynamic control X^κ. The dynamic control is a right continuous with left limits process $X^\kappa = (X^\kappa(t)\colon 0 \leqslant t \leqslant T)$ taking values in $\mathbb{R}^n_+$. Here, $X^\kappa(t) = (X^\kappa_1(t), \ldots, X^\kappa_n(t))$, where $X^\kappa_j(t)$ is interpreted as the number of servers engaged in activity j in the κth system at time $t \in [0, T]$. We denote the sequence of staffing vectors and dynamic control policies by $\{b^\kappa\}$ and $\{X^\kappa\}$, respectively. We next define the class of admissible policies.

Definition 1. A sequence of dynamic controls $\{X^\kappa\}$ is said to be admissible with respect to a given sequence of staffing vectors $\{b^\kappa\}$, if for each κ, the dynamic control X^κ is admissible with respect to the staffing vector b^κ, i.e., there exist processes Z^κ and Q^κ, both having time domain $[0, T]$, both necessarily unique, both taking values in $\mathbb{R}^m_+$, that jointly satisfy

$$AX^\kappa(t) \leqslant b^\kappa, \tag{6}$$

$$Q^\kappa(t) = Z^\kappa(t) - BX^\kappa(t) \geqslant 0, \tag{7}$$

$$\begin{aligned} Z^\kappa_i(t) = F^\kappa_i(t) &- N^{(2)}_i\left(\int_0^t (R^\kappa X^\kappa)_i(s)\,ds\right) \\ &- N^{(3)}_i\left(\int_0^t \gamma^\kappa_i Q^\kappa_i(s)\,ds\right) \geqslant 0, \end{aligned} \tag{8}$$

where

$$F^\kappa_i(t) = N^{(1)}_i\left(\int_0^t \Lambda^\kappa_i(s)\,ds\right)$$

for all $i = 1, \ldots, m$ and all $t \in [0, T]$.

We define the total cost for the κth system under the dynamic control X^κ and staffing level b^κ to be

$$\mathcal{J}^\kappa(X^\kappa, b^\kappa) = c^\kappa \cdot b^\kappa + \sum_{i=1}^m p_i N^{(3)}_i\left(\int_0^T \gamma^\kappa_i Q^\kappa_i\,ds\right).$$

Definition 2. A sequence of staffing vectors $\{b^\kappa_*\}$ along with a corresponding sequence of admissible dynamic controls $\{X^\kappa_*\}$ is said to be asymptotically optimal if, for any other admissible sequence of staffing vectors $\{b^\kappa\}$ and corresponding dynamic controls $\{X^\kappa\}$,

$$\limsup_{\kappa\to\infty} \frac{\mathbb{E}[\mathcal{J}^\kappa(b^\kappa_*, X^\kappa_*)]}{\mathbb{E}[\mathcal{J}^\kappa(b^\kappa, X^\kappa)]} \leqslant 1. \tag{9}$$

It will be shown later that a pair $\{b^\kappa_*\}$, $\{X^\kappa_*\}$ is asymptotically optimal if and only if $\mathbb{E}[\mathcal{J}^\kappa(X^\kappa_*, b^\kappa_*)] \sim \alpha f(\kappa)$, as $\kappa \to \infty$, where the constant α is minimal.

Limiting Dynamics. For the purpose of the next proposition, which characterizes the limiting behavior of the sequence of admissible controls and headcount processes, we make the following technical assumption:

$$\frac{\kappa \log \kappa}{f(\kappa)} \to 0 \quad \text{as } \kappa \to \infty. \tag{10}$$

Proposition 1. *Assuming that* (10) *holds, consider any sequence of staffing vectors* $\{b^\kappa\}$ *and corresponding admissible dynamic controls* $\{X^\kappa\}$ *such that*

$$\int_0^t \frac{\kappa X^\kappa(s)\,ds}{f(\kappa)} \to \int_0^t X(s)\,ds \quad \textit{a.s. as } \kappa \to \infty \tag{11}$$

for all $t \in [0, T]$, *where* $X(\cdot)$ *is a* (*random*) *nonnegative Lebesgue integrable function on* $[0, T]$. *Then, for all* $t \in [0, T]$,

$$\int_0^t \frac{\kappa Z^\kappa(s)\,ds}{f(\kappa)} \to \int_0^t Z(s)\,ds \quad \textit{a.s. as } \kappa \to \infty, \tag{12}$$

where

$$Z(t) = \Gamma^{-1}[\Lambda(t) - RX(t)] + BX(t). \tag{13}$$

Remark. Under the conditions of Proposition 1, using the definition of the queue-length process given in (7) and the above result, we also have for all $t \in [0, T]$,

$$\int_0^t \frac{\kappa Q^\kappa(s)\,ds}{f(\kappa)} \to \int_0^t Q(s)\,ds \quad \text{a.s. as } \kappa \to \infty,$$

where

$$Q(t) = \Gamma^{-1}[\Lambda(t) - RX(t)].$$

For any sequence of dynamic controls, the condition in (11) holds for a subsequence. Thus, the condition is not restrictive and is not needed for the results stated in next section.

Qualitative Insights and Comparison to Standard Fluid Limits. Proposition 1 asserts that in the limit, the headcount process "equilibrates instantly" in the sense that its dynamics degenerate to those given in (13). This behavior is a consequence of the two-scale asymptotic in which the abandonments and service completions occur so rapidly that the system instantly "forgets" its recent state. It is illuminating to contrast the results of Proposition 1 with those derived through "standard" fluid scaling where arrival rates are scaled up linearly, specifically $\bar{\Lambda}^\kappa(\cdot) = \kappa\Lambda(\cdot)$, and service rates and abandonment rates are kept constant. Under this scaling, the number of servers should also scale up linearly to "match" demand. We use an overbar to denote this standard fluid scaling. The admissibility conditions are kept the same as in Definition 1. If we consider any admissible sequence of staffing vectors and dynamic controls $\{\bar{X}^\kappa\}$ under this scaling such that $\kappa^{-1}\bar{X}^\kappa(t) \to \bar{X}(t)$, almost surely, as $\kappa \to \infty$ for all $t \in [0, T]$, then there exist $\mathbb{R}^m_+$-valued processes $\bar{Z} = (\bar{Z}(t)\colon 0 \leqslant t \leqslant T)$ and $\bar{Q} = (\bar{Q}(t)\colon 0 \leqslant t \leqslant T)$ such that

$$\kappa^{-1}(\bar{Z}^\kappa(\cdot), \bar{Q}^\kappa(\cdot)) \to (\bar{Z}(\cdot), \bar{Q}(\cdot)) \quad \text{a.s. as } \kappa \to \infty,$$

where $\bar{Z}$ solves

$$\bar{Z}(t) = \int_0^t \Lambda(s)\,ds - \int_0^t R\bar{X}(s)\,ds - \int_0^t \Gamma\bar{Q}(s)\,ds,$$

$$\bar{Z}(0) = 0 \quad \text{and} \quad \bar{Q}(0) = 0$$

for all $t \in [0, T]$. The limiting system dynamics are therefore given by the solution to an ordinary differential equation (with a random "driver" Λ), which can be shown to have a unique solution. Unfortunately, the above limiting system dynamics typically lead to an intractable control problem. In contrast, the scaling we propose in this section gives rise to the tractable limiting dynamics given in Proposition 1. This will be the key to the asymptotic optimality results proved in §4.

4. Main Results

4.1. An Asymptotic Lower Bound on Achievable Performance

In this section, we develop an asymptotic lower bound on the expected cost under any sequence of staffing vectors and admissible controls. This bound states that the expected total cost must grow *at least* at rate $f(\kappa)$, where $f(\kappa)$ is the super-linear function that scales the arrival rates for the κth system. To this end, we define a mapping $\pi\colon \mathbb{R}_+^m \times \mathbb{R}_+^r \mapsto \mathbb{R}$ as follows. For $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^r$, we denote by $\pi(\lambda, b)$ the optimal value of the following LP: choose x in $\mathbb{R}_+^n$,

$$\begin{aligned} &\min \quad p\cdot(\lambda - Rx) \\ &\;\text{s.t.} \quad Rx \leqslant \lambda, \quad Ax \leqslant b, \quad x \geqslant 0, \end{aligned} \tag{14}$$

where R is the unscaled input-output matrix, A is the capacity consumption matrix, and p is the penalty-rate vector. Let $\Phi(\lambda, b)$ denote the optimal solution set of the LP (14); that is, if $x_* \in \mathbb{R}_+^n$ is an optimal solution of the LP, then $x_* \in \Phi(\lambda, b)$. (Formally, Φ is a point-to-set correspondence from (λ, b) to the solution set.) Let $b_* \in \mathbb{R}_+^r$ be a minimizer of

$$\varphi(b) := c\cdot b + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b)\, dt\right], \tag{15}$$

where c and Λ are the unscaled personnel costs and arrival rate, respectively. The function $\varphi(\cdot)$ is convex (cf. Harrison and Zeevi 2005, Proposition 1), and $\varphi(0)$ is finite because $\mathbb{E}[\int_0^T \Lambda_i(s)\, ds] < \infty$ for all $i = 1, \ldots, m$. Thus, the minimization in (15) can be taken over the compact convex set $\{b \in \mathbb{R}_+^r\colon c\cdot b \leqslant \varphi(0)\}$. Because we are minimizing a convex function over this set, the minimum in (15) is achieved by a finite-valued minimizer b_*. The vector b_* is the staffing level recommended by Harrison and Zeevi (2005).

Theorem 1. *For any sequence of staffing vectors $\{b^\kappa\}$ and corresponding admissible dynamic controls $\{X^\kappa\}$,*

$$\begin{aligned} &\liminf_{\kappa\to\infty} f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^\kappa(X^\kappa, b^\kappa)] \\ &\qquad \geqslant c\cdot b_* + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b_*)\, dt\right], \end{aligned} \tag{16}$$

where $\pi(\cdot,\cdot)$ is the optimal value function of the LP (14), *and b_* is the vector that minimizes* (15).

Theorem 1 asserts that the expected total cost grows at least at rate $f(\kappa)$ as the scale factor κ grows large. The asymptotic lower bound on the scaled expected cost is given by the value of a simple stochastic program, the computation of which does not involve any control considerations.

4.2. An Asymptotically Optimal Policy When Λ Is Observable

In this section, we assume that the system manager can observe the arrival rate process, that is, $\Lambda(t)$ is known at each time $t \in [0, T]$. In addition, we assume that services are interruptible. Both assumptions will be relaxed in the following section. Let the staffing vector b_*^κ for the κth system be chosen as follows:

$$b_*^\kappa = \frac{f(\kappa) b_*}{\kappa}, \tag{17}$$

where b_* is defined as in Theorem 1. Fix $t \in [0, T]$, and consider the LP

$$\begin{aligned} &\min \quad p\cdot(\Lambda^\kappa(t) - R^\kappa x) \\ &\;\text{s.t.} \quad R^\kappa x \leqslant \Lambda^\kappa(t), \quad Ax \leqslant b_*^\kappa, \quad x \geqslant 0. \end{aligned} \tag{18}$$

Let $\Phi^\kappa(\Lambda^\kappa(t), b_*^\kappa)$ denote the optimal solution set of the LP (18) as a function of $(\Lambda^\kappa(t), b_*^\kappa)$; that is, if $x_*^\kappa \in \mathbb{R}_+^n$ solves the above LP, then $x_*^\kappa \in \Phi^\kappa$. We note that the LP (18) is identical to (14), with $\Lambda^\kappa(t)$ and b_*^κ substituted for Λ and b in the right-hand side of the constraints, and R^κ replacing R in the left-hand side of the constraints. Thus, Φ^κ can be defined via Φ, the solution set of the "unscaled" LP (14) given in §4.1. The following "selection theorem" establishes the existence of a Lipschitz continuous mapping from (λ, b) to the solution set of the LP (14).

Proposition 2. *There exists a Lipschitz continuous mapping $\phi\colon \mathbb{R}_+^m \times \mathbb{R}_+^r \mapsto \mathbb{R}^n$ such that $\phi(\lambda, b) \in \Phi(\lambda, b)$ for all $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^r$.*

Given the "selected" function ϕ, we now define the function ϕ^κ as follows: for each $t \in [0, T]$, let

$$\phi^\kappa(\Lambda^\kappa(t), b_*^\kappa) := \frac{f(\kappa)}{\kappa}\phi\left(\frac{\Lambda^\kappa(t)}{f(\kappa)}, \frac{\kappa b_*^\kappa}{f(\kappa)}\right).$$

Using the relationship between the LP (14) and the LP (18), we have that $\phi^\kappa(\Lambda^\kappa(t), b_*^\kappa) \in \Phi^\kappa(\Lambda^\kappa(t), b_*^\kappa)$ for each $t \in [0, T]$, each scaled arrival rate vector $\Lambda^\kappa(t)$, and each staffing vector b_*^κ.

For any $t \in [0, T]$, let $X_*^\kappa(t) = \phi^\kappa(\Lambda^\kappa(t), b_*^\kappa)$, so that $X_*^\kappa(t)$ is a *pointwise solution* to the LP (18). The solution X_*^κ prescribes a control which may not meet the admissibility condition (7). To remedy this, we truncate it appropriately.

Definition 3 (Minimal Truncation). Let $\{b^\kappa\}$ be a sequence of staffing vectors and $\{X^\kappa\}$ a sequence of dynamic controls such that $AX^\kappa(t) \leqslant b^\kappa$ for all κ and $t \in [0, T]$. (Note that X^κ need not be admissible with respect to b^κ.) Let $\{\widetilde{X}^\kappa\}$ be a sequence of dynamic control which is admissible with respect to $\{b^\kappa\}$, and let $\{\widetilde{Z}^\kappa\}$ denote the corresponding sequence of headcount processes. We say that $\{\widetilde{X}^\kappa\}$ is a minimal truncation of $\{X^\kappa\}$, if for each time $t \in [0, T]$ and $i \in \{1, \ldots, m\}$,

$$\widetilde{X}^\kappa(t) \leqslant X^\kappa(t), \quad \text{and}$$

$$(B\widetilde{X}^\kappa)_i(t) < \widetilde{Z}_i^\kappa(t) \quad \text{implies } \widetilde{X}_j^\kappa(t) = X_j^\kappa(t) \quad \text{for all } j \text{ such that } i(j) = i.$$

The above definition ensures that the truncated control meets the admissibility condition (7), i.e., the number of servers assigned to each activity is such that the total number of servers allocated to each customer class does not exceed the total headcount in that class. Further, it ensures that the truncation is in some sense the "minimal" one that meets the admissibility condition. Definition 5 in Appendix B.2 describes an example of minimal truncation. From the definition, it is clear that a minimal truncation is not in general unique.

Theorem 2. *If the technical assumption* (10) *holds, then any sequence of dynamic controls obtained by minimal truncation of* $\{X_*^\kappa\}$, *together with the staffing vectors* $\{b_*^\kappa\}$ *defined in* (17), *is asymptotically optimal.*

The above theorem asserts that the properly-scaled solution to the LP (14) essentially prescribes the optimal server allocation, i.e., it generates controls that achieve the asymptotic lower bound.

4.3. An Asymptotically Optimal Tracking Policy When Λ Is Not Observable

In an actual system, the true arrival rates are unknown and unobservable, and the system manager is only able to observe the arrival epochs. In light of the results established in §4.2, in particular Theorem 2, it stands to reason that by suitably *estimating* the arrival rates one might still be able to establish the desired asymptotic optimality. We assume that the true arrival rate vector is unknown at any instant of time, but the distribution of the process Λ is available (e.g., derived from historical data) prior to the planning horizon $[0, T]$, so that the optimization problem in (15) can be solved. In particular, throughout this section we assume that the staffing vector used for the κth system is given in (17). In contrast, the dynamic control at time $t \in [0, T]$ may depend on all the events (including arrivals, service completions, and abandonments) up until that time.

Let us denote the estimator of the arrival rate by $\hat{\Lambda}^\kappa(t) = (\hat{\Lambda}_1^\kappa(t), \ldots, \hat{\Lambda}_m^\kappa(t))$. We restrict attention to estimators that are nonanticipating with respect to the information set generated by arrivals. That is, these estimators are constructed based on past arrival observations, ruling out clairvoyance on the part of the system manager.

We now construct a dynamic control policy which hinges on an arrival rate estimator $\hat{\Lambda}^\kappa(\cdot)$; this class of controls will be referred to as Λ-*tracking controls*. The main idea is to use $\hat{\Lambda}^\kappa(\cdot)$ to derive a "plug-in" estimate of the LP-based policy discussed in the previous section. Specifically, for any $t \in [0, T]$, let $\widehat{X}_*^\kappa(t) = \phi^\kappa(\hat{\Lambda}^\kappa(t), b_*^\kappa)$, where ϕ^κ is the Lipschitz continuous mapping defined in §4.2. Thus, $\widehat{X}_*^\kappa$ denotes the pointwise solution of the LP (18) with $\hat{\Lambda}^\kappa(t)$ substituted for $\Lambda(t)$ in the right-hand side of the constraints. The key property that the arrival rate estimator should satisfy for our proposed "plug in" approach to work is the following.

Definition 4 (Uniform Consistency). An estimator $\hat{\Lambda}^\kappa$ is said to be uniformly consistent if it satisfies

$$\frac{\hat{\Lambda}^\kappa(t)}{f(\kappa)} \to \Lambda(t) \quad \text{a.s. as } \kappa \to \infty, \tag{19}$$

where the convergence is uniform on compact subsets of $(0, T]$.

This notion of consistency ensures that the estimator $\hat{\Lambda}^\kappa(t)$ is uniformly "close" to the actual arrival rate $\Lambda(t)$ for large enough κ, supporting the following result.

Theorem 3. *If the technical assumption* (10) *holds and the estimator used in the* Λ-*tracking policy is uniformly consistent, then any sequence of dynamic controls obtained by minimal truncation of* $\{\widehat{X}_*^\kappa\}$, *together with the staffing vectors* $\{b_*^\kappa\}$ *defined in* (17), *is asymptotically optimal.*

A simple estimator of the arrival rate at time t is one that counts the number of arrivals in a short time window ending at time t, and normalizes this count by the length of the window. Specifically, let $g(\cdot)$ be a nonnegative increasing function, and put

$$\hat{\Lambda}^\kappa(t) = g(\kappa)[F^\kappa(t) - F^\kappa(t - g(\kappa)^{-1})] \tag{20}$$

for $t \in [g(\kappa)^{-1}, T]$, where $F^\kappa(t) = (F_1^\kappa(t), \ldots, F_m^\kappa(t))$ is the vector of the cumulative number of arrivals up until time t in each customer class, and $g(\kappa)^{-1}$ represents the length of the *sliding window* in which arrivals are counted. The next result establishes the uniform consistency of this estimator.

Proposition 3. *If* $g(\kappa) \to \infty$ *and* $f(\kappa)^{-1} g(\kappa)^2 \log \kappa \to 0$ *as* $\kappa \to \infty$, *then the estimator defined in* (20) *is uniformly consistent.*

In the above proposition, $1/g(\kappa)$ represents the length of a sliding window that is used to estimate the arrival rates. The above growth condition ensures that the window length decreases to zero at a slow enough rate so as to ensure consistent estimation of the arrival rate, while still shrinking fast enough so that the arrival rate itself does not change within the window. Assuming that (10) holds, the hypothesis of this proposition can be satisfied, for example, by taking $g(\kappa) = \kappa^\alpha$ for some $\alpha \in (0, 0.5]$.

4.4. A Discrete-Review Λ-Tracking Policy

The Λ-tracking policy described in the previous section suffers from two shortcomings:

- The arrival rate estimator (20) and the LP (18) need to be calculated and re-solved, respectively, at each instant in time. This is clearly not feasible for purposes of implementation.
- The server allocation is given by a solution to an LP and therefore may change frequently and in an abrupt manner, resulting in a significant amount of "chatter" in the controls. This may lead to excessive service interruptions which are not desirable in a call center environment.

To alleviate the deficiencies stated above, we now propose a *discrete-review* implementation of Λ-tracking policies. These controls are also based on the estimation of arrival rates, for which the same window size of $g(\kappa)^{-1}$ is used. However, instead of a sliding window, nonoverlapping windows are used, and the LP is solved only at discrete points in time that mark the ends of these estimation windows. Specifically, we partition the time interval $[0, T]$ into $g(\kappa)T$ review periods of equal length. In the lth review period, $l = 1, \ldots, \lfloor g(\kappa)T \rfloor$, the following estimate of Λ is used:

$$\hat{\Lambda}^{l,\kappa} = g(\kappa)\left[F^\kappa\left(\frac{l-1}{g(\kappa)}\right) - F^\kappa\left(\frac{l-2}{g(\kappa)}\right)\right],$$

where $F^\kappa(t)$ is the vector of cumulative arrivals up until time t in each customer class. Here, $\lfloor x \rfloor$ is the maximum integer less than x.

The dynamic control uses the estimator $\hat{\Lambda}^{l,\kappa}$ in the same manner as in the general class of Λ-tracking policies, i.e., the LP (18) is solved with $\hat{\Lambda}^{l,\kappa}$ as the right-hand side of the constraints, and the optimal solution is minimally truncated to make it admissible. We note that the estimate of the arrival rate is constant over a review period, and the LP (18) is solved only at the beginning of each review period. What we have just described is a Λ-tracking control with the estimator explicitly given by

$$\hat{\Lambda}^\kappa(t) = g(\kappa)\left[F^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}\right) - F^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}\right)\right]. \quad (21)$$

Because the server allocation is constant within each review period, no services are interrupted during this time interval. The only times where any services might be interrupted occur at the beginning of the review periods.

We now modify the discrete review policy to avoid service interruptions altogether. In the beginning of each review period, we let every customer who is being served, referred to as *customers-in-service*, complete her/his service. When all customers-in-service have completed service, server allocation is done based on the solution obtained from the LP (18) with the estimator (21). Because no service is ever interrupted, this is a *nonpreemptive* policy. Let $\hat{X}^\kappa_*(t)$ be the optimal solution of the LP (18) with the estimator (21) in its right-hand side, i.e., $\hat{X}^\kappa_*(t) = \phi^\kappa(\hat{\Lambda}^\kappa(t), b^\kappa_*)$. Let τ^κ_l denote the time elapsed from the beginning of the lth review period in the κth system until all customers-in-service have completed their services.

To summarize, the nonpreemptive discrete review policy is obtained by first dividing the period $[0, T]$ into $g(\kappa)T$ review periods. At the beginning of the lth review period, the arrival rate vector is estimated using $\hat{\Lambda}^{l,\kappa}$, and the LP (18) is solved with this estimator to obtain $\hat{X}^{l,\kappa} = \phi^\kappa(\hat{\Lambda}^{l,\kappa}_*, b^\kappa)$. Then, for a period of length τ^κ_l time units from the commencement of the review period, servers complete the processing of all customers-in-service. From this point in time until the end of the review period, servers are allocated based on the minimal truncation of the dynamic control $\hat{X}^{l,\kappa}_*$. For implementation, we round off the staffing vectors and the controls obtained above to the nearest integer. (We omit this distinction for the purpose of exposition but clearly this modification has no effect on our asymptotic analysis.) The next result establishes that it is possible to choose the number of review periods such that the estimator is uniformly consistent, and cumulative time spent on completing work of customers-in-service at the beginning of review periods is negligible.

Theorem 4. *Suppose that the technical assumption* (10) *holds, that* $\kappa^{-1}\log f(\kappa) \to 0$ *as* $\kappa \to \infty$, *and that* $g(\kappa) = \kappa^\alpha$ *for some* $\alpha \in (0, 0.5]$. *Then, the sequence of discrete review dynamic controls obtained by minimal truncation of* $\{\hat{X}^\kappa_*\}$, *together with the staffing vectors* $\{b^\kappa_*\}$ *as defined in* (17), *is asymptotically optimal.*

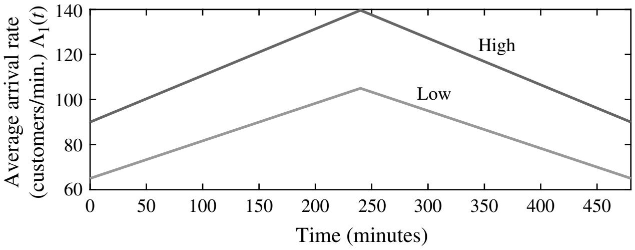
5. Numerical Examples

The key to all our main results is Proposition 1, which is stated in §3, and says essentially the following: if the scale of the system is large, then the headcount process Z, the dynamic control X, and the instantaneous arrival rate vector Λ are approximately linked by the simple relationship (13), that is, one has approximately $Z(t) = \Gamma^{-1}[\Lambda(t) - RX(t)] + BX(t)$ for all $t \in [0, T]$.

To provide a "picture proof" of Proposition 1, we consider a simple system with a single-customer class and a single-server pool (i.e., $m = 1$ and $r = 1$), and take the planning horizon to be one day comprised of $T = 480$ minutes. To illustrate the manner in which the system "equilibrates" and is then governed by the trajectory given by (13), let us focus on the following system parameters. We take the service rates to be $\mu = 1$ customers-per-minute, and the abandonment rate to be $\gamma = 0.5$ customers-per-minute. The arrival rate may be either "high" or "low" with equal probability assigned to the two paths given in Figure 2.

The cost of employing a server for one day is $c = \$240$, and the abandonment penalty is $p = \$2$ per customer. Solving the staffing problem given in (15), we find that $b_* = 115$ servers. Figure 3(a) depicts a sample path of the headcount process for the given system using the obvious dynamic control policy $X(t) = \min(b_*, Z(t))$, $t \geqslant 0$,

Figure 2. Arrival rate pattern for a single-class/single-pool example.



superimposed on the asymptotic path (13) described in Proposition 1 (these results correspond to a "high" realization of the arrival rate). The headcount process in Figure 3(a) does indeed fluctuate around the asymptotic path given in (13).

To illustrate the time it takes an empty system to reach this "equilibrium" behavior, Figure 3(b) "zooms in" on the system dynamics around $t = 0$. The system equilibrates to the limiting path (13) within one to two minutes, after which it follows this path, in spite of the temporal changes in the arrival rate. To summarize, one can say that the limiting system state has essentially no dynamics, it instantly "forgets" its past and its evolution at any time instant only depends on the instantaneous arrival rate and the control policy (which is itself a function of the instantaneous arrival rate).

Next, we illustrate the lower bound on system performance and its achievability (Theorems 1 and 4) by considering an example in which the dynamic routing policy is not trivial or obvious. Specifically, we consider a system with two customer classes ($m = 2$) which is served by two server pools ($r = 2$). There are three processing activities ($n = 3$). Servers in pool 1 can serve only class 1 customers (activity 1), while servers in pool 2 are cross-trained and can serve both class 1 and class 2 customers (activities 2 and 3, respectively). Callers of class 1 and 2 arrive according to a doubly-stochastic Poisson process whose rates are displayed in Figure 4. We take the scaling function to be $f(\kappa) = \kappa^2$. All the services are exponentially distributed with unit rate, that is, $\mu_j = 1$ customers-per-minute for $j = 1, 2, 3$. Customers of class 1 abandon at rate $\gamma_1 = 0.2$ customers-per-minute, whereas customers of class 2 abandon at rate $\gamma_2 = 1$ customers-per-minute. The abandonment penalties for class 1 and class 2 are $p_1 = \$4$ per customer and $p_2 = \$1$ per customer, respectively. The cost of a server in pool 1 is \$600 per day and \$720 per day in pool 2 (where the servers are cross-trained).

Solving the staffing problem in (15) we get $b_* = (50, 50)$. We now simulate the system to obtain estimates of the total expected cost under two policies. The first is the discrete review nonpreemptive policy derived in §4.3. We divide the time horizon into review periods of equal length, and at the beginning of each such review period the arrival rate is estimated based on the number of arrivals in the last review period using (21). With this estimate, we then solve the LP (18) to obtain a routing of customers to servers. As soon as a server finishes the tasks to which s/he was assigned, s/he is allocated a customer based on the new routing decision. In addition, if the solution of the LP does not allocate all the servers in some pools, we allocate them whenever there are customers waiting for service based on the priority rule given by the objective function of this LP. The performance of this policy is evaluated for system scales $\kappa = 10, 20, \ldots, 200$ (with $\kappa = 50$ being the "reference system"), and the number of review periods for the κth system is chosen to be $8\kappa^{0.45}$. The second policy, which serves for comparison purposes, seeks to minimize the value of the objective function in (5) at each time instant (specifically, at each arrival or departure epoch, because we

Figure 3. (a) Simulation of the headcount process and the asymptotic analogue given in Proposition 1. (b) "Relaxation time" to equilibrium: the graph depicts the simulated system dynamic over the first six minutes in Figure (a), and the corresponding asymptotic path given in Proposition 1.

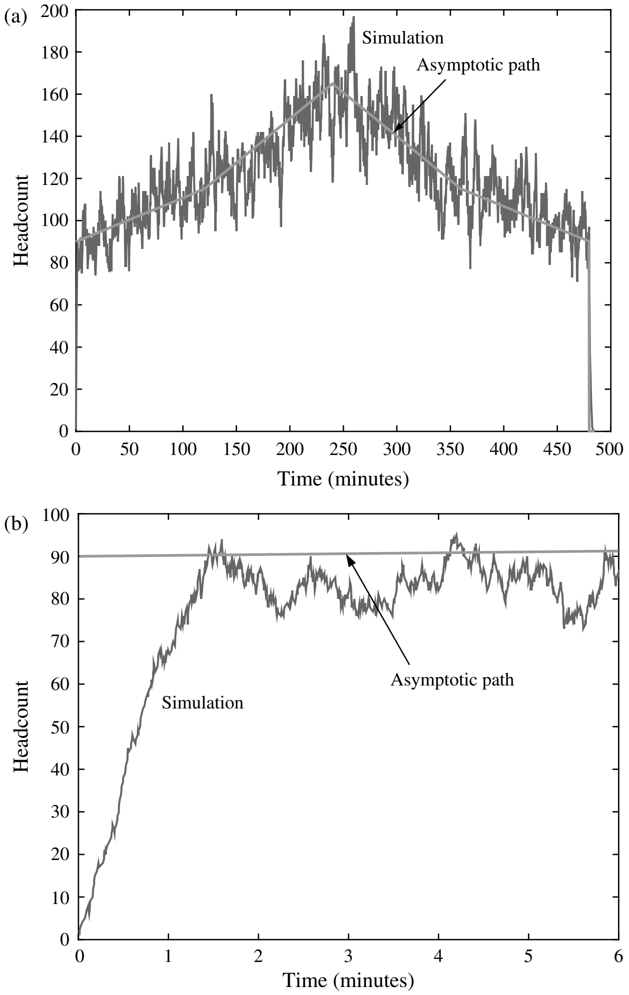
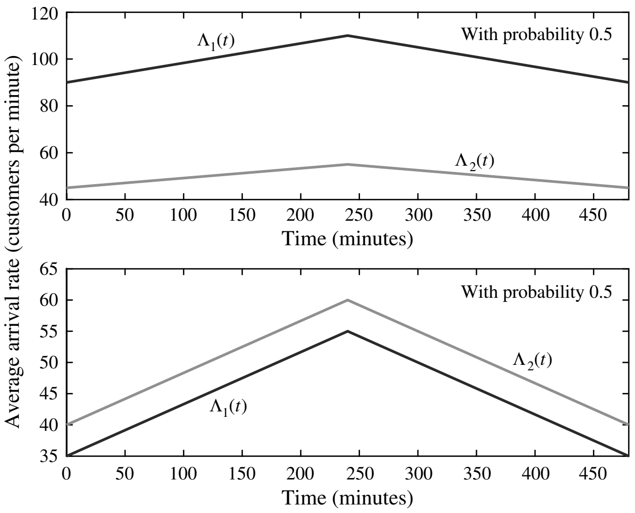


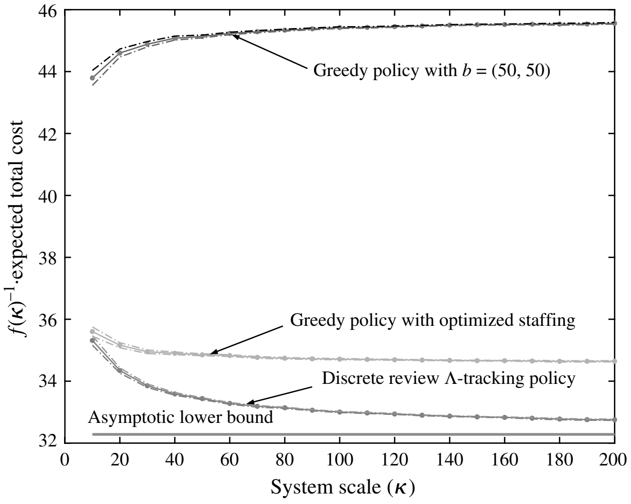
Figure 4. Arrival rates for the two-class/two-pool example.



focus on a nonpreemptive service discipline). In particular, this is a "greedy" policy which gives priority to the class i customer for which the penalty rate $\gamma_i p_i$ is largest. In our example, this simply means that servers in pool 2 give priority to class 2, because $\gamma_2 p_2 > \gamma_1 p_1$. (Recall that servers in pool 1 can only serve class 1.) For our simulation study, we consider two staffing levels for the greedy policy: in the first case we set the staffing level b to be $(50, 50)$, the optimal value given by our LP-based method; in the second case we optimize b given that the greedy policy is to be used for control.

Figure 5 depicts the simulation results for the above policies at various system scales, with the total expected cost scaled by $f(\kappa)^{-1}$. The simulation results use stratified sampling based on the arrival rate processes to reduce variance. This results in a tight confidence interval, because the variance of the estimator from stratified sampling depends on the conditional variance, and the variance of the scaled expected cost conditioned on the arrival rate processes approaches zero as $\kappa \to \infty$. The number of simulation runs for both arrival processes depicted in Figure 4 is either 200 (if the system scale κ is less than 160) or 50 (if the system scale κ exceeds 170). As is evident, the Λ-tracking discrete review policy outperforms the greedy policy with optimized staffing, which in turn outperforms the greedy policy with staffing level $b = (50, 50)$. Moreover, as κ grows large, the cost of the system under the discrete review policy is close to the asymptotic lower bound, differing by about 4% when $\kappa = 50$, as predicted by Theorem 4.

Figure 5. Scaled expected total cost as a function of the system scale (κ) for the two-class two-pool example; dotted lines correspond to a 95% confidence interval for the simulated results.



6. Concluding Remarks

The notion of a planning horizon plays an important part in our problem formulation. The interval $[0, T]$ represents the *smallest* block of time over which the staffing level must be kept constant. Our model assumes that the following holds: staffing decisions are made before the beginning of the planning horizon, and temporal and stochastic variation within the planning horizon are not negligible (or the interval is not short enough to reasonably support such an assumption).

Our model assumes that both service times and "impatience" random variables are exponential. The memoryless property of the exponential distribution allows us to express various system quantities (e.g., cumulative number of abandonments) using a simple time change of a Poisson process that in turn supports a simple state descriptor. In addition, we assume that the arrival process is described as a time change of a Poisson process. While it is important to investigate the robustness of our method relative to these distributional assumptions, we do not attempt such analysis in this paper, leaving it for future work. What we believe to be true is that the exponential assumptions for arrival processes and service times can be relaxed, but the exponential assumption with respect to the "impatience" random variables is crucial to obtain the limiting dynamics given in (13).

The dynamic routing control proposed in this paper is given by a minimal truncation of the solution to an LP. (The minimal truncation effectively projects the solution onto the space of admissible controls.) As such, this control does not explicitly use system state information and hence runs "open loop," except for the tracking of arrival rate. However, the asymptotic regime described in this paper is such that (in the limit) arrival rates translate instantaneously to system state. Hence, the proposed "open loop" control implicitly uses state information encoded in arrival rates. Of course, one can argue that a more refined notion of asymptotic optimality than the one advocated in this paper would require a bona fide closed-loop control rule.

Finally, the model presented in this paper can be extended to include linear holding costs and to allow admission control decisions along with dynamic routing control and staffing. That extension is undertaken in Bassamboo et al. (2005).

Appendix A. Proofs of the Main Results

Let $(\Omega, \mathscr{H}, \mathbb{P})$ be the probability space on which all processes described in §3 are defined. Let $\mathscr{F}_t = \sigma(\Lambda(s)\colon 0 \leqslant s \leqslant t)$ represent the information set generated by the arrival rate process up until time t. In a similar vein, let the information set generated by arrivals, departures, and abandonment up until time t in the κth system be represented by $\mathscr{H}_t^\kappa$ for all $t \in [0,T]$. Let $D[0,T]$ denote the space of functions defined over $[0,T]$ which are right continuous with left limits. In much of what follows, as well as in Appendix B, statements are said to hold almost surely for almost all time $t \in [0,T]$. (Appendix B is available at http://or.pubs.informs.org/Pages/collect.html.) Note that the above is weaker than the assertion that a statement holds for almost all time $t \in [0,T]$, almost surely. This distinction is a consequence of pointwise limits as opposed to functional limits. Finally, proofs of all lemmas cited in this appendix can be found in Appendix B.

Proof of Proposition 1. Consider any sequence of staffing vectors $\{b^\kappa\}$ and corresponding admissible dynamic control policies $\{X^\kappa\}$. For each κ, the dynamics of the headcount process are given by (see (6)–(8))

$$Z_i^\kappa(t) = F_i^\kappa(t) - N_i^{(2)}\left(\int_0^t (R^\kappa X^\kappa)_i(s)\,ds\right) - N_i^{(3)}\left(\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s))_i\,ds\right) \tag{22}$$

for all $i = 1, \ldots, m$ and $t \in [0,T]$. Dividing both sides of the equation by $f(\kappa)$, we now appeal to the following two lemmas that establish the convergence of the rescaled processes in (22) as $\kappa \to \infty$.

Lemma 1. *For any $t \in [0,T]$,*

$$\frac{F_i^\kappa(t)}{f(\kappa)} \to \int_0^t \Lambda_i(s)\,ds \quad \textit{a.s. as } \kappa \to \infty$$

for all $i = 1, \ldots, m$. Further, consider any admissible sequence of dynamic controls $\{X^\kappa\}$ that satisfies condition (11)*. Then, for any $t \in [0,T]$,*

$$\frac{N_i^{(2)}\big(\int_0^t (R^\kappa X^\kappa)_i(s)\,ds\big)}{f(\kappa)} \to \int_0^t (RX)_i(s)\,ds \quad \textit{a.s. as } \kappa \to \infty$$

for all $i = 1, \ldots, m$.

Lemma 2. *Consider any admissible sequence of dynamic controls $\{X^\kappa\}$ which satisfies condition* (11)*. Then,*

$$\text{(i)}\quad \frac{Z_i^\kappa(t)}{f(\kappa)} \to 0, \qquad \text{(ii)}\quad \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)}\,ds \to M_i(t),$$

$$\text{(iii)}\quad \frac{N_i^{(3)}\big(\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s))_i\,ds\big)}{f(\kappa)} \to \gamma_i\left(M_i(t) - \int_0^t (BX(s))_i\,ds\right)$$

almost surely as $\kappa \to \infty$ for all $i = 1, \ldots, m$ and for almost all $t \in [0,T]$.

Also, we have that $\{X^\kappa\}$ satisfies (11) by assumption. Applying Lemmas 1 and 2 to the three terms on the right-hand side of (22), and Lemma 2 to the left-hand side of (22) gives

$$0 = \int_0^t (\Lambda(s) - RX(s))\,ds - \Gamma\left(M(t) - \int_0^t BX(s)\,ds\right), \tag{23}$$

almost surely for almost all $t \in [0,T]$, where $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_m)$, and $M(t) = (M_1(t), \ldots, M_m(t))$. Thus, we have

$$\int_0^t \frac{\kappa Z^\kappa(s)}{f(\kappa)}\,ds \to \int_0^t (\Gamma^{-1}[\Lambda(s) - RX(s)] + BX(s))\,ds \quad \text{a.s. as } \kappa \to \infty$$

for all $t \in [0,T]$. This completes the proof. □

Proof of Theorem 1. Consider any sequence of staffing vectors $\{b^\kappa\}$ and corresponding admissible controls $\{X^\kappa\}$. We shall first prove the result under the assumption that

$$\frac{\kappa b^\kappa}{f(\kappa)} \to b \quad \text{as } \kappa \to \infty, \tag{24}$$

where $b \geqslant 0$. All subsequent probabilistic statements are to be interpreted in the almost sure sense and the term is omitted for brevity. Because $\{f(\kappa)^{-1}\mathscr{J}(X^\kappa, b^\kappa)\}$, $\kappa = 1, 2, \ldots$, is a sequence in $\mathbb{R}_+$, it has a subsequence $\{\kappa_n\colon n = 1, 2, \ldots\}$ which converges to the $\liminf_{\kappa\to\infty} f(\kappa)^{-1}\mathscr{J}(X^\kappa, b^\kappa)$. Further, because X^{κ_n} is admissible, by (6) and assumption (24) we have that $\kappa_n X^{\kappa_n}/f(\kappa_n)$ is uniformly bounded. Next, we state a general result for uniformly bounded nonnegative functions.

Lemma 3. *Given a sequence of uniformly bounded nonnegative functions Y^κ in $D[0,T]$, then for every subsequence there exists a further subsequence Y^{κ_n} and integrable function Y such that $\int_B Y^{\kappa_n}(t)\,dt \to \int_B Y(t)\,dt$ as $n \to \infty$ for any Borel set B of $[0,T]$, where Y is nonnegative for almost all $t \in [0,T]$.*

Appealing to the above lemma, there exists a function $X\colon \Omega \times [0,T] \mapsto \mathbb{R}_+$ defined for almost all $\omega \in \Omega$ and

(Lebesgue) almost all $t \in [0, T]$ and a further subsequence $\{\kappa_{n'}: n' = 1, 2, \ldots\}$ such that

$$\frac{\kappa_{n'}}{f(\kappa_{n'})}\int_0^t X^{\kappa_{n'}}(s)\,ds \to \int_0^t X(s)\,ds \quad \text{as } n' \to \infty$$

for all $t \in [0, T]$. To simplify notation, we shall drop the further subsequence index and assume that the above holds on the initial subsequence. Because Proposition 1 applies to this subsequence, from (23) it follows that

$$\Gamma\biggl(M(T) - \int_0^T BX(s)\,ds\biggr) = \int_0^T (\Lambda(s) - RX(s))\,ds, \quad (25)$$

where

$$M(T) = \lim_{n\to\infty}\int_0^T \frac{\kappa_n Z^{\kappa_n}(s)}{f(\kappa_n)}\,ds.$$

We then have

$$\begin{aligned} &f(\kappa)^{-1}\mathcal{J}^{\kappa_n}(X^{\kappa_n}, b^{\kappa_n}) \\ &\to c\cdot b + p\cdot\Gamma\biggl(M(T) - \int_0^T BX(s)\,ds\biggr) \quad \text{as } n \to \infty \\ &= c\cdot b + \int_0^T p\cdot[\Lambda(t) - RX(t)]\,dt, \qquad (26) \end{aligned}$$

where the limit follows from Lemma 2(iii) (in the proof of Proposition 1), and (24) which implies that $f(\kappa)^{-1}c^{\kappa}\cdot b^{\kappa} \to c\cdot b$ as $\kappa \to \infty$ (the last equality follows from (25)). Next, we show that $p\cdot[\Lambda(t) - RX(t)] \geqslant \pi(\Lambda(t), b)$ for almost all $t \in [0, T]$. Note that $X(t)$ satisfies the constraints of the LP (14). To this end, we have that for almost all $t \in [0, T]$ (relative to Lebesgue measure),

$$\Lambda(t) - RX(t) \geqslant 0,$$

$$AX(t) \leqslant b, \quad \text{and}$$

$$X^{\kappa_n}(t) \geqslant 0 \quad \text{implies } X(t) \geqslant 0,$$

where the first inequality follows from the fact that

$$\frac{\int_0^t \gamma_i(Z^{\kappa_n}(s) - BX^{\kappa_n}(s))_i\,ds}{f(\kappa_n)} \to \int_0^t (\Lambda_i(s) - (RX)_i(s))\,ds \quad \text{as } n \to \infty$$

for all $i = 1, \ldots, m$, and $\int_0^t \gamma_i(Z^{\kappa_n}(s) - BX^{\kappa_n}(s))_i\,ds$ is nondecreasing in t for each κ_n. Thus, we have $\int_0^t(\Lambda(s) - RX(s))_i\,ds$ is nondecreasing in t. Consequently, $\Lambda(t) - RX(t) \geqslant 0$ for almost all $t \in [0, T]$. The second inequality follows using a similar argument and the fact that $AX^{\kappa_n} \leqslant b^{\kappa_n}$ implies $\int_0^t(b^{\kappa_n} - AX_i^{\kappa_n}(s))\,ds$ is nondecreasing in t for each κ_n. The optimality of $\pi(\Lambda(t), b)$, together with the above result and Fatou's Lemma, yields that for any admissible sequence of dynamic controls $\{X^{\kappa}\}$ and staffing vectors $\{b^{\kappa}\}$,

$$\begin{aligned} &\liminf_{\kappa\to\infty} f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^{\kappa}(X^{\kappa}, b^{\kappa})] \\ &\qquad \geqslant c\cdot b + \mathbb{E}\biggl[\int_0^T \pi(\Lambda(t), b)\,dt\biggr]. \qquad (27) \end{aligned}$$

Using the fact that b_* is the minimizer of the right-hand side, we get the desired result under assumption (24).

Now, suppose that the limit of the sequence $\{\kappa b^{\kappa}/f(\kappa)\}$ as $\kappa \to \infty$ does not exists. Because $\{f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^{\kappa}(X^{\kappa}, b^{\kappa})]\}$ is a sequence in $\mathbb{R}_+$, it has a subsequence which converges to its lim inf. Also, there exists a further subsequence to this subsequence on which the limit $\lim_{n'\to\infty}\kappa_{n'}b^{\kappa_{n'}}/f(\kappa_{n'}) = b$ exists. Note that if b is infinite, there is nothing to prove. If b is finite, then for this subsequence the above analysis shows that (27) holds. Further, using the fact that b_* is the minimizer of the right-hand side of (27), we have that

$$\liminf_{\kappa\to\infty} f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^{\kappa}(X^{\kappa}, b^{\kappa})] \geqslant c\cdot b_* + \mathbb{E}\biggl[\int_0^T \pi(\Lambda(t), b_*)\,dt\biggr].$$

This completes the proof. □

Proof of Proposition 2. Using the Lipschitz selection theorem (see Aubin and Frankowska 1990, Theorem 9.4.3), it suffices to show that the correspondence Φ defined in §4.1 is Lipschitz and $\Phi(\lambda, b)$ is a nonempty closed convex set for all $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^r$. For the latter, first observe that $x = 0$ is feasible for the LP (14), and second, nonnegativity of matrices A and R and the fact that each row of A and R has at least one positive entry implies that the feasible region is compact. Thus, the solution set of the LP (14) is nonempty, closed, and convex. Thus, to complete the proof we need only prove that the correspondence Φ is Lipschitz, i.e., there exist constants C_1 and C_2 such that for any $\lambda_1, \lambda_2 \in \mathbb{R}_+^m$, and $b_1, b_2 \in \mathbb{R}_+^r$, the following holds:

$$\mathcal{H}(\Phi(\lambda_1, b_1), \Phi(\lambda_2, b_2)) \leqslant C_1\|\lambda_1 - \lambda_2\| + C_2\|b_1 - b_2\|,$$

where $\mathcal{H}(A, B)$ is the Hausdorff distance between the sets A and B, and $\|\cdot\|$ is the Euclidean norm. Fix $\lambda_1, \lambda_2 \in \mathbb{R}_+^m$ and $b_1, b_2 \in \mathbb{R}_+^r$. Consider any $x_1^* \in \Phi(\lambda_1, b_1)$. Using Schrijver (1986, Theorem 10.5), there exists $x_2^* \in \Phi(\lambda_2, b_2)$ such that $\|x_1^* - x_2^*\| \leqslant C_1\|\lambda_1 - \lambda_2\| + C_2\|b_1 - b_2\|$, where C_1 and C_2 are constants that depend only on the matrices R and A. Thus, we have

$$d(x_1^*, \Phi(\lambda_2, b_2)) \leqslant \|x_1^* - x_2^*\| \leqslant C_1\|\lambda_1 - \lambda_2\| + C_2\|b_1 - b_2\|,$$

where $d(y, B)$ denotes the distance between the point y and set B. Taking the supremum over all points in $\Phi(\lambda_1, b_1)$, we have

$$\sup_{x\in\Phi(\lambda_1, b_1)} d(x, \Phi(\lambda_2, b_2)) \leqslant C_1\|\lambda_1 - \lambda_2\| + C_2\|b_1 - b_2\|.$$

Using a similar argument, we get a bound for $\sup_{x\in\Phi(\lambda_2, b_2)} d(x, \Phi(\lambda_1, b_1))$, and consequently, by definition of the Hausdorff distance, we have that the correspondence Φ is Lipschitz. This completes the proof. □

Proof of Theorem 2. The proof is divided into four steps which will be referenced in the subsequent proofs as well: Step 1 establishes the convergence of appropriately scaled processes to their respective limits; Step 2 establishes that the effects of minimal truncation are asymptotically negligible in a suitable sense; Step 3 derives the pathwise convergence of the scaled cost; and Step 4 concludes by showing that the latter convergence holds in expectation.

Let $X_*^\kappa(t) = \phi^\kappa(\Lambda^\kappa(t), b_*)$ for all $t \in [0,T]$, where ϕ^κ is the Lipschitz selection mapping defined in §4.2. Let $\tilde{X}_*^\kappa$ denote the minimal truncation of X_*^κ, and $\tilde{Z}_*^\kappa$ be the headcount process associated with the admissible dynamic control $\tilde{X}_*^\kappa$.

Step 1. By Theorem 1 and the definition of asymptotic optimality, it suffices to show that

$$\limsup_{\kappa\to\infty} f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^\kappa(\tilde{X}_*^\kappa, b_*^\kappa)] \leqslant c\cdot b_* + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b_*)\,dt\right]. \qquad (28)$$

Using the definition of $\{b_*^\kappa\}$ in (17), we have

$$\lim_{n\to\infty}\frac{1}{f(\kappa_n)}c^{\kappa_n}\cdot b_*^{\kappa_n} = c\cdot b_*.$$

All subsequent probabilistic statements are to be interpreted in the almost sure sense and the term is omitted for brevity. Next, we state the following result for the sequence $\{(\kappa/f(\kappa))Z^\kappa\}$, $k = 1, 2, \ldots$.

Lemma 4. *If assumption* (10) *holds, then for any admissible sequence of controls* $\{X^\kappa\}$,

$$\limsup_{\kappa\to\infty}\sup_{0\leqslant t\leqslant T}\frac{\kappa Z^\kappa(t)}{f(\kappa)} < \infty \quad \textit{a.s.}$$

Consider the subsequence over which the lim sup is achieved for $f(\kappa)^{-1}\mathcal{J}^\kappa(\tilde{X}_*^\kappa, b_*^\kappa)$. Consider a further subsequence $\{\kappa_n: n > 0\}$ of this subsequence over which

$$\int \frac{\kappa_n}{f(\kappa_n)}\tilde{X}_*^{\kappa_n} \quad \text{and} \quad \int \frac{\kappa_n}{f(\kappa_n)}\tilde{Z}_*^{\kappa_n}$$

converge to a limit (the existence of such a subsequence follows from Lemma 3 in the proof of Theorem 1 and Lemma 4). Let

$$M_i(T) = \lim_{n\to\infty}\int_0^T \frac{\kappa_n(\tilde{Z}_*^{\kappa_n}(s))_i}{f(\kappa_n)}\,ds \quad \text{for all } i = 1, \ldots, m,$$

$$\int_0^t (\tilde{X}_*(s))_j\,ds = \lim_{n\to\infty}\int_0^t \frac{\kappa_n(\tilde{X}_*^{\kappa_n}(s))_j}{f(\kappa_n)}\,ds \qquad \text{for all } j = 1, \ldots, n$$

and for all $t \in [0,T]$. Because over this subsequence condition (11) holds, we can appeal to Lemma 2(iii) (in the proof of Proposition 1) and (23) to get that for all $i = 1, \ldots, m$,

$$\begin{aligned}\lim_{n\to\infty}&\frac{N_i^{(3)}\left(\int_0^T \gamma_i^{\kappa_n}(\tilde{Z}_*^{\kappa_n}(s) - B\tilde{X}_*^{\kappa_n}(s))_i\,ds\right)}{f(\kappa_n)}\\ &= \gamma_i\left(M_i(T) - \int_0^T (B\tilde{X}_*(s))_i\,ds\right)\\ &= \int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i\,ds.\end{aligned}$$

Step 2. We now show that the truncation effects are negligible in an appropriate limiting sense.

Lemma 5. *Let $X^\kappa(t)$ be an untruncated control satisfying the admissibility condition* (6) *such that*

$$\frac{\kappa}{f(\kappa)}X^\kappa(t) \to X(t) \quad \textit{a.s. as } \kappa\to\infty,$$

where the convergence is uniform over compact sets of $(0,T]$, and X is a continuous process such that $RX(t) \leqslant \Lambda(t)$ for all $t \in [0,T]$. If $\tilde{X}^\kappa(t)$ is a minimal truncation of $X^\kappa(t)$ and assumption (10) *holds, then for all $i = 1, \ldots, m$,*

$$\begin{aligned}\lim_{\kappa\to\infty}&\frac{1}{f(\kappa)}\int_0^T (\Lambda^\kappa(s) - R^\kappa\tilde{X}^\kappa(s))_i\,ds\\ &= \lim_{\kappa\to\infty}\frac{1}{f(\kappa)}\int_0^T (\Lambda^\kappa(s) - R^\kappa X^\kappa(s))_i\,ds \quad \textit{a.s.}\end{aligned}$$

Let $X_*(t) = \phi(\Lambda(t), b_*)$ for all $t \in [0,T]$. Then, by the definition of ϕ^κ, we have $X_*^\kappa(t) = f(\kappa)X_*(t)/\kappa$ for all $t \in [0,T]$. Also, because ϕ is a Lipschitz continuous mapping and Λ is a continuous process, it follows that X_* is also a continuous process. Thus, appealing to the above lemma, we have

$$\int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i\,ds = \int_0^T (\Lambda(s) - RX_*(s))_i\,ds \qquad \text{for all } i = 1, \ldots, m.$$

Step 3. Combining the analysis in Steps 1 and 2, we have

$$\begin{aligned}\lim_{n\to\infty}&\sum_{i=1}^m p_i\frac{\kappa_n N_i^{(3)}\left(\int_0^T \gamma_i^{\kappa_n}(\tilde{Z}_*^{\kappa_n}(s) - B\tilde{X}_*^{\kappa_n}(s))_i\,ds\right)}{f(\kappa_n)}\\ &= \sum_{i=1}^m p_i\int_0^T (\Lambda(s) - RX_*(s))_i\,ds\\ &= \int_0^T \pi(\Lambda(s), b_*)\,ds,\end{aligned}$$

where π is the mapping defined for the LP (14). Consequently, we have

$$\limsup_{\kappa\to\infty} f(\kappa)^{-1}\mathcal{J}^\kappa(\tilde{X}_*^\kappa, b_*^\kappa) = c\cdot b_* + \int_0^T \pi(\Lambda(s), b_*)\,ds.$$

Step 4. Because $\mathcal{J}^\kappa(\tilde{X}_*^\kappa, b_*^\kappa)$ is nonnegative and bounded, using the reverse Fatou Lemma, we have (28). This completes the proof. □

Proof of Theorem 3. Recall that

$$\hat{\Lambda}_i^\kappa(t) = g(\kappa)[F_i^\kappa(t) - F_i^\kappa(t - g(\kappa)^{-1})] \quad \text{for all } i = 1, \ldots, m,$$

and for all time $t \in [0,T]$. Let $\hat{X}_*^\kappa(t)$ be the optimal solution to the LP (18) with the estimator (20), i.e., $\hat{X}_*^\kappa(t) = \phi^\kappa(\hat{\Lambda}^\kappa(t), b_*^\kappa)$, and let $\tilde{X}_*^\kappa$ denote a minimal truncation of $\widehat{X}^\kappa_*$. Let $\widetilde{Z}^\kappa_*$ denote the headcount process associated with the admissible control $\widehat{X}^\kappa_*$.

Before proving the theorem, we sketch an outline of the proof. First, we express the limiting scaled cost in terms of the scaled processes. Next, we show that truncation effects are asymptotically negligible. Then, using the consistency of the estimator, we establish that the scaled solution of the LP (18) with $\hat{\Lambda}$ in the right-hand side converges to the solution of the LP (14). Finally, we use the reverse Fatou Lemma to get the result in expectation.

Step 1. Convergence of the re-scaled processes to their limiting counterparts follows exactly as in Step 1 of the proof of Theorem 2.

Step 2. Let $X_*(t)=\phi(\Lambda(t),b_*)$ for all $t\in[0,T]$. Because $\Lambda(t)$ is continuous and ϕ is Lipschitz, $X_*(t)$ is also continuous. Consider any compact set $B\subset(0,T]$. Using the definition of the mapping ϕ^κ, we have

$$\frac{\kappa}{f(\kappa)}\widehat{X}^\kappa_*(t)-X_*(t)=\phi\left(\frac{\hat{\Lambda}^\kappa(t)}{f(\kappa)},b_*\right)-\phi(\Lambda(t),b_*).$$

Because the mapping ϕ is Lipschitz continuous, we have

$$\left\|\frac{\kappa}{f(\kappa)}\widehat{X}^\kappa_*(t)-X_*(t)\right\|\leqslant C\left\|\frac{\hat{\Lambda}^\kappa(t)}{f(\kappa)}-\Lambda(t)\right\|$$

for all $t\in B$, where $\|\cdot\|$ is the Euclidean norm. Taking the supremum over $t\in B$ and the limit as $\kappa\to\infty$, and using the fact that the estimator is uniformly consistent, we get

$$\sup_{t\in B}\left\|\frac{\kappa}{f(\kappa)}\widehat{X}^\kappa_*(t)-X_*(t)\right\|\to 0\quad\text{a.s. as }\kappa\to\infty.$$

Thus, $\widehat{X}^\kappa_*$ satisfies the conditions of Lemma 5 (in the proof of Theorem 2). Hence, for $i=1,\ldots,m$,

$$\int_0^T(\Lambda(s)-R\widetilde{X}_*(s))_i\,ds=\int_0^T(\Lambda(t)-RX_*(s))_i\,ds,$$

where $\widetilde{X}^\kappa_*$ is the minimal truncation of $\widehat{X}^\kappa_*$. Repeating Steps 3 and 4 in the proof of Theorem 2 completes the proof. □

Proof of Proposition 3. Fix $i\in\{1,\ldots,m\}$. For the remainder of the proof, we shall focus our attention on the set of ωs for which $\int_0^T\Lambda_i(s)\,ds>0$ (the result is trivially true on the complement set). Using the definition of the estimator, we have

$$\begin{aligned}
&\hat{\Lambda}^\kappa_i(t)-\Lambda^\kappa_i(t)\\
&\quad=g(\kappa)\left[F^\kappa_i(t)-\int_0^t\Lambda^\kappa_i(s)\,ds\right]\\
&\qquad-g(\kappa)\left[F^\kappa_i(t-g(\kappa)^{-1})-\int_0^{t-g(\kappa)^{-1}}\Lambda^\kappa_i(s)\,ds\right]\\
&\qquad+g(\kappa)\int_{t-g(\kappa)^{-1}}^t\Lambda^\kappa_i(s)\,ds-\Lambda^\kappa_i(t)
\end{aligned}$$

for all $t\in[g(\kappa)^{-1},T]$. Fix a compact set $B\subset(0,T]$, and fix κ large enough so that $g(\kappa)^{-1}\leqslant\inf\{s\colon s\in B\}$. Then,

$$\begin{aligned}
&\sup_{t\in B}\frac{|\hat{\Lambda}^\kappa_i(t)-\Lambda^\kappa_i(t)|}{f(\kappa)}\\
&\quad\leqslant 2\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\left|F^\kappa_i(t)-\int_0^t\Lambda^\kappa_i(s)\,ds\right|\\
&\qquad+\sup_{t\in B}\left|g(\kappa)\int_{t-g(\kappa)^{-1}}^t\Lambda_i(s)\,ds-\Lambda_i(t)\right|.
\end{aligned}\tag{29}$$

We shall now show that both terms on the right-hand side go to zero as $\kappa\to\infty$.

Step 1. For each κ, fix $\theta_\kappa>0$, and let $M^\kappa=(M^\kappa(t)\colon 0\leqslant t\leqslant T)$ be defined as

$$M^\kappa(t):=\exp\left(\theta_\kappa F^\kappa_i(t)-(e^{\theta_\kappa}-1)\int_0^t f(\kappa)\Lambda_i(s)\,ds\right).\tag{30}$$

Then, M^κ is a martingale adapted to the filtration $\sigma(\mathscr{H}^\kappa_t\vee\mathscr{F}_T)$. Using Doob's submartingale inequality (cf. Ethier and Kurtz 1986), we have for any $\epsilon>0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\left[F^\kappa_i(t)-\int_0^t\Lambda^\kappa_i(s)\,ds\right]>\epsilon\int_0^T\Lambda_i(s)\,ds\,\middle|\,\mathscr{F}_T\right)\\
&=\mathbb{P}\Bigg(\sup_{0\leqslant t\leqslant T}M^\kappa(t)>\exp\bigg(-f(\kappa)\int_0^T\Lambda_i(s)\,ds\\
&\qquad\cdot\left(e^{\theta_\kappa}-1-\theta_\kappa\left(1+\frac{\epsilon}{g(\kappa)}\right)\right)\bigg)\,\Bigg|\,\mathscr{F}_T\Bigg)\\
&\leqslant\exp\left(f(\kappa)\int_0^T\Lambda_i(s)\,ds\left(e^{\theta_\kappa}-1-\theta_\kappa\left(1+\frac{\epsilon}{g(\kappa)}\right)\right)\right)\quad\text{a.s.}
\end{aligned}$$

By choosing $\theta_\kappa=\log(1+\epsilon/g(\kappa))$, we have that the right-hand side of the above equation is bounded by

$$\exp\left(-\frac{f(\kappa)}{g(\kappa)^2}\epsilon\int_0^T\Lambda(s)\,ds\right).$$

Further, for any positive integer l,

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\left[F^\kappa_i(t)-\int_0^t\Lambda^\kappa_i(s)\,ds\right]\\
&\qquad>\epsilon\int_0^T\Lambda_i(s)\,ds,\text{ i.o.}\,\Bigg|\,\mathscr{F}_T\Bigg)\\
&\qquad\qquad\leqslant\sum_{\kappa=l}^\infty\exp\left(-\frac{f(\kappa)}{g(\kappa)^2}\epsilon\int_0^T\Lambda_i(s)\,ds\right)\quad\text{a.s.,}
\end{aligned}$$

and because the growth condition implies that the summation on the right-hand side is finite, we get

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\left[F^\kappa_i(t)-\int_0^t\Lambda^\kappa_i(s)\,ds\right]\\
&\qquad>\epsilon\int_0^T\Lambda_i(s)\,ds,\text{ i.o.}\,\Bigg|\,\mathscr{F}_T\Bigg)=0\quad\text{a.s.}
\end{aligned}$$

Hence, for any $\epsilon > 0$,

$$\mathbb{P}\biggl(\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\biggl[F_i^\kappa(t)-\int_0^t\Lambda_i^\kappa(s)\,ds\biggr] > \epsilon\int_0^T\Lambda_i(s)\,ds,\text{ i.o.}\biggr)=0,$$

and thus,

$$\sup_{0\leqslant t\leqslant T}\frac{g(\kappa)}{f(\kappa)}\biggl|F_i^\kappa(t)-\int_0^t\Lambda_i^\kappa(s)\,ds\biggr|\to 0 \quad \text{a.s. as } \kappa\to\infty. \tag{31}$$

Step 2. We now consider the second term on the right-hand side of (29). Because the set B is closed and $0\notin B$, we have $0<\beta := \inf\{s\colon s\in B\}$. Further, continuity of Λ_i over $[0,T]$ implies uniform continuity, thus for any $\epsilon>0$, there exists a $\delta>0$ such that for all $t,s\in[0,T]$ with $|t-s|<\delta$, we have $|\Lambda(s)-\Lambda(t)|<\epsilon$. Choose κ_1 such that $g(\kappa)^{-1}<\min(\beta,\delta)$ for all $\kappa>\kappa_1$. Then, for all $\kappa>\kappa_1$ and for all $t\in B$, we have

$$\biggl|g(\kappa)\int_{t-g(\kappa)^{-1}}^t\Lambda_i(s)-\Lambda_i(t)\biggr| \leqslant g(\kappa)\int_{t-g(\kappa)^{-1}}^t|\Lambda_i(s)-\Lambda_i(t)|\,ds\leqslant\epsilon,$$

hence,

$$\sup_{t\in B}\biggl|g(\kappa)\int_{t-g(\kappa)^{-1}}^t\Lambda_i(s)\,ds-\Lambda_i(t)\biggr|\to 0 \quad \text{a.s. as } \kappa\to\infty.$$

This completes the proof. □

Proof of Theorem 4. Recall that

$$\hat\Lambda^\kappa(t)=g(\kappa)\biggl[F^\kappa\biggl(\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}\biggr)-F^\kappa\biggl(\frac{\lfloor tg(\kappa)\rfloor-1}{g(\kappa)}\biggr)\biggr],$$

where $\lfloor x\rfloor$ is the maximum integer less than or equal to x, and $\widehat X_*^\kappa(t)$ is the optimal solution to the LP (18) with the estimator (21), i.e., $\widehat X_*^\kappa(t)=\phi^\kappa(\hat\Lambda^\kappa(t),b_*^\kappa)$. Let $\widetilde X_*^\kappa$ denote the admissible dynamic routing policy obtained from X_*^κ as described in §4.4 and $\widetilde Z_*^\kappa$ the headcount process associated with the admissible control $\widetilde X_*^\kappa$. Let I^κ denote the cumulative time spent on completing the previous period's assigned task in the κth system. By definition,

$$I^\kappa=\bigcup_{i=1}^{g(\kappa)T}\biggl[\frac{i-1}{g(\kappa)},\frac{i-1}{g(\kappa)}+\tau_i^\kappa\biggr], \tag{32}$$

where τ_i^κ is the time required to complete the services of customers-in-service at the beginning of the ith review period in the κth system. Then, for all $t\in[0,T]\backslash I^\kappa$, $\widetilde X_*^\kappa$ is a minimal truncation of X_*^κ.

The key steps in the proof are similar to that of Theorem 3 with the addition that we now need to establish that the time required for customers-in-service to complete their services at the beginning of each review period is asymptotically negligible in a suitable sense.

Step 1. Convergence of the re-scaled processes to their limiting counterparts follows exactly as in Step 1 of the proof of Theorem 2.

Step 2. To establish that the estimator is consistent based on the growth conditions on $f(\kappa)$ and $g(\kappa)$, we use the following result.

Lemma 6. *If $g(\kappa)\to\infty$ and $f(\kappa)^{-1}(g(\kappa)^2\log\kappa)\to 0$ as $\kappa\to\infty$, then the estimator defined in* (21) *is uniformly consistent.*

Using the same argument as in Step 2 of Theorem 3, we have that $\kappa X^\kappa(t)/f(\kappa)\to X(t)$ as $\kappa\to\infty$, where the convergence is uniform over compact sets of $(0,T]$. We now give an analogue of Lemma 5 (in the proof of Theorem 2), which establishes that the minimal truncation effects are asymptotically negligible in a suitable sense.

Lemma 7. *Let $X^\kappa(s)$ be an untruncated control satisfying the admissibility condition* (6), *such that*

$$\frac{\kappa}{f(\kappa)}X^\kappa(s)\to X(s) \quad \textit{a.s. as } \kappa\to\infty,$$

where the convergence is uniform over compact sets of $(0,T]$, and X is a continuous process with $RX(t)\leqslant\Lambda(t)$ for all $t\in[0,T]$. Let $\widetilde X^\kappa(s)$ be a minimal truncation of $X^\kappa(s)$ over the set I^κ defined in (32) *and suppose that assumption* (10) *holds. Then, for all $i=1,\ldots,m$,*

$$\lim_{\kappa\to\infty}\frac{1}{f(\kappa)}\int_0^T\bigl(\Lambda_i^\kappa(s)-(R^\kappa\widetilde X^\kappa(s))_i\mathbb{I}_{\{s\in[0,T]\backslash I^\kappa\}}\bigr)\,ds \leqslant \lim_{\kappa\to\infty}\frac{1}{f(\kappa)}\int_0^T(\Lambda^\kappa(s)-R^\kappa X^\kappa(s))_i\,ds \quad \textit{a.s.}$$

Now, by the definition of $(\kappa_n/f(\kappa_n))\widetilde X_*^{\kappa_n}$,

$$\int_0^T(\Lambda(s)-R\widetilde X_*(s))_i\,ds = \lim_{n\to\infty}\int_0^T\biggl(\Lambda(s)-\frac{\kappa_n}{f(\kappa_n)}R\widetilde X_*^{\kappa_n}(s)\biggr)_i\,ds \quad \text{a.s.}$$

for $i=1,\ldots,m$. Further, we have

$$\lim_{n\to\infty}\int_0^T\biggl(\Lambda(s)-\frac{\kappa_n}{f(\kappa_n)}R\widetilde X_*^{\kappa_n}(s)\biggr)_i\,ds \overset{\text{(a)}}{\leqslant} \lim_{n\to\infty}\int_0^T\biggl(\Lambda_i(t)-\frac{\kappa_n}{f(\kappa_n)}(R\widetilde X_*^{\kappa_n}(s))_i\mathbb{I}_{\{s\in[0,T]\backslash I^{\kappa_n}\}}\biggr)\,ds \overset{\text{(b)}}{=} \int_0^T(\Lambda(s)-RX_*(s))_i\,ds \quad \text{a.s.}$$

for $i=1,\ldots,m$, where $X_*(t)=\phi(\Lambda(t),b_*)$ for all $t\in[0,T]$. Inequality (a) follows from nonnegativity of the process $\widetilde X_*^\kappa$, and equality (b) follows from Lemma 7 because $\widehat X_*^\kappa$ satisfies the conditions of the lemma. Finally, repeating Steps 3 and 4 in the proof of Theorem 2 completes the proof. □

Acknowledgment

The research of Assaf Zeevi was supported in part by NSF grant DMI-0447652.

References

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems Theory Appl.* **51** 287–329.

Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52** 271–292.

Armony, M., I. Gurvich, A. Mandelbaum. 2005. Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers. Working paper, New York University, New York.

Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.* **14** 1084–1134.

Aubin, J.-P., H. Frankowska. 1990. *Set-Valued Analysis*. Birkhäuser, Boston, MA.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems Theory Appl.* (*QUESTA*) **51** 249–285.

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

Bremaud, P. 1981. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, New York.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Chen, B., S. Henderson. 2001. Two issues in setting call centre staffing levels. *Ann. Oper. Res.* **108** 175–192.

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

Gans, N., G. van Ryzin. 1997. Optimal control of a multi-class, flexible queueing system. *Oper. Res.* **45** 677–693.

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

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

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

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

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

Harrison, J. M. 1998. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policy. *Ann. Appl. Probab.* **8** 822–848.

Harrison, J. M., M. J. Lopez. 1999. Heavy traffic resource polling in parallel-server systems. *Queueing Systems* **33** 339–368.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* **52**(2) 243–257.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7** 20–36.

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

Kurtz, T. G. 1978. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Their Appl.* **6** 223–240.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52**(6) 836–855.

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

Massey, W. A., W. Whitt. 1998. Uniform acceleration expansions for Markov chains with time-varying rates. *Ann. Appl. Probab.* **8** 1130–1155.

Schrijver, A. 1986. *Theory of Linear and Integer Programming*. John Wiley and Sons, New York.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* **7** 276–294.

Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci.* **37** 307–314.

Whitt, W. 1992. Understanding the efficiency of multi-server systems. *Management Sci.* **38** 708–723.

Whitt, W. 2001. *Stochastic-Process Limits*. Springer-Verlag, New York.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.