Queueing Systems 51, 249–285, 2005

Dynamic Routing and Admission Control in High-Volume Service Systems: Asymptotic Analysis via Multi-Scale Fluid Limits

ACHAL BASSAMBOO achalb@stanford.edu
J. MICHAEL HARRISON harrison_michael@gsb.stanford.edu
Graduate School of Business, Stanford University

ASSAF ZEEVI assaf@gsb.columbia.edu
Graduate School of Business, Columbia University

Received 1 October 2004; Revised 17 March 2005

Abstract. Motivated by applications in telephone call centers, we consider a service system model with m customer classes and r server 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 dynamic control are considered: customers may be either blocked or accepted at the time of their arrival, and then accepted customers of each class must be routed, either immediately upon acceptance or after some period of waiting, to a server pool that is qualified to handle that class. Customers who are made to wait before commencement of their service are liable to defect. The objective is to minimize the expected sum of blocking costs, waiting costs and defection costs over a fixed and finite planning horizon. We consider an asymptotic parameter regime in which (i) the arrival rates, service rates and defection rates are uniformly accelerated by a large factor κ, then (ii) arrival rates are increased by an additional factor $g(\kappa)$, and the number of servers in each pool is increased by $g(\kappa)$ as well. This produces a separation of time scales, justifying a pointwise stationary stochastic fluid approximation for our original system model. In the stochastic fluid approximation, optimal admission control and routing decisions are determined by a simple linear program that uses the current arrival rate vector λ as data. We explain how to implement the fluid model's optimal control policy in our original service system context, and prove that the proposed implementation is asymptotically optimal in the first-order sense.

Keywords: call centers, queueing, admission control, dynamic routing, fluid limits, doubly stochastic, asymptotic analysis, performance bounds, abandonments

AMS subject classification: 60K30, 90B15, 90B36

1. Introduction

Motivated by applications in telephone call centers, we consider in this paper a two-level problem of dynamic control for large-scale service systems. The two levels of the problem are dynamic admission control, whereby some arrivals are accepted for service and others are "blocked," and dynamic routing of customers to servers. In a call center context the former type of control is achieved by means of "busy signals," and the latter type of control is referred to as *skills-based routing*.

Our model of a service system has multiple customer classes and multiple server pools. Each pool consists of identical servers whose skills determine which customer classes those servers can process, and the rates at which such services can be delivered. We allow the arrival rates for the customer classes to vary both temporally and stochastically. Upon arrival a customer is either blocked or admitted into the system. Customers who are admitted but not served immediately are stored in infinite-capacity (possibly virtual) buffers. We assume that customers of any given class will defect if forced to wait too long before the commencement of their service. (In a call center context such defections are referred to as *abandoned* calls.)

As the title of this paper indicates, the model that we analyze has potential applications in service contexts other than call centers, such as systems for processing loan applications, or "customer contact centers" where agents handle a mix of telephone calls, e-mail correspondence and "web chat." However, our model was formulated with call centers in mind, and exclusive use of context-neutral language makes for a stilted, artificial exposition. Thus, vivid call center terms like "busy signal" and "abandoned call" will be used frequently hereafter; readers who are interested in other service contexts should have no trouble substituting appropriate synonyms.

We treat pool sizes as exogenously determined parameters, so personnel costs are uncontrollable, but three types of congestion-related cost are included in our model. First, there is a *blocking cost* for each customer class. This penalizes the system manager for denying access to customers. Second, there is an *abandonment cost* for each customer class. This captures the penalty associated with customer defections. Finally, a (linear) *holding cost* is incurred at a class-specific rate while customers wait for commencement of their service, or until they abandon, whichever comes first. The system manager's objective is to minimize the sum of these three operating costs.

The dynamic routing problem is as follows. First, whenever a customer is accepted and there exist several idle servers who can handle that customer's class, the system manager must either route the customer to one of them immediately or else have 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.

Admission control and routing decisions are made based on information available at the decision epoch, which includes the number of customers waiting in the various buffers and the number of idle servers in the various pools.

Our analysis of the problem described above is in many regards a direct extension of work reported earlier in Bassamboo, Harrison and Zeevi [1]. Following the pattern established there, we focus on a *multi-scale* asymptotic regime where (i) the arrival, service and abandonment processes are "uniformly accelerated" by a large factor *k,* then

(ii) arrival rates are increased by an additional large factor $g(\kappa)$, and the number of servers in each pool is increased by a factor $g(\kappa)$ as well. Roughly speaking, the uniform acceleration in (i) justifies a *pointwise stationary* approximation of system behavior; and the additional scale-up in (ii) justifies a *fluid* approximation. The separation of time-scales that is characteristic of this regime is discussed in Section 3.

To express this in a slightly different way, the asymptotic regime considered here and in Bassamboo et al. [1] has the following key features. First, it involves letting the number of servers grow without bound, so it is a *many-server regime*. Second, the stochasticity of the arrival rate process dominates all other sources of stochastic variability, which leads to a *stochastic fluid approximation*. Finally, in the limit regime that we consider, the system, "equilibrates instantly," which leads to a pointwise stationary approximation. The main contributions of this paper are as follows.

(a) With regard to dynamic control, we extend the analysis in [1], which focussed on dynamic routing, to service systems with admission control. As in Bassamboo et al. [1], we develop an asymptotic lower bound on achievable expected total cost that is valid for any admissible control policy (see Theorem 1 in Section 4.1). We then show that a threshold-based policy for admission control, together with a server allocation policy based on linear programming, achieves the lower bound and hence is asymptotically optimal (see Theorem 2 in Section 4.2). The proposed policy estimates arrival rates "on the fly," and does not require prior knowledge of these functions.

(b) With regard to mathematical methods, we make heavy use of strong approximations in proving our limit theorems. In comparison with the approach taken in Bassamboo et al. [1], our current approach is better aligned with the contemporary literature on asymptotic analysis in applied probability.

(c) As suggested above, the upshot of our limit theory is to motivate or justify a pointwise stationary fluid approximation to the traditional queueing model with which we start. Section 5 recapitulates the approximating fluid model and explains how to mechanically derive admission control and server allocation policies via direct analysis of that model. This provides what might be called a *fluid-based calculus* for service system control, which can be mastered and applied without reference to the limit theory that supports it.

Literature review. The asymptotic regime described in this paper can be viewed as being a hybrid of many-server fluid limits and the pointwise stationary approximations that have been developed previously in the literature of applied probability. The many-server regime for a birth and death process describing a single-class, single-pool system in heavy traffic was first made rigorous by Halfin and Whitt [5]. Fluid and diffusion approximations for non-stationary Markovian queueing networks with many servers were developed by Mandelbaum, Massey and Reiman [12], and Whitt [18] has recently developed fluid approximations for a non-Markovian single-class/singlepool system. Pointwise stationary approximations for simple Markovian queueing models with non-stationary arrivals were first introduced by Green and Kolesar [4] and subsequently made rigorous by Whitt [16]; for further refinements see Massey and Whitt [13]. 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. Recent work on joint admission control and routing/sequencing in a multi-class/single-pool system includes that of Plambeck, Kumar and Harrison [14] which uses heavy-traffic limits, and that of Maglaras and Van Mieghem [11] which uses fluid models; see also references therein. Other antecedent literature relevant to this paper has been thoroughly surveyed in Harrison and Zeevi [8] and Bassamboo et al. [1].

The remainder of this paper is structured as follows. Section 2 lays out our service system model, including a specification of its economic objective. Section 3 describes the asymptotic parameter regime on which we focus, and the separation of time-scales that it involves. Section 4 states the main results and provides a simple numerical example. Section 5 develops the fluid-based calculus referred to in (c) above. Those readers interested only in the approximation and not in the asymptotic analysis that supports it may wish to jump directly from Section 2 to Section 5, at least on initial reading. In certain respects, our fluid-based calculus provides only a crude description of a dynamic control policy; Section 5 further describes desirable refinements that are the subject of continuing research. In Section 6, following the template provided in Harrison and Zeevi [7] and Bassamboo et al. [1], we explain how asymptotically optimal *staffing plans* for the various server pools can be developed based on this paper's analysis of dynamic control policies. Proofs of the main results are given in Appendix A, while Appendix B contains the proofs of auxiliary results.

2. Problem formulation

2.1. *Preliminaries*

There are m customer classes and r server pools in our general call center model. Server pool k consists of b_k identical servers ($k = 1, \ldots, r$); a call center model with $m = 3$ and $r = 2$ is portrayed schematically in figure 1. Customers of various classes arrive randomly over time. Upon arrival a customer may be admitted to the system or may be blocked. Customers who are blocked leave immediately. Blocked calls and abandoned calls are represented by different sets of dashed lines in Figure 1.

Several different server pools may be capable of handling a given customer class. By the same token, servers in a given pool may be *cross-trained* to handle customers from different classes. To describe the server capabilities more precisely, we shall use the notion of "activities," described in Harrison and Lopez [6]. There are n processing activities, each of which corresponds to servers from one particular pool serving customers of one particular class. (The activities are denoted by solid arrows connecting buffers to server pools in figure 1.) For each activity $j = 1, \ldots, n$ we denote by $i(j)$ and $k(j)$ the customer class being served and the server pool involved, respectively. We

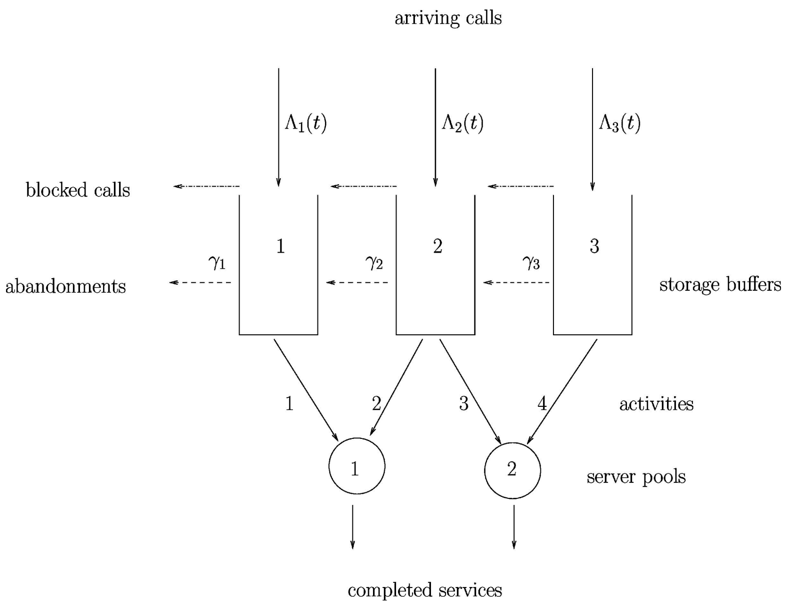

Figure 1. Schematic representation of a call center with three customer classes, two agent pools and four activities.

assume that the service times associated with activity j are exponentially distributed with rate μ_j, and that the service times are independent of arrival processes and of one another. It is important to note that we allow the service time of a customer to depend on both the customer's class and server pool where s/he receives service.

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*: 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*: its (k, j)th element is 1 if activity j draws on the capacity of server pool k and is zero otherwise. (The matrices R and A are exactly as in Harrison and Lopez [6].) 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.

An important feature of our model is the potential for customer abandonments. Each class i customer is endowed with an exponentially distributed "impatience" random variable τ with mean $1/\gamma_i$, independent of the impatience random variables of other customers, and of service times and arrival processes. A customer abandons when her/his waiting time in queue (exclusive of her/his own service time) exceeds τ time units. Let $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_m)$ denote the *abandonment rate matrix*.

We shall now spell out the probabilistic structure of arrival processes of the various customer classes. To this end, we consider a complete probability space $(\Omega, \mathcal{H}, \mathbb{P})$ on which are defined 3 m mutually independent unit rate Poisson processes denoted by

$N_i^{(\ell)} = (N_i^{(\ell)}(t) : 0 \leq t \leq \infty)$ for $i = 1, \ldots, m$ and $\ell = 1, 2, 3$. On the same space are defined m continuous, non-negative arrival rate processes $\Lambda_i = (\Lambda_i(t) : 0 \leq t \leq T)$ satisfying $\mathbb{E}[\int_0^T \Lambda_i(t)dt] < \infty$ for $i = 1, \ldots, m$, independent of the Poisson processes $N_i^{(\ell)}$. We shall use $N^{(1)} = (N_1^{(1)}, \ldots, N_m^{(1)})$ to construct arrivals. Specifically, let

$$F_i(t) = N_i^{(1)}\left(\int_0^t \Lambda_i(s)\, ds \right), \tag{1}$$

where $F_i(t)$ represents the cumulative number of class i arrivals up to time t. This is a standard construction of a doubly stochastic Poisson process (cf. Bremaud, [1]). Put $F = (F(t) : 0 \leq t \leq T)$ where $F(t) = (F_1(t), \ldots, F_m(t))$. The construction of completed services and abandonments under a given control will be done in an analogous manner using the Poisson processes $N^{(2)}$ and $N^{(3)}$, see (5) below.

2.2. *Control formulation*

As in Bassamboo et al. [1], we shall adopt a general formulation that allows services to be interrupted at any time without penalty, and further allows control decisions to be based on information about the future. That is, our definition of an "admissible control" is overly generous, but that apparent defect simply strengthens the conclusions eventually reached. This issue will be revisited in the discussion that concludes this section.

In the current context a *dynamic control* is defined as a pair of stochastic processes (U, X), where $U = (U(t) : 0 \leq t \leq T)$ takes values in $\mathbb{R}_+^m$ and has sample paths that are nondecreasing and right-continuous with left limits, and $X = (X(t) : 0 \leq t \leq T)$ takes values in $\mathbb{R}_+^n$ and has sample paths that are right-continuous with left limits and are Lebesgue integrable. Writing $U(t) = (U_1(t), \ldots, U_m(t))$ for the *admission control*, and $X(t) = (X_1(t), \ldots, X_n(t))$ for the *routing control*, we interpret $U_i(t)$ as the cumulative number of blocked class i customers up until time t, and $X_j(t)$ as the number of servers engaged in activity j at time t. Given the latter interpretation, it would perhaps be more natural to use a term like "server allocation policy" in describing the second element of a dynamic control (U, X), but the matching of servers and customers is invariably described as "call routing" in the literature of call center management.

A dynamic control (U, X) is said to be admissible if there exist processes Z and Q, both taking values in $\mathbb{R}_+^m$, both having time domain $[0, T]$ and both necessarily unique, that jointly satisfy conditions (2–5) below for $0 \leq t \leq T$. $Z_i(t)$ represents the number of class i customers in the system at time t, and $Q_i(t)$ represents the number of class i customers who are waiting for service at time t. We call Z and Q the *headcount* process and *queue length* process, respectively. The relationships that (U, X, Z, Q) must jointly satisfy for all $t \in [0, T]$ are the following:

$$U(t) - U(s) \leq F(t) - F(s) \quad \text{for all } s \in [0, t], \tag{2}$$

$$
\begin{aligned}
&AX(t) \le b, && (3)\\
&Q(t) = Z(t) - BX(t) \ge 0, && (4)\\
&Z_i(t) = F_i(t) - N_i^{(2)}\left(\int_0^t (RX)_i(s)ds\right) - N_i^{(3)}\left(\int_0^t \gamma_i Q_i(s)ds\right) - U_i(t)\\
&\text{for } i = 1, \ldots, m. && (5)
\end{aligned}
$$

Condition (5) is the system dynamics equation: the second term on the right-hand side represents the cumulative number of service completions up to time t, the third term represents the cumulative number of abandonments up to time t, and the last term represents the cumulative number of blocked calls up to time t. The instantaneous service rates and abandonment rates for class i are $(RX)_i$ and $\gamma_i Q_i$, respectively. The first admissibility constraint (2) requires that the number of blocked customers be less than the number of arrivals during any time interval for each class. The second constraint (3) requires that the number of servers in a given pool who are engaged in processing activities at a given time not exceed the total number of servers available in that pool. In our third constraint, (4), $BX(t)$ is a vector whose components represent the number of servers allocated to each customer class, and the constraint thus prohibits allocating to a given class more servers than the headcount in that class. Given a dynamic control (U, X), one can view the headcount process Z and the queue length process Q as the unique solution of (4) and (5), which can be constructed jump-to-jump starting from time zero. Since the primitive processes $N_i^{(\ell)}$ are independent Poisson processes, the probability of simultaneous jumps is zero, and hence there almost surely exists a pair (Z, Q) satisfying the aforementioned relationship. This and other features of the model formulation are discussed at greater length in Bassamboo et al. [1].

2.3. *Economic objective*

Let $p^a = (p_1^a, \ldots, p_m^a)$ be the *abandonment cost vector*, where p_i^a is the cost associated with a class i customer not being served due to abandonment. Let $p^b = (p_1^b, \ldots, p_m^b)$ be the *blocking cost vector*, where p_i^b is the cost associated with blocking a class i customer. Finally, let $h = (h_1, \ldots, h_m)$ be the *holding cost vector*, where h_i is the cost of holding for one unit of time a class i customer who is waiting for his/her service to commence. The total cost of the system under an admissible control (U, X) is given by

$$\mathcal{J}(U, X) := \left[\sum_{i=1}^{m}\left(p_i^b U_i(T) + \int_0^T h_i Q_i(s)ds + p_i^a N_i^{(3)}\left(\int_0^T \gamma_i Q_i(s)ds\right)\right)\right], \quad (6)$$

which represents the sum of holding costs and abandonment and blocking penalties for the various customer classes. The objective of the system manager is to choose an admissible dynamic control (U, X) that minimizes the expected total cost $\mathbb{E}[\mathcal{J}(U, X)]$.

2.4. **Discussion**

Our definition of an admissible control allows services to be interrupted at any time and resumed later (possibly by a different server) without penalty, and also does not rule out clairvoyance on the part of the system manager. It turns out that the asymptotic lower bound on achievable performance derived in Section 4 applies to this broad family of controls. Moreover, we will subsequently construct a family of LP-based policies that achieve this lower bound and are both non-preemptive and non-anticipating. Thus, in the asymptotic regime that we consider, the system manager cannot significantly improve system performance even by interrupting services or "looking into the future." The reader should further note that integrality constraints are relaxed in our formulation of both the admission control and routing problem. The asymptotic regime we define in Section 3 is one where the non-zero components of (X, U) are large, so the distinction between integer and non-integer values is not significant. With regard to the probabilistic assumptions pertaining to arrivals, service completions and abandonments, the reader is referred to Harrison and Zeevi [8] and Bassamboo et al. [1]; for further discussion in the context of call center management, see Gans, Koole and Mandelbaum [3]. In the above problem formulation the staffing is fixed exogenously; see Section 6 for discussion of optimal staffing.

3. An asymptotic parameter regime

As explained in the discussion that concludes this section, the asymptotic parameter regime described immediately below is the same one considered in Bassamboo et al. [1], except for trivial distinctions to be noted. The current formulation is slightly more convenient mathematically, and it is more aligned with standard practice in the literature of applied probability, cf. [17].

3.1. A parametric family of system models

We consider a sequence of system models indexed by $\kappa \in \mathbb{N}$. The planning horizon is κT in the κth system, and the arrival process is doubly stochastic with rate

$$\Lambda^{\kappa}(\kappa t) = g(\kappa)\Lambda(t) \tag{7}$$

for all $t \in [0, T]$, where $g(\cdot)$ is a non-negative function such that $g(\kappa) \to \infty$ as $\kappa \to \infty$. The service rates and the abandonment rates remain fixed. Since the arrival rate is scaled up by a factor $g(\kappa)$, we also scale the number of servers by a factor of $g(\kappa)$; that is, the number of servers in the κ^{th} system is

$$b^{\kappa} = g(\kappa)b. \tag{8}$$

For each system in the sequence indexed by κ, the system manager chooses a dynamic control (U^κ, X^κ) that meets all the restrictions spelled out in Section 2. In the obvious way, a process or quantity associated with the κ^{th} system is indicated by appending a superscript κ to notation established earlier in Section 2. For example, $\mathcal{J}^\kappa(U^\kappa, X^\kappa)$ is the total cost incurred over the interval $[0, \kappa T]$ when the dynamic control (U^κ, X^κ) is employed in the κ^{th} system.

Definition 1 (first-order asymptotic optimality)**.** A sequence of admissible controls $\{(U_*^\kappa, X_*^\kappa)\}$ is said to be asymptotically optimal to first order if for any other admissible sequence of controls $\{(U^\kappa, X^\kappa)\}$,

$$\limsup_{k\to\infty} \frac{\mathbb{E}[\mathcal{J}^\kappa(U_*^\kappa, X_*^\kappa)]}{\mathbb{E}[\mathcal{J}^\kappa(U^\kappa, X^\kappa)]} \leq 1. \tag{9}$$

3.2. Limiting dynamics

In this section we characterize the limiting system behavior under an admissible control, assuming that $g(\kappa)$ satisfies the following growth condition:

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

The above condition is purely technical required for the proofs. Further, we define the following scaled quantities for $t \in [0, T]$:

$$\bar{Z}^\kappa(t) = \frac{Z^\kappa(\kappa t)}{g(\kappa)}, \bar{Q}^\kappa(t) = \frac{Q^\kappa(\kappa t)}{g(\kappa)}, \bar{X}^\kappa(t) = \frac{X^\kappa(\kappa t)}{g(\kappa)}, \bar{U}^\kappa(t) = \frac{U^\kappa(\kappa t)}{\kappa g(\kappa)}. \tag{11}$$

Proposition 1. Assume that (10) holds and consider any sequence of admissible dynamic controls $\{(U^\kappa, X^\kappa)\}$ such that for all $t \in [0, T]$,

$$\left(\int_0^t \bar{X}^\kappa(s)ds, \bar{U}^\kappa(t)\right) \to \left(\int_0^t X(s)ds, U(t)\right) \ a.s. \ as\ \kappa \to \infty, \tag{12}$$

where $X(\cdot)$ is a non-negative Lebesgue integrable function on $[0, T]$, and $U(\cdot)$ is a non-negative nondecreasing function on $[0, T]$. Then there exist non-negative Lebesgue integrable functions $V(\cdot)$ and $Z(\cdot)$ on $[0, T]$ such that for all $t \in [0, T]$,

$$U(t) = \int_0^t V(s)ds \tag{13}$$

and

$$\int_0^t \bar{Z}^\kappa(s)ds \to \int_0^t Z(s)ds \; a.s. \; as \; \kappa \to \infty, \tag{14}$$

where

$$Z(t) = \Gamma^{-1}[\Lambda(t) - RX(t) - V(t)] + BX(t). \tag{15}$$

Discussion. Our parametric family of models could be specified equivalently as follows. First, the planning horizon is fixed at T, independent of κ, but the arrival rate process for the κ^{th} system is $\Lambda^\kappa(t) = \kappa g(\kappa)\Lambda(t)$, $0 \leq t \leq T$. Second, all service rates and abandonment rates are scaled up by a factor of κ in the κ^{th} system (that is, $R^\kappa = \kappa R$ and $\Gamma^\kappa = \kappa\Gamma$). Finally, the vector of holding cost rates in the κ^{th} system is $h^\kappa = \kappa h$. This is precisely the asymptotic regime considered in Bassamboo et al. [1], except that holding costs were not explicitly considered in that earlier paper, and the notation $f(\kappa)$ was used for the quantity here denoted $\kappa g(\kappa)$.

Readers can easily verify the equivalence claimed in the preceding paragraph, which is more or less obvious from the scaling in (11). However, a few more words about holding costs may be in order. It was observed in section 6 of Harrison and Zeevi [8] that a model with abandonment rates h_i, abandonment penalties p_i^a and holding cost rates h_i is economically equivalent to one with zero holding costs and modified abandonment penalties $p_i^a + h_i/\gamma_i$ $(i = 1, \ldots, m)$. The preceding paragraph described a κ^{th} system with abandonment rates $\kappa\gamma_i(i = 1, \ldots, m)$, as κ grows large, holding costs will remain balanced with abandonment penalties if and only if the holding costs grow linearly with κ; otherwise, one of those two cost components will dominate the other as $\kappa \to \infty$.

In the fixed-time-horizon view of our limit regime, one can identify three distinct "time scales" that separate as $\kappa \to \infty$: changes in the arrival rate process Λ (these might be described as "demand shifts") occur over time intervals of order 1; individual service times and abandonment times are of order $1/\kappa$; and inter-arrival times are of order $1/\kappa g(k)$. Increasing pool sizes via (8) restores the original degree of balance between total demand and total service capacity, and by accelerating all of the arrival, service and abandonment processes we obtain a limiting fluid model that "equilibrates instantly" in response to demand shifts. Equation (15) is the mathematical expression of that last phenomenon.

4. Main results

In this section, we propose a sequence of dynamic controls that is asymptotically optimal in the sense of Definition 1. We first develop an asymptotic lower bound on expected total cost under any admissible sequence of controls, and then describe a policy that asymptotically achieves this lower bound.

4.1. *Lower bound on achievable performance*

For $\lambda \in \mathbb{R}^m_+$ and $b \in \mathbb{R}^r_+$, let $\pi(\lambda, b)$ denote the optimal value of the following linear program (LP): choose $x \in \mathbb{R}^n$, $q \in \mathbb{R}^m$ and $v \in \mathbb{R}^m$ to

$$\begin{aligned} &\text{minimize } p^a \cdot \Gamma q + h \cdot q + p^b \cdot v \\ &\text{subject to } \lambda = Rx + \Gamma q + v, \\ &\qquad Ax \leq b, \\ &\qquad x \geq 0, q \geq 0, v \geq 0\,. \end{aligned} \tag{16}$$

Here R is the input-output matrix, A is the capacity consumption matrix, Γ is the abandonment matrix, h is the holding cost vector and p^a and p^b are the abandonment penalty vector and blocking penalty vector, respectively. The above LP provides a *local fluid approximation* to the system manager's objective, seeking to minimize the (fluid-scale) cost rate associated with abandonments, holding costs and blocking costs. The first constraint represents the limiting dynamics obtained via the multi-scale fluid limit in Proposition 1. The second and third constraints follows from the admissibility conditions given in (3–4) in Section 2.

To simplify the LP (16), we define $p = (p_1, \ldots, p_m)$ to be

$$p_i := \min\left(p^b_i, p^a_i + \frac{h_i}{\gamma_i}\right) \tag{17}$$

for all $i = 1, \ldots, m$. Elements of the vector p will be referred to as *effective loss penalties*, for the following reason. First, the net penalty associated with an abandonment of a class i customer is given by the sum of the abandonment penalty cost and the cost of holding a customer until s/he abandons (which, in expectation, takes $1/\gamma_i$ time units). Thus the net penalty associated with abandonment is $p^a_i + h_i/\gamma_i$. In the asymptotic regime that we consider the system manager can effectively choose whether lost customers will be blocked or will abandon their calls. Thus, the effective loss penalty for class i customers is the minimum of the net penalty associated with abandonment and the blocking penalty p^b_i. Now consider the following LP: choose $x \in \mathbb{R}^n$ to

$$\begin{aligned} &\text{minimize } p \cdot (\lambda - Rx) \\ &\text{subject to } Rx \leq \lambda, Ax \leq b, x \geq 0. \end{aligned} \tag{18}$$

As the following proposition shows, our original LP (16) is essentially equivalent to (18).

Proposition 2. For any $\lambda \in \mathbb{R}^m_+$ and $b \in \mathbb{R}^r_+$, let x_* be any optimal solution of LP (18), and let (q_*, v_*) be defined as follows:

$$(q_*)_i = \begin{cases} (\lambda - Rx_*)_i/\gamma_i & \text{if } p_i = p^a_i + \frac{h_i}{\gamma_i} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

and

$$(v_*)_i = \begin{cases} 0 & \text{if } p_i = p_i^a + \frac{h_i}{\gamma_i} \\ \lambda_i - (Rx_*)_i & \text{otherwise} \end{cases} \tag{20}$$

for all $i = 1, \ldots, m$. Then, (x_*, q_*, v_*) solves LP (16), and the optimal value of LP (18) is equal to $\pi(\lambda, b)$, the optimal value of LP (16).

For each $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^n$, let $\Phi(\lambda, b)$ denote the optimal solution set of LP (18); that is, $\Phi(\lambda, b)$ consists of all optimal solutions x_* for LP(18). The following is immediate from Proposition 2 of Bassamboo et al. [1].

Proposition 3. There exists a Lipschitz continuous mapping $\phi : \mathbb{R}_+^m \times \mathbb{R}_+^r \mapsto \mathbb{R}_+^n$ such that $\phi(\lambda, b) \in \Phi(\lambda, b)$ for all $\lambda \in \mathbb{R}_+^m$ and $b \in \mathbb{R}_+^r$.

Theorem 1. For any sequence of admissible controls $\{(U^\kappa, X^\kappa)\}$,

$$\liminf_{\kappa \to \infty} (\kappa g(\kappa))^{-1} \mathbb{E}[\mathcal{J}^\kappa(U^\kappa, X^\kappa)] \geq \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b)dt\right], \tag{21}$$

where $\pi(\cdot, \cdot)$ is the optimal value function of the LP (18), and b is the constant vector appearing in (8).

Theorem 1 shows that as κ grows large, the expected total cost must grow at least at the rate $\kappa g(\kappa)$ under any admissible sequence of controls.

4.2. An asymptotically optimal policy

Our main focus in this section is on joint routing and admission control that will achieve the asymptotic lower bound derived in the previous section. We assume that the system manager cannot directly observe the arrival rate process; that is, $\Lambda(t)$ is not known at any instant of time. We estimate the arrival rate at time t by counting the number of arrivals in a short window of time ending at t, and normalizing this by the length of the window. Specifically, we use an estimator of the form

$$\hat{\Lambda}^\kappa(t) = l(\kappa)^{-1}[F^\kappa(t) - F^\kappa(t - l(\kappa))], \tag{22}$$

where $l(\cdot)$ is a non-negative increasing function. Fix $t \in [0, \kappa T]$ and consider the following LP: choose $x \in \mathbb{R}^n$ to

$$\begin{aligned} &\text{minimize} && p \cdot (\hat{\Lambda}^\kappa(t) - Rx) \\ &\text{subject to} && Rx \leq \hat{\Lambda}^\kappa(t), Ax \leq b^\kappa, x \geq 0, \end{aligned} \tag{23}$$

where b^κ is defined in (8). Let ϕ^κ be defined as follows:

$$\phi^\kappa(\lambda, b) := g(\kappa)\phi\left(\frac{\lambda}{g(\kappa)}, \frac{b}{g(\kappa)}\right), \tag{24}$$

where ϕ is the Lipschitz continuous mapping described in Proposition 3. Using the relationship between (18) and (23), we have that $\phi^\kappa(\hat{\Lambda}^\kappa(t), b^\kappa)$ solves LP (23).

For any $t \in [0, T]$ let

$$X_*^\kappa(t) = \phi^\kappa(\hat{\Lambda}^\kappa(t), b^\kappa), \tag{25}$$

so that $X_*^\kappa(t)$ is a *pointwise solution* to LP (23). The solution X_*^κ prescribes a control which may not be admissible, specifically the solution may violate the admissibility constraint $BX_*^\kappa(t) \leq Z^\kappa(t)$. To remedy this, we truncate it. The following definition was introduced in Bassamboo et al. [1].

Definition 2 (minimal truncation)**.** Let $\{X^\kappa\}$ be a sequence of dynamic controls such that $AX^\kappa(t) \leq b^\kappa$ for all κ and all $t \in [0, T]$. (Note that X^κ need not be admissible.) Let $\{\tilde{X}^\kappa\}$ be a sequence of dynamic controls which is admissible with respect to $\{b^\kappa\}$, and let $\{\tilde{Z}^\kappa\}$ denote the corresponding sequence of headcount processes. We say that $\{\tilde{X}^\kappa\}$ is a minimal truncation of $\{X^\kappa\}$ if, for each time $t \in [0, T]$ and $i \in \{1, \ldots, m\}$,

$$\tilde{X}^\kappa(t) \leq X^\kappa(t).$$

and

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

For the purpose of admission control, we partition the customer classes into two sets S_a and S_b defined as follows:

$$\begin{aligned} \mathcal{S}_a &= \left\{ i \in \{1, \ldots, m\} : p_i = p_i^a + \frac{h_i}{\gamma_i} \right\} \\ \mathcal{S}_b &= \{1, \ldots, m\} \backslash \mathcal{S}_a. \end{aligned}$$

Proposition 2 suggests that it is optimal not to block customers from classes belonging to the set $\mathcal{S}_a$, because for such customers the blocking penalty is more than the net abandonment penalty. Thus the first property of our proposed admission control policy is that

$$(U_*^\kappa)_i(t) = 0 \quad \text{for all } t \in [0, \kappa T] \text{ and } i \in S_a. \tag{26}$$

For $i \in \mathcal{S}_b$ the system manager should use blocking rather than allowing customers to abandon, but in doing so should also keep enough customers in the system to avoid server idleness. Therefore, to implement the admission control policy in the set $\mathcal{S}_b$, we consider an appropriate threshold function. Let $L(\kappa)$ be such that $L(\kappa)/g(\kappa) \to 0$ and $\log \kappa / L(\kappa) \to 0$ as $\kappa \to \infty$. In our proposed policy, the system manager blocks calls whenever $Q_i^\kappa(t) > L(\kappa)$ and $i \in \mathcal{S}_b$. That is, the optimal admission control $(U_*^\kappa)_i$ for $i \in \mathcal{S}_b$ is the maximal non-decreasing process which satisfies, for all $i \in \mathcal{S}_b$,

$$\begin{aligned} &\int_0^{\kappa T} \mathbb{I}_{\{Q_i^\kappa(t)\}} d(U_*^\kappa)_i(t) = 0 \\ \text{and} \quad &U_i^\kappa(t) - U_i^\kappa(s) \le F_i^\kappa(t) - F_i^\kappa(s) \text{ for all } 0 \le s \le t \le \kappa T \ . \end{aligned} \tag{27}$$

Condition (27) ensures that customers are blocked only when the queue length equals or exceeds $L(\kappa)$. (Note that such a process satisfies the admissibility conditions.) Here $L(\kappa)$ can be viewed as "safety stocks" that keep server utilization sufficiently high. The above growth condition on $L(\kappa)$ ensures that it increases at a slow enough rate so that holding costs do not increase substantially, while at the same time it increases at a fast enough rate to keep server utilization sufficiently high.

Our second main result is the following.

Theorem 2. Assume (10) holds, let $\kappa^{-1}\log(g(\kappa)) \to 0$ as $\kappa \to \infty$, and $l(\kappa) = \kappa^a$ for some $\alpha \in [0.5, 1)$. For each $\kappa \in \mathbb{N}$ let $\tilde{X}^\kappa$ be any routing control obtained by minimal truncation of the process X_*^κ defined in (25), and let U_*^κ be the admission control defined in (26)-(27). Then $\{(U^\kappa, \tilde{X}_*^\kappa)\}$ is asymptotically optimal.

In (22), $l(\kappa)$ 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.

The policy described in Theorem 2 is non-anticipating but may cause service interruptions. To alleviate this deficiency, one can modify the above policy using ideas in Section 4.4 of Bassamboo et al. [1] to get a non-preemptive *discrete-review* implementation of the above policy. These controls are also based on the estimation of arrival rates, using the same window size of $l(\kappa)$. However, instead of a sliding window, non-overlapping 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 $[0, \kappa T]$ into review periods of lengths $l(\kappa)$. In each review period the arrival rate is estimated based on the arrivals in the last review period. The dynamic control then uses this estimator, instead of the one in (22). For further discussion see Section 4.4 of Bassamboo et al. [1].

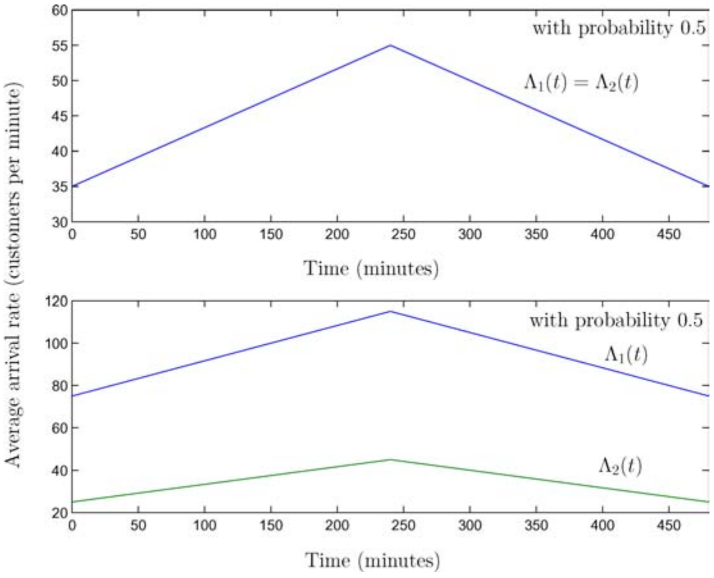

Figure 2. Arrival rates for the two-class/two-pool example.

4.3. *A numerical example*

In this section we illustrate via a numerical example the lower bound on system performance and its achievability, described in Theorems 1 and 2. We consider a service system with two customer classes ($m = 2$) that are served by two server pools ($r = 2$). There are three processing activities ($n = 3$). Server pool 1 can serve only class 1 customers (activity 1), whereas pool 2 can serve both class 1 (activity 2) and class 2 customers (activity 3). The arrival rate processes Λ_i are specified in figure 2. The number of servers in each pool is 50, *i.e.*, the staffing vector is $b = (50, 50)$. For simplicity we take $\mu_j = 1$ for $j = 1, 2, 3$. Abandonment rates for the two customer classes, expressed in customers per minute, are $\gamma_1 = 1/3$ and $\gamma_2 = 1/2$. The abandonment costs associated with class 1 and class 2 are $p_1^a = \$1.50$ per customer and $p_2^a = \$0.50$ per customer. The holding costs associated with class 1 and class 2 are $h_1 = \$0.50$ per customer per minute and $h_2 = \$0.25$ per customer per minute. The cost of blocking customers of both classes is $p_1^b = p_2^b = \$2$ per customer. The effective loss penalty vector as defined in (17) is $p = (2, 1)$, and we have the sets $S_a = \{2\}$ and $S_b = \{1\}$. Consequently, under our proposed policy the system manager will not block any class 2 customer.

We now simulate the system to obtain estimates of expected total cost under the proposed policy. We take the scaling function to be $g(\kappa) = \kappa$. For estimation of the arrival rate, we consider non-overlapping windows of length $l(\kappa)$ instead of a sliding window. In particular, we divide the time horizon into review periods of length $l(\kappa) = 0.2 * \kappa^{0.55}$. At the beginning of each review period we estimate the arrival rate using (22), and solve LP (23) with this estimator to obtain the optimal routing vector $X_* = (X_1, X_2, X_3)$,

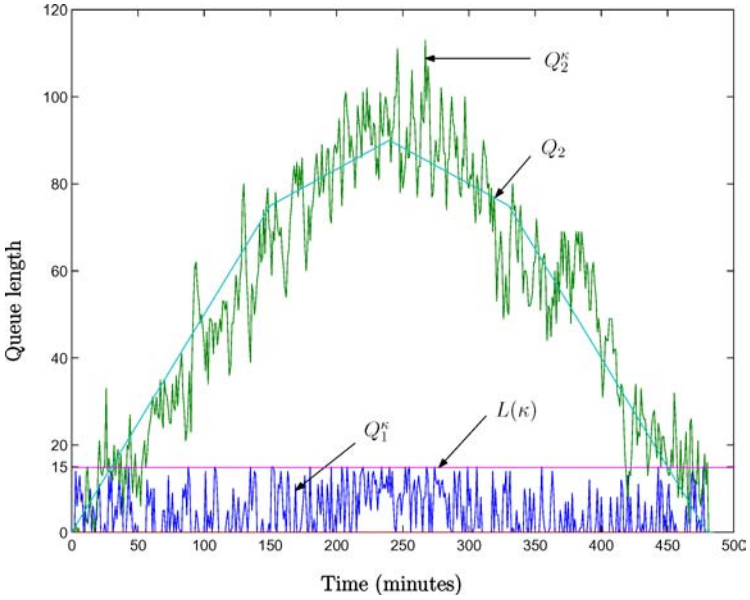

Figure 3. Comparison of the simulated queue length for class 1 and class 2, and the limiting fluid queue length given by (15).

each coordinate in the vector designates the number of servers assigned to the respective activity. This nominal allocation of servers to activities is held fixed until the end of the review period. Each arriving customer of class 1 is assigned to a server in pool 1 if one is available. If all servers in pool 1 are busy and the number of servers in pool 2 that are currently processing class 1 customers is less than the nominal allocation X_2, then the arriving customer is assigned a server in pool 2. Otherwise, the arriving customer is placed in the class 1 queue. Similar logic applies to an arriving class 2 customer. Specifically, if the number of servers in pool 2 processing customers of class 2 is below the nominal allocation to activity 3, X_3, then the arriving customer of class 2 is assigned a server in pool 2; otherwise that customer is placed in the class 2 queue. If $X_2 + X_3 < b_2$, then $(b_2 - X_2 - X_3)$ servers of pool 2 are used as "flexible servers." That is, any arriving customer which is to be placed in the queue by the assignment logic mentioned above is processed by one of these server if one is available. The admission control policy does not block customers of class 2, and it blocks customers of class 1 when the queue length of class 1 exceeds $L(\kappa) = (\log \kappa)^2$. (These scaling functions adhere to the conditions articulated in Theorem 2.) We shall refer to the case $\kappa = 50$ as our "reference system"; that is, the system data provided above corresponds to $\kappa = 50$. The performance of the policy is evaluated for system scales of $\kappa = 10, \ldots, 200$. Figure 3 depicts a sample path of the simulated actual queue lengths for the system with $\kappa = 50$, juxtaposed with the theoretical limiting dynamics described in (15); the arrival rate used to generate this plot corresponds to the bottom graph in figure 2. The admission control policy prescribes

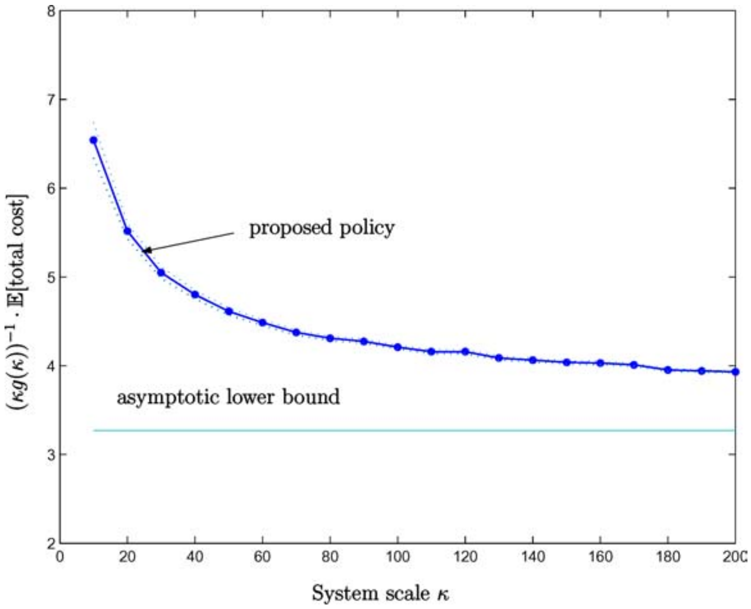

Figure 4. Scaled expected total cost as a function of the system scale κ for the 2-class/2-pool example; dotted lines correspond to 95% confidence intervals for the simulated results.

blocking of class 1 customers whenever $Q_1^\kappa > L(\kappa)$, (thus, $Q_1^\kappa \to 0$ asymptotically). For $\kappa = 50$, the threshold $L(\kappa)$ is 15 customers. The dynamic behavior of the class 2 queue length is dictated by the dynamic routing policy, and this follows the limiting fluid trajectory $Q_2 = (\Lambda_2(t) - X_2(t))/\gamma_2$. We observe that the simulated queue length process fluctuates around its limiting trajectory.

Figure 4 depicts the performance of the proposed policy relative to the asymptotic lower bound given in Theorem 1, where the total cost is scaled by $(\kappa g(\kappa))^{-1}$. We used stratified sampling with respect to the arrival rate to reduce the variance of the total cost estimate. The number of simulation runs for both arrival rates shown in Figure 2 was 100 for $\kappa = 10, \ldots, 100$, and 50 for $\kappa = 110, \ldots, 200$. As κ grows, the total cost under the proposed policy tends towards the asymptotic lower bound as announced in Theorem 2. The convergence rate of the cost under the proposed policy to the asymptotic lower bound is visibly quite slow; we expect this rate to be $O(1/\sqrt{\kappa g(\kappa)})$. Thus, in the above example the rate of convergence is $O(1/\kappa)$ and the noticeable gap we observe may be attributed in part to the constant present in the "big-oh" term.

5. A fluid-based calculus for service system control

Section 2 described a conventional stochastic model of service system dynamics, denoting by U and X the two elements of the system manager's control policy, and by Z and Q the associated headcount and queue length processes. Equation (15) in Section 3 defines a pointwise stationary fluid model (PSFM) that is vastly simpler than the original stochastic model but, according to the limit theory developed in this paper, provides a good approximation in a parameter regime of practical interest. Proposition 2 in Section 4 shows how to compute an optimal control policy for the PSFM via linear programming. In this section we recapitulate both the specification and solution of the PSFM, making no reference to the limit theory that supports or justifies it, and interpret the solution obtained. Notation introduced in Section 2 will be re-used with essentially the same meaning.

Let b, R, A, B, Γ, p^a, p^b and h be the vectors and matrices defined in Section 2, and let p be the m-vector defined in terms of p^a, p^b and h via (17). Proceeding as if the m-dimensional arrival rate process $\Lambda = (\Lambda(t) : 0 \leq t \leq T)$ and were directly observable, we define an admissible control for the PSFM as a pair of processes V and X, taking values in $\mathbb{R}_+^m$ and $\mathbb{R}_+^n$ respectively, that jointly satisfy

$$RX(t) + V(t) \leq \Lambda(t) \tag{28}$$

and

$$AX(t) \leq b \tag{29}$$

for all $t \in [0, T]$. Writing $V(t) = (V_1(t), \ldots, V_m(t))$, we interpret $V_i(t)$ as the rate at which the customers of class i are blocked at time t. 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. We associate with an admissible control (V, X) a triple of processes (U, Q, Z) defined as follows:

$$U(t) = \int_0^t V(s)\,ds, \tag{30}$$

$$Q(t) = \Gamma^{-1}[\Lambda(t) - RX(t) - V(t)], \tag{31}$$

and

$$Z(t) = BX(t) + Q(t) \tag{32}$$

for $0 \leq t \leq T$. The total cost associated with an admissible control (V, X) is defined to be

$$\mathcal{J}(V, X) + \int_0^T [p^b \cdot V(t) + h \cdot Q(t) + p^a \cdot \Gamma Q(t)]dt. \tag{33}$$

The system manager's objective is to minimize $\mathbb{E}[\mathcal{J}(V, X)]$, but in fact the optimal policy identified in the next paragraph minimizes $\mathcal{J}(V, X)$ with probability 1, not just in expectation.

As in Section 4.1 we denote by $\pi(\lambda, b)$ the optimal objective value of the following LP, where $\lambda \in \mathbb{R}^m_+$ is arbitrary: choose x to

$$\begin{aligned} &\text{minimize} && p \cdot (\lambda - Rx) \\ &\text{subject to} && Rx \le \lambda,\ Ax \le b,\ x \ge 0. \end{aligned} \tag{34}$$

Also as in Section 4.1, let ϕ be a Lipschitz continuous mapping $\mathbb{R}^m_+ \times \mathbb{R}^r_+ \mapsto \mathbb{R}^n_+$ such that $\phi(\lambda, b)$ is an optimal solution of the LP (34). Now let the subsets $\mathcal{S}_a$ and $\mathcal{S}_b$, of $\{1, \dots, m\}$ be defined as in Section 4.2, and consider the admissible control (V_*, X_*) defined as follows for each $t \in [0, T]$:

$$X_*(t) = \phi(\Lambda(t), b) \tag{35}$$

and for $i = 1, \dots, m$

$$(V_*(t))i = \begin{cases} \Lambda_i(t) - (RX_*(t))i & \text{if } i \in \mathcal{S}_b, \\ 0 & \text{if } i \in \mathcal{S}_a. \end{cases} \tag{36}$$

Proposition 2 of Section 4.1 shows that (V_*, X_*) minimizes the integrand on the right side of (33) for every $t \in [0, T]$ with probability 1 (that is, for every $t \in [0, T]$ and every possible realization of Λ), the associated total cost being

$$\mathcal{J}(V_*, X_*) = \int_0^T \pi(\Lambda(t), b)\, dt. \tag{37}$$

The definition (36) of our optimal admission control V_* forbids the blocking of customer classes $i \in \mathcal{S}_a$, because for these classes it is less expensive to let customers abandon of their own accord than to deny them access. For each $i \in \mathcal{S}_b$ and each possible arrival rate λ, our Definition (36) specifies the fraction of class i arrivals who are to be blocked, that fraction being $(\lambda_i - (R\phi(\lambda, b))i)/\lambda_i$. Customers from classes $i \in \mathcal{S}_b$ who are not blocked are to be served immediately upon arrival. This control is essentially impossible to implement in a given system, but we devise a policy where the waiting times in queue for each class $i \in \mathcal{S}_b$ are very small, and as the system grows large those waiting times tend to zero. Hence, in the limit, the customers get served immediately upon arrival. The definition (35) of X_* specifies, for each arrival rate vector λ that might be observed, how the servers in each pool k should be allocated to the various activities for which pool k is responsible. If the total number of servers thus allocated is less than b_k, then the remaining servers in pool k are simply to be idle.

Of course, these "interpretations" of our optimal solution (35)–(36) for the fluid control problem do not really provide an implementable plan of action, for several reasons. First, the arrival rate vector $\Lambda(t)$ is not actually observable, so one must use as input in the LP computations an estimator $\hat{\Lambda}(t)$ of $\Lambda(t)$. In Bassamboo et al. [1] and again in Section 4 of this paper, we have described estimation schemes (based either on sliding windows or on non-overlapping windows) that are adequate for proving asymptotic optimality, but other methods may be preferable in practical applications.

Roughly speaking, the length of the window used for estimating the arrival rate should be large enough to ensure an accurate estimate of the arrival rate, yet it should also be small enough so that the arrival rate itself does not change within the window.

A second impediment to literal implementation of our PSFM solution is that the recommended server allocations $X_*(t)$ may change rapidly as $\hat{\Lambda}(t)$ changes, so one must "smooth" the control X_* to avoid service interruptions and undesirable disruption of the operating environment. In Bassamboo et al. [1] and again in Section 4 of this paper, reference has been made to discrete-review policies that avoid such problems but still ensure asymptotic optimality. A separate criticism of our fluid-based "dynamic routing" control $X_*(t)$ is that its server allocations at time t are based solely on the demand estimate $\hat{\Lambda}(t)$, without any consideration of the current system state $Z(t)$. A noteworthy feature of the asymptotic parameter regime studied here is that asymptotic optimality can be achieved by such a crude control policy. However, one can undoubtedly achieve lower cost for any given system by using a more refined policy of state-feedback form. In fact, such a refinement can be systematically developed within the framework of our PSFM, as we plan to show in future research.

Finally, note that the threshold-based implementation of the optimal admission control V_* that we described in Section 4 is of state-feedback form. Roughly speaking, the threshold level should be "small enough" so that holding costs do not increase substantially, and "large enough" to keep server utilization high. Our proof of Theorem 2 shows that the threshold-based implementation does achieve asymptotically the blocking fractions prescribed in (36), and such an approach is almost certainly preferable to open-loop enforcement of the blocking fractions derived from (36).

6. A fluid-based staffing method

Having restricted attention thus far to dynamic control issues, we conclude this paper with a brief consideration of the higher-level staffing problem. Exactly as in Harrison and Zeevi [7] and Bassamboo et al. [1], we suppose that the system manager can choose any capacity vector $b \in \mathbb{R}^r_+$ for use over the time interval $[0, T]$. The associated personnel cost is $c \cdot b$ where $c = (c_1, \dots, c_r)$ and c_k is the cost vector of employing one type k server over the entire planning horizon. The vector b must be chosen at $t = 0$, before any actual demand is observed, and by assumption it cannot be changed before time T. (The latter assumption is essentially a [1] of the "planning horizon" T.) Of course, the choice of b constraints dynamic control decisions during the planning period, and (37) provides an estimate of the minimum achievable cost in the control phase based on our PSFM. Thus we are led to the following optimization problem: choose $b \geq 0$ to

$$\text{minimize} \quad c \cdot b + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b)\, dt\right], \tag{38}$$

where $\pi(\lambda, b)$ is the value of the LP (34).

This problem is of the type considered in Harrison and Zeevi [8] except that the LP that gives rise to the function $\pi(\lambda, b)$ is slightly more complicated in our current setting. Harrison and Zeevi [8] showed that a problem of the form (38) reduces to a standard stochastic program that is readily solvable, even for systems of realistic scale, by a mixture of linear programming and Monte Carlo simulation. Bassamboo et al. [1] showed that the stochastic programming solution b_* is asymptotically optimal in a setting without admission control, and that proof extends with virtually no change to the more general setting of this paper.

Next we consider a more elaborate capacity planning problem. Let the planning horizon $[0, T]$ be divided into L intervals. The system manager chooses a *staffing schedule* that consists of staffing vectors for all the L intervals. Let b^ℓ denote the staffing vector used in the ℓ^{th} interval. The staffing cost associate with staffing schedule $b^1, \ldots, b^L$ can be modeled as

$$\sum_{\ell=1}^{L} c^\ell \cdot b^\ell + \sum_{\ell=1}^{L-1} (b^{\ell+1} - b^\ell) \cdot P^\ell (b^{\ell+1} - b^\ell), \tag{39}$$

where $c^\ell = c = (c_1^\ell, \ldots, c_r^\ell)$ and c_k^ℓ is the cost vector of employing one type k server during the ℓ^{th} interval. Here $P^\ell = diag(p_1^\ell, \ldots, p_r^\ell)$ is a diagonal matrix and p_k^ℓ is the penalty associated with changing the staffing level in pool k by one server between the ℓ^{th} and $(\ell + 1)^{st}$ interval. Again, (37) provides an estimate of the minimum achievable cost in the control phase based on our PSFM. Thus, we are led to the following optimization problem: choose $b^\ell \in \mathbb{R}_+^r$ for $\ell = 1, \ldots, L$ to

$$\text{minimize} \sum_{\ell=1}^{L} c^\ell \cdot b^\ell + \sum_{\ell=1}^{L-1} (b^{\ell+1} - b^\ell) \cdot P^\ell (b^{\ell+1} - b^\ell) + \sum_{\ell=1}^{L} \mathbb{E}\left[\int_{\frac{(\ell-1)T}{L}}^{\frac{\ell T}{L}} \pi(\Lambda(t), b^\ell)\, dt \right],$$

where $\pi(\Lambda(t), b)$ is the optimal solution of the routing LP (34). It can be verified that the objective function is convex and there exists a finite-valued optimal solution. One can extend the asymptotic optimality proof in the current paper to show that the proposed staffing schedule and dynamic PSFM-based control is asymptotically optimal.

A Proof of the main results

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the probability space on which all processes described in Section 3 are defined. Let $\mathcal{F}_t = \sigma(\Lambda(s) : 0 \leq s \leq t)$ represent the information set generated by the arrival rate processes up until time 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 $t \in [0, T]$. 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 admissible control policies $\{(U^\kappa, X^\kappa)\}$ satisfying (10). Fix a time $t \in [0, T]$ and $i \in \{1, \dots, m\}$. For each κ, the dynamics of the headcount process is given by

$$\begin{aligned} Z_i^\kappa(\kappa t) =: {} & F_i^\kappa(\kappa t) - N_i^{(2)}\left(\int_0^{\kappa t} (RX^\kappa)_i(s)\,ds\right) - N_i^{(3)}\left(\int_0^{\kappa t} \gamma i(Z^\kappa(s) - BX^\kappa(s))_i ds\right) \\ & - U_i^\kappa(\kappa t), \end{aligned} \tag{40}$$

for all $t \in [0, T]$, where $F_i^\kappa(\kappa t)$ denotes the number of class i arrivals until time κt. We now use the following strong approximation result from Kurtz [10], which follows directly from Komlós, Major and Tusnady [9].

Proposition 4 (Kurtz (1978), Lemma 3.1)**.** A standard (unit rate) Poisson process $(N(t) : t \geq 0)$ can be realized on the same probability space as a standard Brownian motion $(W(t) : t \geq 0)$ in such a way that

$$\xi := \sup_{t \geq 0} \frac{|N(t) - t - W(t)|}{\log(\max\{2, t\})}$$

has a finite moment generating function in a neighborhood of the origin.

Using the above proposition, there exist Brownian motions $W_i^{(\ell)}$ for $\ell = 1, 2, 3$ such that

$$\begin{aligned} Z_i^\kappa(\kappa t) = {} & \left[\int_0^{\kappa t} \Lambda_i^\kappa(s)ds - \int_0^{\kappa t} (RX^\kappa)_i(s)ds - \int_0^{\kappa t} \gamma i(Z^\kappa(s) - BX^\kappa(s))_i ds - U_i^\kappa(\kappa t)\right] \\ & + \left[W_i^{(1)}\left(\int_0^{\kappa t} \Lambda_i^\kappa(s)ds\right) - W_i^{(2)}\left(\int_0^{\kappa t} (RX^\kappa)_i(s)ds\right) - W_i^{(3)}\left(\int_0^{\kappa t} \gamma_i(Z^\kappa(s) - BX^\kappa(s))_i ds\right)\right] \\ & + \left[O\left(\log\left(\int_0^{\kappa t} \Lambda_i^\kappa(s)ds\right)\right) + O\left(\log\left(\int^{\kappa t} (RX^\kappa)_i(s)ds\right)\right) + O\left(\log\left(\int_0^{\kappa t} \gamma_i(Z^\kappa(s) - BX^\kappa(s))_i ds\right)\right)\right] \\ = {} & I_{i,1}^\kappa(t) + I_{i,2}^\kappa(t) + I_{i,3}^\kappa(t). \end{aligned} \tag{41}$$

where $f^\kappa(t) = O(g^\kappa(t))$ a.s. if $\limsup_{\kappa \to \infty} |f^\kappa(t)|/|g^\kappa(t)| < \infty$ a.s. We need the following lemma, which states that $\bar{Z}_i^\kappa(t) = Z_i^\kappa(\kappa t)/g(\kappa)$ is uniformly bounded.

Lemma 1. If assumption (10) holds, then for any admissible sequence of controls $\{U^\kappa, X^\kappa\}$

$$\limsup_{\kappa \to \infty} \sup_{0 \leq t \leq T} \bar{Z}_i^\kappa(t) \leq M < \infty \; a.s.,$$

for all $i = 1, \dots, m$, where M is an $\mathcal{F}_T$-measurable r.v.

The above proposition implies that

$$\frac{Z_i^\kappa(\kappa t)}{\kappa g(\kappa)} \to 0 \quad a.s. \text{ as } \kappa \to \infty.$$

Dividing both sides of (41) by $\kappa g(\kappa)$, we appeal to the following lemma, which establishes the convergence of the second and third terms as $\kappa \to \infty$.

Lemma 2. If assumption (10) holds, then for any admissible sequence of controls $\{(U^\kappa, X^\kappa)\}$

$$\frac{I_{i,2}^\kappa(t) + I_{i,3}^\kappa(t)}{\kappa g \kappa} \to \quad 0 \quad a.s. \quad as \quad \kappa \to \infty,$$

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

Thus, we get

$$\frac{I_{i,1}^\kappa(t)}{\kappa g(\kappa)} \to \quad 0 \quad a.s. \quad \text{as} \quad \kappa \to \infty,$$

for all $i = 1, \ldots, m$. Using the Definition of Λ^κ in (7) and assumption (10) we have

$$\int_0^t \bar{Z}_i^\kappa(u)du \to \int_0^t (\gamma_i^{-1}[\Lambda_i(s) - (R\bar{X}(s))_i] + (B\bar{X}(s))_i)ds - \gamma_i^{-1}U_i(t) \quad a.s. \quad (42)$$

as $\kappa \to \infty$. Since $\bar{Z}_i^\kappa(s) - (B\bar{X}(s))_i \geq 0$, (42) implies that $U_i(t) - U_i(s) \geq \int_s^t \Lambda_i(u)du$, for all $0 \leq s \leq t \leq T$. Hence, $U_i(t)$ is Lipschitz and thus there exists (a.s.) an integrable function V such that $U_i(t) = \int_0^t V_i(s)ds$ for all $t \in [0, T]$. Substituting $\int_0^t V_i(s)ds$ for $U_i(t)$ in (42) completes the proof.

Proof of Proposition 2. Consider any optimal solution x_* of the LP (18) and let (q_*, v_*) be defined as in the statement of the proposition. Then (x_*, q_*, v_*) is feasible for LP (16) and $p^a \cdot \Gamma q_* + h \cdot q_* + p^b \cdot v_* = p \cdot (\lambda - Rx_*)$. Let (x', q', v') be an optimal solution to LP (16). Construct the vector (q'', v'') from x' using (19) and (20). Then, we have $p^a \cdot \Gamma q' + h \cdot q' + p^b \cdot v' \geq p^a \cdot \Gamma q'' + h \cdot q'' + p^b \cdot v''$. Since (x', q'', v'') is feasible, it is optimal for LP (16). Further, note that $p^a \cdot \Gamma q'' + h \cdot q'' + p^b \cdot v'' = p \cdot (\lambda - Rx')$. Since x' is feasible for LP (18), using the optimality of x_* we get that $p \cdot (\lambda - Rx') \geq p \cdot (\lambda - Rx_*)$. Hence $p^a \cdot \Gamma q'' + h \cdot q'' + p^b \cdot v'' \geq p^a \cdot \Gamma q_* + h \cdot q_* + p^b \cdot v_*$. Consequently, (x_*, q_*, v_*) is an optimal solution of LP (16), which completes the proof. □

Proof of Theorem 1. Consider any sequence of admissible controls $\{(U^\kappa, X^\kappa)\}$. All subsequent probabilistic statements are to be interpreted in the almost sure sense, and the term is omitted for brevity. Since $\{(\kappa g(\kappa))^{-1}\mathcal{J}^\kappa(U^\kappa, X^\kappa) : \kappa = 1, 2, \ldots\}$ is a sequence in $\mathbb{R}_+$, it has a subsequence $\{\kappa_n : n = 1, 2, \ldots\}$ which converges to the $\liminf_{\kappa\to\infty}(\kappa g(\kappa))^{-1}\mathcal{J}^\kappa(U^\kappa, X^\kappa)$. Further, since $\{(U^{\kappa_n}, X^{\kappa_n})\}$ is admissible, by (3) we have that $\bar{X}^{\kappa_n}$ is uniformly bounded. Appealing to Lemma 4 from Bassamboo et al. [1] [which is similar to Lemma 3 of this paper], there exists a function $X : \Omega \times [0, T] \mapsto \mathbb{R}_+$ 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

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

for all $t \in [0, T]$ and $j \in \{1, \ldots, n\}$. Using the fact that Λ is continuous and satisfies $\mathbb{E}[\int_0^T \Lambda(t)dt] < \infty$, along with the admissibility condition (2), it follows that

$$\limsup_{\kappa_n \to \infty} \bar{U}_i^{\kappa_n}(t) \le \int_0^T \Lambda_i(s)\,ds,$$

for all $t \in [0, T]$ and $i \in \{1, \ldots, m\}$. Next, we state a general result for uniformly bounded non-negative functions.

Lemma 3. Let $\{Y^\kappa\}$ be a sequence of uniformly bounded, non-negative, non-decreasing functions in $D[0, T]$. Then, for every subsequence there exists a further subsequence $Y^{\kappa n}$ and a non-decreasing function Y, such that $Y^{\kappa_n}(t) \to Y(t)$ as $n \to \infty$ for almost all $t \in [0, T]$.

Appealing to the lemma above, there exists a non-decreasing function $U : \Omega \times [0, T] \mapsto \mathbb{R}_+$ 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

$$\bar{U}_i^{\kappa_{n''}}(t) \to U_i(t) \quad \text{as} \quad n'' \to \infty$$

for all $t \in [0, T]$ and $i \in \{1, \ldots, m\}$. To simplify notation we shall drop the index of this further subsequence and assume that the above holds on the initial subsequence. Since Proposition 1 applies to this subsequence, from (42) it follows that for all $i \in \{1, \ldots, m\}$

$$U_i(T) + \int_0^T \gamma_i(Z(s) - BX(s))_i ds = \int_0^T (\Lambda(s) - RX(s))_i ds \quad \text{a.s.,} \tag{43}$$

where $\int_0^T Z_i(s)ds = \lim_{n\to\infty} \int_0^T \bar{Z}_i^{\kappa_n}(s)ds$. We then have,

$$\begin{aligned}
(\kappa g(\kappa))^{-1}\mathcal{J}^{\kappa_n}(X^{\kappa_n}, b^{\kappa_n}) &\to p^a \cdot \Gamma\left(M(T) - \int_0^T BX(s)\,ds\right) \\
&\quad + h \cdot \left(M(T) - \int_0^T BX(s)\,ds\right) + p^b \cdot U(T), \quad \text{as} \quad n \to \infty \qquad (44) \\
&\overset{(a)}{\ge} p \cdot \Gamma\left(M(T) - \int_0^T BX(s)\,ds\right) + p \cdot U(T) \\
&\overset{(b)}{=} \int_0^T p \cdot [\Lambda(t) - RX(t)]\,dt \quad \text{a.s.,}
\end{aligned}$$

where: the limit follows using the same strong approximation arguments used in the proof of Proposition 1; the inequality (a) follows from the Definition of p; and (b) follows from (43). Next, we show that $p \cdot [\Lambda(t) - RX(t)] \geq \pi(\Lambda(t), b)$ for almost all $t \in [0, T]$. Note that $X(t)$ satisfies the constraints of LP (18). To this end, we have that for almost all $t \in [0, T]$ (with respect to Lebesgue measure), $\Lambda(t) - RX(t) \geq 0$, $AX(t) \leq b$, and $X(t) \geq 0$. The first inequality follows from the fact that

$$\int_0^t \gamma_i(\bar{Z}^{\kappa_n}(s) - B\bar{X}^{\kappa_n}(s))_i ds + \bar{U}_i^{\kappa_n}(t) \to \int_0^t (\Lambda(s) - RX(s))_i ds \quad \text{a.s.} \quad \text{as} \quad n \to \infty,$$

for all $i = 1, \ldots, m$, and $\int_0^t \gamma_i(\bar{Z}^{\kappa_n}(s) - B\bar{X}^{\kappa_n}(s))_i ds$, and $\bar{U}^{\kappa_n}(t)$ are non-decreasing in t for each κ_n. Thus, we have that $\int_0^t(\Lambda(s) - RX(s))_i ds$ is non-decreasing in t. Consequently, $\Lambda(t) - RX(t) \geq 0$ for almost all $t \in [0, T]$. The second inequality follows using a similar argument, and since $AX^{\kappa_n} \leq b^{\kappa_n}$ implies that $\int_0^T(b^{\kappa_n} - AX^{\kappa_n}(s))ds$ is non-decreasing in t for each κ_n. (The last inequality follows from the condition $X^{\kappa_n}(t) \geq 0$.) 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 $\{(U^\kappa, X^\kappa)\}$

$$\liminf_{\kappa \to \infty} (\kappa g(\kappa))^{-1} \mathbb{E}[\mathcal{J}^\kappa(U^\kappa, X^\kappa)] \geq c \cdot b + \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b) dt\right]. \tag{45}$$

This completes the proof. □

Proof of Theorem 2. Let $\hat{X}_*^\kappa(t)$ be the optimal solution to the LP (23) with the estimator (22), *i.e.*, $\hat{X}_*^\kappa(t) = \phi^\kappa(\hat{\Lambda}^\kappa(t), b_*^\kappa)$ where ϕ^κ is the Lipschitz continuous mapping defined in (24). Let $\tilde{X}_*^\kappa$ denote a minimal truncation of $\hat{X}_*^\kappa$. Recall that U_*^κ satisfies (26) and (27). Let Z_*^κ denote the headcount process associated with the admissible control $(U_*^\kappa, \tilde{X}_*^\kappa)$.

By Theorem 1 and the Definition of asymptotic optimality, it suffices to show that

$$\limsup_{\kappa \to \infty} (\kappa g(\kappa))^{-1} \mathbb{E}[\mathcal{J}^\kappa(U_*^\kappa, \tilde{X}_*^\kappa)] \leq \mathbb{E}\left[\int_0^T \pi(\Lambda(t), b) dt\right]. \tag{46}$$

Consider the subsequence over which the lim sup is achieved for $(\kappa g(\kappa))^{-1}\mathcal{J}^\kappa(U_*^\kappa, \tilde{X}_*^\kappa)$. Consider a further subsequence $\{\kappa_n : n > 0\}$ of this sequence over which both $\int_0^{\kappa_n t} g(\kappa_n)^{-1}\tilde{X}_*^{\kappa_n}(s)ds$, $\int_0^{\kappa_n t} g(\kappa_n)^{-1}\tilde{Z}_*^{\kappa_n}(s)ds$ and $(\kappa_n g(\kappa_n))^{-1}U_*^{\kappa_n}(\kappa_n t)$ converge to a limit. Let

$$\begin{aligned}
\int_0^t (Z_*(s))_i ds &= \lim_{n \to \infty} \int_0^{\kappa_n t} \frac{(\tilde{Z}_*^{\kappa_n}(s))_i}{g(\kappa_n)} ds \quad \text{for all} \quad i = 1, \ldots, m, \\
\int_0^t (\tilde{X}_*(s))_i ds &= \lim_{n \to \infty} \int_0^{\kappa_n t} \frac{(\tilde{X}_*^{\kappa_n}(s))_i}{g(\kappa_n)} ds \quad \text{for all} \quad i = 1, \ldots, n, \\
U_*(t) &= \lim_{n \to \infty} \frac{U^{\kappa_n t}(\kappa_n t)}{\kappa_n g(\kappa_n)},
\end{aligned}$$

for all $t \in [0, T]$. Since over this common subsequence condition (12) holds, we have for all $i = 1, \dots m$,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{\kappa_n g(\kappa_n)} \Bigg[& p_i^a N_i^{(3)} \left(\int_0^{\kappa_n T} \gamma_i (\tilde{Z}_*^{\kappa_n}(s) - B\tilde{X}_*^{\kappa_n}(s))_i ds \right) \\ & + h_i \int_0^{\kappa_n T} (Q_*^{\kappa_n}(t))_i dt + p^b (U_*^{\kappa_n}(\kappa_n T))_i \Bigg] \\ & = (p_i^a \gamma_i + h_i) \left(\int_0^T (Z_*(s) - B\tilde{X}_*(s))_i ds \right) + p_i^b (U_*(T))_i. \end{aligned} \tag{47}$$

Let $X_*(t) = \phi(\Lambda(t), b)$ for all $t \in [0, T]$. Since $\Lambda(t)$ 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

$$\frac{\hat{X}_*^\kappa(\kappa t)}{g(\kappa)} - X_*(t) = \phi\left(\frac{\hat{\Lambda}^\kappa(\kappa t)}{g(\kappa)}, b \right) - \phi(\Lambda(t), b).$$

Since the mapping ϕ is Lipschitz continuous, there exists a finite constant C such that

$$\left\| \frac{\hat{X}_*^\kappa(\kappa t)}{g(\kappa)} - X_*(t) \right\| \le C \left\| \frac{\hat{\Lambda}^\kappa(t)}{g(\kappa)} - \Lambda(t) \right\|,$$

for all $t \in B$, where $\|\cdot\|$ is the Euclidean norm. Taking supremum over $t \in B$, the limit as $\kappa \to \infty$, and using the fact that the estimator is uniformly consistent [see Bassamboo et al. ([1], Proposition 3)], we get

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

Thus, $\hat{X}_*^\kappa$ satisfies the conditions of the following two Lemmas.

Lemma 4. Let $X^\kappa(t)$ be an untruncated control satisfying the admissibility condition (3) such that $g(\kappa)^{-1} X^\kappa(\kappa t) \to X(t) \quad a.s. \quad \text{as} \quad \kappa \to \infty$, where the convergence is uniform over compact sets of (0, T], and $X : [0, T] \mapsto \mathbb{R}_+^n$ is continuous and such that $RX(t) \le \Lambda(t)$ for all $t \in [0, T]$. Fix an $i \in \{1, \dots, m\}$ and let $U_i^\kappa(T) = 0$ for all κ. If $\tilde{X}^\kappa(t)$ is a minimal truncation of $X^\kappa(t)$ and assumption (10) holds, then

$$\lim_{\kappa\to\infty} \frac{1}{\kappa g(\kappa)} \int_0^{\kappa T} (\Lambda^\kappa(s) - R\tilde{X}^\kappa(s))_i ds = \lim_{\kappa\to\infty} \frac{1}{\kappa g(\kappa)} \int_0^{\kappa T} (\Lambda^\kappa(s) - RX^\kappa(s))_i ds \quad a.s.$$

Lemma 5. Let $X^\kappa(t)$ be an untruncated control satisfying the admissibility condition (3) such that $g(\kappa)^{-1} X^\kappa(\kappa t) \to X(t) \quad a.s. \quad \text{as} \quad \kappa \to \infty$, where the convergence is uniform over compact sets of (0,T], and $X : [0, T] \mapsto \mathbb{R}^n_+$ is continuous and such that $RX(t) \leq \Lambda(t)$ for all $t \in [0, T]$. Fix an $i \in \{1, \ldots, m\}$ and let U^κ_i be as defined in (27), and $L(\kappa)$ satisfies the following technical condition

$$\frac{L(\kappa)}{g(\kappa)} \to 0, \quad \text{and} \quad \frac{\log(\kappa)}{L(\kappa)} \to 0 \quad as \quad \kappa \to \infty. \tag{48}$$

If $\tilde{X}^\kappa(t)$ is a minimal truncation of $X^\kappa(t)$ and assumption (10) holds, then

$$\lim_{\kappa\to\infty} \frac{U^\kappa_i(\kappa T)}{\kappa g(\kappa)} = \lim_{\kappa\to\infty} \frac{1}{\kappa g(\kappa)} \int_0^{\kappa T} (\Lambda^\kappa(s) - RX^\kappa(s))_i ds \quad a.s.,$$
$$\text{and} \int_0^{\kappa T} \frac{Z^\kappa_i(s) - (B\tilde{X}^\kappa)_i}{\kappa g(\kappa)} ds \to 0 \quad a.s. \quad \text{as} \quad \kappa \to \infty.$$

For $i \in \mathcal{S}_a$, where $\mathcal{S}_a$ is defined in Section 4.2 to be the set of customer classes for which no customers are blocked under the proposed policy, i.e., $U_i(T) = 0$, we appeal to Proposition 1 and the Definition of p to get that

$$\begin{aligned}(p^a_i \gamma_i + h_i)\left(\int_0^T (Z_*(s) - B\tilde{X}_*(s))_i ds\right) + p^b_i (U_*(T))_i &= p_i \int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i \, ds \\ &= p_i \int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i ds \quad a.s., \end{aligned} \tag{49}$$

where $\tilde{X}^\kappa_*$ is a minimal truncation of $\hat{X}^\kappa_*$, and the second equality above follows from Lemma 4. Similarly, appealing to Lemma 5 and Proposition 1, we have for all $i \in S_b$, where S_b is defined in Section 4.2 to be the set of customer classes for which customers are blocked based on a threshold,

$$\begin{aligned}\left(p^a_i \gamma_i + h_i\right)\left(\int_0^T (Z_*(s) - B\tilde{X}_*(s))_i ds\right) + p^b_i (U_*(T))_i &= p_i \int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i ds \\ &= p_i \int_0^T (\Lambda(s) - R\tilde{X}_*(s))_i ds \quad a.s. \end{aligned} \tag{50}$$

Using (47), (49) and (50) we have

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathcal{J}^{\kappa_n}(U^{\kappa_n}_*, \tilde{X}^{\kappa_n}_*)}{\kappa g(\kappa)} &= \sum_{i=1}^m p_i \int_0^T (\Lambda(s) - RX_*(s))_i ds \\ &= \int_0^T \pi(\Lambda(s), b) ds \quad a.s.,\end{aligned}$$

where π is the mapping defined for LP (18). Consequently, we have

$$\limsup_{\kappa\to\infty}(\kappa g(\kappa))^{-1}\mathcal{J}^{\kappa}(U_{*}^{\kappa}, \tilde{X}_{*}^{\kappa}) = \int_0^T \pi(\Lambda(s), b)\, ds \quad a.s.$$

Since $\mathcal{J}^{\kappa}(U_{*}^{\kappa}, \tilde{X}_{*}^{\kappa})$ is non-negative and bounded, using the reverse Fatou lemma we get (46). This completes the proof. □

B auxiliary results

Proof of Lemma 1. The proof follows straightforwardly from Bassamboo et al. [1], Lemma 3) who establish the result for the same system considered here only without admission control; the added blocking can only decrease the headcount. □

Proof of Lemma 2. Fix $i \in \{1, \ldots, m\}$. Using the Definition of Λ^{κ} in (7) and the fact $\mathbb{E}\left[\int_0^T \Lambda_i(t)dt\right] < \infty$ we have

$$\int_0^{\kappa t} \Lambda_i^{\kappa}(s)ds = \kappa g(\kappa)\int_0^t \Lambda_i(s)ds \leq M_1\kappa g(\kappa), \tag{51}$$

for some M_1 such that $M_1 < \infty$ a.s. Also since the admissible control $\{X^{\kappa}\}$ satisfies $AX^{\kappa} \leq b^{\kappa}$, using the Definition of b^{κ} in (8) we have

$$\int_0^{\kappa t} (RX^{\kappa})_i(s)ds \leq M_2\kappa g(\kappa) \tag{52}$$

for some $M_2 < \infty$. Lastly, using Lemma 1, we also have for κ large

$$\int_0^{\kappa t} \gamma_i(Z^{\kappa}(s) - BX^{\kappa}(s))_i ds \leq M_3\kappa g(\kappa) \tag{53}$$

for some M_3 such that $M_3 < \infty$ a.s. Using (51),(52) and (53) along with the fact that

$$\frac{|W(t)|}{t} \to 0 \quad a.s. \quad \text{as } t \to \infty,$$

where $(W(s) : 0 \leq s \leq T)$ is a standard Brownian motion, we get the desired result. This completes the proof. □

Proof of Lemma 3. Omitted. □

Proof of Lemma 4. Since $U_i^{\kappa}(T) = 0$ for all κ, no customer is blocked. The proof then follows directly from Bassamboo et al. ([1], Lemma 5). □

Proof of Lemma 5. Fix an $i \in \mathcal{S}_b$. Since any continuous function on a compact set can be approximated to arbitrary accuracy from above or below by a piecewise constant function with finite number of discontinuities, we can approximate X and Λ as follows. Given an $\epsilon > 0$, there exists $N < \infty, 0 = t_0 < t_1 < \cdots < t_N = T$ and constants $X_1, \ldots, X_N, \Lambda_1, \ldots, \Lambda_N$ such that

$$\epsilon < X(t) - X_\ell < 2\epsilon e, \epsilon < \Lambda_\ell - \Lambda(t) < 2\epsilon e \quad \text{for all} \quad t \in [t_{\ell-1}, t_\ell) \tag{54}$$

for all $\ell = 1, \ldots, N$, where e is a vector of ones in $\mathbb{R}_n$. Let $Y(t)$ and $\bar{\Lambda}(t)$ be defined as follows

$$\begin{aligned} Y(t) &= X_\ell \quad \text{for all} \quad t \in [t_{\ell-1}, t_\ell), \\ \bar{\Lambda}(t) &= \Lambda_\ell \quad \text{for all} \quad t \in [t_{\ell-1}, t_\ell). \end{aligned}$$

Put $\bar{\Lambda} = (\bar{\Lambda}(t) : 0 \le t \le T)$ and $Y = (Y(t) : 0 \le t \le T)$, and consider a sequence of systems referred to as System I's. The κ^{th} System I has an arrival rate $g(\kappa)\bar{\Lambda}_i(\kappa^{-1}t)$ and rate of service completion given by

$$\begin{pmatrix} \text{net rate of service completion} \\ \text{for class } i \text{ customer at time } t \end{pmatrix} = \begin{cases} g(\kappa)(RY(\kappa^{-1}t))_i & \text{if } \bar{\zeta}_i^\kappa \ge g(\kappa)(BY(\kappa^{-1}t) \\ & \qquad +2\epsilon Be)_i + L(\kappa) \\ 0 & \qquad \text{otherwise,} \end{cases} \tag{55}$$

for all $t \in [0, \kappa T]$, where $\bar{\zeta}_i^\kappa(t)$ represents the headcount process in this system. The net abandonment rate at time $t \in [0, \kappa T]$ is $\gamma_i(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t) - 2g(\kappa)\epsilon Be - L(\kappa))_i^+$, where $x^+ = \max\{0, x\}$.

With regard to the admission control, the system manager blocks an arriving job at time t if $\bar{\zeta}_i^\kappa(t) > g(\kappa)(BY(\kappa^{-1}t) + 2\epsilon Be)_i + 2L(\kappa)$. By Lemma 1, there exists a finite M which is measurable with respect to $\mathcal{F}_T$ and such that $\limsup_{\kappa\to\infty} \sup_{0\le s\le \kappa T}\{g(\kappa)^{-1} Z_i^\kappa(s)\} \le M$. For each time κt_ℓ, $l = 1, \ldots, N$, we increase the headcount in buffer i for this system to $M^\kappa = Mg(\kappa)$.

Let Z_i^κ, U_i^κ be the headcount process and admission control in the original system for class i and let $\bar{\Psi}_i^\kappa(t)$ be the number of customers blocked in this alternate system up until time t. The following lemma asserts that $Z_i^\kappa(t)$ is dominated by $\bar{\zeta}_i^\kappa(t)$. □

Lemma 6. There exists a construction of a sequence of System I's on the same probability space as the original system, such that for K sufficiently large

$$Z_i^\kappa(t) \le \bar{\zeta}_i^\kappa(t) \quad a.s.,$$

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

Using the above lemma we have, for κ sufficiently large, that

$$\frac{\int_0^{\kappa T}(Z^\kappa(t) - B\tilde{X}^\kappa(t))_i dt}{\kappa g(\kappa)} \le \frac{\int_0^{\kappa T}(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ dt}{\kappa g(\kappa)} \quad a.s.$$

Thus, we have

$$\limsup_{\kappa\to\infty} \frac{\int_0^{\kappa T}(Z^\kappa(t) - B\tilde{X}^\kappa(t))_i dt}{\kappa g(\kappa)} \le \limsup_{\kappa\to\infty} \frac{\int_0^{\kappa T}(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ dt}{\kappa g(\kappa)} \quad a.s.$$

We next appeal to the following lemma which gives a bound on the right-hand-side above.

Lemma 7. For the sequence of System I's described above we have for all $i = 1, \ldots, m$ that

$$\limsup_{\kappa\to\infty} \frac{\int_0^{\kappa T}(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ dt}{\kappa g(\kappa)} \le 2\epsilon(Be)_i T\,.$$

Using Lemma 7 and letting ϵ go to zero we have

$$\limsup_{\kappa\to\infty} \int_0^{\kappa T} \frac{(Z^\kappa(t) - B\tilde{X}^\kappa(t))_i dt}{\kappa g(\kappa)} = 0\,.$$

Now, we consider another sequence of systems referred to as-System II's. The κ^{th} System II has an arrival rate $\bar{\Lambda}$ and rate of service completion given by

$$\begin{pmatrix}\text{net rate of service completion}\\ \text{for class } i \text{ customer at time } t\end{pmatrix} = \left\{\begin{matrix} g(\kappa)(R(Y(\kappa^{-1}t) - 2\epsilon e))_i^+ & \text{if } \bar{\zeta}_i^\kappa \ge g(\kappa)(BY(\kappa^{-1}t))_i \\ 0 & \text{otherwise}\end{matrix}\right\}, \tag{56}$$

for all $t \in [0, \kappa T]$, where $\hat{\zeta}_i^\kappa(t)$ represents the headcount process in this system. There are no abandonments in System II. In addition, the system manager blocks a job if $\hat{\zeta}_i^\kappa > g(\kappa)(BY(t)+L(\kappa))$. Again by Lemma 1, there exists a finite M which is measurable with respect to $\mathcal{F}_T$ such that $\limsup_{\kappa\to\infty}\sup_{0\le s\le\kappa T}\{g(\kappa)^{-1}Z_i^\kappa(s)\} < M$. For each time κt_ℓ, $l = 1, \ldots, N$, we increase the headcount in buffer i for this system to $M^\kappa = Mg(\kappa)$. Let $\hat{\Phi}_i^\kappa(t)$ be the number of customers blocked in this system up until time t. We now use the following two lemmas.

Lemma 8. There exists a construction of a sequence of System II's on the same probability space as the original system, such that for K sufficiently large,

$$U_i^\kappa(\kappa T) \le \hat{\Psi}_i^\kappa(\kappa T) \quad a.s.,$$

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

Lemma 9. For the sequence of System II's defined above we have

$$\limsup_{\kappa\to\infty} \frac{\hat{\Psi}_i^\kappa(\kappa T)}{\kappa g(\kappa)} = \sum_{\ell=1}^{N} [(\bar{\Lambda}_\ell - RX_\ell - 2\epsilon Re)_i^+][t_{\ell-1} - t_\ell] \quad a.s.,$$

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

Thus, we have the following inequalities

$$\begin{aligned}\limsup_{\kappa\to\infty} \frac{U_i^\kappa(\kappa T)}{\kappa g(\kappa)} &\geq \sum_{\ell=1}^{N} \big[(\bar{\Lambda}_\ell - RX_\ell - 2\epsilon Re)_i^+\big][t_{\ell-1} - t_\ell] \\ &= \int_0^T (\bar{\Lambda}(t) - R(Y(t) - 2\epsilon e))_i^+ dt \\ &\geq \int_0^T (\Lambda(t) - RX(t))_i dt + M_2 \epsilon T \quad a.s.,\end{aligned}$$

where M_2 is a constant independent of ϵ. Letting ϵ go to zero, completes the proof.

Proof of Lemma 6. Using uniform convergence of $g(\kappa)^{-1}X^\kappa(\cdot)$ to $X(\cdot)$ and (54), we have for κ sufficiently large that

$$g(\kappa)Y(\kappa^{-1}s) \geq X^\kappa(s) \geq g(\kappa)Y(\kappa^{-1}s) + 2\epsilon e \tag{57}$$

for all $s \in [0, \kappa T]$. Fix $\ell \in \{1, \ldots, N\}$. We shall prove the assertion for the interval $[\kappa t_{\ell-1}, \kappa t_\ell)$. First note that for all $t \in [\kappa t_{\ell-1}, \kappa t_\ell)$ we have $\bar{\zeta}_i^\kappa(t) \geq g(\kappa)(BY(\kappa^{-1}t) + 2\epsilon Be)_i + L(\kappa)$. Define

$$t^*_{\ell,\kappa} = \min\{\inf\{t \geq \kappa t_{\ell-1} : Z_i^\kappa(t) \leq g(\kappa)(BY(\kappa^{-1}t) + 2\dot{\epsilon} Be)_i + 2L(\kappa)\}, \kappa t_\ell\}\,.$$

Then we have $Z_i^\kappa(t) \leq g(\kappa)(BY(\kappa^{-1}t) + 2\epsilon Be)_i + L(\kappa)$ for all $t \in [t^*_{\ell,\kappa}, \kappa t_\ell)$. Thus, the result holds trivially for the interval $[t^*_{\ell,\kappa}, \kappa t_\ell)$. Now, we consider the interval $[\kappa t_{\ell-1}, t^*_{\ell,\kappa})$, using the Definition of minimal truncation for the routing control, and the Definition of $t^*_{\ell,\kappa}$, we have $\tilde{X}_j^\kappa(t) = X_j^\kappa(t)$ for $[\kappa t_{\ell-l}, t^*_{\ell,\kappa})$ and $i(j) = i$. We shall construct the original system and System I on the same space in the following manner: if $Z_i^\kappa(t) = \bar{\zeta}_i^\kappa(t)$ we use the same Poisson processes to generate the next arrival, service completion and abandonment; otherwise we let them evolve independently. (Similar constructions are discussed in Whitt [15] and Appendix B.2 of Bassamboo et al. [1]). Now consider any time $t \in [\kappa t_{\ell-1}, t^*_{\ell,\kappa})$ at which $Z_i^\kappa(t) = \bar{\zeta}_i^\kappa(t)$. The arrival rate in System I is higher than the arrival rate in the original system, i.e., $g(\kappa)\bar{\Lambda}_i(\kappa^{-1}t) \geq g(\kappa)\Lambda_i(\kappa^{-1}t)$. Using (57), we have for sufficiently large κ the service rate in System I is less than the service rate in the original system, i.e., $g(\kappa)(RY(\kappa^{-1}t))_i \leq (RX^\kappa t))_i$, and finally the abandonment rate in System I is less than that in the original system, i.e.,

$$\gamma_i(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t) - 2g(\kappa)\epsilon Be - L(\kappa))_i^+ \leq \gamma_i(Z^\kappa(t) - BX^\kappa(t) - L(\kappa))_i^+.$$

Since the arrival rates are higher, and the service rates and abandonment rates are both smaller for System I compared to the original system when $Z_i^\kappa(t) = \bar{\zeta}_i^\kappa(t)$, and the same Poisson processes are used for generating arrivals and service times and abandonments, we have that Z^κ will have a downward jump before $\bar{\zeta}_i^\kappa$ whenever $Z_i^\kappa(t) = \bar{\zeta}_i^\kappa(t)$. Since for κ sufficiently large $\bar{\zeta}_i^\kappa(\kappa t_{\ell-1}) \geq Z_i^\kappa(\kappa t_{\ell-1})$, the result holds for $t \in [0, \kappa T]$. This completes the proof. □

Proof of Lemma 7. Fix $\ell \in \{1, \dots, N\}$. Consider the interval $[\kappa t_{\ell-1}, \kappa t_\ell)$. It suffices to show

$$\limsup_{\kappa\to\infty} \frac{\int_{\kappa t_{\ell-1}}^{\kappa t_\ell} (\bar{\zeta}^\kappa(t) - g(\kappa) BY(\kappa^{-1}t))_i^+ \, dt}{\kappa g(\kappa)} \leq 2\epsilon(Be)_i(t_\ell - t_{\ell-1}) \text{ a.s.}$$

Define

$$t_{\ell,\kappa}^* = \min\{\inf\{t \geq \kappa t_{\ell-1} : \bar{\zeta}_i^\kappa(t) \leq g(\kappa)(BY(\kappa^{-1}t) + 2\epsilon Be)_i + 2L(\kappa)\}, \kappa t_\ell\}.$$

Note that for $t \in [t_{\ell,\kappa}^*, \kappa t_\ell]$ we have $(\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ \leq 2g(\kappa)\epsilon(Be)_i + 2L(\kappa)$. Note that all customers arriving in $[\kappa t_{\ell-1}, t_{\ell,\kappa}^*]$ are blocked and $\bar{\zeta}_i^\kappa(\kappa t_\ell) = Mg(\kappa)$. Thus,

$$\frac{\int_{\kappa t_{\ell-1}}^{\kappa t_\ell} (\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ \, dt}{\kappa g(\kappa)} \leq 2\epsilon(Be)_i(t_\ell - t_{\ell-1}) + \frac{2L(\kappa)}{g(\kappa)} + \frac{M(t_{\ell,\kappa}^* - t_{\ell-1})}{g(\kappa)} \tag{58}$$

Taking limsup as $\kappa\to\infty$ of both sides of (58) and using the growth condition $L(\kappa)/g(\kappa) \to 0$ as $\kappa \to \infty$ for the second term on the right-hand-side we get

$$\begin{aligned} \limsup_{\kappa\to\infty} \frac{\int_{\kappa t_{\ell-1}}^{\kappa t_\ell} (\bar{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i^+ \, dt}{\kappa g(\kappa)} &\leq 2\epsilon(Be)_i(t_\ell - t_{\ell-1}) \\ &+ \limsup_{\kappa\to\infty} \frac{Mg(\kappa)(t_{\ell,\kappa}^* - t_{\ell-1})}{\kappa g(\kappa)} \text{ a.s.} \end{aligned}$$

To complete the proof we shall prove

$$\limsup_{\kappa\to\infty} \frac{t_{\ell,\kappa}^* - \kappa t_\ell}{\kappa} = 0 \text{ a.s.}$$

For this we shall consider two cases as follows:
Case I: Let $(RY(t_{\ell-1}))_i > 0$. Consider a r.v. S_ℓ^κ which is the sum of $Mg(\kappa)$ exponentials with rate $g(\kappa)(RY(t_{l-1}))_i$. S_ℓ^κ can be constructed on the same probability space as the sequence of System I's such that $t_{\ell,\kappa}^* - \kappa t_\ell \leq S_\ell^\kappa$. We also have the following relations

$$\mathbb{E}\left[\frac{S_\ell^\kappa}{\kappa}\right] = \frac{M}{\kappa(RY(t_{\ell-1}))_i}, \quad Var\left[\frac{S_\ell^\kappa}{\kappa}\right] = \frac{M}{\kappa^2 g(\kappa)(RY(t_{\ell-1}))_i^2}.$$

Thus, by the Chebychev bound we have for any $\delta > 0$

$$\sum_{\kappa=1}^{\infty} \mathbb{P}\left(\left|\frac{S_\ell^\kappa}{\kappa} - \frac{M}{\kappa(RY(t_{\ell-1}))_i}\right| > \delta\right) \leq \sum_{\kappa=1}^{\infty} \frac{M}{\delta\kappa^2 g(\kappa)(RY(t_{\ell-1}))_i^2} < \infty.$$

Hence using Borel-Cantelli we have

$$\left|\frac{S_\ell^\kappa}{\kappa} - \frac{M}{\kappa(RY(t_{\ell-1}))_i}\right| \to 0 \text{ a.s., as } \kappa \to \infty.$$

Since $M < \infty$ a.s. and $(RY(t_{\ell-1}))_i > 0$ we have $M/(RY(t_{\ell-1}))_i < \infty$ a.s., thus we have

$$\limsup_{\kappa\to\infty} \frac{t_{\ell,\kappa}^* - \kappa t_\ell}{\kappa} \leq \limsup_{\kappa\to\infty} \frac{S_\ell^\kappa}{\kappa} = 0 \text{ a.s.}$$

Case II: Let $(RY(t_{\ell-1}))_i = 0$. Consider a r.v. $\hat{S}_\ell^\kappa$ which is the maximum of $Mg(\kappa)$ exponentials with rate γ_i. $\hat{S}_\ell^\kappa$ can be constructed on the same probability space as the sequence of System I's such that $t_{\ell,\kappa}^* - \kappa t_\ell \leq \hat{S}_\ell^\kappa$. We also have the following relations

$$\mathbb{E}\left[\frac{\hat{S}_\ell^\kappa}{\kappa}\right] \leq \frac{C_1 \log(g(\kappa)M)}{\kappa\gamma_i}, \quad Var\left[\frac{\hat{S}_\ell^\kappa}{\kappa}\right] = \frac{C_2}{\kappa^2\gamma_i},$$

where C_1 and C_2 are constants. Thus, by the Chebychev bound we have for any $\delta > 0$

$$\sum_{\kappa=1}^{\infty} \mathbb{P}\left(\left|\frac{\hat{S}_\ell^\kappa}{\kappa} - \mathbb{E}\left[\frac{\hat{S}_\ell^\kappa}{\kappa}\right]\right| > \delta\right) \leq \sum_{\kappa=1}^{\infty} \frac{C_2}{\delta\kappa^2} < \infty.$$

Hence using Borel-Cantelli we have

$$\left|\frac{\hat{S}_\ell^\kappa}{\kappa} - \mathbb{E}\left[\frac{\hat{S}_\ell^\kappa}{\kappa}\right]\right| \to 0 \text{ a.s. as } \kappa \to \infty.$$

Since $M < \infty$ a.s. and $\kappa^{-1}\log(g(\kappa)) \to 0$ as $\kappa \to \infty$ we have $(\kappa\gamma_i)^{-1}(C_1 \log(g(\kappa)M))$ $\to 0$ a.s. as $\kappa \to \infty$, which in turn implies

$$\limsup_{\kappa\to\infty} \frac{t_{\ell,\kappa}^* - \kappa t_\ell}{\kappa} \leq \limsup_{\kappa\to\infty} \frac{\hat{S}_\ell^\kappa}{\kappa} = 0 \text{ a.s.}$$

This completes the proof. □

Proof of Lemma 8. We shall prove the result by considering following modification of System II defined in the proof of lemma 5. (The modified system is referred to as System III.) All parameters are identical to System II except that the system manager rejects a customer if $\check{\zeta}_i^\kappa(s) > (BX^\kappa(s))_i + L(\kappa)$, where $\check{\zeta}_i^\kappa(s)$ is the headcount in System III. Let $\check{\Psi}_i^\kappa$ be the admission control for System III. We shall construct the original system and System III on the same space in the following manner: we use same Poisson process to generate arrivals; and if $Z_i^\kappa(t) = \bar{\zeta}_i^\kappa(t)$ we use the same Poisson processes to generate the service completion and abandonment; otherwise we let the service completion and abandonments occur independently. We note that if $Z_i^\kappa(s) = \bar{\zeta}_i^\kappa(s)$ at some time instant, then the arrival rate into System III, $g(\kappa)\bar{\Lambda}_i(\kappa^{-1}t)$, is higher than rate of arrival into the original system, $g(\kappa)\Lambda_i(\kappa^{-1}t)$, *i.e.*, $g(\kappa)\bar{\Lambda}_i(\kappa^{-1}t) \geq g(\kappa)\Lambda_i(\kappa^{-1}t)$. Further, for κ sufficiently large, the service rate in the original system is greater than that in System III since

$$(R\tilde{X}^\kappa(t))_i \geq g(\kappa)(RY(\kappa^{-1}t) - 2\epsilon e)_i^+ \text{ if } Z_i^\kappa(t) > g(\kappa)(BY(t))_i, \tag{59}$$

for all $t \in [0, \kappa T]$. The above follows from the fact that $g(\kappa)^{-1}X^\kappa(\kappa^{-1}t)$ converges to $X(t)$ uniformly. Also, there are no abandonments in System III. Since the arrival rate is higher and the service rate and abandonment rates are smaller for System III compared to the original system whenever $Z_i^\kappa(s) = \check{\zeta}_i^\kappa(s)$, then there exists a construction of System III on the same probability space such that $Z_i^\kappa(s)$ jumps downward before $\check{\zeta}_i^\kappa(s)$ if $Z_i^\kappa(s) = \check{\zeta}_i^\kappa(s)$. Hence if $Z_i^\kappa(s) = \check{\zeta}_i^\kappa(s)$ then for all $t > s$ we have $Z_i^\kappa(s) \leq \check{\zeta}_i^\kappa(s)$. Since for κ large $Z_i^\kappa(0) \leq \check{\zeta}_i^\kappa(0) = Mg(\kappa)$, these systems are constructed on the same probability space such that $Z_i^\kappa(t) \leq \check{\zeta}_i^\kappa(t)$ for all $t \in [0, \kappa T]$ and any arrival to the original system also correspond to an arrival to System III. Since any arrival in the original system that is blocked implies that the corresponding arrival in System III is also blocked, we have that

$$U_i^\kappa(T) \leq \check{\Psi}_i^\kappa(T) \text{ a.s.} \tag{60}$$

Next we shall prove that there exists a construction of System II on the same probability space such $\hat{\Psi}_i^\kappa(T) \geq \check{\Psi}_i^\kappa(T)$. For this we shall describe a construction such that

$$\hat{\Psi}_i^\kappa(t) + \hat{\zeta}_i^\kappa(t) \geq \check{\Psi}_i^\kappa(t) + \check{\zeta}_i^\kappa(t), \tag{61}$$

$$\text{and} \quad \hat{\zeta}_i^\kappa(t) \leq \check{\zeta}_i^\kappa(t) \text{ a.s.}, \tag{62}$$

for all $t \in [0, \kappa T]$. We use the same Poisson processes to generate arrivals and services times in System II and III. It is easy to verify that the relationship described in (61–62) holds at time $t = 0$. To show that (62) holds at all times, consider any time t at which $\hat{\zeta}_i^\kappa(t) = \check{\zeta}_i^\kappa(t)$. The arrival rate and the service rate for both systems are identical, and since the buffer size in System III is larger than System II we have that any arrival blocked by System II will also be blocked by System III. Thus, the stated inequality $\hat{\zeta}_i^\kappa(s) \leq \check{\zeta}_i^\kappa(s)$ holds for all $s > t$. For (61) we note that any arrival to the system will increase either the headcount or the number of customers blocked, hence an arrival maintains the inequality in (61). Further, since any time instant corresponding to a service completion in System II also corresponds to a service completion in

System III, we have that the inequality $\hat{\Psi}_i^\kappa(t) + \hat{\zeta}_i^\kappa(t) \geq \check{\Psi}_i^\kappa(t) + \check{\zeta}_i^\kappa(t)$ holds for all $t \in [0, \kappa T]$. Thus, we have $U_i^\kappa(T) \leq \check{\Psi}_i^\kappa(T) \leq \hat{\Psi}_i^\kappa(T)$, and the proof is complete. □

Proof of Lemma 9. Fix an $\ell \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, m\}$. It suffices to show that

$$\limsup_{\kappa \to \infty} \frac{\hat{\Psi}_i^\kappa(t)(kt_\ell) - \hat{\Psi}_i^\kappa(t)(kt_{\ell-1})}{\kappa g(\kappa)} \leq (\Lambda_\ell - R(X_\ell - 2\epsilon e))_i^+ [t_\ell - t_{\ell-1}] \text{ a.s.}$$

Note the numerator of the left-hand-side is the number of class i customers blocked during $[\kappa t_{\ell-1}, \kappa t_\ell]$. We assume $R(X_\ell - 2\epsilon e))_i^+ > 0$, otherwise all the jobs are simply blocked and the result then follows from the strong approximation result given in Proposition 4. Consider the process $(\hat{\zeta}^\kappa(t) - g(\kappa)BY(\kappa^{-1}t))_i$. The dynamics of this process is the same as an $M/M/1$ queue with finite buffer of size $L(\kappa)$ having arrival rate $g(\kappa)(\Lambda_\ell)_i$ and service rate is $g(\kappa)(R(X_\ell - 2\epsilon e))_i^+$. Consider the following modified M/M/1 queue where the initial number of customers is $(Mg(\kappa) - g(\kappa)BY(\kappa^{-1}t_{\ell-1}))_i^+$ and all the jobs are blocked up until the time the queue becomes empty. We denote the time when the queue becomes empty by $\hat{t}_{\ell,\kappa}$. For all $t \in [\hat{t}_{\ell,\kappa}, \kappa t_\ell]$, this system operates as the aforementioned $M/M/1$ queue. Let the number of customers blocked during the time interval $[\kappa t_{\ell-1}, \kappa t_\ell]$ be represented by G_i^κ. Since all the jobs are blocked during the interval $[\kappa t_{\ell-1}, \hat{t}_{\ell,\kappa}]$, we have

$$\hat{\Psi}_i^\kappa(kt_\ell) - \hat{\Psi}_i^\kappa(t)(kt_{\ell-1}) \leq G_i^\kappa \text{ a.s.}$$

Using arguments similar to that in Lemma 7, we have

$$\limsup_{\kappa \to \infty} \frac{\hat{t}_{\ell,\kappa} - \kappa t_{\ell-1}}{\kappa g(\kappa)} \to 0 \text{ a.s.}$$

Next, combining the strong approximation result stated in Proposition 4 with the above, we have

$$\frac{N_i^{(1),\kappa}\left(\int_{\kappa t_{\ell-1}}^{\hat{t}_{\ell,\kappa}} g(\kappa)\bar{\Lambda}_i(\kappa^{-1}s)ds\right)}{\kappa g(\kappa)} \to 0 \text{ a.s. as } \kappa \to \infty, \tag{63}$$

where $N_i^{(1),\kappa}$ is the unit rate Poisson process that generates arrivals for System II. Fix a $\delta > 0$. Next we consider a system consisting of two buffers: Buffer A and Buffer B. For Buffer A, the arrival rate is $g(\kappa)((\bar{\Lambda}_\ell)_i, (R(X_\ell - 2\epsilon e))_i - \delta)^+$ and service rate $g(\kappa)(R(X_\ell - 2\epsilon e))_i^+$ and the buffer size is $L(\kappa)$. Let $Z_A^\kappa(t)$ be the headcount in Buffer A. We initialize Buffer A with its steady-state distribution. Note that Buffer A is an $M/M/1$ queue with finite buffer size and traffic intensity $\rho < 1$ independent of κ. Buffer B has arrival rate of $g(\kappa)((\bar{\Lambda}_\ell - R(X_\ell - 2\epsilon e))_i - \delta)^+$ and all the customers are blocked. Let U_A^κ and U_B^κ represent the customers blocked at Buffer A and B during the time $[\kappa t_{\ell-1}, \kappa t_\ell]$. Then using (63) and the fact that the number of blocked customers in this two-buffer system is greater than $(G_i^\kappa - N_i^{(1),\kappa}(\int_{\kappa t_{\ell-1}}^{\hat{t}_{\ell,\kappa}} \bar{\Lambda}_i(\kappa^{-1}s)ds))$, we have that

$$\limsup_{\kappa\to\infty} \frac{\check{\Psi}_i^\kappa}{\kappa g(\kappa)} \leq \limsup_{\kappa\to\infty} \frac{U_A^\kappa + U_B^\kappa}{\kappa g(\kappa)} \text{ a.s.}$$

Since all the customers arriving to buffer B are blocked, then U_B^κ is given by the number of jumps in a Poisson process with rate $g(\kappa)((\bar{\Lambda}_\ell - R(X_\ell - 2\epsilon e))_i - \delta)^+$ in time interval $[\kappa t_{\ell-1}, \kappa t_\ell]$. Using the strong approximation result stated in Proposition 4 we have that

$$\frac{U_B^\kappa}{\kappa g(\kappa)} \to [t_\ell - t_{\ell-1}]((\Lambda_\ell - R(X_\ell - 2\epsilon e))_i - \delta)^+ \text{ a.s. as } \kappa \to \infty.$$

Next, since Buffer A is in steady-state, we have

$$\mathbb{E}\left[\frac{U_A^\kappa}{\kappa g(\kappa)}\right] = \frac{\rho^{L(\kappa)}(1 - \rho^{L(\kappa)+1})}{1-\rho}\Lambda_\ell.$$

Thus, for any $\nu > 0$ using the growth condition $L(\kappa)/(\log \kappa)$ as $\kappa \to \infty$ and Markov's inequality, we have

$$\sum_{\kappa=1}^{\infty} \mathbb{P}\left(\frac{U_A^\kappa}{\kappa g(\kappa)} > \nu\right) \leq \sum_{\kappa=1}^{\infty} \frac{C\rho^{L(\kappa)}}{\nu} < \infty,$$

for some constant C, given $\rho < 1$ independent of κ. Using Borel-Cantelli we have that

$$\frac{U_A^\kappa}{\kappa g(\kappa)} \to 0 \text{ a.s. as } \kappa \to \infty.$$

Thus, we have

$$\limsup_{\kappa\to\infty} \frac{\hat{\Psi}_i^\kappa(t)(\kappa t_\ell) - \hat{\Psi}_i^\kappa(\kappa t_{\ell-1})}{\kappa g(\kappa)} \leq [t_\ell - t_{\ell-1}]((\Lambda_\ell - R(X_\ell - 2\epsilon e))_i - \delta)^+ \text{ a.s.}$$

We now get the desired result by letting $\delta \to 0$. This completes the proof. □

References

[1] A. Bassamboo, J.M. Harrison and A. Zeevi, Design and control of a large call center: Asymptotic analysis of an LP-based method, Operations Research (To appear 2005).

[2] P. Bremaud, *Point Processes and Queues: Martingale Dynamics*, (Springer Verlag, New York, 1981).

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

[4] L. Green and P. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals, Management Science 37 (1991) 84–97.

[5] S. Halfin and W. Whitt, Heavy-traffic limits for queues with many exponential servers, Operations Research 29 (1981) 567–588.

[6] J.M. Harrison and M. Lopez, Heavy traffic resource pooling in parallel-server systems, Queueing Systems 33 (1999) 339–368.

[7] J.M. Harrison and A. Zeevi, Dynamic scheduling of a multi-class queue in the Halfin-Whitt heavy traffic regime, Operations Research 52 (2004) 243–257.

[8] J.M. Harrison and A. Zeevi, A method for staffing large call centers based on stochastic fluid models, Manufacturing & Service Operations Management 7 (2005) 20–36.

[9] J. Komlós, P. Major and G. Tusnady, An approximation of partial sums of independent random variables and the sample distribution I, Z. Wahr. und Verw. Gebiete 32 (1975) 111–131.

[10] T. Kurtz, Strong approximation theorems for density dependent Markov chains, Stochastic Processes and Their Applications 6 (1978) 223–240.

[11] C. Maglaras and J. Van Mieghem, Admission and sequencing control under delay constraints with applications to GPS and GLQ. To appear in the European Journal of Operations Research (2004).

[12] A. Mandelbaum, W. Massey and M. Reiman, Strong approximations for Markovian service networks, Queueing Systems, Theory and Applications (QUESTA) 30 (1998) 149–201 .

[13] W.A. Massey and W. Whitt, Uniform acceleration expansions for Markov chains with time-varying rates, Annals of Applied Probabability 8(4) (1998) 1130–1155.

[14] E. Plambeck, S. Kumar and J.M. Harrison, A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls, Queueing Systems 39 (2001) 23–54.

[15] W. Whitt, Comparing counting processes and queues, Advances in Applied Probability 13 (1981) 207–220.

[16] W. Whitt, The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase, Management Science 37 (1991) 307–314.

[17] W. Whitt, *Stochastic-Process Limits*, (Springer-Verlag, New York, 2002).

[18] W. Whitt, Fluid models for many-server queues with abandonments, Working paper (2004). To appear in Operations Research.