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

Achal Bassamboo* Stanford University
J. Michael Harrison† Stanford University
Assaf Zeevi‡ Columbia University

Submitted: June 8, 2004, Revised: June 21, 2005
To appear in *Operations Research*

Abstract

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

1 Introduction

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

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

*Graduate School of Business, e-mail: achalb@stanford.edu
†Graduate School of Business, e-mail: harrison_michael@gsb.stanford.edu
‡Graduate School of Business, e-mail: assaf@gsb.columbia.edu

particular customer class, and conversely, there can be more than one customer class that is served by a particular agent pool.

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

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

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

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

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

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

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

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

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

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

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

Existing analytical approaches and related work. As indicated by Gans et al. (2003, §5.1), both staffing and dynamic routing problems in multi-class/multi-pool call centers are essentially outside the reach of exact analytical methods [for exceptions, see, *e.g.*, Gans and Zhou (2003) and Yahalom and Mandelbaum (2005)]. Thus most research on these problems has focused on various forms of approximations, a particularly prominent role being played by two asymptotic regimes.

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

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

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

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

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

The remainder of the paper. Section 2 provides a precise description of the call center model and economic objective. Section 3 describes the asymptotic parameter regime used in later analysis. Section 4 gives the main results, and section 5 presents "picture proofs" of these results by means of simulation experiments. Section 6 concludes with some remarks and directions for future research. Proofs of the main results are given in Appendix A, while Appendix B contains proofs of auxiliary results.

2 Problem Formulation

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

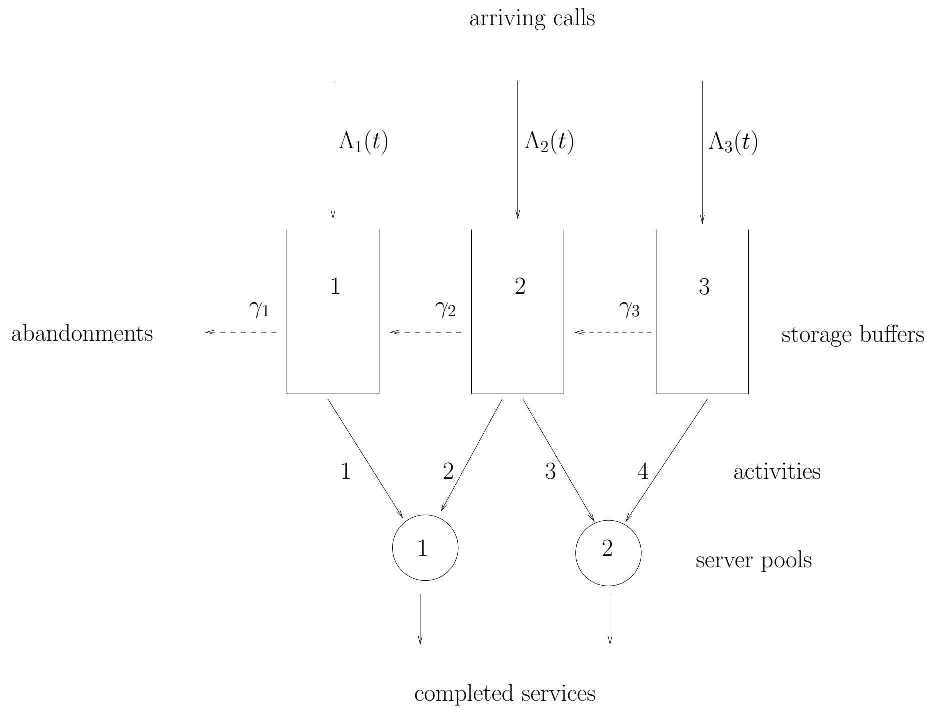


Figure 1: A call center with three customer classes, two agent pools and four activities.

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

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

Gans et al. (2003).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 An Asymptotic Parameter Regime

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

where

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

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

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

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

where $\bar{Z}$ solves

$$\begin{aligned}\bar{Z}(t) &= \int_0^t \Lambda(s)ds - \int_0^t R\bar{X}(s)ds - \int_0^t \Gamma\bar{Q}(s)ds,\\ \bar{Z}(0) &= 0 \ \text{ and } \ \bar{Q}(0) = 0\end{aligned}$$

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

4 Main Results

4.1 An asymptotic lower bound on achievable performance

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

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

where R is the unscaled input-output matrix, A is the capacity consumption matrix and p is the penalty-rate vector. Let $\Phi(\lambda, b)$ denote the optimal solution set of the LP (14); that is, if $x_* \in \mathbb{R}^n_+$

is an optimal solution of the LP, then $x_* \in \Phi(\lambda, b)$. (Formally, Φ is a point-to-set correspondence from (λ, b) to the solution set.) Let $b_* \in \mathbb{R}^r_+$ be a minimizer of

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

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

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

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

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

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

4.2 An asymptotically optimal policy when Λ is observable

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

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

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

$$\begin{aligned} \min \quad & p \cdot (\Lambda^\kappa(t) - R^\kappa x) \\ \text{s.t.} \quad & R^\kappa x \le \Lambda^\kappa(t),\ Ax \le b^\kappa_*,\ x \ge 0. \end{aligned} \tag{18}$$

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

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

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

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

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

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

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

$$\widetilde{X}^\kappa(t) \leq X^\kappa(t), \text{ and}$$
$$(B\widetilde{X}^\kappa)_i(t) < \widetilde{Z}^\kappa_i(t) \text{ implies } \widetilde{X}^\kappa_j(t) = X^\kappa_j(t) \text{ for all } j \text{ such that } i(j) = i.$$

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

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

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

4.3 An asymptotically optimal tracking policy when Λ is not observable

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

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

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

Definition 4 (uniform consistency) *An estimator $\widehat{\Lambda}^\kappa$ is said to be uniformly consistent if it satisfies*

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

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

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

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

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

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

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

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

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

4.4 A discrete-review Λ-tracking policy

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

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

To alleviate the deficiencies stated above, we now propose a *discrete-review* implementation of Λ-tracking policies. These controls are also based on the estimation of arrival rates, for which the same window size of $g(\kappa)^{-1}$ is used. However, instead of a sliding window, 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 $[0, T]$ into $g(\kappa)T$ review periods of equal length. In the ℓ^{th} review period, $\ell = 1, \ldots, \lfloor g(\kappa)T \rfloor$, the following estimate of Λ is used:

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

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

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

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

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

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

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

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

5 Numerical Examples

The key to all our main results is Proposition 1, which was stated in Section 3 and says essentially the following: if the scale of the system is large, then the headcount process Z, the dynamic control

X, and the instantaneous arrival rate vector Λ are approximately linked by the simple relationship (13), that is, one has approximately $Z(t) = \Gamma^{-1}[\Lambda(t) - RX(t)] + BX(t)$ for all $t \in [0, T]$.

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

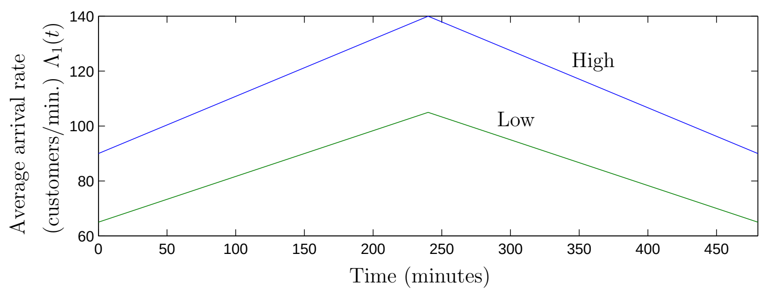
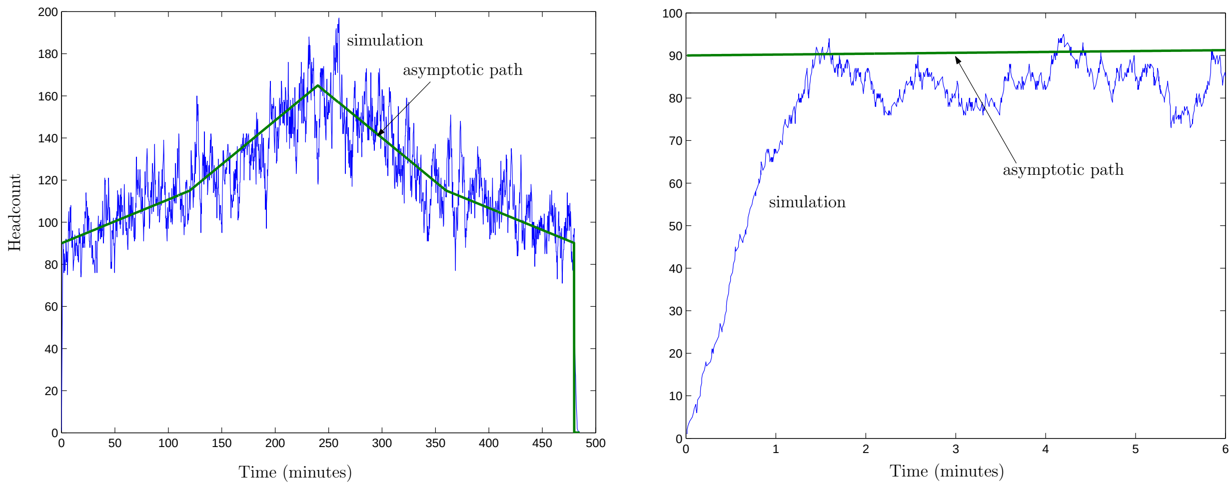


Figure 2: Arrival rate pattern for a single-class/single-pool example.

The cost of employing a server for one day is $c = \$240$, and the abandonment penalty is $p = \$2$ per customer. Solving the staffing problem given in (15), we find that $b_* = 115$ servers. Figure 3(a) depicts a sample path of the headcount process for the given system using the obvious dynamic control policy $X(t) = \min(b_*, Z(t))$, $t \geq 0$, superimposed on the asymptotic path (13) described in Proposition 1 (these results correspond to a "High" realization of the arrival rate). The headcount process in Figure 3(a) does indeed fluctuate around the asymptotic path given in (13).

To illustrate the time it takes an empty system to reach this "equilibrium" behavior, Figure 3(b) "zooms in" on the system dynamics around $t = 0$. The system equilibrates to the limiting path (13) within 1 to 2 minutes, after which it follows this path, in spite of the temporal changes in the arrival rate. To summarize, one can say that the limiting system state has essentially no dynamics, it instantly "forgets" its past and its evolution at any time instant only depends on the instantaneous arrival rate and the control policy (which is itself a function of the instantaneous arrival rate).

Next we illustrate the lower bound on system performance and its achievability (Theorems 1 and 4), by considering an example in which the dynamic routing policy is not trivial or obvious. Specifically, we consider a system with two customer classes ($m = 2$) which is served by two server pools ($r = 2$). There are three processing activities ($n = 3$). Servers in pool 1 can serve only class 1 customers (activity 1), while servers in pool 2 are cross-trained and can serve both class 1 and class



(a) (b)

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

2 customers (activities 2 and 3, respectively). Callers of class 1 and 2 arrive according to a doubly-stochastic Poisson process whose rates are displayed in Figure 4. We take the scaling function to be $f(\kappa) = \kappa^2$. All the services are exponentially distributed with unit rate, that is, $\mu_j = 1$ customers-per-minute for $j = 1, 2, 3$. Customers of class 1 abandon at rate $\gamma_1 = 0.2$ customers-per-minute, whereas customers of class 2 abandon at rate $\gamma_2 = 1$ customers-per-minute. The abandonment penalties for class 1 and class 2 are $p_1 = \$4$ per customer and $p_2 = \$1$ per customer, respectively. The cost of a server in pool 1 is \$600 per day and \$720 per day in pool 2 (where the servers are cross-trained).

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

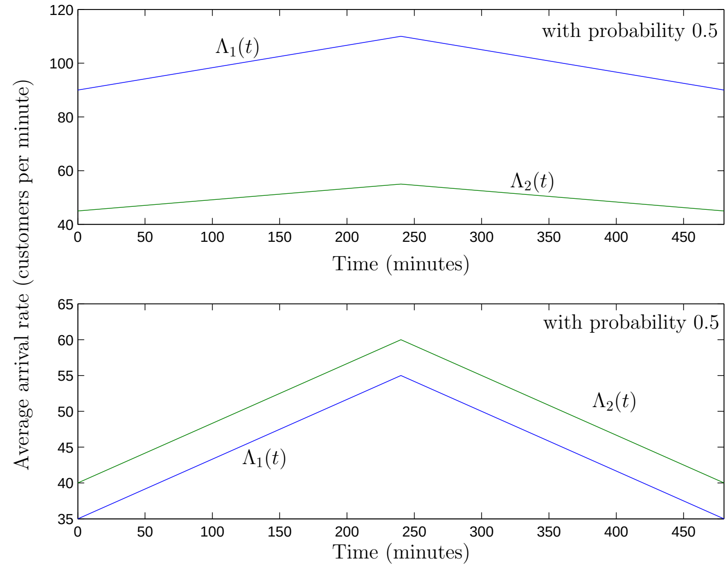


Figure 4: Arrival rates for the two-class/two-pool example.

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

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

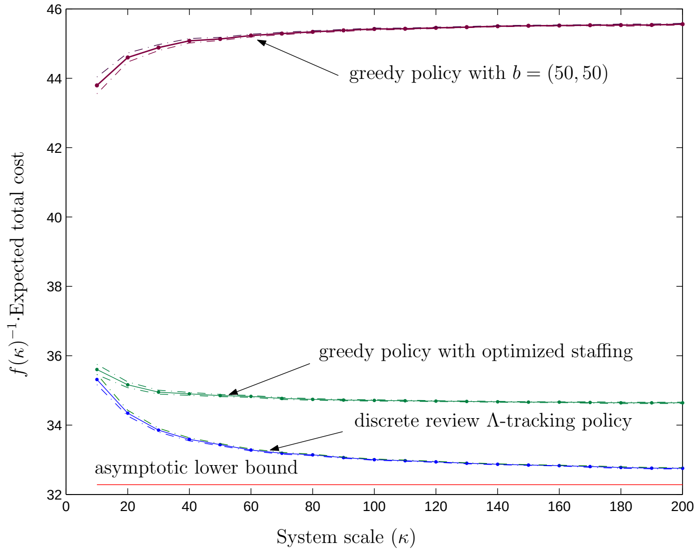


Figure 5: Scaled expected total cost as a function of the system scale (κ) for the 2-class 2-pool example; dotted lines correspond to 95% confidence interval for the simulated results.

6 Concluding Remarks

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

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

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

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

A Proofs 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 $\mathcal{F}_t = \sigma(\Lambda(s) : 0 \le s \le t)$ represent the information set generated by the arrival rate process up until time t. In the similar vein, let the information set generated by arrivals, departures and abandonment up until time t in the κ^{th} system be represented by $\mathcal{H}_t^\kappa$ for all $t \in [0,T]$. Let $D[0,T]$ denote the space of functions defined over $[0,T]$ which are right-continuous with left limits. In much of what follows, as well as in Appendix B, statements are said to hold almost surely for almost all time $t \in [0,T]$. Note that the above is weaker than the assertion that a statement holds for almost all time $t \in [0,T]$, almost surely. This distinction is a consequence of pointwise limits as opposed to functional limits. Finally, proofs of all lemmas cited in this appendix can be found in Appendix B.

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

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

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

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

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

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

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

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

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

$$\text{(i)}\ \frac{Z_i^\kappa(t)}{f(\kappa)} \to 0,\ \text{(ii)}\ \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)}ds \to M_i(t),$$
$$\text{(iii)}\ \frac{N_i^{(3)}(\int_0^t \gamma_i^\kappa(Z^\kappa(s) - BX^\kappa(s))_i ds)}{f(\kappa)} \to \gamma_i \left(M_i(t) - \int_0^t (BX(s))_i ds \right)$$

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

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

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

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

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

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

■

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

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

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

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

Appealing to the above lemma there exists a function $X : \Omega \times [0,T] \mapsto \mathbb{R}_+$ defined for almost all $\omega \in \Omega$ and (Lebesgue) almost all $t \in [0,T]$ and a further subsequence $\{\kappa_{n'} : n' = 1, 2, \ldots\}$ such that

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

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

$$\Gamma\left(M(T) - \int_0^T BX(s)ds\right) = \int_0^T (\Lambda(s) - RX(s))ds \tag{25}$$

where

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

We then have,

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

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

$$\begin{aligned} &\Lambda(t) - RX(t) \geq 0,\\ &AX(t) \leq b, \quad \text{and}\\ &X^{\kappa_n}(t) \geq 0 \text{ implies } X(t)\geq 0, \end{aligned}$$

where the first inequality follows from the fact that

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

for all $i = 1,\ldots,m$, and $\int_0^t \gamma_i(Z^{\kappa_n}(s) - BX^{\kappa_n}(s))_i ds$ is non-decreasing in t for each κ_n. Thus, we have $\int_0^t (\Lambda(s) - RX(s))_i\,ds$ 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 the fact $AX^{\kappa_n} \leq b^{\kappa_n}$ implies $\int_0^t (b^{\kappa_n} - AX_i^{\kappa_n}(s))\,ds$ is non-decreasing in t for each κ_n. The optimality of $\pi(\Lambda(t), b)$, together with the above result and Fatou's lemma yields that for any admissible sequence of dynamic controls $\{X^\kappa\}$ and staffing vectors $\{b^\kappa\}$,

$$\liminf_{\kappa\to\infty} f(\kappa)^{-1}\mathbb{E}[\mathcal{J}^\kappa(X^\kappa, b^\kappa)] \geq c\cdot b + \mathbb{E}[\int_0^T \pi(\Lambda(t), b)dt] \tag{27}$$

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

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

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

This completes the proof. ■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Consider the subsequence over which the lim sup is achieved for $f(\kappa)^{-1}\mathcal{J}^\kappa(\widetilde{X}_*^\kappa, b_*^\kappa)$. Consider a further subsequence $\{\kappa_n : n > 0\}$ of this subsequence over which $\int \frac{\kappa_n}{f(\kappa_n)} \widetilde{X}_*^{\kappa_n}$ and $\int \frac{\kappa_n}{f(\kappa_n)} \widetilde{Z}_*^{\kappa_n}$ converge to a limit [the existence of such a subsequence follows from Lemma 3 in the proof of Theorem 1 and Lemma 4]. Let

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

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

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

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

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

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

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

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

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

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

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

Step 3: Combining the analysis in Step 1 and 2 we have

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

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

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

Step 4: Since $\mathcal{J}^\kappa(\widetilde{X}_*^\kappa, b_*^\kappa)$ is non-negative and bounded, using the reverse Fatou lemma we have (28). This completes the proof. ■

Proof of the Theorem 3. Recall that

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

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

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

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

Step 2: Let $X_*(t) = \phi(\Lambda(t), b_*)$ for all $t \in [0, T]$. 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

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

Since the mapping ϕ is Lipschitz continuous, we have

$$\left\| \frac{\kappa}{f(\kappa)} \widehat{X}_*^\kappa(t) - X_*(t) \right\| \leq C \left\| \frac{\widehat{\Lambda}^\kappa(t)}{f(\kappa)} - \Lambda(t) \right\|,$$

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

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

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

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

where $\widetilde{X}_*^\kappa$ is the minimal truncation of $\widehat{X}_*^\kappa$. Repeating Step 3 and 4 in the proof of Theorem 2 completes the proof. ■

Proof of Proposition 3. Fix $i \in \{1, \ldots, m\}$. For the remainder of the proof we shall focus our attention on the set of ω's for which $\int_0^T \Lambda_i(s)ds > 0$ (the result is trivially true on the complement set). Using the definition of the estimator, we have

$$\begin{aligned} \widehat{\Lambda}_i^\kappa(t) - \Lambda_i^\kappa(t) &= g(\kappa)\left[F_i^\kappa(t) - \int_0^t \Lambda_i^\kappa(s)ds\right] - g(\kappa)\left[F_i^\kappa(t - g(\kappa)^{-1}) - \int_0^{t-g(\kappa)^{-1}} \Lambda_i^\kappa(s)ds\right] \\ &\quad + g(\kappa)\int_{t-g(\kappa)^{-1}}^t \Lambda_i^\kappa(s)ds - \Lambda_i^\kappa(t) \end{aligned}$$

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

$$\sup_{t \in B} \frac{|\widehat{\Lambda}_i^\kappa(t) - \Lambda_i^\kappa(t)|}{f(\kappa)} \leq 2 \sup_{0 \leq t \leq T} \frac{g(\kappa)}{f(\kappa)} \left| F_i^\kappa(t) - \int_0^t \Lambda_i^\kappa(s)ds \right| + \sup_{t \in B} \left| g(\kappa) \int_{t-g(\kappa)^{-1}}^t \Lambda_i(s)ds - \Lambda_i(t) \right| \quad (29)$$

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

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

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

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

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

By choosing, $\theta_\kappa = \log\left(1+\frac{\epsilon}{g(\kappa)}\right)$, we have that the right-hand-side of the above equation is bounded by $\exp\left(-\frac{f(\kappa)}{g(\kappa)^2}\epsilon\int_0^T \Lambda(s)ds\right)$. Further, for any positive integer ℓ

$$\mathbb{P}\left(\sup_{0\le t\le T}\frac{g(\kappa)}{f(\kappa)}\left[F_i^\kappa(t)-\int_0^t \Lambda_i^\kappa(s)ds\right] > \epsilon\int_0^T \Lambda_i(s)ds, \text{ i.o.}\middle|\mathcal{F}_T\right) \le \sum_{\kappa=\ell}^{\infty}\exp\left(-\frac{f(\kappa)}{g(\kappa)^2}\epsilon\int_0^T \Lambda_i(s)ds\right),$$

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

$$\mathbb{P}\left(\sup_{0\le t\le T}\frac{g(\kappa)}{f(\kappa)}\left[F_i^\kappa(t)-\int_0^t \Lambda_i^\kappa(s)ds\right] > \epsilon\int_0^T \Lambda_i(s)ds, \text{ i.o.}\middle|\mathcal{F}_T\right) = 0 \text{ a.s.}$$

Hence for any $\epsilon > 0$,

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

and thus

$$\sup_{0\le t\le T}\frac{g(\kappa)}{f(\kappa)}\left|F_i^\kappa(t)-\int_0^t \Lambda_i^\kappa(s)ds\right| \to 0 \text{ a.s. as } \kappa\to\infty. \tag{31}$$

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

$$\left|g(\kappa)\int_{t-g(\kappa)^{-1}}^t \Lambda_i(s)-\Lambda_i(t)\right| \le g(\kappa)\int_{t-g(\kappa)^{-1}}^t |\Lambda_i(s)-\Lambda_i(t)|\,ds \le \epsilon,$$

hence

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

This completes the proof. ■

Proof of Theorem 4. Recall that

$$\widehat{\Lambda}^\kappa(t) = g(\kappa)\left[F^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}\right) - F^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor-1}{g(\kappa)}\right)\right],$$

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

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

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

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

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

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

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

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

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

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

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

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

Now, by definition of $\frac{\kappa_n}{f(\kappa_n)} \widetilde{X}_*^{\kappa_n}$,

$$\int_0^T (\Lambda(s) - R\widetilde{X}_*(s))_i ds = \lim_{n\to\infty} \int_0^T \left(\Lambda(s) - \frac{\kappa_n}{f(\kappa_n)} R\widetilde{X}_*^{\kappa_n}(s)\right)_i ds \quad \text{a.s.}$$

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

$$\begin{aligned} \lim_{n\to\infty} \int_0^T (\Lambda(s) - \frac{\kappa_n}{f(\kappa_n)} R\widetilde{X}_*^{\kappa_n}(s))_i ds \quad &\overset{\text{(a)}}{\leq} \quad \lim_{n\to\infty} \int_0^T \left(\Lambda_i(t) - \frac{\kappa_n}{f(\kappa_n)} (R\widetilde{X}_*^{\kappa_n}(s))_i \mathbb{I}_{\{s\in[0,T]\setminus I^{\kappa_n}\}}\right) ds \\ &\overset{\text{(b)}}{=} \quad \int_0^T (\Lambda(s) - RX_*(s))_i ds \quad \text{a.s.} \end{aligned}$$

for $i = 1, \ldots, m$, where $X_*(t) = \phi(\Lambda(t), b_*)$ for all $t \in [0,T]$. The inequality (a) follows from non-negativity of the process $\widetilde{X}_*^\kappa$, and the equality (b) follows from the Lemma 7 since $\widehat{X}_*^\kappa$ satisfies the conditions of the Lemma. Finally, repeating Steps 3 and 4 in the proof of Theorem 2 completes the proof. ∎

B Auxiliary Results

B.1 Proofs of lemmas from Appendix A

Proof of Lemma 1. Since $N_i^{(1)}(t)$ is a Poisson process with unit rate, it follows that $N_i^{(1)}(t)/t \to 1$ almost surely as $t \to \infty$ for all $i = 1, \ldots, m$. Now, for a given $t \in [0, 1]$ define $\Delta_i(t) := \int_0^t \Lambda_i(s)ds \geq 0$. We assume $\Delta_i(t) > 0$ as the case $\Delta_i(t) = 0$ is trivial. Using the fact that $\lim_{\kappa\to\infty} f(\kappa) = \infty$, we get

$$\lim_{\kappa\to\infty} \frac{N_i^{(1)}(f(\kappa)\Delta_i(t))}{f(\kappa)} = \Delta_i(t) \quad \text{a.s.} \tag{33}$$

hence the result follows from the scaling of $\Lambda^\kappa(\cdot)$.

For the second result, we use the following fact which we state without proof. Consider a function $d(k)$ with $d(k) \to \infty$ as $k \to \infty$. If $S(t)$ and $c(k)$ satisfy $S(t)/t \to 1$ as $t \to \infty$ and $c(k)/d(k) \to \gamma > 0$ as $k \to \infty$, then

$$\frac{S(c(k))}{c(k)} \to 1 \quad \text{as } k \to \infty.$$

Note that for admissible X^κ satisfying (11),

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

Thus, the convergence of the scaled and time changed Poisson process follows from (33). This completes the proof. ∎

Proof of Lemma 2. (i) Fix a $i \in \{1, \ldots, m\}$. Suppose there exists a $t \in [0, T]$ such that

$$\limsup_{\kappa\to\infty} \int_0^t \frac{Z_i^\kappa(s)}{f(\kappa)} ds > 0 \tag{34}$$

on a set of ω's with positive probability. Dividing both sides of (22) by $f(\kappa)$ and taking the lim sup as $\kappa \to \infty$ we have

$$\begin{aligned}
&\limsup_{\kappa\to\infty} \left\{ \frac{Z_i^\kappa(s)}{f(\kappa)} + \frac{N_i^{(3)}(\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s))_i ds)}{f(\kappa)} \right\} \\
&= \limsup_{\kappa\to\infty} \left\{ \frac{N_i^{(1)}(\int_0^t \Lambda^\kappa(s)ds)}{f(\kappa)} - \frac{N_i^{(2)}(\int_0^t (R^\kappa X^\kappa)_i(s)ds)}{f(\kappa)} \right\} \\
&= \int_0^t (\Lambda(s) - RX(s))ds < \infty
\end{aligned} \tag{35}$$

where the second equality follows from Lemma 1. Now, since $\{X^\kappa\}$ satisfies (11) we have

$$\lim_{\kappa\to\infty} \frac{\int_0^t \gamma_i^\kappa (BX^\kappa)_i(s)ds}{f(\kappa)} = \int_0^t \gamma_i (BX)_i(s)ds < \infty. \tag{36}$$

This together with the assumption (34) and the scaling of γ^κ gives

$$\limsup_{\kappa\to\infty} \frac{\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s))_i ds}{f(\kappa)} = \infty.$$

Using the same reasoning as in the second part of the proof of Lemma 1, we conclude that the left-hand-side of (35) is infinite on a set of ω's with positive probability, in contradiction. Hence for all $t \in [0,T]$

$$\limsup_{\kappa\to\infty} \int_0^t \frac{Z_i^\kappa(s)}{f(\kappa)} ds = 0 \quad \text{a.s.},$$

and hence for all $t \in [0,T]$, $\int_0^t Z_i^\kappa(s)/f(\kappa)ds \to 0$ a.s. as $\kappa \to \infty$. Since the integrand is nonnegative, the assertion in the lemma follows.

(ii) Fix an $i \in \{1,\dots,m\}$. Suppose there exist a set $C \subseteq [0,T]$ with positive Lebesgue measure such that for all $t \in C$

$$\underline{M}(t) := \liminf_{\kappa\to\infty} \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)} ds < \limsup_{\kappa\to\infty} \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)} ds =: \overline{M}(t),$$

on a set of ω's with positive probability. Then, there exist two subsequences $\{\kappa_n : n = 1,2,\dots\}$ and $\{\kappa_{n'} : n' = 1,2,\dots\}$ on which $\underline{M}$ and $\overline{M}$ are achieved, respectively. Appealing to the strong law of large numbers for $N_i^{(3)}(t)$, and the proof of Lemma 1, (36) implies that on these subsequences

$$\begin{aligned}
\lim_{n\to\infty} \frac{N_i^{(3)}(\int_0^t \gamma_i^{\kappa_n}(Z^{\kappa_n}(s) - BX^{\kappa_n}(s))_i ds)}{f(\kappa_n)} &= \gamma_i \left(\underline{M}(t) - \int_0^t (BX(s))_i ds \right) \\
\lim_{n'\to\infty} \frac{N_i^{(3)}(\int_0^t \gamma_i^{\kappa_{n'}}(Z^{\kappa_{n'}}(s) - BX^{\kappa_{n'}}(s))_i ds)}{f(\kappa_{n'})} &= \gamma_i \left(\overline{M}(t) - \int_0^t (BX(s))_i ds \right)
\end{aligned}$$

Now, by (22) and Lemma 1, we have that

$$\lim_{\kappa\to\infty} \frac{N_i^{(3)}(\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s))_i ds)}{f(\kappa))}$$

exists almost surely for almost all $t \in [0,T]$, in contradiction. Thus, we conclude that

$$\liminf_{\kappa\to\infty} \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)} ds = \limsup_{\kappa\to\infty} \int_0^t \frac{\kappa Z_i^\kappa(s)}{f(\kappa)} ds =: M_i(t)$$

almost surely, for almost all $t \in [0,T]$.

(iii) Combining (36), the definition of $M_i(t)$ and the scaling of γ^κ we have

$$\frac{\int_0^t \gamma_i^\kappa (Z^\kappa(s) - BX^\kappa(s)_i ds)}{f(\kappa))} \to \gamma_i \left(M_i(t) - \int_0^t (BX(s))_i ds \right) \quad \text{a.s. as } \kappa \to \infty \tag{37}$$

for all $i = 1,\dots,m$ and almost all $t \in [0,T]$. The result then follows directly from the strong law of large numbers for $N_i^{(3)}(t)$, (37) and the proof of Lemma 1. ∎

Proof of Lemma 3. It suffices to prove that there exists a subsequence for which the desired result holds. Consider the sequence $Y^\kappa = (Y^\kappa(t) : t \in [0,T])$ of uniformly bounded nonnegative functions in $D[0,T]$. In particular, suppose that $Y^\kappa(t) \le M$ for all $t \in [0,T]$ and for all $\kappa > 0$. Consider an extension of these functions to $D[0,2T]$ such that

$$Y^\kappa(t) = \begin{cases} Y^\kappa(t) & \text{if } t \in [0,T] \\ 2M - \frac{1}{T}\int_0^T Y^\kappa(s)ds & \text{if } t \in (T,2T] \end{cases}$$

Since Y^κ is a nonnegative function which is right-continuous with left-limits, we can view it as the density of a measure μ^κ on $[0,2T]$, where

$$\mu^\kappa(A) := \frac{1}{2MT}\int_A Y^\kappa(t)dt$$

for all measurable sets A in $[0,2T]$. Since the sequence of measures $\{\mu^\kappa : \kappa > 0\}$ is tight, Prohorov's theorem [cf. Billingsley (1999)] states that these measures are relatively compact. Thus, there exists a subsequence $\{\mu_{\kappa_\ell}\}$ which converges weakly to a measure μ. Then, there exists a Lebesgue decomposition of this measure given by

$$\mu(A) = \int_A Y(s)ds + \mu_s(A)$$

for all measurable sets A in $[0,2T]$, where μ_s is singular with respect to Lebesgue measure on $[0,2T]$, and $Y(\cdot)$ is the Radon-Nikodym derivative of the measure μ with respect to Lebesgue measure. Since μ_s is singular with respect to Lebesgue measure, there exists a set A with Lebesgue measure zero such that $\mu_s(A^c) = 0$. We shall next show that μ is absolutely continuous with respect to Lebesgue measure, *i.e.*, the singular part is zero. Given any $\epsilon > 0$, there exists an open set A^ϵ such that $A \subseteq A^\epsilon$ and Lebesgue measure of A^ϵ is less than ϵ. Since

$$\mu^\kappa(A^\epsilon) = \int_{A^\epsilon} Y^\kappa(s)ds \le 2M\epsilon,$$

and A^ϵ is open, we have $\mu_s(A) \le \mu(A^\epsilon) \le \liminf_{n\to\infty} \mu^{\kappa_n}(A^\epsilon)$, by the Portmanteau Theorem. Since ϵ can be made arbitrarily small, we have $\mu_s = 0$. This completes the proof. ■

Proof of Lemma 4. Consider the following system, constructed on the same probability space as the original one. All nominal service rates, for each activity in the original system, are now taken to be α, and all nominal abandonment rates are set to be α. The system is driven is driven by the same unit rate Poisson processes as the original one. Put $\alpha := \min\{\{\mu_j : 1 \le j \le n\} \cup \{\gamma_i : 1 \le i \le m\}\}$. Fix a κ and consider the κ^{th} system in which all customer classes get served and abandon at the rate $\kappa\alpha$, in particular the time-in-system of any customer has the distribution of an exponential random variable with rate $\kappa\alpha$. Consider another queuing system constructed in the same manner, with infinite servers and service times being exponentially distributed with rate $\kappa\alpha$. Let the number-in-system process at time t in this system be denoted by $Y^\kappa(t)$. Clearly, the headcount process Z^κ in the original system under any admissible control X^κ is (pointwise) dominated by Y^κ. Fix an

$i \in \{1, \dots, m\}$. It is clear that Y_i^κ conditioned on $\mathcal{F}_T$ is simply the headcount of an $M_t^\kappa/M^\kappa/\infty$ queue (a queue whose arrival stream is given by a nonhomogeneous Poisson process with rate $\Lambda_i^\kappa(t)$, the service times are exponential with rate $\kappa\alpha$, and the number of servers is infinite). This, in turn, can be expressed as the number of points in a certain region generated by an associated nonhomogeneous Poisson process [see Taylor and Karlin (1998, Theorem 6.1)]. In particular, the intensity of this nonhomogeneous Poisson process for the i^{th} customer class in our case would be $\Lambda_i^\kappa(t)\kappa\alpha e^{-\kappa\alpha y}dydt, \quad$ for all $(y,t) \in [0,\infty) \times [0,T]$.

To prove the assertion of the lemma, it suffices to establish that

$$\limsup_{\kappa\to\infty} \sup_{0\le t\le T} \frac{\kappa Y_i^\kappa(t)}{f(\kappa)} < \infty \text{ a.s.}$$

Divide the time $[0,T]$ into κT intervals of length κ^{-1} and consider the n^{th} interval $[\kappa^{-1}(n-1), \kappa^{-1}nT]$. Let $Y_{i,n}^\kappa$ denotes the maximum number-in-system in this interval. Thus $Y_{i,n}^\kappa$ is bounded by the number of points in the region, $[0,\infty] \times [0, nT/\kappa] \setminus B_n^\kappa$, generated by the two dimensional Poisson process where $B_n^\kappa = \{(y,t) : t + y \le \kappa^{-1}(n-1), t \in [0,T], y \in [0,\infty]\}$. Since $\Lambda_i(t)$ is continuous, it is uniformly bounded (let the uniform bound be M' which is $\mathcal{F}_T$-measurable). Further, since $\mathbb{E}[\int_0^T \Lambda_i(t)dt] < \infty$, we have $M' < \infty$ almost surely. The number of points in the aforementioned region is given by a Poisson r.v. whose mean is bounded above by $Mf(\kappa)/\kappa$ where $M = M'(1+\alpha^{-1})$. Using the Chernoff bound we get that

$$\mathbb{P}\left(Y_n^\kappa \ge 2M\frac{f(\kappa)}{\kappa} \middle| \mathcal{F}_T \right) \le \exp\left(-2(1-\log 2)M\frac{f(\kappa)}{\kappa} - 1 \right) \text{ a.s.}$$

Summing the inequality over all n and κ, we get

$$\sum_{\kappa=1}^{\infty}\sum_{n=1}^{\kappa T} \mathbb{P}\left(Y_n^\kappa \ge 2M\frac{f(\kappa)}{\kappa} \middle| \mathcal{F}_T \right) < \infty \tag{38}$$

by the growth condition on $f(\kappa)$. Then, by (38) we have $\mathbb{P}\left(\sup_{0\le t\le T} Y^\kappa(t) \ge 2M\frac{f(\kappa)}{\kappa}, \text{ i.o.} \middle| \mathcal{F}_T \right) = 0$ a.s. which further implies that $\mathbb{P}\left(\sup_{0\le t\le T} Y^\kappa(t) \ge 2M\frac{f(\kappa)}{\kappa}, \text{ i.o.} \right) = 0$. Thus,

$$\limsup_{\kappa\to\infty} \sup_{0\le t\le T} \frac{\kappa Z^\kappa(t)}{f(\kappa)} \le 2M := 2M'(1+\alpha) < \infty \text{ a.s.} \tag{39}$$

This completes the proof. ■

Proof of Lemma 5. The proof is given in several steps to ease referencing in other results that follow a similar proof pattern.

Step 1: Fix an $i \in \{1, \dots, m\}$. Using the definition of minimal truncation we have

$$\liminf_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(s) - R^\kappa \widetilde{X}^\kappa(s))_i ds \ge \lim_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(s) - R^\kappa X^\kappa(s))_i ds, \text{ a.s.}$$

All subsequent probabilistic statements are to be interpreted in the almost sure sense and the term is omitted for brevity. Since any continuous function on a compact set can be approximated from above or below by piecewise constant function with finite number of discontinuities, we can approximate X and Λ as follows. Given an $\epsilon > 0$, there exists $N < \infty$, $0 = t_0 < t_1 < \ldots. < t_N = T$ and constants $X_1, \ldots, X_N$, $\Lambda_1, \ldots, \Lambda_N$ such that

$$\epsilon < X(t) - X_\ell < 2\epsilon e, \ \epsilon < \Lambda_\ell - \Lambda(t) < 2\epsilon e \ \text{ for all } t \in (t_{\ell-1}, t_\ell) \tag{40}$$

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

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

Put $\bar{\Lambda} = (\bar{\Lambda}(t) : 0 \leq t \leq T)$ and $Y = (Y(t) : 0 \leq t \leq T)$, and consider a sequence of systems with (untruncated) control $Y^\kappa = f(\kappa)Y/\kappa$ and arrival rate $\bar{\Lambda}$. Let $\widetilde{Y}^\kappa(t)$ be a minimal truncation of $Y^\kappa(t)$ with priority given to the slower servers (see Definition 5 in Appendix B.2). By (39) in Lemma 4, there exists a finite M which is measurable with respect to $\mathcal{F}_T$ and such that $\limsup_{\kappa\to\infty} \sup_{0\leq s\leq T}\{f(\kappa)^{-1}\kappa Z_i^\kappa(s)\} < M$. For each time t_ℓ, $\ell = 1, \ldots, N$, we increase the headcount in buffer i to $M^\kappa = Mf(\kappa)/\kappa$. Let Z^κ be the headcount process in the original system and $\overline{Z}^\kappa$ be the headcount process in the modified system. Using the uniform convergence of X^κ to X and (40), there exists $\kappa_1 < \infty$ such that

$$Y^\kappa(t) \leq X^\kappa(t) \text{ for all } t \in [0, 1], \ \kappa > \kappa_1.$$

Consider a subsequence over which the lim sup of $\frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(s) - R^\kappa \widetilde{X}^\kappa(s))_i ds$ is achieved. Since $\{f(\kappa)^{-1}\kappa Z^\kappa\}$, $\{f(\kappa)^{-1}\kappa\widetilde{X}^\kappa\}$, $\{f(\kappa)^{-1}\kappa\overline{Z}^\kappa\}$ and $\{f(\kappa)^{-1}\kappa\widetilde{Y}^\kappa\}$ are bounded, using Lemma 3 there exists a subsequence $\{\kappa_n : n > 0\}$ over which their integrals over time converge pointwise.

Step 2: First observe that $0 \leq X^\kappa(t) - Y^\kappa(t) \leq 3\kappa^{-1}f(\kappa)\epsilon e$ for κ large enough, and also the arrival rate $\bar{\Lambda}$ is larger than $\Lambda(t)$. Using Lemma 9 (Appendix B.2) we can construct sequence of systems on the same probability space such that

$$\gamma_i^{\kappa_n}\left(Z_i^{\kappa_n}(t) - B\widetilde{X}^{\kappa_n}(t)\right)_i \leq \gamma_i^{\kappa_n}\left(\overline{Z}_i^{\kappa_n}(t) - B\widetilde{Y}^{\kappa_n}(t)\right)_i + 3\kappa^{-1}f(\kappa)\epsilon(Be)_i$$

for all $t \in [0, T]$ and n large enough. The second term on the right-hand-side appears since the controls can change over time which can cause an increase in queuelength not due to arrivals. Integrating over $[0, T]$, dividing both sides of the above inequality by $f(\kappa_n)$ and appealing to Proposition 1 we get

$$\lim_{n\to\infty} \frac{1}{f(\kappa_n)} \int_0^T (\Lambda^{\kappa_n}(s) - R^\kappa \widetilde{X}^{\kappa_n}(s))_i ds \leq \lim_{n\to\infty} \frac{1}{f(\kappa_n)} \int_0^T (\bar{\Lambda}^{\kappa_n}(s) - R^\kappa \widetilde{Y}^{\kappa_n}(s))_i ds.$$

Step 3: Now using Lemma 10 (Appendix B.2) and the fact that $|M| < \infty$, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^{\kappa_n}(s) - R^{\kappa_n}\widetilde{Y}^{\kappa_n}(s))_i ds &= \sum_{\ell=1}^{N} (t_{\ell+1} - t_\ell)(\overline{\Lambda}_{\ell+1} - RY_{\ell+1})_i^+ ds \\
&= \lim_{n\to\infty} \frac{1}{f(\kappa_n)} \int_0^T (\overline{\Lambda}^{\kappa_n}(s) - R^{\kappa_n} Y^{\kappa_n}(s))_i^+ ds \\
&\leq \lim_{n\to\infty} \frac{1}{f(\kappa_n)} \int_0^T (\Lambda^{\kappa_n}(s) - R^{\kappa_n} X^{\kappa_n}(s))_i ds + M_2 \epsilon T,
\end{aligned}$$

where $(x)^+ := \max\{x, 0\}$, $M_2 = C\|R\|$ for some constant C, and $\|\cdot\|$ denotes sup norm, and the last inequality follows from (40).

Step 4: Combining Steps 2 and 3 we have that for any $\epsilon > 0$,

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(s) - R^\kappa \widetilde{X}^\kappa(s))_i ds \leq \lim_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(s) - R^\kappa X^\kappa(s))_i ds + M_2 \epsilon T,$$

hence taking $\epsilon \to 0$, we get the desired result. This completes the proof. ■

Proof of Lemma 6. Fix a $t \in [0, T]$ and an $i \in \{1, \ldots, m\}$. For the remainder of the proof we shall focus our attention on the set of ω's for which $\int_0^T \Lambda_i(s) ds > 0$ (the result is trivially true on the complement set). Put

$$A_i^\kappa(t) := \left[F_i^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}\right) - F_i^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}\right)\right].$$

Then,

$$\begin{aligned}
&\widehat{\Lambda}_i^\kappa(t) - \Lambda_i^\kappa(t) \\
&= g(\kappa)\left[F_i^\kappa\left(\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}\right) - \int_0^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}} \Lambda_i^\kappa(s) ds\right] - g(\kappa)\left[F_i^\kappa(\left(\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}\right) - \int_0^{\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}} \Lambda_i^\kappa(s) ds\right] \\
&\quad + g(\kappa) \int_{\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}}^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}} \Lambda_i^\kappa(s) ds - \Lambda_i^\kappa(t).
\end{aligned}$$

Fix a compact set $B \subset (0, T]$, and κ such that $g(\kappa)^{-1} \leq \inf\{s : s \in B\}$. Then, we have

$$\sup_{t\in B} \frac{|\widehat{\Lambda}_i^\kappa(t) - \Lambda_i^\kappa(t)|}{f(\kappa)} \leq 2 \sup_{0\leq t\leq T} \frac{g(\kappa)}{f(\kappa)} \left|F_i^\kappa(t) - \int_0^t \Lambda_i^\kappa(s) ds\right| + \sup_{t\in B} \left| g(\kappa) \int_{\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}}^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}} \Lambda_i(s) ds - \Lambda_i(t)\right|, \quad (41)$$

since $\Lambda_i^\kappa = f(\kappa)\Lambda_i$. Now, the first term on the right-hand-side goes to zero almost surely as $\kappa \to \infty$ by (31) in the proof of Proposition 3. Let us now consider the second term on the right-hand-side of (41). First note that since the set B is closed and $0 \notin B$, we have $0 < \beta := \{s : s \in B\}$. Further, $\Lambda_i(s)$ is uniformly continuous on $[0, T]$, thus for any $\epsilon > 0$ there exists a $\delta > 0$ such that for all $t, s \in [0, T]$ such that $|t - s| < \delta$ we have $|\Lambda(s) - \Lambda(t)| < \epsilon$. Choose κ_1 such that $g(\kappa)^{-1} < \min\{\beta, \delta/2\}$ for all $\kappa > \kappa_1$. Then, for all $\kappa > \kappa_1$ and for all $t \in B$, we have

$$\left| g(\kappa) \int_{\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}}^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}} \Lambda_i(s) ds - \Lambda_i(t)\right| \leq g(\kappa) \int_{\frac{\lfloor tg(\kappa)\rfloor - 1}{g(\kappa)}}^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}} |\Lambda_i(s) - \Lambda_i(t)|\, ds \leq \epsilon.$$

Thus,

$$\sup_{t\in B}\left|g(\kappa)\int_{\frac{\lfloor tg(\kappa)\rfloor-1}{g(\kappa)}}^{\frac{\lfloor tg(\kappa)\rfloor}{g(\kappa)}}\Lambda_i(s)ds-\Lambda_i(t)\right|\to 0 \ \text{ a.s. as } \kappa\to\infty\ .$$

This completes the proof. ■

Proof of Lemma 7. The proof mimics that of Lemma 5 and follows in a similar stepwise fashion; below we indicate the necessary changes. Fix an $i\in\{1,\ldots,m\}$.

Step 1: We construct approximating processes to the original system as in Step 1 of the proof of Lemma 5 with the following modifications. $\widetilde{Y}^\kappa(t)$ is the minimal truncation giving priority to slower servers of $Y^\kappa(t)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}$ (see Definition 5 in Appendix B.2). $Z^\kappa(t)$ is the headcount associated with the dynamic control $\widetilde{X}^\kappa(t)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}$.

Step 2: Since $0\le (X^\kappa(t)-Y^\kappa(t))\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}\le 3\kappa^{-1}f(\kappa)\epsilon e$ for κ large, and the arrival rate $\bar\Lambda(t)$ is larger than the $\Lambda(t)$, using Lemma 9 we have (almost surely)

$$\gamma_i^{\kappa_n}\left(Z_i^{\kappa_n}(t)-B\widetilde{X}^{\kappa_n}(t)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}\right)_i\le\gamma_i^{\kappa_n}\left(\overline{Z}_i^{\kappa_n}(t)-B\widetilde{Y}^{\kappa_n}(t)\right)_i+3\kappa^{-1}f(\kappa)\epsilon e,$$

for all $t\in[0,T]$ and n large enough (the subsequence is the one over the which the integral of the scaled processes converge). Divide the above equation by $f(\kappa_n)$ and take a limit as $n\to\infty$. Using Proposition 1 we have that (almost surely)

$$\lim_{n\to\infty}\frac{1}{f(\kappa_n)}\int_0^T(\Lambda^{\kappa_n}(s)-R^\kappa\widetilde{X}^{\kappa_n}(s)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}})_i ds\le\lim_{n\to\infty}\frac{1}{f(\kappa_n)}\int_0^T(\bar\Lambda^{\kappa_n}(s)-R^\kappa\widetilde{Y}^{\kappa_n}(s))_i ds.$$

Step 3: Using Lemma 12 (Appendix B.2) and the fact that $\int_0^T\mathbb{I}_{\{s\in I^\kappa\}}ds\to 0$ a.s. as $\kappa\to\infty$ which follows from Lemma 13 (Appendix B.2), we get a similar upper bound as in Step 3 of the proof of Lemma 5.

Finally repeating Step 4 as in the proof of Lemma 5 completes the proof. ■

B.2 Side Lemmas

The following "comparison lemma" is similar in flavor to the results in Whitt (1981).

Lemma 8 *Consider two systems each with m customer classes and arrival rates Λ^1 and Λ^2 satisfying $\Lambda^1(\cdot)\le\Lambda^2(\cdot)$; input-output matrices R^1 and R^2; routing matrices B^1 and B^2; and abandonment rate matrices Γ^1 and Γ^2. Suppose for the system with arrival rate Λ^ℓ, the admissible dynamic control X^ℓ is used, and let Z^ℓ be the associated headcount process for $\ell=1,2$. Assume that*

$$R^1X^1(t)+\Gamma^1[Z^1(t)-B^1X^1(t)]\ge R^2X^2(t)+\Gamma^2[Z^2(t)-B^1X^2(t)],\qquad(42)$$

whenever $Z^1(t)=Z^2(t)$. Then, there exists a common probability space on which both systems can be constructed, and if $Z^1(0)\le Z^2(0)$ then $Z^1(t)\le Z^2(t)$ for all time t with probability 1.

Proof. Fix an $i \in \{1,\ldots,m\}$. We now illustrate the construction for the i^{th} customer class. The system headcount process for the i^{th} customer class is given by

$$Z_i^\ell(t) = Z_i^\ell(0) + N_\ell^{(1)}\left(\int_0^t \Lambda_i^\ell(u)du\right) - N_\ell^{(2)}\left(\int_0^t \Big((R^\ell X^\ell(u))_i + \gamma_i^\ell(Z^\ell(u) - B^\ell X^\ell(u))_i\Big)\,du\right).$$

for $\ell = 1,2$ and all $t \in [0,T]$. Now consider the difference process

$$\begin{aligned} Z_i^1(t) - Z_i^2(t) &= Z_i^1(s) - Z_i^2(s) - N_{2,s}^{(1)}\left(\int_s^t \Lambda_i^2(u)du\right) \\ &\quad + N_{2,s}^{(2)}\left(\int_s^t \left((R^2X^2(u))_i + \gamma_i^2(Z^2(u) - B^2X^2(u))_i\right)du\right) \\ &\quad + N_{1,s}^{(1)}\left(\int_s^t \Lambda_i^1(u)du\right) - N_{1,s}^{(2)}\left(\int_s^t \left((R^1X^1(u))_i + \gamma_i^1(Z^1(u) - B^1X^1(u))_i\right)du\right) \end{aligned} \tag{43}$$

where

$$\begin{aligned} N_{\ell,s}^{(1)}(t) &:= N_\ell^{(1)}\left(t + \int_0^s \Lambda_i^\ell(u)du\right) - N_\ell^{(1)}\left(\int_0^s \Lambda_i^\ell(u)du\right) \\ N_{\ell,s}^{(2)}(t) &:= N_\ell^{(2)}\left(t + \int_0^t (R^\ell X^\ell(u))_i + \int_0^t \gamma_i^\ell(Z^\ell(u) - B^\ell X^\ell(u))_i du\right) \\ &- N_\ell^{(2)}\left(\int_0^t \Big(R^\ell X^\ell(u))_i + \gamma_i^\ell(Z^\ell(u) - B^\ell X^\ell(u))_i\Big)\,du\right) \end{aligned}$$

Note that $N_{\ell,s}^{(\mathrm{i})}(\cdot)$ is a unit rate Poisson process which drives the arrival process ($i=1$), or the departure process ($i=2$) due to a service completion or abandonment. To describe the dependence between the two systems we shall now explain the dependence between these Poisson processes. If $Z_i^1(s) = Z_i^2(s)$ then $N_{1,s}^{(r)}(t) = N_{2,s}^{(r)}(t) = N_s^{(r)}(t)$ for $r = 1,2$ and $t \in [0,\tau_*]$, where τ_* is defined as follows. Let

$$\begin{aligned} \tau_a &= \inf\left\{t \ge s:\ N_s^{(1)}\left(\int_s^t \Lambda_i^2(u)du\right) \ge 1\right\}, \\ \tau_s &= \inf\left\{t \ge s:\ N_s^{(2)}\left(\int_s^t (R^1X^1(u))_i + \gamma_i^1(Z_i^1(u) - B^1X^1(u))du\right) \ge 1\right\}. \end{aligned}$$

Let $\tau_1 = \min(\tau_a, \tau_b)$, then define

$$\tau_* = \begin{cases} \int_s^{t_1} \Lambda_i^2(u)du & \text{if } \tau_1 = \tau_a \\ \int_s^{t_1} \left(R^1X^1(u))_i + \gamma_i^1(Z^1(u) - B^1X^1(u)\right)_i du & \text{otherwise.} \end{cases}$$

Let $\tau_2 = \inf\{t > \tau_1 : Z_i^1(t) = Z_i^2(t)\}$, then $N_{\ell,t_1}^1(t)$ for $t \in [0, \int_{t_1}^{t_2} \Lambda_i^\ell(s)ds]$ and $\ell = 1,2$, and $N_{\ell,t_1}^2(t)$ for $t \in [0, \int_0^t \left((R^\ell X^\ell(u))_i + \gamma_i^\ell(Z^\ell(u) - B^\ell X^\ell(u))_i\right)du]$ and $\ell = 1,2$ are four mutually independent Poisson processes. From τ_2, onwards we repeat the above the construction. This completes the construction of the two systems.

We now verify that the above construction has the desired property. Let us define $\widetilde{\tau} = \inf\{t : Z_i^2(t) < Z_i^1(t)\}$, and assume $\mathbb{P}(\widetilde{\tau} < \infty) > 0$. Fix an $\omega \in \{\omega : \widetilde{\tau}(\omega) < \infty\}$. Since $Z_i^2(t)$ and $Z_i^1(t)$ are discrete and with probability one jump up or down by at most one unit at any time, we have, with probability one, that $Z_i^1(\widetilde{\tau}) - Z_i^2(\widetilde{\tau}) = 1$. Further, there exists $\delta > 0$ such that

$$\begin{aligned} &Z_i^2(t) = Z_i^1(t) = Z_i(t) \text{ for all } t \in [\widetilde{\tau} - \delta, \widetilde{\tau}] \\ &N^{(1)}_{\ell,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}-\epsilon} \Lambda_i^\ell(u)du\right) = 0, \\ &N^{(2)}_{1,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}-\epsilon} \Big((R^\ell X^\ell(u))_i + \gamma_i^\ell (Z^\ell(u) - B^\ell X^\ell(u))_i\Big)\right) = 0, \end{aligned}$$

for all $0 < \epsilon < \delta$ and $\ell = 1, 2$. Using (43), with $s = \widetilde{\tau} - \delta$ and $t = \widetilde{\tau}$ we also have

$$\begin{aligned} &N^{(1)}_{1,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}} \Lambda_i^1(u)du\right) - N^{(2)}_{1,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}} \big(R^1 X^1(u))_i + \gamma_i^1 (Z^1(u) - B^1 X^1(u)\big)_i\, du\right) \\ -&N^{(2)}_{1,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}} \Lambda_i^2(u)du\right) + N^{(2)}_{2,\widetilde{\tau}-\delta}\left(\int_{\widetilde{\tau}-\delta}^{\widetilde{\tau}} \big(R^2 X^2(u))_i + \gamma_i^2 (Z^2(u) - B^2 X^2(u)\big)_i\, du\right) = 1 \end{aligned}$$

Thus, we get the contradiction as either the first term or the last term on the left-hand-side is 1 and all other terms are zero. This follows since the construction of these processes along with (42) and the assumption $\Lambda_i^1(\cdot) \le \Lambda_i^2(\cdot)$ implies that if the first (fourth) term is positive, then the third (second) term must be negative. This completes the proof. ■

Definition 5 (minimal truncation with priority to the slower servers) *Assume without loss of generality that $\mu_j \le \mu_{j+1}$ for all $j = 1, \ldots, n-1$. We say that a sequence of admissible controls $\{\widetilde{X}^\kappa\}$ is a minimal truncation of the sequence of controls $\{X^\kappa\}$ with priority given to the slower servers if*

$$\widetilde{X}_j^\kappa(t) = \min\left\{X_j^\kappa(t), Z_{i(j)}^\kappa(t) - \sum_{j' \in S_j} \widetilde{X}_{j'}^\kappa(t)\right\}$$

for all $j = 1, \ldots, n$ and all $t \in [0, T]$, where $\{Z^\kappa\}$ is the sequence of the headcount processes associated with the sequence of admissible controls $\{\widetilde{X}^\kappa\}$, and $S_j = \{j' : j' < j,\ i(j') = i(j)\}$.

Lemma 9 *Consider two systems, labelled System 1 and System 2, which are identical except they have different arrival rates Λ^1 and Λ^2, respectively, which satisfy $\Lambda^1(\cdot) \le \Lambda^2(\cdot)$. Consider two controls X^1 and X^2 such that $0 \le X^1(t) - X^2(t) \le \delta e$, where $\delta > 0$ and e is vector of ones in $\mathbb{R}^n$. For the system with arrival rate $\Lambda^1(t)$, the dynamic control $\widetilde{X}^1$ which is a minimal truncation of $X^1(t)$ is used, and let Z^1 be the associated headcount process. For the system with arrival rate $\Lambda^2(t)$, the dynamic control $\widetilde{X}^2$ which is a minimal truncation of X^2 with priority given to the slower servers is used, and let Z^2 be the associated headcount process. Then there exists a common probability space on which both systems can be constructed, and if $Z^1(0) \le Z^2(0)$ then*

$$Z^1(t) - B\widetilde{X}^1(t) \le Z^2(t) - B\widetilde{X}^2(t) + \delta Be \text{ a.s.}$$

for all $t \in [0, T]$.

Proof. Consider another system referred to henceforth as System 3, which is identical to the system with arrival rate Λ^2 except it has one additional server pool which serves all the classes at rate zero and has $n\delta$ servers. Let Z^3 be the headcount of this system and $Z^3(0) = Z^2(0) + \delta Be$. Let the dynamic control used for this system be $X^3(t) = (\widetilde{X}^2(t), n\delta)$ where (x, y) represents a vector in $\mathbb{R}^{n+1}$ whose first n coordinates are the vector $x \in \mathbb{R}^n$, and the $(n+1)^{th}$ coordinate is $y \in \mathbb{R}$. Consider any probability space on which System 3 and an independent Poisson process are constructed. It is easy to verify that we can add System 2 to this probability space such that $Z^3(t) = Z^2(t) + \delta Be$ for all $t \in [0, T]$. Using Definitions 3 and 5 of minimal truncation, and minimal truncation with priority given to the slower servers, we have that if $Z^1(t) = Z^3(t)$ then $R\widetilde{X}^1(t) \leq R\widetilde{X}^2(t)$. Further, using the assumption $X^1(t) - X^2(t) \leq \delta e$, we have that the condition (42) is satisfied for System 1 and System 3. Appealing to Lemma 8, we can construct Systems 1, 2 and 3 on a a common probability space such that $Z^1(t) \leq Z^3(t) = Z^2(t) + \delta Be$ for all $t \in [0, T]$. Fix an $i \in \{1, \ldots, m\}$ and $t \in [0, T]$. Suppose $(Z^1)_i(t) - (B\widetilde{X}^1)_i(t) > 0$. Then, using the definition of minimal truncation we have $(B\widetilde{X}^1)_i(t) = (BX^1)_i(t)$; if $Z^1_i(t) - (B\widetilde{X}^1)_i(t) = 0$ the result holds trivially. Now, to complete the proof we consider two cases:
Case I: If $Z^2_i(t) - (B\widetilde{X}^2)_i(t) > 0$, then using the definition of minimal truncation we have $(B\widetilde{X}^2)_i(t) = (BX^2)_i(t)$. Using the fact that $0 \leq X^1(t) - X^2(t)$ we have the claimed result.
Case II: If $Z^2(t)_i - (B\widetilde{X}^2)_i(t) = 0$, then using the definition of minimal truncation, we have

$$Z^1(t)_i \leq (BX^2(t) + \delta Be)_i \leq (BX^1(t) + \delta Be)_i \text{ a.s.}$$

This completes the proof. ■

Lemma 10 *Let Λ^κ and X^κ, the arrival rate and (untruncated) control for the κ^{th} system, satisfy*

$$X^\kappa(t) = \frac{f(\kappa)}{\kappa}x, \;\; \Lambda^\kappa(t) = \frac{f(\kappa)}{\kappa}\lambda$$

for all $t \in [0, T]$ and all κ, where $x \in \mathbb{R}^m_+$ and $\lambda \in \mathbb{R}^m_+$. Let $Z^\kappa(0) := \kappa^{-1} f(\kappa) Me$ where $M \in \mathbb{R}_+$, and e is vector of ones in $\mathbb{R}^m$. If assumption (10) holds and $\widetilde{X}^\kappa(s)$ is the minimal truncation of X^κ with priority to the slower servers then

$$\frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(t) - R^\kappa \widetilde{X}^\kappa(t))_i dt \to T(\lambda - Rx)^+_i \quad a.s. \;\; as \; \kappa \to \infty$$

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

Proof. Fix an $i \in \{1, \ldots, m\}$. All subsequent probabilistic statements are to be interpreted in the almost sure sense and the term is omitted for brevity.

Step 1: By the properties of minimal truncation and the definition of the arrival rate and control

$$\int_0^T (\Lambda^\kappa(t) - R^\kappa \widetilde{X}^\kappa(t))_i dt \geq T(\lambda^\kappa - R^\kappa x^\kappa)_i \quad \text{a.s}$$

where $x^\kappa = xf(\kappa)/\kappa$ and $\lambda^\kappa = f(\kappa)\lambda$. Dividing by $f(\kappa)$ and taking $\liminf$ as $\kappa \to \infty$ we get

$$\liminf_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T (\Lambda^\kappa(t) - R^\kappa \widetilde{X}^\kappa(t))_i dt \geq T(\lambda - Rx)_i \quad \text{a.s.}$$

Step 2: Let activities $1, \ldots, \ell$ serve customer class i such that $\mu_1 \leq \ldots \leq \mu_\ell$, (relabelling activities if necessary). Thus, the minimal truncation for the κ^{th} system is given by

$$\begin{aligned} \widetilde{X}_1^\kappa(t) &= \min\{X_1^\kappa(t), Z_i^\kappa(t)\} \\ \widetilde{X}_j^\kappa(t) &= \min\left\{ X_j^\kappa(t), Z_i^\kappa(t) - \sum_{j'=1}^{j-1} \widetilde{X}_{j'}^\kappa(t) \right\} \end{aligned}$$

for all $j = 1, \ldots, \ell$ and $t \in [0, T]$, where Z_i^κ is the headcount process of customer class i associated with the admissible control $\widetilde{X}^\kappa$. Put $V_i^\kappa = \sum_{j=1}^{\ell} x_j^\kappa$.

Step 3: Since $f(\kappa)$ grows superlinearly there exists κ_1 such that for all $\kappa > \kappa_1$, $(R^\kappa x^\kappa)_i > \gamma_i^\kappa$. Fix a $\kappa > \kappa_1$. Consider a system with V_i^κ servers such that all servers except one process work at rate zero, and one server has service rate $(R^\kappa x^\kappa)_i$. The arrival rate and abandonment rate for the system is that corresponding to customer class i. Suppose the system gives priority to the servers with zero service rate and let the headcount in this system be denoted by $\overline{Z}_i^\kappa(t)$. If the original headcount in this system is $\max\{Mf(\kappa)/\kappa, V_i^\kappa\}$, then using Lemma 8 we have $Z_i^\kappa(t) \leq \overline{Z}_i^\kappa(t)$ for all $t \in [0, T]$. Note that the dynamics of $\overline{Z}_i$ are closely related to an M/M/1 queue with abandonments. Specifically, consider an M/M/1 queue with arrival rate λ_i^κ, service rate $(R^\kappa x^\kappa)_i$, abandonment rate γ_i^κ, and let the headcount at time t be denoted by $Y_i^\kappa(t)$. Then, if $Y_i^\kappa(0) = \overline{Z}^\kappa(0) - V_i^\kappa + 1$ and $\overline{Z}^\kappa(0) \geq V_i^\kappa - 1$, it follows that $Y_i^\kappa(t) = \overline{Z}^\kappa(t) - V_i^\kappa + 1$ for all $t \in [0, T]$. Thus, we have

$$Z_i^\kappa(t) - V_i^\kappa \leq Y_i^\kappa(t) \quad \text{a.s.}$$

where $Y_i^\kappa(t)$ is the headcount at time $t \in [0, T]$ of the M/M/1 queue defined above with the initial state $Y_i^\kappa(0) = \max\{Mf(\kappa)/\kappa, V_i^\kappa\} - V_i^\kappa + 1$.

Step 4: Fix an $\epsilon > 0$. Consider another system (referred to henceforth as "System 2") consisting of two buffers labelled Buffer I and II. Buffer I has an arrival rate equal to $\min\{\lambda_i^\kappa, (R^\kappa x^\kappa)_i(1-\epsilon)\}$ and has one server working at rate $(R^\kappa x^\kappa)_i$. There are no abandonments from this buffer. Let the traffic intensity be denoted by $\eta = \min\{\lambda_i/(Rx)_i, 1-\epsilon\} < 1$, and note η is independent of κ. Buffer II is a "no server" system with arrival rate $(\lambda^\kappa - R^\kappa x^\kappa(1-\epsilon))_i^+$, no servers, and abandonment rate γ_i^κ. Let the headcount of this system at time t be represented by the vector $(Y_i^{\kappa,\mathrm{I}}(t), Y_i^{\kappa,\mathrm{II}}(t))$ where $Y_i^{\kappa,\mathrm{I}}(t)$ and $Y_i^{\kappa,\mathrm{II}}(t))$ is to be interpreted as the headcount (including the customers in service) in

Buffer I and Buffer II respectively. Using an argument similar to that used in the proof of Lemma 8 together with the condition $(R^\kappa x^\kappa)_i > \gamma_i^\kappa$, we have that if $Y_i^{\kappa,\mathrm{I}}(0) + Y_i^{\kappa,\mathrm{II}}(0) \geq Y_i^\kappa(0)$, then $Y_i^{\kappa,\mathrm{I}}(t) + Y_i^{\kappa,\mathrm{II}}(t) \geq Y_i^\kappa(t)$ a.s., where $Y_i^\kappa(t)$ is the headcount of the M/M/1 system defined in Step 3. Let $(Y_i^{\kappa,\mathrm{I}}(0), Y_i^{\kappa,\mathrm{II}}(0)) = (Y_{i,\pi}^{\kappa,\mathrm{I}}, (Mf(\kappa)/\kappa - V_i^\kappa - Y_{i,\pi}^{\kappa,\mathrm{I}})^+ + 1)$, where $Y_{i,\pi}^{\kappa,\mathrm{I}}$ is a random variable drawn from the stationary distribution of the headcount in Buffer I. So, for all $\kappa > \kappa_1$, we have

$$(Z_i^\kappa(t) - V_i^\kappa)^+ \leq Y_i^{\kappa,\mathrm{I}}(t) + Y_i^{\kappa,\mathrm{II}}(t) \ \text{ a.s.},$$

for all $t \in [0, T]$. Consequently,

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa) dt \leq \limsup_{\kappa\to\infty} \left[\frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{I}}(t) dt + \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(t) dt\right]. \quad (44)$$

We shall now bound both terms on right-hand-side of (44).

Step 5: First note that the headcount $Y_i^{\kappa,\mathrm{I}}(t)$ for any $t \in [0, T]$ has the stationary distribution of an M/M/1 queue with traffic intensity η. Choose any $\theta \in (0, \frac{-\log \eta}{T})$, then we have

$$\begin{aligned} \mathbb{E}\left[\exp\left(\theta \int_0^T Y_i^{\kappa,\mathrm{I}}(t) dt\right)\right] &\overset{(a)}{\leq} \mathbb{E}\left[\int_0^T \frac{\exp\left(\theta T Y_i^{\kappa,\mathrm{I}}(t)\right) dt}{T}\right] \\ &\overset{(b)}{=} \int_0^T \frac{\mathbb{E}\left[\exp\left(\theta T Y_i^{\kappa,\mathrm{I}}(0)\right)\right]}{T} dt \\ &\overset{(c)}{=} \frac{e^{\theta T}(1-\eta)}{T(1-\eta e^{\theta T})} =: L. \end{aligned}$$

The inequality (a) follows from Jensen's inequality on Lebesgue measure; the equality (b) follows from Fubini's theorem and stationarity; and equality (c) follows from the stationary distribution of the headcount process. Given any $\epsilon > 0$, using a Chernoff bound we get that

$$\mathbb{P}\left(\frac{\int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{I}}(t)}{f(\kappa)} > \epsilon\right) \leq L e^{-\epsilon \frac{f(\kappa)}{\kappa \gamma_i}}$$

Under assumption (10), we have that the right-hand-side of the above is summable. Thus, by Borel-Cantelli we have that the first term on right-hand-side of (44) is zero.

Step 6: The dynamics at Buffer II at System 2 are given by

$$Y_i^{\kappa,\mathrm{II}}(t) = Y_i^{\kappa,\mathrm{II}}(0) + \widetilde{N}_i^{(1),\kappa}\left(((\lambda^\kappa - R^\kappa x^\kappa (1-\epsilon))_i)^+ t\right) - \widetilde{N}_i^{(3),\kappa}\left(\int_0^t \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(s) ds\right) \quad (45)$$

for all time $t \in [0, T]$, where $\widetilde{N}_i^{(1),\kappa}$ and $\widetilde{N}_i^{(3),\kappa}$ are independent unit rate Poisson processes associated with arrivals and abandonments at Buffer II. (Note these Poisson processes are indexed by κ as we use Lemma 8 to construct System 2 for each κ separately.) To prove the result we need the following side lemma whose proof is deferred to the end of the current proof.

Lemma 11 *Consider Buffer II described above. If assumption (10) holds, then for all* $0 \le \epsilon \le 1$

$$\text{(i)} \frac{1}{f(\kappa)} \left| \widetilde{N}_i^{(3),\kappa} \left(\int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt \right) - \int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt \right| \to 0 \quad a.s. \ as\ \kappa \to \infty,$$

$$\text{(ii)} \frac{\widetilde{N}_i^{(1),\kappa} \left(\lambda_i^\kappa - (R^\kappa x^\kappa)_i (1-\epsilon))^+ T \right)}{f(\kappa)} \to (\lambda_i - (Rx)_i(1-\epsilon))^+ T \quad a.s. as\ \kappa \to \infty,$$

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

Appealing to part (i) of the lemma above, we have

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \widetilde{N}^{(3),\kappa} \left(\int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt \right) = \limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt.$$

Using (45) we further have that

$$\begin{aligned}\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \widetilde{N}^{(3),\kappa} \left(\int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt \right) &\le \limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \left(Y_i^{\kappa,\text{II}}(0) + \widetilde{N}^{(1),\kappa} \left((\lambda^\kappa - R^\kappa x^\kappa (1-\epsilon))_i^+ T \right) \right) \\ &= (\lambda - Rx(1-\epsilon))_i^+ T,\end{aligned}$$

where the last equality follows from Lemma 11(ii). Plugging back in to (44) we have

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ dt \le (\lambda - Rx(1-\epsilon))_i^+ T.$$

Since the above holds for all $\epsilon > 0$, we get

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ dt \le (\lambda - Rx)_i^+ T.$$

The desired result now follows from a result similar to Proposition 1 which can proved in an analogus manner as Lemma 11. This completes the proof. ■

Proof of Lemma 11. Proof of (i): Fix $i \in \{1, \dots, m\}$, and let $\upsilon_i^\kappa = \int_0^T \gamma_i^\kappa Y_i^{\kappa,\text{II}}(t)dt$ and $\psi_i = (\lambda - Rx(1-\epsilon))_i^+$ for all κ. Using the system dynamics in (45) we have

$$\begin{aligned}\upsilon_i^\kappa &\le \gamma_i^\kappa T \left(Y_i^{\kappa,\text{II}}(0) + \widetilde{N}^{(1),\kappa}(f(\kappa)\psi_i T) \right) \\ &\le \gamma_i^\kappa T \left(\frac{M f(\kappa)}{\kappa} + \widetilde{N}^{(1),\kappa}(f(\kappa)\psi_i T) \right) =: \overline{\upsilon}_i^\kappa,\end{aligned}$$

where M is the scaled headcount of class i customers at time 0. Note that the bound $\overline{\upsilon}^\kappa$ is independent of the poisson process $\widetilde{N}^{(3),\kappa}$ (this is crucial in what follows). We now use the following strong approximation result from Kurtz (1978) which follows directly from Komlós, Major and Tusnady (1975).

Proposition 4 (Kurtz (1978), Lemma 3.1) *A standard (rate one) Poisson process* $(N(t) : t \ge 0)$ *can be realized on the same probability space as a standard Brownian motion* $(W(t) : t \ge 0)$ *in such a way that the positive random variable* ξ *given by*

$$\xi := \sup_{t>0} \frac{|N(t) - t - W(t)|}{\log(2 \vee t)} < \infty$$

has a finite moment generating function in a neighborhood of the origin, where $a \vee b := max(a, b)$.

Using the above proposition, there exist Brownian motions W_i^κ and random variables ξ_i^κ such that

$$\begin{aligned} |\widetilde{N}_i^{(3),\kappa}(\upsilon_i^\kappa) - \upsilon_i^\kappa| &\leq \xi_i^\kappa \log(2 \vee \upsilon_i^\kappa) + |W_i^\kappa(\upsilon^\kappa)| \\ &\leq \xi_i^\kappa \log(2 \vee \overline{\upsilon}_i^\kappa) + \sup_{0 \leq s \leq \overline{\upsilon}_i^\kappa} |W_i^\kappa(s)|. \end{aligned} \tag{46}$$

To complete the proof, we shall show that as $\kappa \to \infty$, the right-hand-side scaled by $f(\kappa)$ goes to zero. We define

$$U_i^\kappa := \mathbb{E}[\overline{\upsilon}_i^\kappa] = f(\kappa)MT\gamma_i + \kappa\gamma_i T^2 f(\kappa)\psi_i \tag{47}$$

$$V_i^\kappa := Var[\overline{\upsilon}_i^\kappa] = \kappa^2\gamma_i^2 T^3 f(\kappa)\psi_i. \tag{48}$$

Now, to bound the first term note that

$$\log(2 \vee \overline{\upsilon}_i^\kappa) \leq \log(2 + \overline{\upsilon}_i^\kappa) = \log U_i^\kappa + \log(1 + \frac{2 + \overline{\upsilon}_i^\kappa - U_i^\kappa}{U_i^\kappa}) \leq \log U_i^\kappa + \frac{2 + \overline{\upsilon}_i^\kappa - U_i^\kappa}{U_i^\kappa}.$$

Squaring both sides and taking expectations we have

$$\mathbb{E}[\log^2(2 \vee \overline{\upsilon}_i^\kappa)] \leq (\log U_i^\kappa)^2 + \frac{4}{(U_i^\kappa)^2} + \frac{V_i^\kappa}{(U_i^\kappa)^2} + \frac{4\log U_i^\kappa}{U_i^\kappa}.$$

Now using the assumption (10) and the bound stated above we have

$$\sum_{\kappa=1}^{\infty} \frac{\mathbb{E}[\log^2(2 \vee \overline{\upsilon}_i^\kappa)]}{f(\kappa)^2} < \infty.$$

For any $\delta > 0$, using a Chebechev bound we have

$$\mathbb{P}(\xi_i^\kappa \log(2 \vee \overline{\upsilon}_i^\kappa) \geq f(\kappa)\delta) \leq \frac{\mathbb{E}[(\xi_i^\kappa)^2 \log^2(2 \vee \overline{\upsilon}_i^\kappa)]}{\delta^2 f(\kappa)^2} = \mathbb{E}[\xi_i^\kappa]^2 \frac{\mathbb{E}[\log^2(2 \vee \overline{\upsilon}_i^\kappa)]}{\delta^2 f(\kappa)^2}.$$

The above equality follows from the independence of $\overline{\upsilon}_i^\kappa$ and ξ_i^κ which in turn follows from independence of $\overline{\upsilon}_i^\kappa$ and $\widetilde{N}^{(3),\kappa}$ and the fact that ξ_i^κ depends only on the process $\widetilde{N}^{(3),\kappa}$. Since the random variables $\{\xi_i^\kappa\}$ are identically distributed, and their moment generating function exists in the neighborhood of origin (by Proposition 4), we have $\sup_\kappa E[(\xi_i^\kappa)^2] < \infty$. Thus, we have

$$\sum_{\kappa=1}^{\infty} \mathbb{P}(\xi_i^\kappa \log(2 \vee \overline{\upsilon}_i^\kappa) \geq f(\kappa)\delta) < \infty \text{ a.s.,}$$

and using Borel-Cantelli we get

$$\frac{\xi_i^\kappa \log(2 \vee \overline{\upsilon}_i^\kappa)}{f(\kappa)} \to 0 \text{ a.s. as } \kappa \to \infty.$$

For the second term on the right-hand-side of (46), fix a $\delta > 0$, and consider the following sequence of inequalities

$$\begin{aligned} \mathbb{P}\left(\sup_{0 \leq s \leq \overline{\upsilon}_i^\kappa} |W_i^\kappa(s)| \geq f(\kappa)\delta\right) &\leq 2\mathbb{P}\left(\sup_{0 \leq s \leq \overline{\upsilon}_i^\kappa} W_i^\kappa(s) \geq f(\kappa)\delta\right) \\ &= 4\mathbb{P}(W_i^\kappa(\overline{\upsilon}_i^\kappa \geq f(\kappa)\delta) \\ &\leq 4\exp(-\theta_i^\kappa f(\kappa)\delta)\mathbb{E}[\exp(\theta_i^\kappa W_i^\kappa(\overline{\upsilon}_i^\kappa))], \end{aligned}$$

where the last inequality follows from a Chernoff bound for any $\theta_i^\kappa > 0$. Next, using the moment generating function of a Gaussian and Poisson r.v. we have that

$$4\exp(-\theta_i^\kappa f(\kappa)\delta)\mathbb{E}[\exp(\theta_i^\kappa W_i^\kappa(\overline{v}_i^\kappa))] = 4\exp\left(f(\kappa)\left[-\theta_i^\kappa\delta + (\theta_i^\kappa)^2\gamma_i MT + (e^{\frac{\gamma_i\kappa T(\theta_i^\kappa)^2}{2}} - 1)\psi_i T\right]\right).$$

Define $C_1 := (\theta_i^\kappa)^2\gamma_i MT$ and $C_2 = 0.9\gamma_i\kappa T^2\psi_i$. Choosing $\theta_i^\kappa = \frac{\delta}{2(C_1+\kappa C_2)}$ and using the fact $e^x \leq 1 + 1.8x$ for $x \in (0,1)$ we have

$$4\exp\left(f(\kappa)\left[-\theta_i^\kappa\delta + (\theta_i^\kappa)^2\gamma_i MT + (e^{\frac{\gamma_i\kappa T(\theta_i^\kappa)^2}{2}} - 1)\psi_i T\right]\right) \leq 4\exp\left(\frac{-C_3 f(\kappa)\delta^2}{\kappa}\right),$$

for κ large, where C_3 is a constant. Using the assumption (10) we have

$$\sum_{\kappa=1}^{\infty}\mathbb{P}\left(\sup_{0\leq s\leq\overline{v}_i^\kappa}|W_i^\kappa(s)| \geq f(\kappa)\delta\right) < \infty.$$

Hence, using Borel-Cantelli lemma we have

$$\sup_{0\leq s\leq\overline{v}_i^\kappa}\frac{|W_i^\kappa(s)|}{f(\kappa)} \to 0 \quad \text{a.s. as} \quad \kappa\to\infty$$

This completes the proof of (i).

Proof of (ii): Using an argument similar to that used in part (i), we have

$$\limsup_{\kappa\to\infty}\frac{1}{f(\kappa)}\left|\widetilde{N}^{(1),\kappa}\left((\lambda_i^\kappa - (R^\kappa x^\kappa)_i(1-\epsilon))^+T\right) - (\lambda_i^\kappa - (R^\kappa x^\kappa)_i(1-\epsilon))^+T\right| = 0 \quad \text{a.s.}$$

from which the result follows immediately using the scaling of λ^κ, R^κ and x^κ. This completes the proof. ■

Lemma 12 *Let Λ^κ and X^κ, the arrival rate and (untruncated) control for the κ^{th} system, satisfy*

$$X^\kappa(t) = \frac{f(\kappa)}{\kappa}x, \;\; \Lambda^\kappa(t) = \frac{f(\kappa)}{\kappa}\lambda$$

for all $t \in [0,T]$ and all κ, where $x \in \mathbb{R}_+^m$ and $\lambda \in \mathbb{R}_+^m$. Let $Z^\kappa(0) := \kappa^{-1}f(\kappa)Me$ where $M \in \mathbb{R}_+$, and e is vector of ones in $\mathbb{R}^m$. Let $\widetilde{X}^\kappa(t)$ be the minimal truncation with priority to the slower servers, of $X^\kappa(t)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}$. If assumption (10) holds and $\int_0^T \mathbb{I}_{\{s\in I^\kappa\}}ds \to 0$ almost surely as $\kappa\to\infty$, then

$$\frac{1}{f(\kappa)}\int_0^T (\Lambda^\kappa(s) - R^\kappa\widetilde{X}^\kappa(s))_i ds \to T(\lambda - Rx)_i^+ \quad \text{a.s. as } \kappa\to\infty$$

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

Proof. The proof mimics that of Lemma 10 and follow in a similar stepwise fashion; below we indicate the necessary modifications. Fix an $i \in \{1,\ldots,m\}$.

Step 1: Using same argument as in Step 1 of the proof of Lemma 10 the inequality for lim inf is established.

Step 2: We construct the dynamic control as in Step 2 of the proof of Lemma 10 with the modification that $X^\kappa(t)$ is a minimal truncation with priority to the slower servers of $X^\kappa(t)\mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}$.

Step 3: We reduce the system to an M/M/1 queuing system as in Step 3 of the proof of Lemma 10 with the following modification: the server at the M/M/1 queue can only serve at times $t \in [0,T]\setminus I^\kappa$.

Step 4: Fix an $\epsilon > 0$. Consider another system (referred to henceforth as "System 2") consisting of two buffers labelled Buffer I and II. Buffer I has an arrival process which is a doubly stochastic Poisson process with rate given by

$$\Lambda_i^{\kappa,\mathrm{I}}(t) = \begin{cases} \min\{\lambda_i^\kappa, (R^\kappa x^\kappa)_i(1-\epsilon)\} & \text{if } t \in [0,T] \setminus I^\kappa \\ 0 & \text{if } t \in I^\kappa, \end{cases} \tag{49}$$

It has one server that servers only at $t \in [0,T] \setminus I^\kappa$ with service rate $(R^\kappa x^\kappa)_i$, and there are no abandonments from this buffer. Note that for all $t \in I^\kappa$ dynamics at Buffer I "freezes" and the headcount stays the same. For further use, let the traffic intensity for this buffer when the server is working be $\eta = \min\{\lambda_i/(Rx)_i, 1-\epsilon\} < 1$, and note that η is independent of κ. As in Lemma 10, Buffer II is a "no service" system with arrival rate given by

$$\Lambda_i^{\kappa,\mathrm{II}}(t) = \begin{cases} (\lambda^\kappa - R^\kappa x^\kappa(1-\epsilon))_i^+ & \text{if } t \in [0,T] \setminus I^\kappa \\ \lambda_i^\kappa & \text{if } t \in I^\kappa. \end{cases} \tag{50}$$

and abandonment rate γ_i^κ. Let the headcount of this system at time t be $(Y_i^{\kappa,\mathrm{I}}(t), Y_i^{\kappa,\mathrm{II}}(t))$ where $Y_i^{\kappa,\mathrm{I}}(t)$ and $Y_i^{\kappa,\mathrm{II}}(t)$ represent the headcount (including the customer-in-service) in Buffer I and Buffer II respectively. Using a similar argument to that used in the proof of Lemma 8 together with the condition $(R^\kappa x^\kappa)_i > \gamma_i^\kappa$, we have that if $Y_i^{\kappa,\mathrm{I}}(0)+Y_i^{\kappa,\mathrm{II}}(0) \geq Y_i^\kappa(0)$ then $Y_i^{\kappa,\mathrm{I}}(t)+Y_i^{\kappa,\mathrm{II}}(t) \geq Y_i^\kappa(t)$ a.s., where $Y_i^\kappa(t)$ is the headcount of the M/M/1 system defined in Step 3. Let $(Y_i^{\kappa,\mathrm{I}}(0), Y_i^{\kappa,\mathrm{II}}(0)) = (Y_{i,\pi}^{\kappa,\mathrm{I}}, (Mf(\kappa)/\kappa - V_i^\kappa - Y_{i,\pi}^{\kappa,\mathrm{I}})^+ + 1)$, where $Y_{i,\pi}^{\kappa,\mathrm{I}}$ is random variable drawn from stationary distribution of the headcount in an M/M/1 queue with traffic intensity η. So, we have

$$(Z_i^\kappa(t) - V_i^\kappa)^+ \leq Y_i^{\kappa,\mathrm{I}}(t) + Y_i^{\kappa,\mathrm{II}}(t) \quad \text{a.s.}$$

for all $t \in [0,T]$. Since above holds for all $\kappa > \kappa_1$ we have that (almost surely)

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ \mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}} dt \leq \limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Y_i^{\kappa,\mathrm{I}}(t) + Y_i^{\kappa,\mathrm{II}}(t)) dt \tag{51}$$

Step 5: We use the fact that the headcount process at Buffer I, $Y_i^{\kappa,\mathrm{I}}(t)$ has stationary distribution of an M/M/1 queue with traffic intensity given by η, and then follow exactly Step 5 in the proof of Lemma 10.

Step 6: The dynamics at Buffer II at System 2 are given by

$$Y_i^{\kappa,\mathrm{II}}(t) \;=\; Y_i^{\kappa,\mathrm{II}}(0) + \widetilde{N}_i^{(1),\kappa}\left(\int_0^t \Lambda_i^{\kappa,\mathrm{II}}(s)ds\right) - \widetilde{N}_i^{(3),\kappa}\left(\int_0^t \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(s)ds\right) \tag{52}$$

for all time $t \in [0,T]$, where $\widetilde{N}_i^{(1),\kappa}$ and $\widetilde{N}_i^{(3),\kappa}$ are independent unit Poisson process associated with arrivals and abandonments at Buffer II. Essentially repeating the proof of Lemma 11(i), we have

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(t)dt = \limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \widetilde{N}_i^{(3),\kappa}\left(\int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(t)dt\right).$$

Further, using the dynamics (52) at Buffer II we have

$$\begin{aligned}
\limsup_{\kappa\to\infty} \frac{\widetilde{N}_i^{(3),\kappa}\left(\int_0^T \gamma_i^\kappa Y_i^{\kappa,\mathrm{II}}(t)dt\right)}{f(\kappa)} &\le \limsup_{\kappa\to\infty} \frac{\left(Y_i^{\kappa,\mathrm{II}}(0) + \widetilde{N}_i^{(1),\kappa}\left(\int_0^T \Lambda_i^{\kappa,\mathrm{II}}(t)dt\right)\right)}{f(\kappa)} \\
&\overset{\text{(a)}}{\le} \limsup_{\kappa\to\infty} \frac{\left(Y_i^{\kappa,\mathrm{II}}(0) + \widetilde{N}_i^{(1),\kappa}\left((\lambda_i^\kappa - R^\kappa x^\kappa(1-\epsilon))_i^+(1-\epsilon)T + \lambda_i^\kappa \epsilon T\right)\right)}{f(\kappa)} \\
&\overset{\text{(b)}}{=} (\lambda - Rx(1-\epsilon))_i^+(1-\epsilon)T + \lambda_i \epsilon T \quad \text{a.s.}
\end{aligned}$$

Here we have used the assumption that $\int_0^T \mathbb{I}_{\{s\in I^\kappa\}}ds \to 0$ almost surely as $\kappa\to\infty$, and for each $\epsilon > 0$ there exist κ_2 such that

$$\int_0^T \Lambda_i^{\kappa,\mathrm{II}}(t)\mathbb{I}_{\{t\in I^\kappa\}}dt \le \lambda_i^\kappa \epsilon,$$

for all $\kappa > \kappa_2$. Thus inequality (a) follows. The equality (b) in a similar manner to that of to Lemma 11(ii). Thus, we have

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ \mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}dt \le (\lambda - Rx(1-\epsilon))_i^+(1-\epsilon)T + \lambda_i\epsilon T \quad \text{a.s.}$$

Since the above holds for all $\epsilon > 0$, we get

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ \mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}dt \le (\lambda - Rx)_i^+ T \quad \text{a.s.}$$

Using Lemma 4 and the assumption that $\int_0^T \mathbb{I}_{\{s\in I^\kappa\}}ds \to 0$ a.s. as $\kappa\to\infty$ we also have (almost surely)

$$\limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa)^+ \mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}}dt = \limsup_{\kappa\to\infty} \frac{1}{f(\kappa)} \int_0^T \gamma_i^\kappa (Z_i^\kappa(t) - V_i^\kappa \mathbb{I}_{\{t\in[0,T]\setminus I^\kappa\}})^+ dt.$$

The desired result now follows as in the proof of Lemma 10. This completes the proof. ■

Lemma 13 *If $f(\kappa)$ and $g(\kappa)$ satisfy*

$$\sum_{\kappa=1}^\infty \frac{g(\kappa)f(\kappa)}{\kappa} \exp\left(-\frac{\epsilon\kappa}{g(\kappa)}\right) < \infty \tag{53}$$

for all $\epsilon > 0$, then $\sum_{\ell=1}^{g(\kappa)T} \tau_\ell^\kappa \to 0$ almost surely, as $\kappa\to\infty$.

Proof. Let μ be defined as $\mu = \min_{1\le j\le n}\{\mu_j\}$. Let the total servers in the unscaled system be $\overline{b}$, *i.e.*, $\overline{b} = \sum_{k=1}^r b_k$. Note that τ_ℓ^κ is the maximum residual service time among all non-idle servers at the start of the ℓ^{th} review period. As the service time is assumed to be exponential, the residual service times are also exponential. Since $\tau_\ell^\kappa > 0$ it is clear that

$$\liminf_{\kappa\to\infty} \sum_{\ell=1}^{g(\kappa)T} \tau_\ell^\kappa \geq 0 \quad \text{a.s.}$$

Let $\widetilde{\tau}_\ell^\kappa$ be the maximum of the service times assuming all the servers are working at the minimum service rate μ. We construct $\widetilde{\tau}_\ell^\kappa$ on the same probability space in the following manner, for all the non-idle servers we scale the service time by $\mu_j/\mu \geq 1$ and for all idle servers we generate exponential service times with rate μ. Thus, it follows that $\widetilde{\tau}_\ell^\kappa \geq \tau_\ell^\kappa$ a.s. Now, fix $\epsilon > 0$ and note that

$$\begin{aligned} \mathbb{P}\left(\sum_{\ell=1}^{g(\kappa)T} \widetilde{\tau}_\ell^\kappa > \epsilon\right) &\leq g(\kappa)T\mathbb{P}\left(\widetilde{\tau}_\ell^\kappa > \frac{\epsilon}{g(\kappa)}\right) \\ &\overset{\text{(a)}}{\leq} g(\kappa)T\frac{f(\kappa)\bar{b}}{\kappa}\mathbb{P}\left(EXP(\mu\kappa) > \frac{\epsilon}{g(\kappa)}\right) \\ &= \frac{g(\kappa)f(\kappa)\bar{b}T}{\kappa}\exp\left(-\frac{\epsilon\mu\kappa}{g(\kappa)}\right), \end{aligned} \tag{54}$$

where $EXP(\mu\kappa)$ represents a r.v. with exponential distribution and rate $\mu\kappa$ and the inequality in (a) follows from the union bound. Since the right-hand-side of (54) is summable by the growth condition (53), using the Borel-Cantelli lemma we have

$$\mathbb{P}\left(\sum_{\ell=1}^{g(\kappa)T} \widetilde{\tau}_\ell^\kappa > \epsilon, \quad \text{i.o.}\right) = 0.$$

Since, the above equation holds for any $\epsilon > 0$, it follows that

$$\limsup_{\kappa\to\infty} \sum_{\ell=1}^{g(\kappa)T} \tau_\ell^\kappa \leq \lim_{\kappa\to\infty} \sum_{l=1}^{g(\kappa)T} \widetilde{\tau}_\ell^\kappa = 0 \quad \text{a.s.}$$

This completes the proof. ■

References

Armony, M. (2005), 'Dynamic routing in large-scale service systems with heterogeneous servers'. Working paper.

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

Armony, M. and Maglaras, C. (2004), 'On customer contact centers with a call-back option: customer decisions, routing rules, and system design', *Operations Research* **52**, 271–292.

Armony, M. and Mandelbaum, A. (2005), 'Staffing of large service systems: The case of a single customer class and multiple server types'. Working paper.

Atar, R., Mandelbaum, A. and Reiman, M. (2004), 'Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy-traffic', *The Annals of Applied Probability* **14**, 1084–1134.

Aubin, J.-P. and Frankowska, H. (1990), *Set-Valued Analysis*, Birkhäuser, Boston.

Bassamboo, A., Harrison, J. M. and Zeevi, A. (2005), 'Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits', *Queueing Systems, Theory and Applications (QUESTA)* . To appear.

Bell, S. and Williams, R. (2001), 'Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: asymptotic optimality of a continuous review threshold policy', *The Annals of Applied Probability* **11**, 608–649.

Billingsley, P. (1999), *Convergence of Probability Measures*, John Wiley and Sons Inc., New York.

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

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005), 'Statistical analysis of a telephone call center: a queueing science perspective', *JASA* . To appear.

Chen, B. and Henderson, S. (2001), 'Two issues in setting call centre staffing levels', *Annals of Operations Research* **108**, 175–192.

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

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

Gans, N. and van Ryzin, G. (1997), 'Optimal control of a multi-class, flexible queueing system', *Operations Research* **45**, 677–693.

Gans, N. and Zhou, Y. (2003), 'A call-routing problem with service-level constraints.', *Operations Research* **51**, 255–271.

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

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

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

Harrison, J. M. (1998), 'Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policy', *The Annals of Applied Probability* **8**, 822–848.

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

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

Harrison, J. M. and Zeevi, A. (2005), 'A method for staffing large call centers based on stochastic fluid models', *Manufacturing & Service Operations Management* . To appear.

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

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

Kurtz, T. G. (1978), 'Strong approximation theorems for density dependent Markov chains', *Stochastic Processes and Their Applications* **6**, 223–240.

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

Mandelbaum, A. and Stolyar, A. L. (2004), 'Scheduling flexible servers with convex delay costs: heavy traffic optimality of the generalized $c\mu$-rule', *Operations Research* . To appear.

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

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

Taylor, H. and Karlin, S. (1998), *An Introduction to Stochastic Modeling*, Academic Press, San Diego.

Wallace, R. B. and Whitt, W. (2005), 'A staffing algorithm for call centers with skill-based routing'. Working paper.

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

Whitt, W. (1991), 'The pointwise stationary approximation for $M_t/M_t/s$ queues is asy mptotically correct as the rates increase', *Management Science* **37**, 307–314.

Whitt, W. (1992), 'Understanding the efficiency of multi-server systems', *Management Science* **38**, 708–723.

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

Whitt, W. (2005), 'Fluid models for many-server queues with abandonments', *Operations Research* . To appear.

Yahalom, T. and Mandelbaum, A. (2005), 'Optimal control of queueing systems with multi-class customers and multiple servers: V- and N-designs'. Working paper.