A service system with on-demand agent invitations

Guodong Pang
Harold and Inge Marcus Department
of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802
gup3@psu.edu

Alexander L. Stolyar
Department of Industrial
and Systems Engineering
Lehigh University
Bethlehem, PA 18015
stolyar@lehigh.edu

September 25, 2014

Abstract

We consider a service system where agents are invited on-demand. Customers arrive exogenously as a Poisson process and join a customer queue upon arrival if no agent is available. Agents decide to accept or decline invitations after some exponentially distributed random time, and join an agent queue upon invitation acceptance if no customer is waiting. A customer and an agent are matched in the order of customer arrival and agent invitation acceptance under the non-idling condition, and will leave the system simultaneously once matched (service times are irrelevant here).

We consider a feedback-based adaptive agent invitation scheme, which controls the number of pending agent invitations, depending on the customer and/or agent queue lengths and their changes. The system process has two components – 'the difference between agent and customer queues' and 'the number of pending invitations', and is a countable continuous-time Markov chain.

For the case when the customer arrival rate is constant, we establish fluid and diffusion limits, in the asymptotic regime where the customer arrival rate goes to infinity, while the agent response rate is fixed. We prove the process stability and fluid-scale limit interchange, which in particular imply that that both customer and agent waiting times in steady-state vanish in the asymptotic limit. To do this we develop a novel (multi-scale) Lyapunov drift argument; it is required because the process has non-trivial behavior on the state space boundary. When the customer arrival rate is time-varying, we present a fluid limit for the processes in the same asymptotic regime. Simulation experiments are conducted to show good performance of the invitation scheme and accuracy of fluid limit approximations.

Keywords: service systems, call centers, special agents, knowledge workers, on-demand agent invitation, fluid limit, diffusion limit, stability, interchange of limits

1 Introduction

The model in this paper is primarily motivated by applications to call/contact centers, where customer requests (arriving as calls, chat sessions, etc.), are answered and processed by agents. (The model has broader potential applications, for example, to inventory models as discussed later.) In such systems, some of the callers/customers need to be served not by regular agents, who are always available (either working on a call or standing by) during their work hours, but different agents who are highly skilled and/or have knowledge that typical agents may not possess. These agents with special expertise/knowledge, are referred to as *special agents*, or *knowledge workers* [24]. The time of knowledge workers is usually very valuable (and expensive). As a result, it is not rational, or even feasible, to have a pool of knowledge workers constantly available. Instead, they are invited on-demand (usually via remote online access). This mode of operation,

however, poses a challenge, because (a) knowledge workers that accepted an invitation should not wait for an extended (expensive) time before they actually start processing a call, and (b) the callers should not have long waiting time either, because some of the valuable calls will be lost and the service level objectives of the call center will not be achieved. Thus, effective mechanisms must be designed to assure that both these objectives are achieved.

Therefore, a critical part of the operation procedure for such a system, is an algorithm to decide when to invite knowledge workers and how many to invite in order to minimize/stabilize both customer and agent delays. (From now on we will often say "agent" instead of "knowledge worker", as this is the only agent type that we consider in this paper.) One "naive" scheme would be to invite an agent for each arrived customer – this is grossly inefficient, because an agent responds after (often large) delay, and may not respond at all. Another "naive" approach could be that agents are invited at the rate that is equal to the rate of customers arrivals. However, *even if the agent arrival rate is known*, it can be easily checked that the variance of the difference of the number of customer arrivals and agents invitations increases linearly as time evolves. Let alone the fact that the agent arrival rate is usually not known exactly in advance and may change over time. Thus, this scheme is not desirable.

A feedback invitation scheme is proposed for the management of knowledge workers in [24]. The idea is to keep a dynamic target level of invited agents, which changes according to the current system state and its "derivative". The scheme allows to stabilize the system and keep waiting times of customers and agents low. In this paper we consider a stylized model of that invitation scheme – it is such that the stochastic process describing the system dynamics is a continuous-time Markov chain (CTMC). Simulation experiments confirm that our stylized model's behavior is indeed very close to that of the (more practical) feedback invitation scheme in [24]; see section 6. Our analysis of the model shows that waiting times of both customers and agents are stable and asymptotically negligible, as the system scale (the customer arrival rate) goes to infinity.

Specifically, in the stylized model, agents are invited to serve customers, and decide to accept or reject the invitations after some random delay. Customers join a customer queue upon arrival if no agent is available, and are served in the first-come-first-serve discipline. Agents join an agent queue upon acceptance of invitations if no customer is waiting, and serve customers in the order of their acceptance of invitations. Once a customer and an agent are matched, they leave the system immediately. (Service times are irrelevant in our model.) Agent invitations are issued at the event times of customer arrivals and agents acceptance of invitations and at the independent Poisson event times driven by the difference of agent and customer queues. We assume that customer arrival process is Poisson and the invited agent response times are i.i.d. exponential. (We discuss later how these assumptions can be relaxed.) The system state can be described by two variables. One tracks the difference of the agent queue and the customer queue. (Only one of those queues can be positive at any time – we assume that the non-idling condition is in force.) Another variable is the target level of invited agents, which is also the actual number of invited agents. (This is made possible by assuming that agent invitations can be issued or revoked instantaneously if needed.) Under our assumptions, the evolution of the vector ('queue difference', 'target level') is a CTMC.

Exact analysis of the model is prohibitively hard, and thus, we will analyze it in an asymptotic regime where the customer arrival rate becomes large while the distribution of an agent response times is fixed. (Recall that service times do not matter here.) This scaling regime is like the so-called many-server asymptotic regime. In particular, the agent response process is modeled as an infinite-server model with an "arrival" process controlled by the invitation scheme. Our main focus is the case when the customer arrival rate is constant. In this case, we study the process under the fluid and diffusion scaling. On the fluid scale, we show convergence to the fluid limit and uniform global stability of fluid limits (Theorems 1); in addition we prove the process stochastic stability and the limit-interchange property (Theorem 2) – the sequence of fluid-scaled stationary distributions converges to the distribution concentrated on a single point corresponding to zero queues. The key technical challenge in the fluid-scale analysis stems from the fact that the target-level variable has to stay non-negative, which creates a non-trivial boundary behavior. Then, on the diffusion scale, we prove the convergence to the diffusion limit process (Theorem 3) and present the tightness and limit-interchange result (Theorem 4). (The latter can be obtained by adopting the approach in [25, 26]. We give a high level sketch, but not all the lengthy details, in this paper, because believe that the fluid-scale results are of greater

interest and importance for our model.) When the customer arrival rate is time-varying, we give a fluid limit result (Theorem 12). Our simulation experiments show good performance of the feedback scheme, as well as accuracy of the fluid limit approximation.

1.1 Contributions and comparisons

Our research contributes to workforce management in call centers, where both customers and agents must be managed properly in order to minimize operational costs [1, 12]. There has been extensive study on the management of regular agents in the literature; however, there is a lack of stochastic models and analysis for the management of special agents/knowledge workers, which can be engaged dynamically, in response to actual demand. A recently introduced (data-driven) Erlang "S" model [3] (more motivated from the management of regular agents), is related to ours in the general sense that it allows the dependence of agent-availability process on the customer queue length.

Our model also relates to some extent to the literature on matching (double-ended) queues; see, e.g., [15, 17, 19]. In our system, however, customer demand is matched by agents invited through a feedback control mechanism, driven by the system state. This is very different from the standard matching (double-ended) queues since they assume that the entities to be matched arrive exogenously.

It may appear that our model has some similarity with the classical make-to-stock (MTS) queueing model in inventory theory; see, e.g., [13]. In the MTS model, demand of goods arrives as a Poisson process, and items are produced at a single-server or multi-server factory. A standard control algorithm for this system is as follows. Upon the arrival of an order, a "signal" is sent to the factory floor (FF) to produce one item, and an item is delivered to the customer from the finished goods (FG) inventory if it is available, and otherwise, the backlog is increased by one. This algorithm simply means that total inventory (FF inventory plus FG inventory minus backlog) is kept at a constant level. In comparison with our model, the backlog is like our "customer" queue, the FG inventory is like our "agent" queue, the net inventory level (FG inventory minus backlog) is like our "queue difference", and the FF inventory is like our "pending agents". Our model and the algorithm differ from those for the MTS system in several aspects.
1) The key model difference is that we have infinite potential agent pool ("production capacity"), which is used to match any demand.
2) The "invitation" schemes are completely different, far beyond the fact that they apply to different models. The fundamental difference is that in the MTS algorithm the total inventory is kept at a constant level, which is computed as a function of the demand rate; our invitation scheme automatically adjusts to *any* demand rate (as long as it is sufficiently large) – it does not need to be known or explicitly estimated. An "analog" of our scheme in the MTS context would be an algorithm that automatically and dynamically adjusts to any demand rate.
3) Our results are, of course, different. In particular, they show that the invitation scheme is asymptotically optimal in the sense that both customer and agent steady-state waiting times vanish as the system scale (customer arrival rate) becomes large.
4) We also note the robustness of our scheme. The asymptotic optimality is achieved for any setting of the algorithm parameters from a wide range, defined by certain simple conditions. In practice, it is easy to make sure that those conditions hold – it suffices to have only a rough idea about the system parameters (as opposed to exact or even approximate knowledge).
Finally, our model, almost as is, might be useful for some MTS inventory systems, e.g., in settings where there are many "small" producers that may be activated and produce inventory after a delay. In such settings, our asymptotic regime is natural.

As far as techniques are concerned, our work is relevant to the literature on the fluid and diffusion scale tightness and limit interchange in many-server asymptotic regime; see, e.g., [9–11, 25, 26] for an overview. In our model, the process has a nontrivial boundary behavior. This presents new challenges for proving stability and fluid-scale limit-interchange. We have developed a novel approach to establish these properties, via a Lyapunov drift argument that involves multiple time scales. (It deals with the situation where, roughly

speaking, the scale of the Lyapunov function decrease rate is very different on the state space boundary and away from it.) That is the main technical contribution of this paper (see Section 4). We believe that this technical approach can be also used to study other stochastic networks for which the processes may have on-trivial boundary behavior.

1.2 Organization of the paper

The remainder of the paper is organized as follows. We will finish this section with basic notation and conventions below. In section 2, we first describe the model and assumptions in detail. In section 3, we state the main fluid- and diffusion-scale results for the case when the customer arrival rate is constant. These results are proved in sections 4 and 5, respectively. Simulation experiments for the constant arrival rate case are provided in section 6. In section 7, we present a fluid limit result and simulations for the time-varying arrival rate case. We will conclude and discuss future work in section 8.

1.3 Basic notation and conventions

Sets of real and real non-negative numbers (d-dimensional vectors) are denoted by $\mathbb{R}$ and $\mathbb{R}_+$ ($\mathbb{R}^d$ and $\mathbb{R}^d_+$), respectively. $\mathbb{N}$ denotes the set of positive integers. The standard Euclidean norm of a vector $x \in \mathbb{R}^n$ is denoted $\|x\|$. Vectors are viewed as row vectors. For a vector a or matrix A, we write their transposes as a^T or A^T. We often write $x(\cdot)$ to mean the function (or random process) $(x(t),\ t \geq 0)$. Let $D^k = D([0,\infty), \mathbb{R}^k)$ denote the space of $\mathbb{R}^k$-valued functions defined on $[0,\infty)$ that are right continuous with left limits. For a real-valued function $x(\cdot) : \mathbb{R}_+ \to \mathbb{R}$, we use either $x'(t)$ or $(d/dt)x(t)$ to denote the derivative with respect to t, and for $x(\cdot) : \mathbb{R}_+ \to \mathbb{R}^d$, we write $(d/dt)x(t) = (x_1'(t), ..., x_d'(t))$. For a function $x(\cdot)$, we use $x(\infty)$ to denote its limit as $t \to \infty$ if it exists. For a process $x(\cdot)$, we use $x(\infty)$ to denote a random variable having a steady state distribution of the process. We use $\mathbf{1}_A$ to denote an indicator function for a set A, where $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ if $x \notin A$. For a real number x, let $x^+ = \max\{x, 0\}$ and $x^- = -\min\{x, 0\}$, and let $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = 0$ if $x = 0$ and $\mathrm{sgn}(x) = -1$ if $x < 0$. The abbreviation *w.p.*1 means *with probability* 1. Abbreviation *u.o.c.* means *uniform on compact sets* convergence of functions, with the argument (usually in $[0,\infty)$) determined by the context. We write $x^r \to x \in \mathbb{R}^n$ to denote ordinary convergence in $\mathbb{R}^n$, and $x^r \Rightarrow x$ to denote convergence in distribution of random variables taking values in space $\mathbb{R}^n$ equipped with the Borel σ-algebra. Weak convergence of probability measures (on some Polish space) μ_n to μ will also be denoted as $\mu_n \Rightarrow \mu$. For a finite set of scalar functions $f_n(t),\ t \geq 0,\ n \in \mathbb{N}$, a point t is called *regular* if for any subset $\mathbb{N}_o \subseteq \mathbb{N}$ the derivatives

$$\frac{d}{dt} \max_{n \in \mathbb{N}_o} f_n(t) \quad \text{and} \quad \frac{d}{dt} \min_{n \in \mathbb{N}_o} f_n(t)$$

exist. (To be precise, we require that each derivative is proper: both left and right derivatives exist and are equal.) We use the familiar big-O and small-o notations for deterministic functions: for two real-valued functions f and g, we write $f(x) = O(g(x))$ if $\limsup_{x\to\infty} |f(x)/g(x)| < \infty$ and $f(x) = o(g(x))$ if $\limsup_{x\to\infty} |f(x)/g(x)| = 0$.

2 Model and algorithm

Consider a customer contact center where agents are invited on demand and work at their distant "homes". Customers arrive according to a Poisson process of rate $\Lambda > 0$, and join a "customer" queue waiting for an available agent and are served in the order of their arrival. Agents are invited to serve customers according to some scheme specified below. Once being invited, an agent will decide to accept or reject the invitation after some exponentially distributed random time. Let $\beta > 0$ and $\tilde{\beta} > 0$ be the rates at which an invited agent accepts or declines the invitation, respectively. Agents who accept their invitations will join an "agent"

queue waiting for a customer to arrive and serve customers in the order of their acceptance of the invitations. Once a customer and an agent are matched, they will leave the system simultaneously. This happens at the instant of either a customer arrival or an agent invitation acceptance. Thus, we do not consider service times in this model. Let $X(t)$ be the number of pending agents that have been invited but not decided to accept or decline the invitations at time t. Let $Q_c(t)$ be the number of customers in the customer queue at time t and $Q_a(t)$ be the number of agents in the agent queue at time t. Define $Y(t) := Q_a(t) - Q_c(t)$, as the difference of the agent queue and customer queue at time t. We assume that the non-idling condition holds, that is, agents do not idle when there are customers waiting in the customer queue, which implies that at each time t, either the customer queue or the agent queue must be empty. Figure 1 depicts such an agent invitation system.

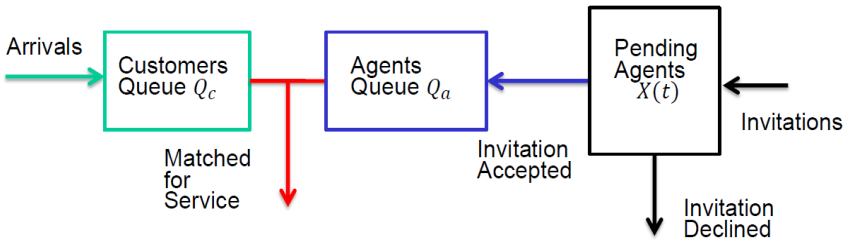


Figure 1: An Agent Invitation System

The feedback invitation scheme in [24], let us label it as *Scheme A*, works as follows. The scheme maintains a "target" $X_{target}(t)$ for the number of invited agents $X(t)$. The target $X_{target}(t)$ is changed by $\Delta X_{target}(t) = [-\gamma \Delta Y(t) - \epsilon Y(t)\Delta(t)]$ at each time t when $Y(t)$ changes by $\Delta Y(t)$ (which can be either +1 or -1), where $\gamma > 0$ and $\epsilon > 0$ are the algorithm parameters and $\Delta(t)$ is the time duration from the previous change of Y. New agents are invited if and only if $X(t) < X_{target}(t)$, where $X(t)$ is the actual number of invited (pending) agents; therefore, $X(t) \geq X_{target}(t)$ holds at all times. In addition, the target $X_{target}(t)$ is not allowed to go below zero, $X_{target}(t) \geq 0$; i.e. if an update of $X_{target}(t)$ makes it negative, its value is immediately reset to 0. Note that $X_{target}(t)$ is *not* necessarily an integer.

To simplify our theoretical analysis, we consider a "stylized" version of Scheme A, which has the same basic dynamics, but keeps $X_{target}(t)$ integer and assumes that $X(t) = X_{target}(t)$ at all times; the latter is equivalent to assuming that not only agent invitations can be issued instantly, but they can also be withdrawn at any time. (In reality, it might not be feasible or desirable to withdraw the invitations. However, our simulations will confirm that the performances of Scheme A and its stylized version are essentially same. See section 6.) Given these assumptions, when pending agents decline invitations, it has no impact on the system state, because $X(t)$ is immediately "replenished" by inviting another agent; therefore, in the analysis of stylized scheme, the events of declined invitations can be ignored (and parameter $\tilde{\beta}$ has no relevance).

Formally, the stylized scheme, which we label as *Scheme B*, is defined as follows. There are three types of mutually independent, and independent of the past, events that affect the dynamics of $X(t)$ and $Y(t)$ in a small time interval $[t, t + dt]$: a customer arrival with probability Λdt, an agent acceptance with probability $\beta X(t)dt$, and an additional event with probability $\epsilon |Y(t)|dt$.

The changes at these event times are described as follows:
(i) Upon a customer arrival, $Y(t)$ changes by $\Delta Y(t) = -1$, and $X(t)$ increases by the average of $\gamma > 0$. For example, if $\gamma = 2.6$ and $\Delta Y(t) = -1$, then $\Delta X(t) = 3$ with probability 0.6 and $\Delta X(t) = 2$ with probability 0.4. To simplify the exposition, we assume from now on that γ is an integer.
(ii) Upon the acceptance of an invitation, it causes the change $\Delta Y(t) = 1$ of Y, which in turn causes the change in $X(t)$ by $\Delta X(t) = -(\gamma \wedge X(t))$, that is, the change is by $-\gamma$ but $X(t)$ is kept to be nonnegative.

(iii) Upon the third type of event, if $X(t) \geq 1$, the change $\Delta X(t) = -\operatorname{sgn}(Y(t))$ occurs; and if $X(t) = 0$, the change $\Delta X(t) = 1$ occurs if $Y(t) < 0$ and $\Delta X(t) = 0$ if $Y(t) \geq 0$.

All theoretical results in this paper concern Scheme B, which we consider in the rest of the paper, unless explicitly stated otherwise. Under this scheme, the two-dimensional process (Y, X) describing the system dynamics is a CTMC. We further assume that the parameters ϵ, γ and β satisfy

$$0 < \epsilon < \gamma^2\beta/4. \tag{1}$$

We see that, informally, the scheme dynamics can be described as

$$(d/dt)X = -\gamma(d/dt)Y - \epsilon Y, \quad (d/dt)Y = \beta X - \Lambda,$$

so that X changes according to a negative feedback with respect to both the current value and increments of Y, while Y changes naturally, according to agents' invitation acceptances and new customer arrivals.

We remark that both Schemes A and B, of course, apply even when the customer arrival rate is time-varying. While our analysis is primarily focused on the case of constant customer arrival rates (sections 3-6), we will provide some results for the time-varying case in section 7. Also, the assumption of Poisson arrivals is not essential for our results and can be easily relaxed; we believe the results can be generalized to allow non-exponential distribution of agent response times as well, but this is less straightforward.

3 Main Results

In this section, we state the main results of the paper. We consider a sequence of systems, indexed by the scaling parameter $r \in \mathbb{R}_+$ and let $r \to \infty$. In the r-th system, the arrival rate is λr, while the parameters β, ϵ, γ are constant. The corresponding r-th process is (Y^r, X^r), where $Y^r = (Y^r(t),\ t \geq 0)$, $X^r = (X^r(t),\ t \geq 0)$. Define fluid-scaled processes with centering

$$(\bar{Y}^r, \bar{X}^r) := r^{-1}(Y^r, X^r - \lambda r/\beta), \tag{2}$$

and diffusion-scaled processes

$$(\hat{Y}^r, \hat{X}^r) := \sqrt{r}(\bar{Y}^r, \bar{X}^r) = r^{-1/2}(Y^r, X^r - \lambda r/\beta). \tag{3}$$

We first state a fluid limit result for $(\bar{Y}^r, \bar{X}^r)$ below. Note that for any r, $\bar{X}^r(t) \geq -\lambda/\beta$ for all t.

Theorem 1. *Consider a sequence of processes $(\bar{Y}^r, \bar{X}^r)$, $r \to \infty$, with deterministic initial states such that $(\bar{Y}^r(0), \bar{X}^r(0)) \to (y(0), x(0))$ for some fixed $(y(0), x(0)) \in \mathbb{R}^2$, $x(0) \geq -\lambda/\beta$. Then, these processes can be constructed on a common probability space, so that the following holds. There exists a unique locally Lipschitz trajectory (y, x), such that, w.p.*1*,*

$$(\bar{Y}^r, \bar{X}^r) \to (y, x) \quad \textit{u.o.c.} \quad \textit{as} \quad r \to \infty, \tag{4}$$

where

$$x(t) \geq -\lambda/\beta, \quad t \geq 0, \tag{5}$$

and at any regular point $t \geq 0$ (all points $t \geq 0$ are regular, except a subset of zero Lebesgue measure), the following holds: if $x(t) > -\lambda/\beta$,

$$\begin{aligned} y'(t) &= \beta x(t), \\ x'(t) &= -\gamma\beta x(t) - \epsilon y(t), \end{aligned} \tag{6}$$

and if $x(t) = -\lambda/\beta$,

$$\begin{aligned} y'(t) &= -\lambda, \\ x'(t) &= [\gamma\lambda - \epsilon y(t)] \vee 0. \end{aligned} \tag{7}$$

The unique limit trajectory (y, x) specified in Theorem 1 will be called a *fluid limit* starting from $(y(0), x(0))$.

Note that the second equation in (6) can also be written as

$$x'(t) = -\gamma y'(t) - \epsilon y(t). \tag{8}$$

When the trajectory is away from the boundary, the ODE (6) can be written as

$$(d/dt)(y, x) = (y, x)A, \tag{9}$$

where

$$A = \begin{bmatrix} 0 & -\epsilon \\ \beta & -\gamma\beta \end{bmatrix}. \tag{10}$$

The assumption (1) on the parameters guarantees that the matrix A has two different negative eigenvalues. When the fluid limit trajectory hits the boundary $x(t) = -\lambda/\beta$ at some time t, y will decrease at rate λ, that is, $y' = -\lambda$ until y hits the value $\gamma\lambda/\epsilon$, and afterwards, the fluid limit will follow the trajectory of the ODE in (6) again. These observations imply, in particular, that the unique stable point of a fluid limit is $(0, 0)$.

We then show the following stability and fluid-scale tightness results.

Theorem 2. *For all sufficiently large r, the system is stable, i.e., the Markov process (Y^r, X^r) is positive recurrent. The sequence of stationary distributions of the fluid-scaled processes $(\bar{Y}^r, \bar{X}^r)$ converges to the Dirac measure concentrated at $(0, 0)$.*

The following result describes transient behavior on the diffusion scale.

Theorem 3. *Suppose the sequence of deterministic initial states is such that $(\hat{Y}^r(0), \hat{X}^r(0)) \to (\hat{Y}(0), \hat{X}(0))$, where $(\hat{Y}(0), \hat{X}(0))$ is a fixed vector in $\mathbb{R}^2$. Then,*

$$(\hat{Y}^r, \hat{X}^r) \Rightarrow (\hat{Y}, \hat{X}) \quad as \quad r \to \infty, \tag{11}$$

where $(\hat{Y}, \hat{X})$ is the unique solution to the SDE

$$\begin{aligned} \hat{Y}(t) &= \hat{Y}(0) + \beta \int_0^t \hat{X}(s)ds - \sqrt{2\lambda}W(t), \\ \hat{X}(t) &= \hat{X}(0) - \beta\gamma \int_0^t \hat{X}(s)ds - \epsilon \int_0^t \hat{Y}(s)ds + \gamma\sqrt{2\lambda}W(t), \end{aligned} \tag{12}$$

and W is a standard Brownian motion. The distribution of $(\hat{Y}(t), \hat{X}(t))$ is Gaussian with mean function $m(t)$ and covariance function $V(t)$ being unique solutions to the following ODEs, respectively:

$$\begin{aligned} \dot{m}(t) &= m(t)A, \\ \dot{V}(t) &= V(t)A + A^T V(t) + \sigma^T\sigma, \end{aligned} \tag{13}$$

where the matrix A is given in (10) and $\sigma := (-\sqrt{2\lambda}, \gamma\sqrt{2\lambda})$. The stationary distribution of $(\hat{Y}, \hat{X})$ is Gaussian with mean $(0, 0)$ and covariance matrix

$$V(\infty) = \begin{bmatrix} \frac{\lambda}{\beta\gamma} & -\frac{\lambda}{\beta} \\ -\frac{\lambda}{\beta} & \frac{\lambda(\beta\gamma^2+\epsilon)}{\beta^2\gamma} \end{bmatrix}. \tag{14}$$

Finally, we state the following diffusion-scale limit interchange result. We will not provide its detailed proof – it can be done, using Theorems 2 and 3 as the starting point, and then following the approach given (for a different setting) in [25, 26].

Theorem 4. *The sequence of stationary distributions of the diffusion-scaled processes $(\hat{Y}^r, \hat{X}^r)$ is tight. Consequently, given Theorem 3, the limit-interchange holds: the limit of stationary distributions of the diffusion-scaled processes $(\hat{Y}^r, \hat{X}^r)$ is equal to the stationary distribution of the limit diffusion process $(\hat{Y}, \hat{X})$.*

4 Fluid scale analysis

4.1 Fluid models

We start studying properties of fluid limits by first considering *fluid models*, which are defined as locally Lipschitz continuous trajectories (y, x), satisfying conditions (5)-(7). In other words, if a fluid limit exists (we do not even claim this yet), it is necessarily a fluid model. (The converse is also true, as we will see, but we do not yet claim this either.)

Lemma 5. *For any initial state $(y(0), x(0))$, there is a unique fluid model starting from it. Moreover, uniformly on the initial states from a given compact set,*

$$(y(t), x(t)) \to (0, 0), \quad t \to \infty,$$

and

$$\max_{t \geq 0} \|(y(t), x(t))\| \quad \textit{is bounded.}$$

Let us construct a fluid model starting from a given initial state $(y(0), x(0))$. When the trajectory is away from the boundary, the fluid model (y, x) evolves according to ODE (9)-(10), and recall that the assumption on the parameters (1) guarantees that there are two different negative eigenvalues of A, namely $-\infty < -\nu_2 < -\nu_1 < 0$.

We choose the two corresponding eigenvectors to be $v_i = (\beta/\nu_i, -1)$, $i = 1, 2$. Switching to the basis v_1, v_2, transforms any vector u to uB^{-1}, where the matrix B has v_1 and v_2 as its first and second rows. We will use norm $\|u\|_* = \|uB^{-1}\|$; in other words, $\|u\|_* = (\alpha_1^2 + \alpha_2^2)^{1/2}$, where α_1 and α_2 are the coordinates of u in the basis v_1, v_2, i.e., $u = \alpha_1 v_1 + \alpha_2 v_2$.

If we consider just the ODE (9) with initial state $(y(0), x(0)) = \alpha_1(0)v_1 + \alpha_2(0)v_2$, without any boundaries, then the solution is

$$(y(t), x(t)) = \sum_{i=1,2} \alpha_i(0) e^{-\nu_i t} v_i. \tag{15}$$

When/if the fluid model trajectory hits the boundary $x = -\lambda/\beta$, say at time t, this can only happen if $y(t) \geq \gamma\lambda/\epsilon$; if so, then x stays on the boundary ($x = -\lambda/\beta$, $x' = 0$) and $y' = -\lambda$ until y hits value $\gamma\lambda/\epsilon$; at that time the fluid model starts obeying the ODE (9) again. Logically, it is possible that a fluid model trajectory can hit the boundary multiple times. However, this cannot actually happen.

Denote $a_i = \beta/\nu_i$, so that $v_i = (a_i, -1)$. This notation is used in the following two lemmas.

Lemma 6. *Consider the solution to the ODE (9) starting from point $(y(0), x(0)) = (b_1, -b_2)$, where $b_1, b_2 > 0$. Then, for any $t > 0$ it is impossible to have $x(t) = -b_2$ and $y(t) \geq b_1$.*

Proof. Denote $\alpha_i(t) = \alpha_i(0)e^{-\nu_i t}$; see (15). We have

$$\alpha_1(0) + \alpha_2(0) = b_2, \quad \alpha_1(0)a_1 + \alpha_2(0)a_2 = b_1.$$

Suppose that for some $t > 0$,

$$\alpha_1(t) + \alpha_2(t) = b_2, \quad \alpha_1(t)a_1 + \alpha_2(t)a_2 \geq b_1.$$

Since $a_1 > a_2$, we must have for some $\delta \geq 0$ and some $t > 0$,

$$\alpha_1(t) = \alpha_1(0) + \delta, \quad \alpha_2(t) = \alpha_2(0) - \delta. \tag{16}$$

The case $\delta = 0$ is impossible, because it would imply that $\alpha_1(0) = \alpha_2(0) = 0$. Therefore, $\delta > 0$. Then, $\alpha_1(0) < 0$, because $\alpha_1(t)$ has the same sign as $\alpha_1(0)$ and smaller absolute value. Then, $\alpha_2(0) = b_2 - \alpha_1(0) > 0$, and $|\alpha_2(0)| > |\alpha_1(0)|$. Then (16) implies

$$|\alpha_1(t)/\alpha_1(0)| < |\alpha_2(t)/\alpha_2(0)|.$$

This, however, is impossible because $\nu_1 < \nu_2$. Contradiction completes the proof. □

As a corollary of this lemma, we see that a solution to the ODE (9), starting from point $(y(0), x(0)) = (\gamma\lambda/\epsilon, -\lambda/\beta)$ cannot ever reach a point such that $x(t) = -\lambda/\beta$ and $y(t) \geq \gamma\lambda/\epsilon$. Therefore, the fluid model trajectory that we are constructing, will hit the boundary at most once. If that happens, it spends a finite time on the boundary, reaches the point $(\gamma\lambda/\epsilon, -\lambda/\beta)$, and then follows the ODE thereafter. All claims of Lemma 5 easily follow.

Next we show that, along a fluid model trajectory, the norm $\|(y(t), x(t))\|_*$ is decreasing when it is large.

Lemma 7. *There exist $\eta > 0$ and $C > 0$ such that, for any fluid model, at any regular point t, $\|(y(t), x(t))\|_* \geq C$ implies*

$$(d/dt)\|(y(t), x(t))\|_* \leq -\eta.$$

Proof. When $(y(t), x(t))$ has a large norm and is away from the boundary (and then satisfies the ODE (9)), the statement is obvious from (15). When/if a trajectory moves on the domain boundary, it has the form $(y(t), -\lambda/\beta)$ where $y(t) > 0$ and $y'(t) = -\lambda$. It also can be written as $(y(t), -\lambda/\beta) = \alpha(t)v_1 - (\alpha(t) - \lambda/\beta)v_2$, where

$$\alpha(t) = \frac{y(t) - (\lambda/\beta)a_2}{a_1 - a_2}.$$

If $(y(t), x(t))$ has a large norm, then $y(t)$ is large positive. Then so is $\alpha(t)$, and moreover $\alpha'(t) = -\lambda/(a_1 - a_2)$. The desired property follows. □

4.2 Proof of Theorem 1

Throughout this section, we are under the assumptions of Theorem 1. Given properties of the fluid models that we have already established, in order to prove Theorem 1, it suffices to show that w.p.1 from any subsequence of r we can choose a further subsequence, along which a u.o.c. convergence to a fluid model holds.

Given the initial state $(Y^r(0), X^r(0))$, we construct the processes (Y^r, X^r), for all r, on the same probability space via a common set of independent Poisson processes as follows:

$$Y^r(t) = Y^r(0) + N_2\left(\beta \int_0^t X^r(s)ds\right) - N_1(\lambda rt), \tag{17}$$

$$X^r(t) = Z^r(t) + \left(-\min_{0 \leq s \leq t} Z^r(s)\right) \vee 0, \tag{18}$$

$$\begin{aligned} Z^r(t) \quad = \quad & X^r(0) + \gamma N_1(\lambda rt) - \gamma N_2\left(\beta \int_0^t X^r(s)ds\right) \\ & + N_3\left(\epsilon \int_0^t (Y^r(s))^- ds\right) - N_4\left(\epsilon \int_0^t (Y^r(s))^+ ds\right), \end{aligned} \tag{19}$$

and $N_i(\cdot)$, $i = 1, ..., 4$, are mutually independent unit-rate Poisson processes. W.p.1, for any r, relations (17)-(19) uniquely define the realization of (Y^r, X^r) via the realizations of the driving processes $N_i(\cdot)$. Relation (18) – the "reflection" at zero – corresponds to the property that $X^r(t)$ cannot become negative.

The functional strong law of large numbers (FSLLN) holds for each Poisson process N_i:

$$N_i(rt)/r \to t, \ r \to \infty, \quad \text{u.o.c.}, \quad \text{w.p.1.} \tag{20}$$

We consider the sequence of associated fluid-scaled processes $(\bar{Y}^r, \bar{X}^r)$ as defined in (2). (Note that the processes $\bar{X}^r$ are centered.) Let a constant $m > \|(y(0), x(0))\|$ be fixed. For each r, on the same probability

space as $(\bar{Y}^r, \bar{X}^r)$, let us define a modified fluid-scaled process $(\bar{Y}^r_m, \bar{X}^r_m)$ as follows. Let $(\bar{Y}^r_m, \bar{X}^r_m)$ start from the same initial state as $(\bar{Y}^r, \bar{X}^r)$, i.e., $(\bar{Y}^r_m(0), \bar{X}^r_m(0)) = (\bar{Y}^r(0), \bar{X}^r(0))$. The modified process $(\bar{Y}^r_m, \bar{X}^r_m)$ follows the same path as $(\bar{Y}^r, \bar{X}^r)$ until the first time that $\|(\bar{Y}^r(t), \bar{X}^r(t))\| \geq m$. Denote this time by τ^r_m. We then freeze the process $(\bar{Y}^r_m, \bar{X}^r_m)$ at the value $(\bar{Y}^r(\tau^r_m), \bar{X}^r(\tau^r_m))$, i.e., $(\bar{Y}^r_m(t), \bar{X}^r_m(t)) = (\bar{Y}^r(\tau^r_m), \bar{X}^r(\tau^r_m))$ for all $t \geq \tau^r_m$.

The proof of the convergence of fluid-scaled processes $(\bar{Y}^r, \bar{X}^r)$ will be in two steps, which are roughly as follows. First, we show the convergence of $(\bar{Y}^r_m, \bar{X}^r_m)$ to a limit trajectory that behaves like a fluid model as long as the state norm is away from m. (Here we will use the fact that the modified processes $(\bar{Y}^r_m, \bar{X}^r_m)$ are uniformly bounded for all r and $t \geq 0$ by construction.) Second, for a given initial state, we choose the constant m large enough, so that the limit trajectory never reaches norm level m, and therefore it is the unique fluid model; this implies that on any finite time interval, w.p.1, for all large r, $(\bar{Y}^r_m, \bar{X}^r_m)$ coincides with $(\bar{Y}^r, \bar{X}^r)$, and therefore the latter converges to the fluid model.

Lemma 8. *Fix $(y(0), x(0))$ and a finite constant $m > \|(y(0), x(0))\|$. The following holds w.p.1. From any subsequence of r, we can find a further subsequence, along which $(\bar{Y}^r_m, \bar{X}^r_m)$ converges u.o.c. to a Lipschitz continuous trajectory (y_m, x_m), which satisfies properties (5) − (7) at any regular time t such that $\|(y_m(t), x_m(t))\| < m$.*

Proof. For the modified fluid-scaled processes $(\bar{Y}^r_m, \bar{X}^r_m)$, we define the associated counting processes for upward and downward jumps: for $t \leq \tau^r_m$,

$$\begin{aligned}
\bar{Y}^{r,\uparrow}_m(t) &= r^{-1} N_2\left(r\beta \int_0^t [\bar{X}^r_m(s) + \lambda/\beta] ds\right), \\
\bar{Y}^{r,\downarrow}_m(t) &= r^{-1} N_1(\lambda r t), \\
\bar{X}^{r,\uparrow}_m(t) &= r^{-1}\gamma N_1(\lambda r t) + r^{-1} N_3\left(r\epsilon \int_0^t \left(\bar{Y}^r_m(s)\right)^- ds\right), \\
\bar{X}^{r,\downarrow}_m(t) &= r^{-1}\gamma N_2\left(r\beta \int_0^t [\bar{X}^r_m(s) + \lambda/\beta] ds\right) + r^{-1} N_4\left(r\epsilon \int_0^t \left(\bar{Y}^r_m(s)\right)^+ ds\right),
\end{aligned} \tag{21}$$

and for $t > \tau^r_m$, all these counting processes are frozen at their values at time τ^r_m, that is,

$$\bar{Y}^{r,\uparrow}_m(t) = \bar{Y}^{r,\uparrow}_m(\tau^r_m), \quad \bar{Y}^{r,\downarrow}_m(t) = \bar{Y}^{r,\downarrow}_m(\tau^r_m), \quad \bar{X}^{r,\uparrow}_m(t) = \bar{X}^{r,\uparrow}_m(\tau^r_m), \quad \bar{X}^{r,\downarrow}_m(t) = \bar{X}^{r,\downarrow}_m(\tau^r_m), \quad t \geq \tau^r_m. \tag{22}$$

Using this notation, relations (17)-(19), and the fact that for $0 \leq t \leq \tau^r_m$ the original and modified processes, $(\bar{Y}^r, \bar{X}^r)$ and $(\bar{Y}^r_m, \bar{X}^r_m)$, coincide, we obtain for all $t \geq 0$:

$$\bar{Y}^r_m(t) = \bar{Y}^r(0) + \bar{Y}^{r,\uparrow}_m(t) - \bar{Y}^{r,\downarrow}_m(t), \tag{23}$$

$$\bar{X}^r_m(t) = \bar{Z}^r_m(t) + \left(-\lambda/\beta - \min_{0 \leq s \leq t} \bar{Z}^r_m(s)\right) \vee 0, \tag{24}$$

$$\bar{Z}^r_m(t) = \bar{X}^r(0) + \bar{X}^{r,\uparrow}_m(t) - \bar{X}^{r,\downarrow}_m(t). \tag{25}$$

The counting processes $\bar{Y}^{r,\uparrow}_m, \bar{Y}^{r,\downarrow}_m, \bar{X}^{r,\uparrow}_m, \bar{X}^{r,\downarrow}_m$ are non-decreasing. Using the FSLLN (20) and the fact that the processes $\bar{Y}^r_m$ and $\bar{X}^r_m$ are uniformly bounded by construction, we see that, w.p.1. for any subsequence of r, there exists a further subsequence along which the set of trajectories $(\bar{Y}^{r,\uparrow}_m, \bar{Y}^{r,\downarrow}_m, \bar{X}^{r,\uparrow}_m, \bar{X}^{r,\downarrow}_m)$ converges u.o.c. to a set of non-decreasing Lipschitz continuous functions $(y^\uparrow_m, y^\downarrow_m, x^\uparrow_m, x^\downarrow_m)$. But then the u.o.c. convergence of $(\bar{Y}^r_m, \bar{X}^r_m, \bar{Z}^r_m)$ to a set of Lipschitz continuous functions (y_m, x_m, z_m) holds, where

$$y_m(t) = y(0) + y^\uparrow_m(t) - y^\downarrow_m(t), \tag{26}$$

$$x_m(t) = z_m(t) + \left(-\lambda/\beta - \min_{0 \leq s \leq t} z_m(s)\right) \vee 0, \tag{27}$$

$$z_m(t) = x(0) + x^\uparrow_m(t) - x^\downarrow_m(t). \tag{28}$$

Using this, and again the FSLLN (20), we can take the limit in (21) to obtain:

$$\begin{aligned} y_m^{\uparrow}(t) &= \beta \int_0^t (x_m(s) + \lambda/\beta) ds, \\ y_m^{\downarrow}(t) &= \lambda t, \\ x_m^{\uparrow}(t) &= \gamma\lambda t + \epsilon \int_0^t y_m^-(s) ds, \\ x_m^{\downarrow}(t) &= \gamma\beta \int_0^t (x_m(s) + \lambda/\beta) ds + \epsilon \int_0^t y_m^+(s) ds, \end{aligned} \tag{29}$$

It is easy to verify that properties $(5) - (7)$ hold for the trajectory (y_m, x_m). This completes the proof. □

Conclusion of the proof of Theorem 1. For the given $(y(0), x(0))$, consider the corresponding (unique) fluid model (y, x). Let us choose $m > \max_{t\geq 0} \|(y(t), x(t))\|$. Now let us apply Lemma 8. W.p.1, from any subsequence of r we can choose a further subsequence along which $(\bar{Y}_m^r, \bar{X}_m^r)$ converges u.o.c. to a Lipschitz continuous trajectory (y_m, x_m), which satisfies properties (5) - (7), as long as $\|(y_m(t), x_m(t))\| < m$. But, as long as $\|(y_m(t), x_m(t))\| < m$, (y_m, x_m) coincides with the fluid model (y, x). By the choice of m, this means that $(y_m, x_m) = (y, x)$. Moreover, along the chosen subsequence, for any fixed $T > 0$, for all sufficiently large r, $(\bar{Y}_m^r, \bar{X}_m^r)$ and $(\bar{Y}^r, \bar{X}^r)$ coincide in the interval $[0, T]$. We see that, along the chosen subsequence, u.o.c. convergence of $(\bar{Y}^r, \bar{X}^r)$ to the fluid model (y, x) holds. This means that w.p.1 the u.o.c. convergence of $(\bar{Y}^r, \bar{X}^r)$ to (y, x) holds for the original sequence of r. □

4.3 Proof of Theorem 2

Given the uniform convergence of fluid limits in Lemma 5, to prove Theorem 2, it suffices to prove the following

Lemma 9. *For all sufficiently large r, the process (X^r, Y^r) (and then $(\bar{X}^r, \bar{Y}^r)$) is stable, with a unique stationary distribution. The sequence of stationary distributions of $(\bar{X}^r, \bar{Y}^r)$ is tight.*

Indeed, suppose Lemma 9 holds. Consider stationary versions of the processes $(\bar{Y}^r(\cdot), \bar{X}^r(\cdot))$. Fix arbitrary $\delta > 0$ and then a sufficiently large compact set B such that, uniformly in all (sufficiently large) r, $\mathbb{P}\{(\bar{Y}^r(0), \bar{X}^r(0)) \in B\} \geq 1 - \delta$. Lemma 5 implies that we can choose a sufficiently large $T > 0$ such that, uniformly on all sufficiently large r and initial states $(\bar{Y}^r(0), \bar{X}^r(0)) \in B$, we have

$$\mathbb{P}\{\|(\bar{Y}^r(T), \bar{X}^r(T))\| \leq \delta \mid (\bar{Y}^r(0), \bar{X}^r(0))\} \geq 1 - \delta.$$

Therefore, for all large r, $\mathbb{P}\{\|(\bar{Y}^r(T), \bar{X}^r(T))\| \leq \delta\} \geq (1 - \delta)^2$. Since this is true for arbitrary $\delta > 0$, we obtain the weak convergence of stationary distributions of $(\bar{Y}^r, \bar{X}^r)$ to the Dirac measure concentrated at $(0, 0)$.

The approach we will take to prove Lemma 9 is to first consider an embedded discrete-time Markov chain (DTMC), which is the original continuous-time chain sampled at a sequence of random stopping times, and show stability and the first moment bound for this DTMC using a Lyapunov function drift criterion. Specifically, we use the norm $\|\cdot\|_*$ defined in section 4.1 as the Lyapunov function. We then use the relation between stationary distributions of the embedded and the original Markov chains.

Proof of Lemma 9. Here we denote the fluid-scaled processes $s^r(t) = (\bar{Y}^r(t), \bar{X}^r(t))$, and to simplify the notation, we will drop the index r, so $s(t) = (\bar{Y}(t), \bar{X}(t))$ below is the random process, not the fluid limit.

We consider the embedded Markov chain. Fix constants $\delta > 0$ and $\tau_{max} > 0$. For the process starting from a given state $s = s(0)$, consider the random stopping time $\tau_\delta(s)$, which is the first time t when $| \|s(t)\|_* - \|s\|_* | \geq \delta$; we then define the stopping time $\tau(s) = \tau_\delta(s) \wedge \tau_{max}$. Define a sequence of stopping times $\tau^{(k)}$, $k = 1, 2, ...$ by

$$\tau^{(1)} = \tau(s(0)),$$

$$\tau^{(k)} = \tau^{(k-1)} + \theta_{\tau^{(k-1)}}\tau^{(1)}, \quad k = 2, 3, ...,$$

where $\theta_\cdot$ is the random time shift operator associated with the process. In more detail,

$$\tau^{(k)} = [\tau^{(k-1)} + \tau_{max}] \wedge \inf\{t > \tau^{(k-1)} : \mid \|s(t)\|_* - \|s(\tau^{(k-1)})\|_* \mid \geq \delta\}, \quad k = 2, 3, ...$$

Consider the embedded discrete-time Markov chain $\hat{s}(k),\ k = 0, 1, \ldots$, using $\tau^{(k)}$ as sampling times. Specifically, if $s(t),\ t \geq 0$, is the original continuous time Markov process, then:

$$\hat{s}(0) = s(0), \quad \hat{s}(k) = s(\tau^{(k)}), \quad k = 1, 2, ...$$

Let $\Phi(s) = \|s\|_*$. For the embedded chain $\hat{s}$, we show that, for some $C_1, C_2 > 0$, uniformly in r,

$$\mathbb{E}[\Phi^2(\hat{s}(1)) - \Phi^2(\hat{s}(0)) \mid \hat{s}(0)] \leq -C_1\Phi(\hat{s}(0)) + C_2. \tag{30}$$

Note that $|\Phi(\hat{s}(1)) - \Phi(\hat{s}(0))|$ is uniformly bounded by the definition of $\tau(s)$. Then, to prove (30) it suffices to show the following: for some constant $\delta_7 > 0$, for any sequence $r \to \infty$ and corresponding $\hat{s}(0) = \hat{s}^r(0)$ such that $\|\hat{s}^r(0)\|_* \uparrow \infty$, we have

$$\mathbb{P}\{\Phi(\hat{s}^r(1)) - \Phi(\hat{s}^r(0)) \leq -\delta_7\} \to 1. \tag{31}$$

It suffices to consider a sequence such that the convergence

$$\frac{1}{\|\hat{s}^r(0)\|_*}\hat{s}^r(0) \to \tilde{s}$$

holds, for some vector $\tilde{s}$ with $\|\tilde{s}\|_* = 1$.

We will study the behavior of the continuous time process $s(t)$, with initial state $s(0) = \hat{s}(0)$, on the interval $[0, \tau(s(0))]$. Before we proceed, we introduce some convenient (although somewhat abusive) notation. For any vector $s = (y, x)$ we denote $s' = (y', x')$ where $y' = \beta x$, $x' = -\epsilon y - \gamma\beta x$; in other words, these are the derivatives of the components of a fluid trajectory $s(t)$ when $s(t) = s$. Similarly, let $\|s\|'_*$ denote $(d/dt)\|s(t)\|_*$ when $s(t) = s$.

Suppose first that $\tilde{s} = (\tilde{y}, \tilde{x})$ is such that $\tilde{x} > 0$. Then, (the sequence of processes can be constructed on a common probability space, such that) w.p.1, u.o.c.

$$s(t/\|s(0)\|_*) - s(0) \to \tilde{s}'t \text{ and } \|s(t/\|s(0)\|_*)\|_* - \|s(0)\|_* \to \|\tilde{s}\|'_*t. \tag{32}$$

From here (31) follows. Indeed, we see that $\tau(s(0)) = \tau_\delta(s(0)) \to 0$, and therefore (31) holds with $\delta_7 = \delta$.

Suppose now that $\tilde{x} = 0$. Then, necessarily, $|\tilde{y}| > 0$ and $\tilde{y}' = 0$. If $\tilde{y} < 0$ then $\tilde{x}' > 0$, and this case is treated the same way as the $\tilde{x} > 0$ case. Therefore, it remains to consider the case when $\tilde{y} > 0$ and, consequently, $\tilde{x}' < 0$.

Consider the sub-case when

$$[x(0) - (-\lambda/\beta)]/|\tilde{x}'| \to \infty;$$

then, we easily check that (32) still holds. This is the scenario when the time τ_{hit} for the $x(t)$ to hit boundary $-\lambda/\beta$ is such that $\tau_{hit} \to 0$ and $\tau_{hit}\|s(0)\|_* \to \infty$; therefore, $\|s(t)\|_*$, which decreases at the rate $\|s(0)\|_* \|\tilde{s}\|'_*$, decreases by δ before time τ_{hit}, and (31) again follows.

Finally, consider the sub-case when (along a subsequence of r)

$$[x(0) - (-\lambda/\beta)]/|\tilde{x}'| \to c \in [0, \infty).$$

In this sub-case, $\tau_{hit}\|s(0)\|_* \to c$. Then, we consider the process such that in the interval $[0, \tau_{hit}\|s(0)\|_*]$ it is the process with time slowdown, as in (32), but from time $\tau_{hit}\|s(0)\|_*$ to infinity, the process continues

in actual time, without slowdown. W.p.1. in the limit we obtain the trajectory which satisfies (32) in the interval $[0, c]$, and then in the interval $[c, \infty)$ we have $x(t) = -\lambda/\beta$ and $y'(t) = -\lambda$. In both intervals, the limit trajectory is such that the norm $\|s(t)\|_*$ is decreasing at least at some positive rate. We are done with proving (31) and (30).

From (30), using standard Lyapunov-Foster argument (see, e.g., Chapter 13 in [21]), we conclude that the embedded chain is stable for each sufficiently large r, and therefore has stationary distribution which is easily seen to be unique. Moreover, the stationary distributions are such that, uniformly in (sufficiently large) r,

$$\mathbb{E}\Phi(\hat{s}(\infty)) \leq C_2/C_1. \tag{33}$$

We also observe that, for any fixed $C_3 > 0$, uniformly on all $\|s\|_* \leq C_3$ and all r, $\mathbb{E}\tau(s) \geq C_4 > 0$. Let us choose C_3 large enough, so that for the embedded chain in steady-state, $\mathbb{P}\{\|\hat{s}(\infty)\|_* \leq C_3\} \geq 1/2$.

Now we use the relation between stationary distributions of the original continuous-time process and the sampled chain:

$$\mathbb{P}\{s(\infty) \in A\} = \frac{\mathbb{E}\left[\mathbb{E}\left[\int_0^{\tau(s(0))} I\{s(t) \in A\}dt \;\middle|\; s(0) = \hat{s}(\infty)\right]\right]}{\mathbb{E}[\tau(\hat{s}(\infty))]}$$

(This relation is quite standard. For a proof, in a somewhat more general context, see Lemma 10.1 in [23].) Then we see that our original continuous-time process is stable for each sufficiently large r, and the stationary distributions are such that, uniformly in (sufficiently large) r, we have

$$\mathbb{E}\|s(\infty)\|_* \leq \frac{\mathbb{E}[\|\hat{s}\|_* + 2\delta]\tau_{max}}{C_4/2} \leq C_5\mathbb{E}\|\hat{s}(\infty)\|_* + C_6 \leq C_7. \tag{34}$$

The uniform bound (34) on the expected norm in steady-state implies the tightness of stationary distributions. □

5 Diffusion scale analysis

5.1 Proof of Theorem 3

We will use the following strong approximation of unit-rate Poisson processes (see Lemma 3.1 in [18]).

Lemma 10. *A unit-rate Poisson process $(\Pi(t) : t \geq 0)$ can be realized on the same probability space as a standard Brownian motion $(B(t) : t \geq 0)$ such that the positive random variable ξ, given by*

$$\xi := \sup_{t \geq 0} \frac{|\Pi(t) - t - B(t)|}{\log(2 \vee t)} < \infty$$

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

We will also need the following Lemma 11. Its proof follows from standard arguments (see, e.g., [22]) and thus is omitted.

Lemma 11. *Consider the mapping $\Phi : D^2 \to D^2$ that takes $(\phi_1, \phi_2) \in D^2$ into $(\psi_1, \psi_2) \in D^2$ determined by the integral representation: for each $t \geq 0$,*

$$\begin{aligned} \psi_1(t) &= \phi_1(t) + \beta \int_0^t \psi_2(s)ds, \\ \psi_2(t) &= \phi_2(t) - \beta\gamma \int_0^t \psi_2(s)ds - \epsilon \int_0^t \psi_1(s)ds, \end{aligned} \tag{35}$$

where β, γ, ϵ are constants. Equations (35) determine (ψ_1, ψ_2) uniquely. The mapping Φ is continuous in the topology of u.o.c. convergence.

Proof of Theorem 3. Since $(\hat{Y}^r(0), \hat{X}^r(0)) \to (\hat{Y}(0), \hat{X}(0))$, we know that the fluid scaled processes $(\bar{Y}^r, \bar{X}^r)$ are such that $(\bar{Y}^r(0), \bar{X}^r(0)) \to (0,0)$. Then by Theorem 1, w.p.1, $(\bar{Y}^r(t), \bar{X}^r(t))$ converges u.o.c. to the fluid limit $(y(t), x(t)) \equiv (0,0)$ as $r \to \infty$. Therefore, w.p.1,

$$\bar{Y}^r(t) \to 0, \quad \bar{X}^r(t) \to 0, \quad \int_0^t |\bar{Y}^r(s)| ds \to 0, \quad \text{and} \quad \int_0^t |\bar{X}^r(s)| ds \to 0, \quad \text{u.o.c.} \tag{36}$$

This obviously implies that, w.p.1, on any finite time interval $[0,T]$, the boundary of $\hat{X}^r$ is not hit for sufficiently large r, i.e. $\hat{X}^r(t) > -\lambda\sqrt{r}/\beta$. Using representation (17) – (19) of the processes (Y^r, X^r) (with $Z^r = X^r$) and Lemma 10, after some manipulation, we see that there exist independent standard Brownian motions B_i, $i = 1, 2$, corresponding to the driving unit-rate Poisson processes N_i, $i = 1, 2$, all constructed on the same probability space, such that in the time interval $t \in [0, T]$, w.p.1,

$$\hat{Y}^r(t) = \hat{Y}^r(0) + \int_0^t \beta \hat{X}^r(s) ds + B_2(\lambda t) - B_1(\lambda t) + \Delta_1^r(t), \tag{37}$$

$$\hat{X}^r(t) = \hat{X}^r(0) - \gamma \int_0^t \beta \hat{X}^r(s) ds - \epsilon \int_0^t \hat{Y}^r(s) ds + \gamma B_1(\lambda t) - \gamma B_2(\lambda t) + \Delta_2^r(t), \tag{38}$$

where

$$\sup_{[0,T]} |\Delta_1^r(t)| = o(1), \qquad \sup_{[0,T]} |\Delta_2^r(t)| = o(1).$$

Letting $r \to \infty$ and applying Lemma 11 we obtain the probability 1, u.o.c., convergence of $(\hat{Y}^r, \hat{X}^r)$ to a limit diffusion process $(\hat{Y}, \hat{X})$ which satisfies (12). This implies the claimed convergence in distribution.

Finally, the transient and stationary distributions of the limit diffusion process $(\hat{Y}, \hat{X})$ follow directly from linear SDEs; see Chapter 5.6 in [16]. This completes the proof of Theorem 3. □

5.2 Comments on the proof of Theorem 4

This proof – of the tightness of stationary distributions on diffusion scale – can be obtained by adapting the approach developed in [25, 26]. That approach uses three steps. The first step is to show the fluid-scale (r-scale) tightness of the stationary distributions of the process; in our context, it means proving that the stationary distributions of $r^{-1}(Y^r, X^r - \lambda r/\beta)$ are tight and asymptotically concentrate at $(0,0)$. We have done this step in Theorem 2. The second step establishes the $r^{1/2+\kappa}$-scale tightness, for any $\kappa \in (0, 1/2)$; namely, tightness of stationary distributions of $r^{-1/2-\kappa}(Y^r, X^r - \lambda r/\beta)$. The argument of this step uses fluid-scale tightness as a starting point, and follows that in [25]. The final step, is to show the diffusion-scale ($r^{1/2}$-scale) tightness; it uses $r^{1/2+\kappa}$-scale tightness as a starting point, and follows the argument analogous to that in [26].

6 Numerical Examples

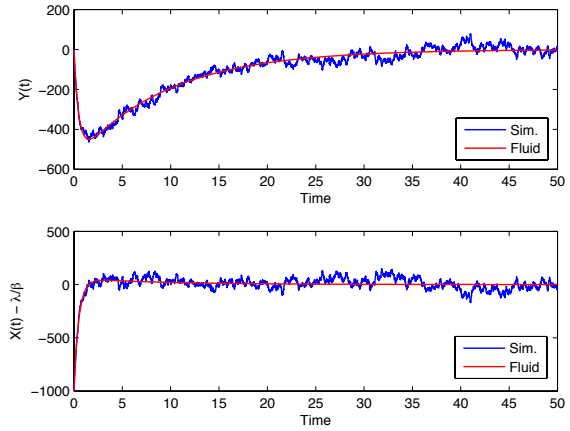
In this section, we present some numerical examples to illustrate the behavior and performance of feedback agent invitation schemes, as well as accuracy of the approximations given by our theoretical results.

Let us provide a general guidance on the setting of algorithm parameters. For the (asymptotic) optimality of Scheme B, this setting does *not* need to be precise. The rule of thumb is that ϵ should be "as large as possible subject to condition $\epsilon < \beta\gamma^2/4$." (This is because the smaller the ϵ, the larger the system convergence time to the stationary point.) For example, we can pick some reasonable value of γ, say 1 or 2. Then, we can use some "ballpark" estimate for β. If this estimate is not very reliable, we set ϵ with large safety margin, say $\epsilon = (1/5)\beta\gamma^2/4$. If the estimate of β is considered more reliable, the choice of ϵ can be more aggressive, say $\epsilon = (1/2)\beta\gamma^2/4$.

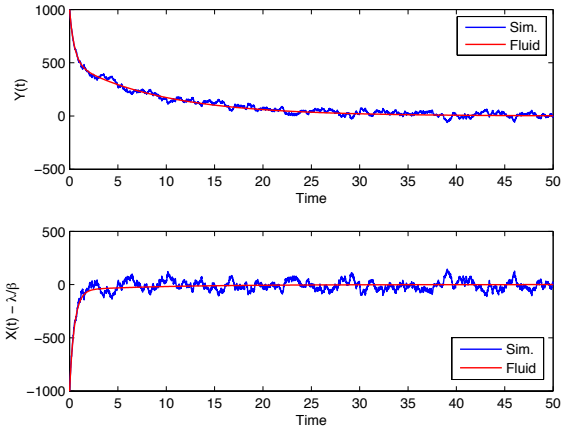
First, we simulate Scheme B. In the numerical examples, we use the following set of parameters:

$$\Lambda = 1000, \quad \beta = 1, \quad \gamma = 2, \quad \epsilon = 0.2,$$

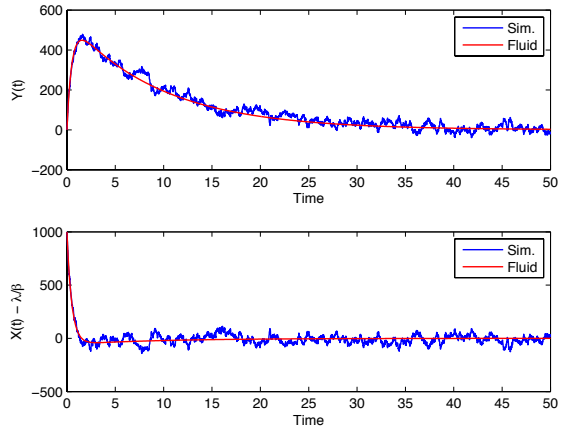
We consider four initial conditions: (i) $(Y(0), X(0)) = (0, 0)$; (ii) $(Y(0), X(0)) = (1000, 0)$; (iii) $(Y(0), X(0)) = (0, 2000)$; and (iv) $(Y(0), X(0)) = (-1000, 2000)$. In each case, we conduct a simulation experiment of the system up to time 50 with these different initial conditions. The comparisons in these cases are shown in Figure 2. We observe that the feedback scheme does bring the system (close) to its desired operating point, and fluid limit provides a very good approximation of the system trajectory.



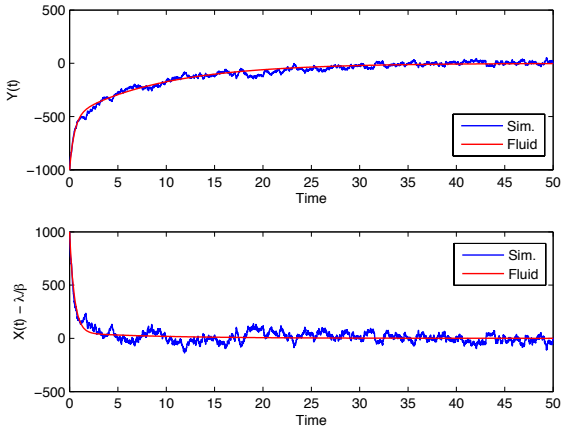
(a) Initial State $(Y(0), X(0)) = (0, 0)$



(b) Initial State $(Y(0), X(0)) = (1000, 0)$



(c) Initial State $(Y(0), X(0)) = (0, 2000)$



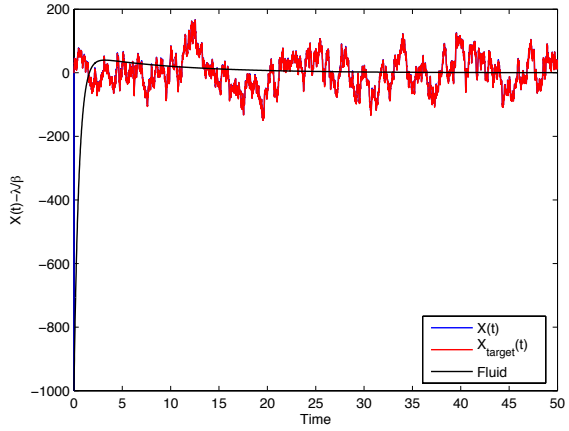
(d) Initial State $(Y(0), X(0)) = (-1000, 2000)$

Figure 2: Scheme B. Comparison of fluid approximations with simulations in one sample path.

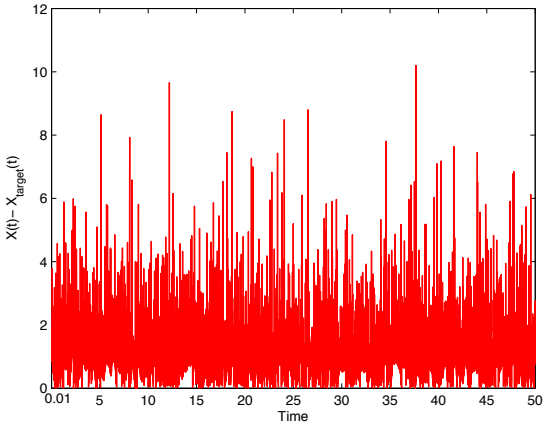
Second, we conduct a simulation experiment for Scheme A with the following parameter values

$$\Lambda = 1000, \quad \beta = 1, \quad \tilde{\beta} = 1, \quad \gamma = 2, \quad \epsilon = 0.2,$$

where $\tilde{\beta}$ is the rejection rate of invitations, and the initial state is $(Y(0), X(0), X_{target}(0)) = (0, 0, 1000)$. The results are shown in Figure 3(a). We see that the magnitude of the difference between X_{target} and the actual number of invited agents X is very small (except at time 0) and can be regarded negligible compared to their scale. This explains why the trajectories (of both X_{target} and X) are well approximated by the fluid trajectory, obtained for the Scheme B. This in turn provides a validation of Scheme B as an approximation of simpler and more practical Scheme A.



(a) Fluid Model v.s. $X(t)$ and $X_{target}(t)$



(b) $X(t) - X_{target}(t))$

Figure 3: Scheme A.

7 Time-Varying Customer Arrivals

So far, we have considered systems in which customers arrive at a constant rate. In this section, we assume that customers arrive according to an inhomogeneous Poisson process. Both Scheme A and Scheme B introduced in section 2 naturally work for time-varying customer arrivals. Here we give a fluid limit result for Scheme B. We consider a sequence of systems (Y^r, X^r), with the arrival rate function in the r-th system being $r\lambda(t),\ t \geq 0$, where $\lambda(\cdot)$ is a locally-bounded piece-wise continuous function with at most finite number of jumps on finite intervals. Define fluid-scaled processes

$$(\tilde{Y}^r, \tilde{X}^r) := r^{-1}(Y^r, X^r). \tag{39}$$

Note that the fluid-scaled process $\tilde{X}^r$ is not centered.

Theorem 12. *Suppose that* $(\tilde{Y}^r(0), \tilde{X}^r(0)) \to (\tilde{y}(0), \tilde{x}(0))$ *for some fixed* $(\tilde{y}(0), \tilde{x}(0)) \in \mathbb{R}^2$, $\tilde{x}(0) \geq 0$. *Then the processes can be constructed on a common probability space, so that the following holds. W.p.1, any subsequence of* r *has a further subsequence, along which*

$$(\tilde{Y}^r, \tilde{X}^r) \to (\tilde{y}, \tilde{x}) \quad u.o.c. \quad as \quad r \to \infty, \tag{40}$$

where $(\tilde{y}, \tilde{x})$ *is a locally Lipschitz trajectory, such that at any regular point* $t \geq 0$ *the following holds: if* $\tilde{x}(t) > 0$,

$$\begin{aligned} \tilde{y}'(t) &= \beta\tilde{x}(t) - \lambda(t), \\ \tilde{x}'(t) &= \gamma\lambda(t) - \gamma\beta\tilde{x}(t) - \epsilon\tilde{y}(t), \end{aligned} \tag{41}$$

and if $\tilde{x}(t) = 0$,

$$\begin{aligned} \tilde{y}'(t) &= -\lambda(t), \\ \tilde{x}'(t) &= [\gamma\lambda(t) - \epsilon\tilde{y}(t)] \vee 0. \end{aligned} \tag{42}$$

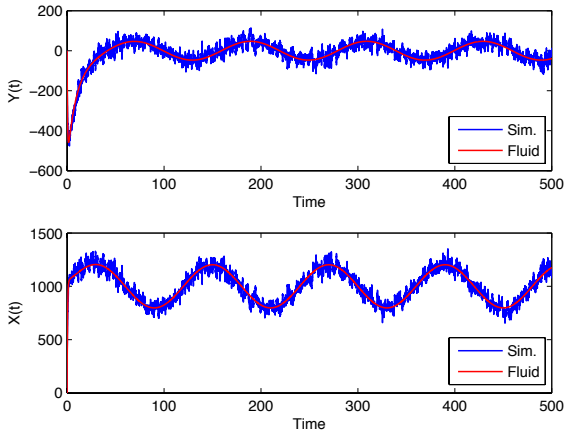
Note that the processes (Y^r, X^r) can be represented from the equations (17) - (19) with $N_1(\lambda rt)$ being replaced by $N_1\left(r\int_0^t \lambda(s)ds\right)$. Given that, the proof of this theorem is a straightforward generalization of the corresponding argument in the proof of Theorem 1. We omit details. Also note that Theorem 12 does not claim uniqueness of the fluid limit trajectory. However, the uniqueness easily follows in many cases of interest. For example, when $\lambda(\cdot)$ is piecewise constant (then the same argument as in the proof of Theorem 1 applies to each of the "pieces"), or when the solution to (41) never hits the $\tilde{x} = 0$ boundary.

7.1 Numerical Example

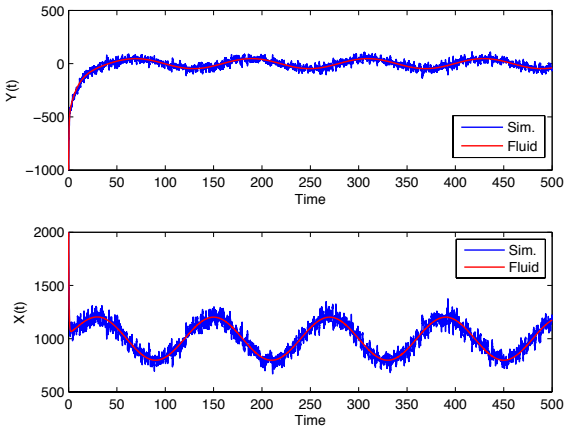
We conduct a simulation experiment with the following set of parameters:

$$\Lambda(t) = 1000 + 200\sin(2\pi t/120), \quad \beta = 1, \quad \gamma = 2, \quad \epsilon = 0.2,$$

and consider two initial conditions: (i) $(Y(0), X(0)) = (0, 0)$ and (ii) $(Y(0), X(0)) = (-1000, 2000)$. In each case, we conduct a simulation experiment of the system up to time 500. See Figure 4 for the comparisons. We observe that the fluid limit trajectory (which is unique in this case) provides a very good approximation of the system dynamics. We also observe that although the values of Y do not converge to zero, they fluctuate around zero at a smaller scale than the system scale (comparing the magnitude of Y, less than 100, with that of the customer arrival rates, 1000).



(a) Initial State $(Y(0), X(0)) = (0, 0)$



(b) Initial State $(Y(0), X(0)) = (-1000, 2000)$

Figure 4: Scheme B. Comparison of fluid approximations with simulations in one sample path when customer arrival rates are time-varying

8 Discussion and further work

We have proposed a new stochastic model and the algorithm for a service system where agents are invited on demand. The key feature of the algorithm is that it is very robust, requires essentially no knowledge of system parameters and easily adapts to parameter changes. It is also very easy to implement. To study its performance we prove the asymptotic results on the fluid and diffusion scale. They, in particular, demonstrate the desirable performance of the scheme: both customer and agent waiting delays vanish as the system scale increases to infinity.

Many more operational challenges for such systems remain open. In this paper, we assumed that there are infinitely many agents available to be invited. This assumption is adequate, if the pool of potential agents is large. In many real cases, however, the agent pool size is not large enough and therefore needs to be taken into account. We also considered the model with single class of customers and single class of agents. When there are multiple classes of customers and/or agent pools, the design of an efficient invitation scheme presents new challenges. Finally, in the model of this paper the service times are irrelevant; however, the service time cannot be ignored the agent pools are finite. Exploring these and other related new problems may be a subject of future work.

In addition, our agent invitation scheme may be useful for the new generation of *cloud-based* call centers. For instance, companies like Arise and LiveOps are providing platforms for businesses to implement their cloud-

based call centers [4, 8, 14, 20]. The management of agents in a cloud-based system presents new operational challenges, in particular, how to guarantee the availability of qualified agents at any time. In some systems, the manager sends invitation requests to qualified agents in order to meet the demand and satisfy the service level agreements. It will be interesting to investigate if our agent invitation scheme can be potentially used in cloud-based services in future work.

Finally, in telemedicine operations, doctors and other medical specialists are valuable (and expensive) resources. Operational guidelines for telehealth services are provided by the American Telemedicine Association [2]. The quality of care delivered via telehealth systems requires timely interactions between health providers and patients - neither health providers nor patients would wait too long. Thus it will also be interesting to study in future work if our "agent" invitation scheme can be potentially useful for telehealth management.

References

[1] Aksin, Z., Armony, M. and Mehrotra, V. (2007) The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management.* 16, 665–688.

[2] American Telemedicine Association. (2014) *Core Operational Guidelines for Telehealth Services Involving Provider-Patient Interactions.* `http://www.americantelemed.org/docs/default-source/standards/core-operational-guidelines-for-telehealth-services.pdf?sfvrsn=6`.

[3] Azriel, D., Feigin, P. D. and Mandelbaum, A. (2014) Erlang S: A data-based model of servers in queueing networks. *Working paper.*

[4] Bengtson, S. (2014) Generating better results with crowdsourcing: Leverage a network of high-quality professionals for customer service. *White paper.* `http://www.arise.com/`

[5] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Stat. Ass.* Vol. 100, No, 469, 36–50 (2005)

[6] Csögö, M. and Horváth, L. (1993) *Weighted Approximations in Probability and Statistics.* Wiley.

[7] Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes: Characterization and Convergence.* Wiley.

[8] Formisano, P. (2014) Flexibility for changing business needs: Improve customer service and drive more revenue with a virtual crowdsourcing solution. *White paper.* `http://www.arise.com/`

[9] Gamarnik, D. and Goldberg, D. (2013) Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. To appear in *Annals of Applied Probability.*

[10] Gamarnik, D. and Momcilovic, P. (2008) Steady-state analysis of a multi server queue in the Halfin-Whitt regime. *Advances in Applied Probability.* 40, 548–577.

[11] Gamarnik, D. and Stolyar, A. L. (2012) Multiclass multi server queueing system in the Halfin-Whitt heavy-traffic regime: asymptotics of the stationary distribution. *Queueing Systems.* 71, 25–51.

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

[13] Gershwin, S. B. (2010) Lecture notes on inventory. `http://ocw.mit.edu/courses/mechanical-engineering/2-854-introduction-to-manufacturing-systems-fall-2010/lecture-notes/MIT2_854F10_inv.pdf`.

[14] Gurvich, I., Lariviere, M., and Moreno-Garcia, A. (2013) Staffing service systems when capacity has a mind of its own. *Working paper.*

[15] Gurvich, I. and Ward, W. (2014) On the dynamic control of matching queues. *Working paper*.

[16] Karatzas, I. and Shreve, S. (1996) *Brownian Motion and Stochastic Calculus.* 2nd edition. Springer.

[17] Kashyap, B. R. K. (1966) The double-ended queue with bulk service and limited waiting space. *Operations Research.* 14(5). 822–834.

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

[19] Liu, X., Gong, Q. and Kulkarni, V. G. (2014) Diffusion models for doubly-ended queues with renewal arrival processes. *Working paper*.

[20] McGee-Smith, S. (2010) Why Companies Are Choosing to Deploy the LiveOps Cloud-Based Contact Center. `http://www.liveops.com/sites/default/files/uploads/lo_wp_mcgee-smith_analytics.pdf`

[21] Meyn, S. P. and Tweedie, R. L. (2009) *Markov Chains and Stochastic Stability.* Cambridge University Press. 2nd edition.

[22] Pang, G., Talreja, R. and Whitt, W. (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys.* **4**, 193–267.

[23] Stolyar, A. L. Control of End-to-End Delay Tails in a Multiclass Network: LWDF Discipline Optimality. *Annals of Applied Probability*, 2003, Vol.13, No.3, pp.1151-1206.

[24] Stolyar, A.L., Reiman, M.I., Korolev, N., Mezhibovsky, V. and Ristock, H. Pacing in Knowledge Worker Engagement. *United States Patent Application 20100266116-A1*, October 2010.

[25] Stolyar, A.L. and Yudovina, E. (2012) Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. *Stochastic Systems*, 2012, Vol.2, No.2, pp.381-408. http://arxiv.org/abs/1201.2978

[26] Stolyar, A.L. (2013) Diffusion scale tightness of invariant distributions of a large-scale flexible service system. *Advances in Applied Probability*, 2015, Vol.47, No.1, to appear. http://arxiv.org/abs/1301.5838

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