A queueing system with on-demand servers: local stability of fluid limits

Lam M. Nguyen · **Alexander L. Stolyar**

Abstract We study a system, where a random flow of customers is served by servers (called agents) invited on-demand. Each invited agent arrives into the system after a random time; after each service completion, an agent returns to the system or leaves it with some fixed probabilities. Customers and/or agents may be impatient, that is, while waiting in queue, they leave the system at a certain rate (which may be zero). We consider the queue-length-based feedback scheme, which controls the number of pending agent invitations, depending on the customer and agent queue lengths and their changes. The basic objective is to minimize both customer and agent waiting times.

We establish the system process fluid limits in the asymptotic regime where the customer arrival rate goes to infinity. We use the machinery of switched linear systems and common quadratic Lyapunov functions to approach the stability of fluid limits at the desired equilibrium point, and derive a variety of sufficient local stability conditions. For our model, we conjecture that local stability is in fact sufficient for global stability of fluid limits; the validity of this conjecture is supported by numerical and simulation experiments. When local stability conditions do hold, simulations show good overall performance of the scheme.

Lam M. Nguyen
Department of Industrial and Systems Engineering,
Lehigh University, Bethlehem, PA 18015, USA
E-mail: LamNguyen.MLTD@gmail.com

Alexander L. Stolyar
Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: stolyar@illinois.edu

punov function

Mathematics Subject Classification 60K25 · 60F17 · 90B15

1 Introduction

Consider a service system where a random flow of customers arrive exogenously. Servers, called *agents*, can be invited on-demand at any time. Invited agents arrive into the system not immediately, but after a random delay. When a customer is matched with an agent, a service occurs. After a service completion the customer certainly leaves the system, while the agent can either leave or, with some fixed probability, return to serve more customers. Customers and/or agents may be impatient, that is, they abandon the system if their wait in queue exceeds some random *patience time*. The basic system control objective is to keep waiting times of both customers and agents small. (The system is schematically shown on Figure 1 in Section 2, where the model is defined formally.) In this paper we make Markov assumptions: the customer arrival process is Poisson; the random agent delays (from invitation to arrival) are i.i.d. with exponential distribution, and so are the service times and patience times for customers and for agents.

This model is a generalized version of that in [13,14]. In [14], there is no abandonment for both queues, and agents always leave the system after service completions. The model in [13] also has no abandonment, but, like in our model, an agent may return to the system after a service completion. Thus, our model is more realistic in many scenarios because customer abandonment is a key factor for call center operations (see e.g. [6,21]).

The model is primarily motivated by call/contact centers (see [20]), where agents that we consider are highly skilled. It is not reasonable to set a fixed working schedule for these agents since their time is very valuable. Instead, they are invited on-demand in real time. The purpose is to design a real-time adaptive agent invitation scheme that minimizes customer and agent waiting times. However, designing an effective, simple and robust agent invitation strategy is non-trivial due to randomness in agent behavior.

We study a feedback-based adaptive scheme of [20,14,13], called *queue-length-based feedback scheme*, which controls the number of pending agent invitations, depending on the customer and/or agent queue lengths and their changes. The key feature of this scheme, making it very attractive for applications, is that it requires *no a priori knowledge of the customer arrival rate* – the scheme adjusts "automatically." (See [14] for a detailed discussion of this feature.)

The algorithm analysis in this paper is substantially more challenging due to greater generality of our model. Just like in [13,14], we consider a "stylized" version of the invitation scheme to make the analysis more tractable. Our simulation experiments in section 7.2 show that the behavior of the stylized

scheme is very close to that of the more practical version of the queue-length-based feedback scheme.

We consider the system in the asymptotic regime where the customer arrival rate goes to infinity while the distributions of the agent response times, the service times and the patience times are fixed. We show convergence of the fluid-scaled process to the fluid limit (Theorem 1), which satisfies a system of differential equations. The key property of interest is the convergence of the fluid limit trajectories to the equilibrium point (at which the queues are zero). This property is referred to as *global stability* of the fluid limits. Establishing global stability appears to be very challenging, due to the fact that fluid limits have complicated behavior – there are two domains where they follow different ODEs, and a "reflecting" boundary. In this paper, we focus on the *local stability* of fluid limits, defined as the stability of the dynamic system which describes fluid limit trajectories away from the boundary.

The **main result** of this paper (Theorem 2) gives sufficient local stability conditions; the proof uses the machinery of switched linear systems and common quadratic Lyapunov functions [10, 19]. Theorem 2 implies many useful sufficient local stability conditions (Corollaries 1 - 12) for special cases, including those where customers never abandon or agents certainly leave the system after service completions. (Some of these corollaries – namely, Corollaries 9, 10 and 12 – strengthen the results in [13] for the non-abandonment system.) These sufficient local stability conditions are robust and easy to achieve in practice.

We **conjecture** that, for our model, local stability is in fact sufficient for global stability, based on a large number of numerical and simulation experiments. In turn, global stability of fluid limits is a crucially important system property – this will be discussed in Section 7.4. Finally, our simulation experiments show good overall performance of the feedback scheme when the local stability conditions do hold.

The model has many applications, or potential applications. For a general discussion of modern call/contact centers and their management, see, e.g. [1, 12]. Another example is telemedicine [2], where "agents" are doctors, invited on-demand to serve patients remotely. The model also arises in other applications, such as crowdsourcing-based customer service (see e.g. [5,3]), taxi-service system, buyers and sellers in a trading market, and assembly systems. The model has relation to classical assemble-to-order models, where customers are orders and "invited agents" are products, which cannot be produced/assembled instantly. The model is also related to "double-ended queues" (see e.g. [9,11]) and matching systems (see e.g. [7]); although in such models arrivals of all types into the system are typically exogenous, as opposed to being controlled.

Paper organization. The rest of the paper is organized as follows. In Section 2, we describe the model and algorithm in detail. Section 3 gives the necessary background on switched linear systems and common quadratic Lyapunov functions (with the necessary results given in Appendix). Section 4 states the main results of the paper, which are proved in Sections 5 and

6. Section 7 provides numerical and simulation experiments; it also contains our conjectures about global and local stability of fluid limits, as well as their implications. A discussion of the results and future work is in Section 8.

Basic notation: Symbols $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{R}_+$ denote the sets of natural, integer, real, real non-negative numbers, respectively. $\mathbb{R}^d$ denotes the d-dimensional vector space. $\mathbb{R}^{d \times d}$ denotes the set of all $d \times d$ real matrices. The standard Euclidean norm of a vector $x \in \mathbb{R}^n$ is denoted $\|x\|$. For a vector a and matrix A, we write their transposes as a^T and A^T, respectively. For a matrix A, we write its inverse and determinant as A^{-1} and $\det(A)$, respectively. We write $x(\cdot)$ to mean the function (or random process) $(x(t), t \geq 0)$. 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, and for $x(\cdot) : \mathbb{R}_+ \to \mathbb{R}^d$, $(d/dt)x(t) = (x'_1(t), \ldots, x'_d(t))$. For $x \in \mathbb{R}$, $x^+ = \max\{x, 0\}$ and $x^- = -\min\{x, 0\}$; and $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = 0$ if $x = 0$, and $\operatorname{sgn}(x) = -1$ if $x < 0$. For $x, y \in \mathbb{R}$, we denote $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. $a \Leftrightarrow b$ means "a is equivalent to b"; $a \Rightarrow b$ means "a implies b". We write $x^r \to x \in \mathbb{R}^n$ to denote ordinary convergence in $\mathbb{R}^n$. 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 $\mathbb{N}_0 \subseteq \mathbb{N}$, the derivatives

$$\frac{d}{dt} \max_{n \in \mathbb{N}_0} f_n(t) \text{ and } \frac{d}{dt} \min_{n \in \mathbb{N}_0} f_n(t)$$

exist. (To be precise, we require that each derivative is proper: both left and right derivatives exist and are equal.)

Abbreviations: *u.o.c.* means *uniform on compact sets* convergence of functions, with the argument determined by the context (usually in $[0, \infty)$); *w.p.1* means *with probability 1*; *i.i.d.* means *independent identically distributed*; RHS means *right hand side*; FSLLN means *functional strong law of large numbers*; *CQLF* means *common quadratic Lyapunov function*; *LTI system* means *linear time-invariant system*.

2 Model and algorithm

2.1 Model

Our model is a generalization of that considered in [13, 14]. 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. There is an infinite pool of 'potential' agents, which can be invited to serve customers. After a potential agent is invited, it becomes a 'pending' agent; we refer to such an event as an *invitation*. A pending agent 'accepts' its invitation and becomes 'active' agent after a random, exponentially distributed, time with mean $1/\beta$; we refer to such an event as an *acceptance*. Upon acceptance events, the new active agents join the (active) agent queue. The customer and agent queues cannot be positive simultaneously: the head-of-the-line customer and agent are immediately matched, leave their queues, and together go to

service. Each service time is an exponentially distributed random variable with mean $1/\mu$; after the service completion, the customer leaves the system, while the corresponding agent either remains active and rejoins the agent queue – this occurs with probability $\alpha \in [0, 1)$ – or leaves the system with probability $1 - \alpha$. Thus, there are two ways in which agents join the queue – an agent becomes active (upon acceptance event) or an already active agent rejoins the queue after service completion. The patience times of customers and agents are independent sequences of i.i.d. exponential random variables with rate $\delta \geq 0$ and $\theta \geq 0$, respectively. When its patience time expires while a customer or server wait in queue, they leave the system. (The model in [13] is a special case of ours, with $\delta = 0$ and $\theta = 0$; in other words, customers and agents certainly wait in their queues until they are matched. The model in [14] is a special case of ours, with $\delta = 0$, $\theta = 0$ and $\alpha = 0$.) Figure 1 depicts such a system.

Let $X(t)$ be the number of pending agents at time t. Let $Y(t) = Q_a(t) - Q_c(t)$ be the difference between the agent and customer queue lengths at time t. (Note that $Q_a(t) = Y^+(t)$ and $Q_c(t) = Y^-(t)$.) Let $Z(t)$ be the number of customers (or agents) in service at time t. The system state at time t is $(X(t), Y(t), Z(t))$.
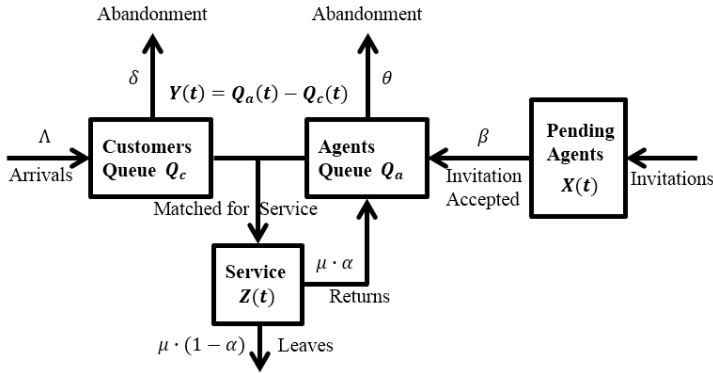

Fig. 1: An agent invitation system

2.2 Algorithm

We study the queue-length-based feedback scheme proposed in [20] and analyzed (for less general systems) in [13,14]. The key feature of this scheme, making it very attractive for applications, is that it requires *no a priori knowledge of the customer arrival rate* – the scheme adjusts "automatically." (See [14,20] for a detailed discussion of this feature.)

The scheme, referred to as the *actual scheme*, maintains a "target" $X_{target}(t)$ for the number of pending agents $X(t)$. $X_{target}(t)$ 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 $\Delta Y(t)$ (+1 or −1), where $\gamma > 0$ and $\epsilon > 0$ are the algorithm parameters and Δt is the time duration from the previous change of Y. New agent invitations occur (i.e., the number of pending agents increases) if and only if $X(t) < X_{target}(t)$, where $X(t)$ is the actual number of pending agents; therefore, $X(t) \geq X_{target}(t)$ holds at all times. In addition, $X_{target}(t) \geq 0$; i.e. if an update of $X_{target}(t)$ makes it negative, its value is immediately reset to zero. Note that $X_{target}(t)$ is not necessarily an integer.

Just like in [13,14], to simplify our theoretical analysis, we consider a "stylized" version of the actual scheme, referred to as the *stylized scheme*, 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 agents can be invited instantly, but pending agents can be removed from the system at any time. Formally, the stylized scheme is defined as follows. There are six types of mutually independent, and independent of the past, events that affect the dynamics of $X(t)$, $Y(t)$ and $Z(t)$ in a small time interval $[t, t + dt]$:

- a customer arrival with probability $\Lambda dt + o(dt)$,
- an acceptance with probability $\beta X(t) dt + o(dt)$,
- an additional event (we will call it a type-3 event) with probability $\epsilon |Y(t)| dt + o(dt)$; unlike other events, it is triggered by the algorithm itself, as opposed to other events triggered by customers' and/or agents' "movement" in the system,
- a service completion with probability $\mu Z(t) dt + o(dt)$,
- an abandonment in the customer queue with probability $\delta Y^-(t) dt + o(dt)$,
- an abandonment in the agent queue with probability $\theta Y^+(t) dt + o(dt)$.

  The changes at these event times are described as follows:

- Upon a customer arrival, if $Y(t) > 0$, $Z(t)$ changes by $\Delta Z(t) = 1$; and if $Y(t) \leq 0$, $Z(t)$ changes by $\Delta Z(t) = 0$. $Y(t)$ changes by $\Delta Y(t) = -1$, and $X(t)$ changes by a random quantity with average $\gamma > 0$. For example, if $\gamma = 1.7$ and $\Delta Y(t) = -1$, then $\Delta X(t) = 2$ with probability 0.7 and $\Delta X(t) = 1$ with probability 0.3. Note that if γ is integer, $\Delta X(t) = \gamma$ w.p.1. To simplify the exposition, we assume that $\gamma > 0$ is an integer.
- Upon an acceptance event, if $Y(t) < 0$, $Z(t)$ changes by $\Delta Z(t) = 1$; and if $Y(t) \geq 0$, $Z(t)$ changes by $\Delta Z(t) = 0$. $Y(t)$ changes by $\Delta Y(t) = 1$, and $X(t)$ changes by $\Delta X(t) = -(\gamma \wedge X(t))$, that is, the change is by $-\gamma$ but $X(t)$ is kept to be nonnegative.
- Upon a type-3 event, if $X(t) \geq 1$, the change $\Delta X(t) = -\text{sgn}(Y(t))$ occurs; and if $X(t) = 0$, the change $\Delta X(t) = 1$ occurs if $Y(t) < 0$ and $\Delta X(t) = 0$ if $Y(t) \geq 0$.
- Upon a service completion, (a) if the agent returns to the agent queue (with probability α), then if $Y(t) < 0$, the change $\Delta Z(t) = 0$ occurs; and if $Y(t) \geq 0$, the change $\Delta Z(t) = -1$ occurs; $Y(t)$ changes by $\Delta Y(t) = 1$, and $\Delta X(t) = -(\gamma \wedge X(t))$. (b) If the agent leaves the system (with probability $1 - \alpha$), then $Z(t)$ changes by $\Delta Z(t) = -1$.

- Upon a customer abandonment, $Y(t)$ changes by $\Delta Y(t) = 1$, and $X(t)$ changes by $\Delta X(t) = -(\gamma \wedge X(t))$.
- Upon an agent abandonment, $Y(t)$ changes by $\Delta Y(t) = -1$, and $X(t)$ changes by $\Delta X(t) = \gamma$.

Let $V(t) = Y^+(t) + Z(t)$ be the total number of agents in the system at time t. Obviously, $(X(t), Y(t), V(t))$ is a random process with states being 3-dimensional integer vectors. However, *very informally*, the basic dynamics of $(X(t), Y(t), V(t))$ under the stylized scheme can be thought of as described by the following ODE

$$\begin{cases} (d/dt)X = -\gamma(d/dt)Y - \epsilon Y \\ (d/dt)Y = \beta X - \Lambda + \alpha\mu Z + \delta Y^- - \theta Y^+ \\ (d/dt)V = \beta X - (1-\alpha)\mu Z - \theta Y^+. \end{cases} \tag{1}$$

ODE (1) is only to provide the basic intuition for the system dynamics – it is not used in the analysis.

3 Switched linear systems and common quadratic Lyapunov functions (CQLF)

This section provides definitions and background on switched linear systems and common quadratic Lyapunov functions, which we need to state our main results. (Some results that will be used in the proofs are given in Appendix.)

Consider a *switched linear system*

$$\Sigma_S : u'(t) = A(t)u(t) \ , \ A(t) \in \mathcal{A} = \{A_1, \ldots, A_m\} \tag{2}$$

where $\mathcal{A}$ is a set of matrices in $\mathbb{R}^{n \times n}$, and $t \to A(t)$ is a mapping from non-negative real numbers into $\mathcal{A}$. (Usually, as in [19], this mapping is required to be piecewise constant with only finitely many discontinuities in any bounded time-interval. In our case this additional condition is not important, because our switched system will have a continuous derivative; see equation (7) below.) For $1 \leq i \leq m$, the ith constituent system of the switched linear system (2) is the *linear time-invariant (LTI) system*

$$\Sigma_{A_i} : u'(t) = A_i u(t). \tag{3}$$

The origin is an *exponentially stable equilibrium* of the switched linear system Σ_S if there exist real constants $C > 0$, $a > 0$ such that $\|u(t)\| \leq Ce^{-at}\|u(0)\|$ for $t \geq 0$, for all solutions $u(t)$ of the system (2) (see [8, 19]).

A symmetric square $n \times n$ matrix M with real coefficients is *positive definite* if $z^T M z > 0$ for every non-zero column vector $z \in \mathbb{R}^n$. A symmetric square $n \times n$ matrix M with real coefficients is *negative definite* if $z^T M z < 0$ for every

non-zero column vector $z \in \mathbb{R}^n$. A square matrix A is called a *Hurwitz matrix* (or *stable matrix*) if every eigenvalue of A has strictly negative real part.

The function $V(u) = u^T P u$ is a *quadratic Lyapunov function* (QLF) for the system $\Sigma_A : u'(t) = Au(t)$ if (i) P is symmetric and positive definite, and (ii) $PA + A^T P$ is negative definite. Let $\{A_1, \ldots, A_m\}$ be a collection of $n \times n$ Hurwitz matrices, with associated stable LTI systems $\Sigma_{A_1}, \ldots, \Sigma_{A_m}$. Then the function $V(u) = u^T P u$ is a *common quadratic Lyapunov function* (CQLF) for these systems if V is a QLF for each individual system (see [10, 19]).

4 Main results

We consider a sequence of systems, indexed by a scaling parameter $r \to \infty$. In the system with index r, the arrival rate is $\Lambda = \lambda r$, while the parameters α, β, μ, δ, θ, ϵ, γ do not depend on r. The corresponding process is $(X^r(t), Y^r(t), Z^r(t)), t \geq 0$. The desired system operating point, at which $(X^r(t), Y^r(t), Z^r(t))$ should be centered is given by $(\lambda r(1-\alpha)/\beta, 0, \lambda r/\mu)$. The explanation of this choice is as follows. If an invitation scheme works as desired, $Y^r(t)$ should be close to 0; the number of customer-agent pairs $Z^r(t)$ should be close to its average value, which is $\lambda r/\mu$, so that the customers leave the system at rate λr; finally, $X^r(t)$ should be close to the value χ, such that the total average rate at which agents join the agent queue, which is $\chi\beta + [(\lambda r)/\mu]\mu\alpha$, is equal to the customer arrival rate λr – this gives $\chi = \lambda r(1-\alpha)/\beta$. However, instead of considering process $(X^r(t), Y^r(t), Z^r(t))$, we will consider process $(X^r(t), Y^r(t), V^r(t))$, which is more convenient for the analysis. (Recall that $Z^r(t) = V^r(t) - (Y^r(t))^+$.) Then the natural centering value for $V^r(t)$ is same as for $Z^r(t)$, namely $\lambda r/\mu$. We define fluid-scaled process with centering as

$$(\bar{X}^r(t), \bar{Y}^r(t), \bar{V}^r(t)) = r^{-1}\left(X^r(t) - \frac{\lambda r(1-\alpha)}{\beta}, Y^r(t), V^r(t) - \frac{\lambda r}{\mu}\right), \quad t \geq 0.$$
(4)

Theorem 1. *Consider a sequence of processes* $(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot))$, $r \to \infty$, *with deterministic initial states such that* $(\bar{X}^r(0), \bar{Y}^r(0), \bar{V}^r(0)) \to (x(0), y(0), v(0))$ *for some fixed* $(x(0), y(0), v(0)) \in \mathbb{R}^3$, $x(0) \geq -\frac{\lambda(1-\alpha)}{\beta}$. *Then, these processes can be constructed on a common probability space, so that the following holds. W.p.1, from any subsequence of* r, *there exists a further subsequence such that*

$$(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot)) \to (x(\cdot), y(\cdot), v(\cdot)) \quad u.o.c. \quad as \quad r \to \infty$$
(5)

where $(x(\cdot), y(\cdot), v(\cdot))$ is a locally Lipschitz trajectory such that at any regular point $t \geq 0$

$$
\begin{cases}
x'(t) = \begin{cases} -\gamma y'(t) - \epsilon y(t), & \text{if } x(t) > -\frac{\lambda(1-\alpha)}{\beta} \\ [-\gamma y'(t) - \epsilon y(t)] \vee 0, & \text{if } x(t) = -\frac{\lambda(1-\alpha)}{\beta} \end{cases} \\
y'(t) = \beta x(t) + \alpha\mu(v(t) - y^+(t)) + \delta y^-(t) - \theta y^+(t) \\
v'(t) = \beta x(t) - (1-\alpha)\mu(v(t) - y^+(t)) - \theta y^+(t).
\end{cases}
\tag{6}
$$

A limit trajectory $(x(\cdot), y(\cdot), v(\cdot))$ specified in Theorem 1 will be called a *fluid limit* starting from $(x(0), y(0), v(0))$.

Remark. Equations (6), which a fluid limit must satisfy, are very natural. They can be thought of as rescaled centered versions of the (informal) equations (1). In addition, (6) includes a "reflection" (or, "regulation") at the boundary $x = -\frac{\lambda(1-\alpha)}{\beta}$, i.e. condition $x(t) \geq -\frac{\lambda(1-\alpha)}{\beta}$ is "enforced" as all times. This additional condition is the centered rescaled version of the condition $X^r(t) \geq 0$, which obviously must hold at all times.

Consider a dynamic system $(x(t), y(t), v(t)) \in \mathbb{R}^3$:

$$
\begin{cases}
x'(t) = -\gamma y'(t) - \epsilon y(t) \\
y'(t) = \beta x(t) + \alpha\mu(v(t) - y^+(t)) + \delta y^-(t) - \theta y^+(t) \\
v'(t) = \beta x(t) - (1-\alpha)\mu(v(t) - y^+(t)) - \theta y^+(t).
\end{cases}
\tag{7}
$$

This system describes the dynamics of fluid limit trajectories when the state is away from the boundary $x = -\frac{\lambda(1-\alpha)}{\beta}$. Note that, although the RHS of (7) is continuous, (7) is a switched linear system (with matrix $A(t)$ taking two different values A_1 and A_2, depending on whether $y \geq 0$ or $y < 0$). System (7) is a generalization of the system considered in [13], referred to as a *non-abandonment system*, which is a special case of ours with $\delta = 0$ and $\theta = 0$.

We say that the fluid limit is *globally stable* if every fluid limit trajectory converges to the equilibrium point $(0, 0, 0)$; and it is *locally stable* if every trajectory of the dynamic system (7) converges to the equilibrium point $(0, 0, 0)$. Note that exponential stability of the system (7) implies local stability.

The following theorem is the main result of this paper. It provides sufficient exponential stability conditions for the system (7).

Theorem 2 (Sufficient exponential stability conditions)**.** *For any set of positive β, μ, ϵ, γ, non-negative δ and θ, and $\alpha \in [0, 1)$, such that either* (i)

$$
\gamma > \max\left\{ \frac{\alpha\mu - \delta}{\beta}, \sqrt{\frac{(2-\alpha)\epsilon\mu + \alpha\epsilon\delta}{\beta\mu}} \right\},
\tag{8}
$$

or (ii)

$$
\gamma > \max\left\{ \frac{\alpha\mu - \delta + \sqrt{(\alpha\mu - \delta)^2 + 4\alpha\mu^2}}{2\beta}, \sqrt{\max\left\{ \frac{\alpha\epsilon(\delta - \mu)}{\beta\mu}, 0 \right\}} \right\}
\tag{9}
$$

holds, a common quadratic Lyapunov function (CQLF) of the system (7) exists, and the system (7) is exponentially stable.

In other words, conditions (8) and (9) are sufficient for local stability of our system. Theorem 2 implies the following useful sufficient local stability conditions (Corollaries 1 - 12) for special cases. Figure 2 depicts the connection between these results.
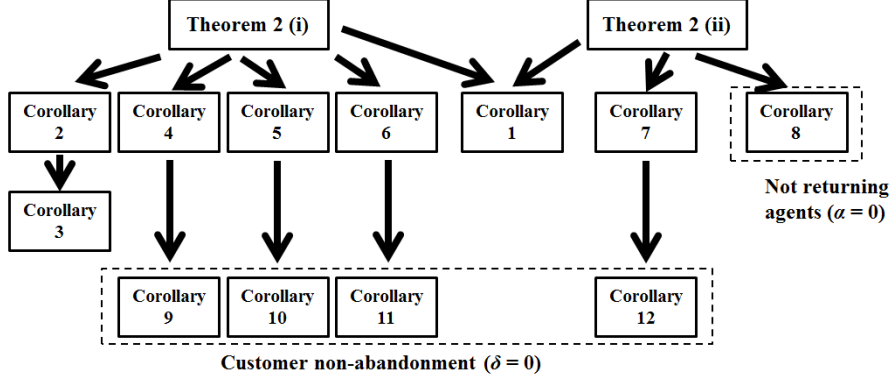

Fig. 2: Result's diagram

Corollary 1. *Given all other parameters are fixed, the system* (7) *is exponentially stable for all sufficiently large γ.*

Corollary 2. *If $\alpha\mu \leq \delta$, then the system* (7) *is exponentially stable under condition*

$$\gamma > \sqrt{\frac{(2-\alpha)\epsilon\mu + \alpha\epsilon\delta}{\beta\mu}}. \tag{10}$$

Corollary 3. *If $\alpha\mu \leq \delta$, then the system* (7) *is exponentially stable for all sufficiently small ϵ.*

Corollary 4. *If $\alpha\mu > \delta$ and $\epsilon \leq \frac{(\alpha\mu-\delta)^2\mu}{(2-\alpha)\mu\beta+\alpha\delta\beta}$, then the system* (7) *is exponentially stable under condition*

$$\gamma > \frac{\alpha\mu - \delta}{\beta}. \tag{11}$$

Corollary 5. *If $\alpha\mu > \delta$ and $\epsilon > \frac{(\alpha\mu-\delta)^2\mu}{(2-\alpha)\mu\beta+\alpha\delta\beta}$, then the system* (7) *is exponentially stable under condition*

$$\gamma > \sqrt{\frac{(2-\alpha)\epsilon\mu + \alpha\epsilon\delta}{\beta\mu}}. \tag{12}$$

Corollary 6. *If $\alpha\mu \geq \delta$, then the system (7) is exponentially stable under condition*

$$\gamma > \frac{\alpha\mu - \delta + \sqrt{(\alpha\mu - \delta)^2 + 8\beta\epsilon}}{2\beta}. \tag{13}$$

Corollary 7. *If $\mu > \delta$, then the system (7) is exponentially stable under condition*

$$\gamma > \frac{\alpha\mu - \delta + \sqrt{(\alpha\mu - \delta)^2 + 4\alpha\mu^2}}{2\beta}. \tag{14}$$

(Note that this condition does not depend on ϵ.)

We also have the following result for the system where agents do not return to the agent queue after service completions.

Corollary 8. *If $\alpha = 0$, then the system (7) is exponentially stable for all positive β, μ, ϵ, γ, and $\delta \geq 0$, $\theta \geq 0$.*

Let us consider a special case when $\delta = 0$, referred to as a *customer non-abandonment system*. Then, Corollaries 4, 5, 6, and 7 imply the following sufficient local stability conditions of the customer non-abandonment system.

Corollary 9. *If $\delta = 0$, $\alpha \in (0,1)$, and $\epsilon \leq \frac{\alpha^2\mu^2}{(2-\alpha)\beta}$, then the system (7) is exponentially stable under condition*

$$\gamma > \frac{\alpha\mu}{\beta}. \tag{15}$$

Corollary 10. *If $\delta = 0$, $\alpha \in (0,1)$, and $\epsilon > \frac{\alpha^2\mu^2}{(2-\alpha)\beta}$, then the system (7) is exponentially stable under condition*

$$\gamma > \sqrt{\frac{(2-\alpha)\epsilon}{\beta}}. \tag{16}$$

Corollary 11. *If $\delta = 0$, and $\alpha \in [0,1)$, then the system (7) is exponentially stable under condition*

$$\gamma > \frac{\alpha\mu + \sqrt{\alpha^2\mu^2 + 8\beta\epsilon}}{2\beta}. \tag{17}$$

Note that if $\theta = 0$, then Corollary 11 is a simpler, equivalent version of the sufficient local stability condition in [13] for the non-abandonment system (Theorem 3 in [13]). Moreover, condition (9) in Theorem 2 implies the following result, which does not depend on ϵ, for the non-abandonment system.

Corollary 12. *If $\delta = 0$, and $\alpha \in [0,1)$, then the system (7) is exponentially stable under condition*

$$\gamma > \frac{(\alpha + \sqrt{\alpha^2 + 4\alpha})\mu}{2\beta}. \tag{18}$$

Having a variety of these sufficient local stability conditions is useful, because some or others may be easier to verify/ensure, depending on the scenario. Note that γ and ϵ are control parameters, while all other parameters are those of the system – they can be potentially measured/estimated in real time. It is not easy to give an intuitive meaning/interpretation of the above local stability conditions. Perhaps Corollary 1 is the easiest to interpret: if the magnitude γ of the system response is large enough, this is sufficient for local stability.

Remark. We note that our *local* stability results apply to more general systems, exhibiting same *local* behavior. For example, suppose the total number of potential agents is not infinite, by finite, scaling with r as κr, where $\kappa > \lambda(1-\alpha)/\beta$. Then, the fluid limits of such system satisfy the same ODE (7) in the vicinity of the origin, and therefore our local stability results apply as is.

5 Proof of Theorem 1

The proof of Theorem 1 is a generalization of the proof of Theorem 1 in [14]. However, it requires additional technical details – we present it here for completeness.

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 limit holds.

Let $N_i(\cdot)$, $i = 1, \ldots, 8$ be mutually independent unit-rate Poisson processes. N_1 is the process which drives customer arrivals. N_2 is the process which drives the acceptance of invitations. N_3 is the process which drives the service completions with agents leaving the system. N_4 is the process which drives the service completions with agents returning the agent queue. N_5 and N_6 are the processes which drive type-3 events, when variable $Y^r(t)$ is negative and positive, respectively. N_7 is the process which drives the abandonment of customers. N_8 is the process which drives the abandonment of agents. Given the initial state $(X^r(0), Y^r(0), V^r(0))$, we construct the process $(X^r(\cdot), Y^r(\cdot), V^r(\cdot))$, for all r, on the same probability space via a common set of independent Poisson process [15] as follows:

$$X^r(t) = G^r(t) + \left(- \min_{0 \le s \le t} G^r(s) \right) \vee 0, \tag{19}$$

$$G^r(t) = X^r(0) + \gamma N_1(\lambda r t) - \gamma N_2 \left(\beta \int_0^t X^r(s) ds \right)$$

$$- \gamma N_4 \left(\alpha \mu \int_0^t (V^r(s) - (Y^r(s))^+) ds \right)$$

$$- \gamma N_7 \left(\delta \int_0^t (Y^r(s))^- ds \right) + \gamma N_8 \left(\theta \int_0^t (Y^r(s))^+ ds \right)$$

$$+ N_5 \left(\epsilon \int_0^t (Y^r(s))^- ds \right) - N_6 \left(\epsilon \int_0^t (Y^r(s))^+ ds \right), \tag{20}$$

$$Y^r(t) = Y^r(0) + N_2\left(\beta\int_0^t X^r(s)ds\right)$$

$$- N_1(\lambda rt) + N_4\left(\alpha\mu\int_0^t (V^r(s) - (Y^r(s))^+)ds\right)$$

$$+ N_7\left(\delta\int_0^t (Y^r(s))^- ds\right) - N_8\left(\theta\int_0^t (Y^r(s))^+ ds\right), \qquad (21)$$

$$V^r(t) = V^r(0) + N_2\left(\int_0^t \beta X^r(s)ds\right)$$

$$- N_3\left(\int_0^t (1-\alpha)\mu(V^r(s) - (Y^r(s))^+)ds\right)$$

$$- N_8\left(\theta\int_0^t (Y^r(s))^+ ds\right). \qquad (22)$$

W.p.1, for any r, relations (19)-(22) uniquely define the realization of $(X^r(\cdot), Y^r(\cdot), V^r(\cdot))$ via the realizations of the driving processes $N_i(\cdot)$. Relation (19), 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:

$$\frac{N_i(rt)}{r} \to t \ , \ r \to \infty \ , \ \text{u.o.c., w.p.1.} \qquad (23)$$

We consider the sequence of associated fluid-scaled processes with centering $(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot))$ as defined in (4). Let a constant $m > \|(x(0), y(0), v(0))\|$ be fixed. For each r, on the same probability space as $(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot))$, let us define a modified fluid-scaled process $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot))$. Let $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot))$ start from the same initial state as $(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot))$, i.e., $(\bar{X}_m^r(0), \bar{Y}_m^r(0), \bar{V}_m^r(0)) = (\bar{X}^r(0), \bar{Y}^r(0), \bar{V}^r(0))$. The modified process $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot))$ follows the same path as $(\bar{X}^r(\cdot), \bar{Y}^r(\cdot), \bar{V}^r(\cdot))$ until the first time t, such that $\|(\bar{X}^r(t), \bar{Y}^r(t), \bar{V}^r(t))\| \geq m$. Denote this time by τ_m^r. We then freeze the process $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot))$ at the value $(\bar{X}^r(\tau_m^r), \bar{Y}^r(\tau_m^r), \bar{V}^r(\tau_m^r))$, i.e. $(\bar{X}_m^r(t), \bar{Y}_m^r(t), \bar{V}_m^r(t)) = (\bar{X}^r(\tau_m^r), \bar{Y}^r(\tau_m^r), \bar{V}^r(\tau_m^r))$ for all $t \geq \tau_m^r$.

Lemma 1. *Fix $(x(0), y(0), v(0))$ and a finite constant $m > \|(x(0), y(0), v(0))\|$. Then, w.p.1 for any subsequence of r, there exists a further subsequence, along which $(\bar{X}_m^r, \bar{Y}_m^r, \bar{V}_m^r)$ converges u.o.c. to a Lipschitz continuous trajectory (x_m, y_m, v_m), which satisfies properties (6) at any regular time $t \geq 0$ such that $\|(x_m(t), y_m(t), v_m(t))\| < m$.*

Proof. For the modified fluid-scaled processes $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot))$, we define the associated counting processes for upward and downward jumps. For

$t \le \tau_m^r$,

$$\bar{X}_m^{r\uparrow}(t) = r^{-1}\gamma N_1(\lambda r t) + r^{-1}\gamma N_8\left(\theta r \int_0^t (\bar{Y}_m^r(s))^+ ds\right)$$

$$+ r^{-1}N_5\left(\epsilon r \int_0^t (\bar{Y}_m^r(s))^- ds\right), \qquad (24)$$

$$\bar{X}_m^{r\downarrow}(t) = r^{-1}\gamma N_2\left(\beta r \int_0^t \left[\bar{X}_m^r(s) + \frac{\lambda(1-\alpha)}{\beta}\right] ds\right)$$

$$+ r^{-1}\gamma N_4\left(\alpha\mu r \int_0^t \left[\bar{V}_m^r(s) + \frac{\lambda}{\mu} - (\bar{Y}_m^r(s))^+\right] ds\right)$$

$$+ r^{-1}\gamma N_7\left(\delta r \int_0^t (\bar{Y}_m^r(s))^- ds\right) + r^{-1}N_6\left(\epsilon r \int_0^t (\bar{Y}_m^r(s))^+ ds\right), \qquad (25)$$

$$\bar{Y}_m^{r\uparrow}(t) = r^{-1}N_2\left(\beta r \int_0^t \left[\bar{X}_m^r(s) + \frac{\lambda(1-\alpha)}{\beta}\right] ds\right)$$

$$+ r^{-1}N_4\left(\alpha\mu r \int_0^t \left[\bar{V}_m^r(s) + \frac{\lambda}{\mu} - (\bar{Y}_m^r(s))^+\right] ds\right)$$

$$+ r^{-1}N_7\left(\delta r \int_0^t (\bar{Y}_m^r(s))^- ds\right), \qquad (26)$$

$$\bar{Y}_m^{r\downarrow}(t) = r^{-1}N_1(\lambda r t) + r^{-1}N_8\left(\theta r \int_0^t (\bar{Y}_m^r(s))^+ ds\right) \qquad (27)$$

$$\bar{V}_m^{r\uparrow}(t) = r^{-1}N_2\left(\beta r \int_0^t \left[\bar{X}_m^r(s) + \frac{\lambda(1-\alpha)}{\beta}\right] ds\right), \qquad (28)$$

$$\bar{V}_m^{r\downarrow}(t) = r^{-1}N_3\left((1-\alpha)\mu r \int_0^t \left[\bar{V}_m^r(s) + \frac{\lambda}{\mu} - (\bar{Y}_m^r(s))^+\right] ds\right)$$

$$+ r^{-1}N_8\left(\theta r \int_0^t (\bar{Y}_m^r(s))^+ ds\right), \qquad (29)$$

and for $t > \tau_m^r$, all these counting processes are frozen at their values at time τ_m^r, that is,

$$\begin{cases} \bar{X}_m^{r\uparrow}(t) = \bar{X}_m^{r\uparrow}(\tau_m^r) , \ \bar{X}_m^{r\downarrow}(t) = \bar{X}_m^{r\downarrow}(\tau_m^r) , \\ \bar{Y}_m^{r\uparrow}(t) = \bar{Y}_m^{r\uparrow}(\tau_m^r) , \ \bar{Y}_m^{r\downarrow}(t) = \bar{Y}_m^{r\downarrow}(\tau_m^r) , \\ \bar{V}_m^{r\uparrow}(t) = \bar{V}_m^{r\uparrow}(\tau_m^r) , \ \bar{V}_m^{r\downarrow}(t) = \bar{V}_m^{r\downarrow}(\tau_m^r). \end{cases} \qquad (30)$$

Using the relations (19)-(22) and the fact that for $0 \le t \le \tau_m^r$ the original process $(\bar{X}^r, \bar{Y}^r, \bar{V}^r)$ and the modified process $(\bar{X}_m^r, \bar{Y}_m^r, \bar{V}_m^r)$ coincide, we have

for all $t \geq 0$,

$$\bar{X}_m^r(t) = \bar{G}_m^r(t) + \left(-\lambda(1-\alpha)/\beta - \min_{0 \leq s \leq t} \bar{G}_m^r(s) \right) \vee 0, \tag{31}$$

$$\bar{G}_m^r(t) = \bar{X}^r(0) + \bar{X}_m^{r\uparrow}(t) - \bar{X}_m^{r\downarrow}(t), \tag{32}$$

$$\bar{Y}_m^r(t) = \bar{Y}^r(0) + \bar{Y}_m^{r\uparrow}(t) - \bar{Y}_m^{r\downarrow}(t), \tag{33}$$

$$\bar{V}_m^r(t) = \bar{V}^r(0) + \bar{V}_m^{r\uparrow}(t) - \bar{V}_m^{r\downarrow}(t). \tag{34}$$

The counting processes $\bar{X}_m^{r\uparrow}(\cdot)$, $\bar{X}_m^{r\downarrow}(\cdot)$, $\bar{Y}_m^{r\uparrow}(\cdot)$, $\bar{Y}_m^{r\downarrow}(\cdot)$, $\bar{V}_m^{r\uparrow}(\cdot)$, $\bar{V}_m^{r\downarrow}(\cdot)$ are non-decreasing. Using FSLLN (23) and the fact that the processes $\bar{X}_m^r(\cdot)$, $\bar{Y}_m^r(\cdot)$, and $\bar{V}_m^r(\cdot)$ 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 $(\bar{X}_m^{r\uparrow}(\cdot), \bar{X}_m^{r\downarrow}(\cdot), \bar{Y}_m^{r\uparrow}(\cdot), \bar{Y}_m^{r\downarrow}(\cdot), \bar{V}_m^{r\uparrow}(\cdot), \bar{V}_m^{r\downarrow}(\cdot))$ converges u.o.c. to a set of non-decreasing Lipschitz continuous functions $(x_m^\uparrow(\cdot), x_m^\downarrow(\cdot), y_m^\uparrow(\cdot), y_m^\downarrow(\cdot), v_m^\uparrow(\cdot), v_m^\downarrow(\cdot))$. But then the u.o.c. convergence of $(\bar{X}_m^r(\cdot), \bar{Y}_m^r(\cdot), \bar{V}_m^r(\cdot), \bar{G}_m^r(\cdot))$ to a set of Lipschitz continuous functions $(x_m(\cdot), y_m(\cdot), v_m(\cdot), g_m(\cdot))$ holds, where

$$x_m(t) = g_m(t) + \left(-\lambda(1-\alpha)/\beta - \min_{0 \leq s \leq t} g_m(s) \right) \vee 0, \tag{35}$$

$$g_m(t) = x(0) + x_m^\uparrow(t) - x_m^\downarrow(t), \tag{36}$$

$$y_m(t) = y(0) + y_m^\uparrow(t) - y_m^\downarrow(t), \tag{37}$$

$$v_m(t) = v(0) + v_m^\uparrow(t) - v_m^\downarrow(t), \tag{38}$$

and the following holds for t before fluid trajectory hits $\|(x_m(t), y_m(t), v_m(t))\| = m$

$$x_m^\uparrow(t) = \gamma\lambda t + \gamma\theta \int_0^t y_m^+(s)ds + \epsilon \int_0^t y_m^-(s)ds, \tag{39}$$

$$x_m^\downarrow(t) = \gamma\beta \int_0^t \left(x_m(s) + \frac{\lambda(1-\alpha)}{\beta} \right) ds + \gamma\alpha\mu \int_0^t \left(v_m(s) + \frac{\lambda}{\mu} - y_m^+(s) \right) ds$$
$$+ \gamma\delta \int_0^t y_m^-(s)ds + \epsilon \int_0^t y_m^+(s)ds, \tag{40}$$

$$y_m^\uparrow(t) = \beta \int_0^t \left(x_m(s) + \frac{\lambda(1-\alpha)}{\beta} \right) ds + \alpha\mu \int_0^t \left(v_m(s) + \frac{\lambda}{\mu} - y_m^+(s) \right) ds$$
$$+ \delta \int_0^t y_m^-(s)ds, \tag{41}$$

$$y_m^\downarrow(t) = \lambda t + \theta \int_0^t y_m^+(s)ds, \tag{42}$$

$$v_m^\uparrow(t) = \beta \int_0^t \left(x_m(s) + \frac{\lambda(1-\alpha)}{\beta} \right) ds, \tag{43}$$

$$v_m^\downarrow(t) = (1-\alpha)\mu \int_0^t \left(v_m(s) + \frac{\lambda}{\mu} - y_m^+(s) \right) ds + \theta \int_0^t y_m^+(s)ds. \tag{44}$$

Hence,

$$
\begin{cases}
x'_m(t) = \begin{cases}
-\gamma\beta x_m(t) - \gamma\alpha\mu(v_m(t) - y_m^+(t)) + \gamma\theta y_m^+(t) - \gamma\delta y_m^-(t) - \epsilon y_m(t), \\
\qquad \text{if } x_m(t) > -\frac{\lambda(1-\alpha)}{\beta} \\
[-\gamma\beta x_m(t) - \gamma\alpha\mu(v_m(t) - y_m^+(t)) + \gamma\theta y_m^+(t) - \gamma\delta y_m^-(t) - \epsilon y_m(t)] \\
\lor 0, \qquad \text{if } x_m(t) = -\frac{\lambda(1-\alpha)}{\beta}
\end{cases} \\
y'_m(t) = \beta x_m(t) + \alpha\mu(v_m(t) - y_m^+(t)) + \delta y_m^-(t) - \theta y_m^+(t) \\
v'_m(t) = \beta x_m(t) - (1-\alpha)\mu(v_m(t) - y_m^+(t)) - \theta y_m^+(t).
\end{cases}
\tag{45}
$$

It is easy to verify that, at any regular time $t \geq 0$ such that $\|(x_m(t), y_m(t), v_m(t))\| < m$, properties (6) hold for the trajectory $(x_m(\cdot), y_m(\cdot), v_m(\cdot))$. $\square$

Conclusion of the proof of Theorem 1. It is easy to see that

$$
\frac{d}{dt}\|(x_m(t), y_m(t), v_m(t))\| \leq C\|(x_m(t), y_m(t), v_m(t))\|
\tag{46}
$$

for any m and some $C > 0$.

From Gronwall's inequality [4], we have

$$
\|(x_m(t), y_m(t), v_m(t))\| \leq \|(x(0), y(0), v(0))\|e^{Ct} \text{ for all } t \geq 0
\tag{47}
$$

For a given $(x(0), y(0), v(0))$, let us fix $T_l > 0$ and choose $m_l > \|(x(0), y(0), v(0))\|e^{CT_l}$. For this $T_l > 0$, there exists a subsequence r^l, along which $(\bar{X}^r, \bar{Y}^r, \bar{V}^r)$ converges uniformly to $(x_{m_l}, y_{m_l}, v_{m_l})$, which satisfies properties (6), at any $t \in [0, T_l]$. The limit trajectory $(x_{m_l}, y_{m_l}, v_{m_l})$ does not hit m_l in $[0, T_l]$. Subsequence $r^l = \{r_1^l, r_2^l, \dots\}$ is such that, w.p.1, for all sufficiently large r along the subsequence r^l, $(\bar{X}^r(t), \bar{Y}^r(t), \bar{V}^r(t)) = (\bar{X}_{m_l}^r(t), \bar{Y}_{m_l}^r(t), \bar{V}_{m_l}^r(t))$ at any $t \in [0, T_l]$.

We consider a sequence $T_1, T_2, \dots, \to \infty$. We construct a subsequence r^* by using Cantor's diagonal procedure [17] from subsequences $r^1, r^2, \dots$ ($r^1 \supseteq r^2 \supseteq \dots$) corresponding to $T_1, T_2, \dots$, respectively (i.e. $r_1^* = r_1^1, r_2^* = r_2^2, \dots$). Clearly, for this subsequence r^*, w.p.1, $(\bar{X}^r, \bar{Y}^r, \bar{V}^r)$ converges u.o.c. to (x, y, v), which satisfies properties (6), at any regular point $t \in [0, \infty)$. $\square$

6 Proof of Theorem 2

In order to prove Theorem 2, it suffices to show that LTI systems of the switched linear system (7) have a CQLF.

The system (7) is a switched linear system with $m = 2$. (Note that $y^+ = y$ if $y \geq 0$ and $y^+ = 0$ if $y < 0$, and $y^- = 0$ if $y \geq 0$ and $y^- = -y$ if $y < 0$.) Namely, for $y \geq 0$,

$$
\begin{cases}
x'(t) = (-\gamma\beta)x(t) + (\gamma\alpha\mu + \gamma\theta - \epsilon)y(t) + (-\gamma\alpha\mu)v(t) \\
y'(t) = (\beta)x(t) + (-\alpha\mu - \theta)y(t) + (\alpha\mu)v(t) \\
v'(t) = (\beta)x(t) + ((1-\alpha)\mu - \theta)y(t) + (-(1-\alpha)\mu)v(t)
\end{cases}
\tag{48}
$$

and for $y < 0$,

$$\begin{cases} x'(t) = (-\gamma\beta)x(t) + (\gamma\delta - \epsilon)y(t) + (-\gamma\alpha\mu)v(t) \\ y'(t) = (\beta)x(t) + (-\delta)y(t) + (\alpha\mu)v(t) \\ v'(t) = (\beta)x(t) + (-(1-\alpha)\mu)v(t) \end{cases} \tag{49}$$

We can rewrite the systems above as two LTI systems $u'(t) = A_1 u(t)$ and $u'(t) = A_2 u(t)$, where $u(t) = (x(t), y(t), v(t))^T$ and

$$A_1 = \begin{pmatrix} -\gamma\beta & \gamma\alpha\mu + \gamma\theta - \epsilon & -\gamma\alpha\mu \\ \beta & -\alpha\mu - \theta & \alpha\mu \\ \beta & (1-\alpha)\mu - \theta & -(1-\alpha)\mu \end{pmatrix} , \ A_2 = \begin{pmatrix} -\gamma\beta & \gamma\delta - \epsilon & -\gamma\alpha\mu \\ \beta & -\delta & \alpha\mu \\ \beta & 0 & -(1-\alpha)\mu \end{pmatrix}. \tag{50}$$

Lemma 2. *Matrix A_1 in (50) is Hurwitz for all positive β, γ, μ, ϵ, $\delta \geq 0$, $\theta \geq 0$ and $\alpha \in [0, 1)$.*

Proof. The characteristic equation of A_1 is

$$\lambda^3 + (\beta\gamma + \mu + \theta)\lambda^2 + (\beta\epsilon + \beta\gamma\mu + \mu\theta)\lambda + \beta\epsilon\mu = 0. \tag{51}$$

By Proposition 1, it suffices to verify that

$$\beta\gamma + \mu + \theta > 0 , \ \beta\epsilon + \beta\gamma\mu + \mu\theta > 0 , \ \beta\epsilon\mu > 0, \text{ and} \tag{52}$$

$$(\beta\gamma + \mu + \theta)(\beta\epsilon + \beta\gamma\mu + \mu\theta) - \beta\epsilon\mu = \beta^2\gamma^2\mu + \beta^2\gamma\epsilon + \beta\gamma\mu^2 + 2\beta\gamma\mu\theta +$$
$$+ \beta\theta\epsilon + \mu^2\theta + \mu\theta^2 > 0. \tag{53}$$

The conditions (52) and (53) are obviously true. $\square$

Lemma 3. *For positive β, γ, μ, ϵ, $\delta \geq 0$, $\theta \geq 0$ and $\alpha \in [0, 1)$, matrix A_2 in (50) is Hurwitz if and only if*

$$\left(\frac{\beta\gamma + \delta}{\mu} + (1-\alpha)\right)\left(\frac{\beta\gamma\mu + \delta\mu(1-\alpha)}{\beta\epsilon} + 1\right) > 1 \tag{54}$$

Proof. The characteristic equation of A_2 is

$$\lambda^3 + (\beta\gamma + \mu(1-\alpha) + \delta)\lambda^2 + (\beta\epsilon + \beta\gamma\mu + \delta\mu(1-\alpha))\lambda + \beta\epsilon\mu = 0. \tag{55}$$

By Proposition 1, it suffices to verify that

$$\beta\gamma + \mu(1-\alpha) + \delta > 0 , \ \beta\epsilon + \beta\gamma\mu + \delta\mu(1-\alpha) > 0 , \ \beta\epsilon\mu > 0, \tag{56}$$

and $(\beta\gamma + \mu(1-\alpha) + \delta)(\beta\epsilon + \beta\gamma\mu + \delta\mu(1-\alpha)) - \beta\epsilon\mu > 0$, which is equivalent to (54) since

$$(\beta\gamma + \mu(1-\alpha) + \delta)(\beta\epsilon + \beta\gamma\mu + \delta\mu(1-\alpha)) - \beta\epsilon\mu > 0$$
$$\Leftrightarrow (\beta\gamma + \delta + \mu(1-\alpha))(\beta\gamma\mu + \delta\mu(1-\alpha) + \beta\epsilon) > \beta\epsilon\mu$$
$$\Leftrightarrow \left(\frac{\beta\gamma + \delta}{\mu} + (1-\alpha)\right)\left(\frac{\beta\gamma\mu + \delta\mu(1-\alpha)}{\beta\epsilon} + 1\right) > 1.$$

The conditions (56) are obviously true. $\square$

It is easy to see that Lemma 3 implies the following result.

Corollary 13. *Matrix A_2 in* (50) *is Hurwitz if*

$$\gamma > \frac{\alpha\mu - \delta}{\beta}. \tag{57}$$

(Note that $\gamma > 0$ by definition.)

Lemma 4. *Matrix A_2 in* (50) *is Hurwitz under the condition either* (8) *or* (9).

    Proof. This easily follows by applying Corollary 13.

Lemma 5. *Matrix product $A_1 A_2$ has no negative eigenvalues under the condition either* (8) *or* (9).

    Proof. With the help of MATLAB symbolic calculation, it can be shown that A_1 is non-singular and

$$A_1^{-1} = \begin{pmatrix} -\frac{\theta}{\beta\epsilon} & -\frac{(\alpha\epsilon - \epsilon + \gamma\theta)}{\beta\epsilon} & \frac{\alpha}{\beta} \\ -\frac{1}{\epsilon} & -\frac{\gamma}{\epsilon} & 0 \\ -\frac{1}{\epsilon} & \frac{(\epsilon - \gamma\mu)}{\epsilon\mu} & -\frac{1}{\mu} \end{pmatrix}. \tag{58}$$

    By Proposition 4, to demonstrate that the product $A_1 A_2$ has no negative eigenvalues, it will suffice to show that $[A_1^{-1} + \tau A_2]$ is non-singular for all $\tau \geq 0$. We have

$$\det[A_1^{-1} + \tau A_2] = [\beta^2\epsilon^2\mu^2\tau^3 + (\beta^2\epsilon^2 + \beta^2\gamma^2\mu^2 - 2\beta\epsilon\mu^2 + \delta\mu^2\theta +$$
$$+\alpha\beta\epsilon\mu^2 + \beta\delta\gamma\mu^2 - \alpha\delta\mu^2\theta + \beta\gamma\mu^2\theta - \alpha\beta\delta\epsilon\mu - \alpha\beta\delta\gamma\mu^2)\tau^2 + (\mu^2 - \alpha\mu^2 +$$
$$+\beta^2\gamma^2 - 2\beta\epsilon + \delta\theta + \beta\delta\gamma + \alpha\delta\mu + \beta\gamma\theta - \alpha\mu\theta - \alpha\beta\gamma\mu)\tau + 1]/(-\beta\epsilon\mu). \tag{59}$$

    To show $\det[A_1^{-1} + \tau A_2] \neq 0$ for all $\tau \geq 0$, it will suffice to show that the numerator of the ratio (59) is strictly positive. We can represent the numerator of the ratio (59) as follows.

    (a) Under the condition (8), the numerator of the ratio (59) is

$$\beta^2\epsilon^2\mu^2\tau^3 + (\beta\epsilon\tau - 1)^2 + [(\beta^2\gamma^2\mu^2 - 2\beta\epsilon\mu^2 - \alpha\beta\delta\epsilon\mu + \alpha\beta\epsilon\mu^2) +$$
$$+\delta\mu^2\theta(1 - \alpha) + \beta\delta\gamma\mu^2(1 - \alpha) + \beta\gamma\mu^2\theta]\tau^2 + [\mu^2(1 - \alpha) + \beta\gamma(\beta\gamma - \alpha\mu + \delta) +$$
$$+\alpha\delta\mu + (\beta\gamma - \alpha\mu + \delta)\theta]\tau \overset{\text{(61)-(62)}}{>} 0, \tag{60}$$

since the condition (8) implies

$$\gamma > \frac{\alpha\mu - \delta}{\beta} \Rightarrow \beta\gamma - \alpha\mu + \delta > 0, \tag{61}$$

and $\gamma > \sqrt{\dfrac{(2 - \alpha)\epsilon\mu + \alpha\epsilon\delta}{\beta\mu}} \Rightarrow \beta^2\gamma^2\mu^2 - 2\beta\epsilon\mu^2 - \alpha\beta\delta\epsilon\mu + \alpha\beta\epsilon\mu^2 > 0.$ (62)

    Hence, the numerator of the ratio (59) is strictly greater than 0 under the condition (8).

(b) Under the condition (9), the numerator of the ratio (59) is

$$(\beta\epsilon\tau - 1)^2\mu^2\tau + (\beta\epsilon\tau - 1)^2 +$$
$$+[(\beta^2\gamma^2\mu^2 - \alpha\beta\delta\epsilon\mu + \alpha\beta\epsilon\mu^2) + \delta\mu^2\theta(1-\alpha) + \beta\delta\gamma\mu^2(1-\alpha) + \beta\gamma\mu^2\theta]\tau^2 +$$
$$+[(\beta^2\gamma^2 - \beta\gamma(\alpha\mu - \delta) - \alpha\mu^2) + \alpha\delta\mu + (\beta\gamma - \alpha\mu + \delta)\theta]\tau \overset{(64)\text{-}(65)}{>} 0, \quad (63)$$

since the condition (9) implies

$$\gamma > \frac{\alpha\mu - \delta + \sqrt{(\alpha\mu - \delta)^2 + 4\alpha\mu^2}}{2\beta} > \frac{\alpha\mu - \delta}{\beta}$$
$$\Rightarrow \beta^2\gamma^2 - \beta\gamma(\alpha\mu - \delta) - \alpha\mu^2 > 0 \text{ and } \beta\gamma - \alpha\mu + \delta > 0 \quad (64)$$
$$\text{and } \gamma > \sqrt{\max\left\{\frac{\alpha\epsilon(\delta - \mu)}{\beta\mu}, 0\right\}} \Rightarrow \beta^2\gamma^2\mu^2 - \alpha\beta\delta\epsilon\mu + \alpha\beta\epsilon\mu^2 > 0. \quad (65)$$

Hence, the numerator of the ratio (59) is strictly greater than 0 under the condition (9). Therefore, A_1A_2 has no negative eigenvalues under the condition either (8) or (9). $\square$

Conclusion of the proof of Theorem 2. By Lemma 2, A_1 is Hurwitz for all positive β, γ, μ, ϵ; $\delta \geq 0$, $\theta \geq 0$; and $\alpha \in [0, 1)$. By Lemma 4, A_2 is Hurwitz under the condition either (8) or (9). It is easy to verify that the difference $A_1 - A_2$ has rank one. By Lemma 5, A_1A_2 has no negative real eigenvalues under the condition either (8) or (9). Hence, by Proposition 3, two LTI systems $u'(t) = A_1u(t)$ and $u'(t) = A_2u(t)$ have a CQLF. Therefore, the system (7) is exponentially stable under the condition either (8) or (9). $\square$

7 Numerical and simulation experiments and conjectures

In this section, we present some numerical and simulation experiments. These results are for both stylized and actual schemes, and all results are for the true system which includes boundary $X \geq 0$. We also put forward some conjectures based on these experiments.

In all simulations, we always assume $r = 1000$, but specify only the actual arrival rate $\Lambda = \lambda r$. On the plots labeled 'fluid', $X(t), Y(y), V(t)$ are replaced by their *fluid approximations*

$$X(t) = rx(t) + \frac{\lambda r(1 - \alpha)}{\beta}, \quad Y(t) = ry(t), \quad V(t) = rv(t) + \frac{\lambda r}{\mu},$$

respectively, where $(x(\cdot), y(\cdot), v(\cdot))$ is the corresponding fluid limit.

7.1 Stylized scheme

Example 1. Consider the following set of parameters, which satisfies condition (8):

$$\Lambda = 2000 \ , \ \alpha = 0.5 \ , \ \beta = 3 \ , \ \mu = 2 \ , \ \gamma = 1 \ , \ \epsilon = 1.5 \ , \ \delta = 1 \ , \ \theta = 0.1$$

with four initial conditions: (a) $(X(0), Y(0), Z(0)) = (0, 0, 0)$; (b) $(X(0), Y(0), Z(0)) = (0, 2000, 0)$; (c) $(X(0), Y(0), Z(0)) = (2000, -2000, 1000)$; (d) $(X(0), Y(0), Z(0)) = (2000, 4000, 1000)$. The red line of the figure is the fluid approximation and the blue one is the simulation experiment. We see the converging trajectories on the Figure 3. Note that Figures 3b and 3d show that the trajectory hits the boundary on X. We also did the numerical/simulation experiments with many different sets of parameters satisfying the condition (8). All results, including those not shown on Figure 3, suggest the global stability of the system.
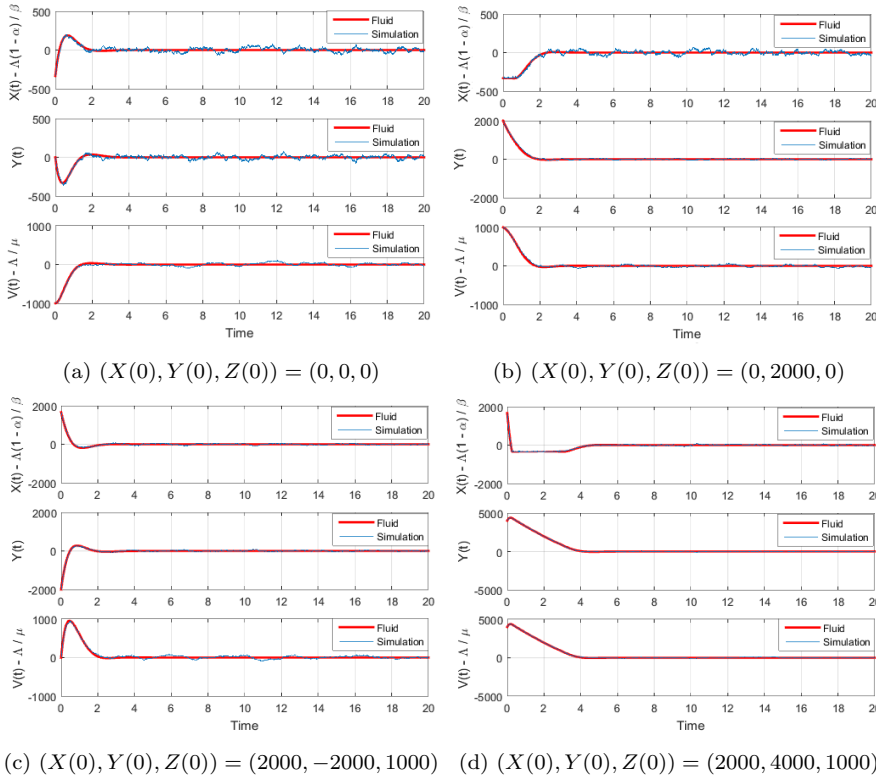

(a) $(X(0), Y(0), Z(0)) = (0, 0, 0)$ (b) $(X(0), Y(0), Z(0)) = (0, 2000, 0)$

(c) $(X(0), Y(0), Z(0)) = (2000, -2000, 1000)$ (d) $(X(0), Y(0), Z(0)) = (2000, 4000, 1000)$

Fig. 3: Stylized scheme: Comparison of fluid approximations with simulations in Example 1

Example 2. We use sets of parameters:

$$\Lambda = 2000 \ , \ \alpha = 0.9 \ , \ \beta = 0.05 \ , \ \mu = 0.5 \ , \ \epsilon = 1 \ , \ \delta = 0.01 \ , \ \theta = 0.01$$

with four different values of γ ($\gamma_1 = 1$, $\gamma_2 = 5$, $\gamma_3 = 10$, and $\gamma_4 = 20$) (Figure 4). The sets of parameters with $\gamma_1 = 1$ and $\gamma_2 = 5$ do not satisfy the condition (8) while the sets of parameters with $\gamma_3 = 10$ and $\gamma_4 = 20$ satisfy the condition (8). We consider an initial condition $(X(0), Y(0), Z(0)) = (1000, 6000, 2000)$. On the Figures 4b, 4c and 4d, we see that the trajectories converge. However, Figure 4a shows the trajectory that never converges under the set of parameters with $\gamma_1 = 1$.
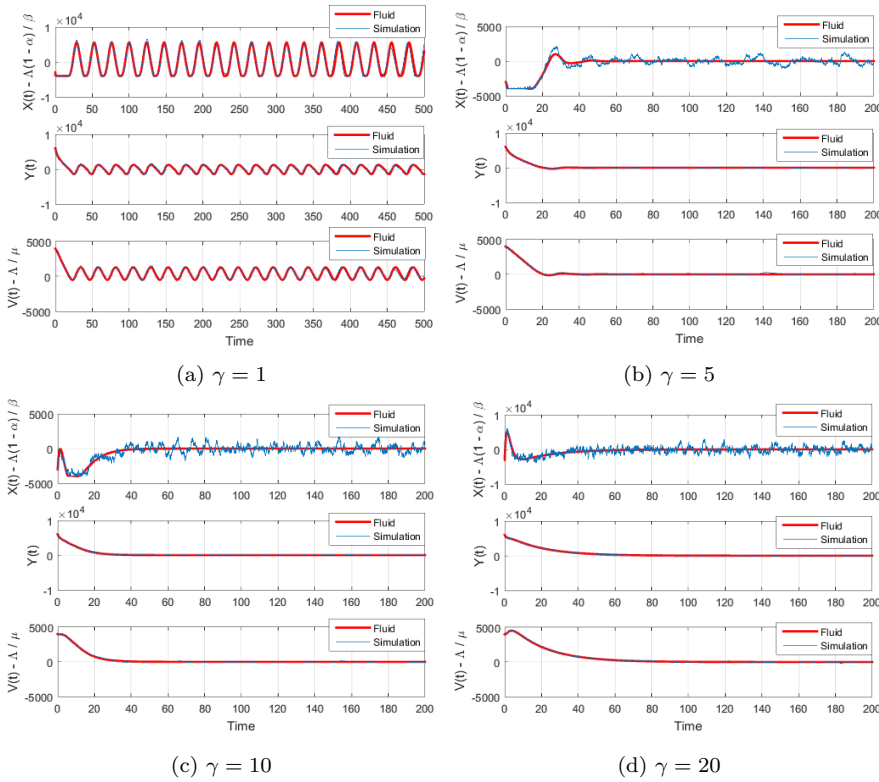

Fig. 4: Stylized scheme: Comparison of fluid approximations with simulations in Example 2

With many numerical/simulation experiments, the results, including those not shown on Figure 4, suggest both local and global stability of the system for all sufficiently large γ.

Our simulation experiments show that the fluid trajectory provides a very good approximation for the behavior of stylized scheme.

7.2 Actual scheme

Example 3. We conduct a simulation experiment for the actual scheme with the same set of parameters as in Example 1:

$$\Lambda = 2000 \ , \ \alpha = 0.5 \ , \ \beta = 3 \ , \ \mu = 2 \ , \ \gamma = 1 \ , \ \epsilon = 1.5 \ , \ \delta = 1 \ , \ \theta = 0.1$$

with two initial conditions $(X(0), Y(0), Z(0), X_{target}(0)) = (0, 0, 0, 0)$ and $(X(0), Y(0), Z(0), X_{target}(0)) = (0, 0, 0, 1000)$. (Note that this set of parameters satisfies the condition (8).) The results are shown in Figures 5 and 6. 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 negligible compared to their values. This explains why the trajectories of X_{target} and X are well approximated by the fluid trajectory, obtained for the stylized scheme.

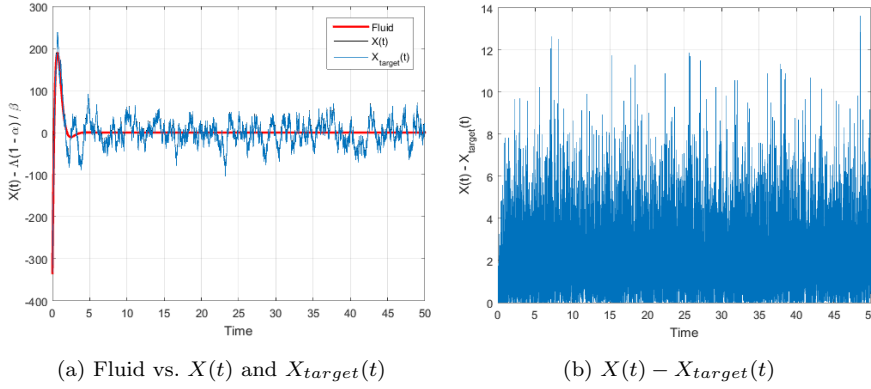(a) Fluid vs. $X(t)$ and $X_{target}(t)$ (b) $X(t) - X_{target}(t)$

Fig. 5: Actual scheme: $(X(0), Y(0), Z(0), X_{target}(0)) = (0, 0, 0, 0)$

For the stylized scheme, the results suggest the global stability of our system for all sufficiently large γ. However, the problem with large γ is that the behavior of the stylized scheme may significantly deviate from the behavior of the actual scheme, as illustrated by the following example.

Example 4. Consider the following set of parameters:

$$\Lambda = 2000 \ , \ \alpha = 0.7 \ , \ \beta = 0.5 \ , \ \mu = 3 \ , \ \epsilon = 1 \ , \ \delta = 1 \ , \ \theta = 2$$

with two values of γ ($\gamma_1 = 10, \gamma_2 = 20$); and an initial condition $(X(0), Y(0), Z(0), X_{target}(0)) = (0, 0, 0, 1000)$ (Figure 7). These results show that the behavior of the actual scheme deviates substantially from the behavior of the fluid trajectory with large γ.

Since $\alpha\mu > \delta$ and $\epsilon \leq \frac{(\alpha\mu - \delta)^2 \mu}{(2-\alpha)\mu\beta + \alpha\delta\beta}$, then we choose $\gamma = 2.3$ such that $\gamma > \frac{\alpha\mu - \delta}{\beta}$ (Corollary 4). We can see that, with a "good" value of γ, the
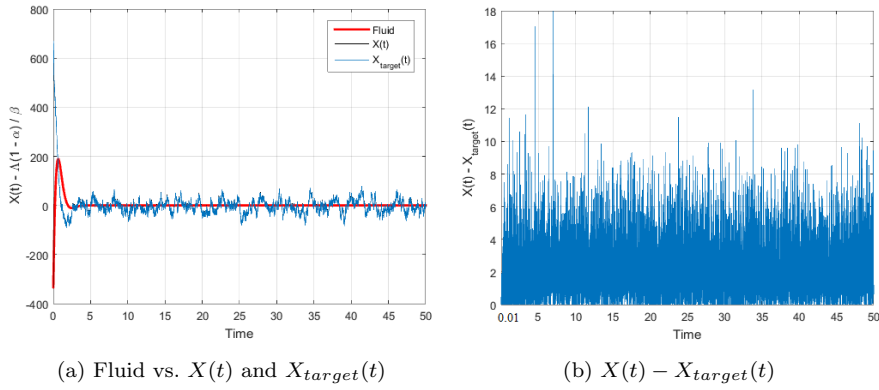
(a) Fluid vs. $X(t)$ and $X_{target}(t)$ (b) $X(t) - X_{target}(t)$

Fig. 6: Actual scheme: $(X(0), Y(0), Z(0), X_{target}(0)) = (0, 0, 0, 1000)$

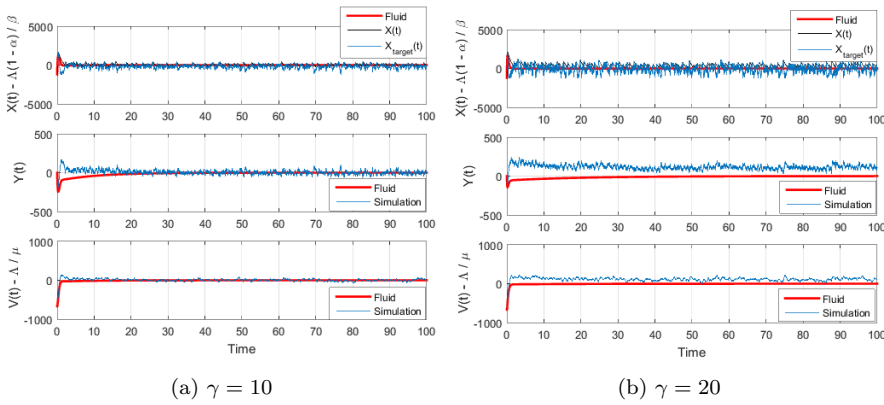(a) $\gamma = 10$ (b) $\gamma = 20$

Fig. 7: Problem with large γ of the actual scheme

behavior of the actual scheme deviates negligibly from the behavior of the fluid trajectory (Figure 8a) and the difference between X_{target} and X is not large compared to their values (Figure 8b).

7.3 Global vs. local stability of fluid limits

In this paper, we have derived some sufficient local stability conditions for the fluid limits. Based on a variety of simulation experiments above for the stylized scheme, we conjecture that local stability is sufficient for global stability of fluid limits for our model. In the next example, we compare the behavior of fluid limits for the system without boundary (given by (7)) with that of the system with boundary (given by (6)).

Example 5. Consider two set of parameters, which satisfy the local stability conditions, so that the trajectory of the system (7) converges to the equilibrium

(a) $\gamma = 2.3$ (b) $X(t) - X_{target}(t)$
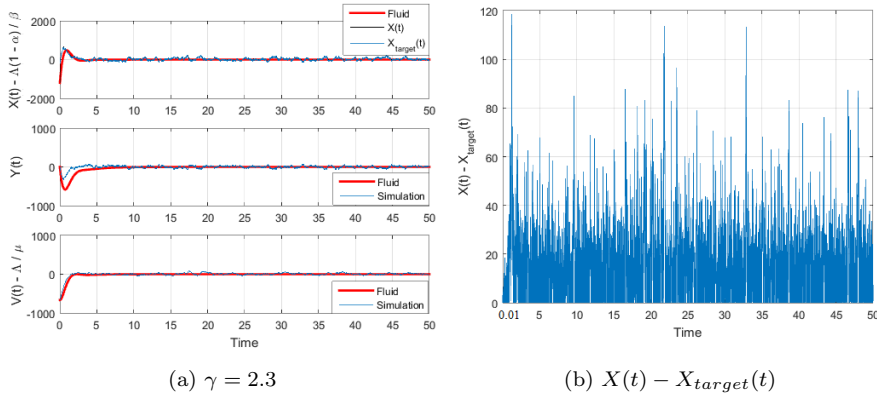
Fig. 8: A "good" value of γ for the actual scheme

point $(0, 0, 0)$ (Figure 9). The red line of the figure is the trajectory of the system (6), which may hit the boundary $X = 0$, and the black one is the trajectory of the system (7), for which there is no boundary.

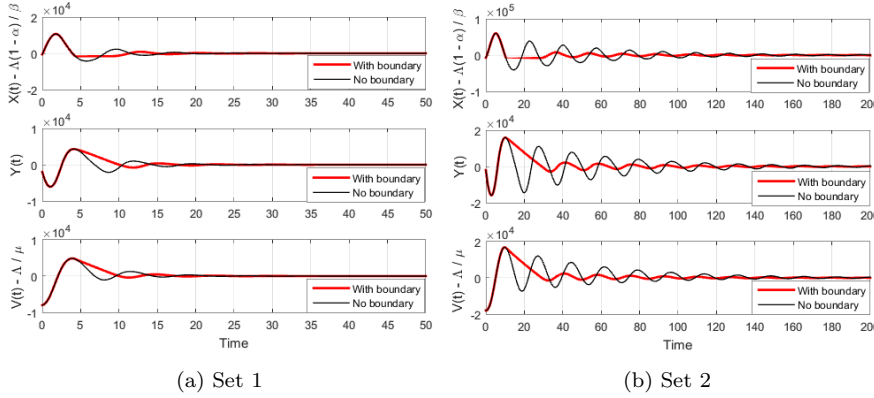(a) Set 1   (b) Set 2

Fig. 9: Fluid trajectories of the systems (6) and (7)

With many experiments, the results, including those not shown in Figure 9, further suggest the global stability of the fluid limits, when the local stability holds.

7.4 Conjectures and their implications.

Based on our numerical and simulation experiments, we make the following conjectures.

Conjecture 1. Our system is globally stable if it is locally stable.

Conjecture 2. Given all other parameters are fixed, our system is globally stable for all sufficiently large γ.

Obviously, Conjecture 1 is stronger than Conjecture 2 because we have proved the local stability when γ is large in this paper. We note again, however, that in a practical application the value of γ should not be made too large, because the stylized scheme behavior, which we studied in this paper, may substantially deviate from the behavior of the actual scheme, where uninviting pending agents are not allowed.

Conjecture 1, if true, is important, because global stability is a key system property, due to the following reasons. If the the global stability does hold, then it is further natural to conjecture that our (more general) model and algorithm possess "nice" properties analogous to those proved in [14] (for the special case where agents do not return and there are no abandonments), namely that the steady-state customer and agent queue lengths scale as $O(\sqrt{r})$ (and then waiting times scales as $O(1/\sqrt{r})$) as the customer arrival rate $\lambda r \to \infty$. On the other hand, if global stability does not hold, then, as our simulations suggest, the system fluid limit oscillates, which results in large, $O(r)$, queue lengths (and $O(1)$ waiting times).

8 Discussion and further work

In this paper, we study a feedback-based agent invitation scheme for a model with randomly behaving agents and possible abandonment of customers and agents. This model is motivated by a variety of existing and emerging applications. The focus of the paper is on the stability properties of the system fluid limits, arising as asymptotic limits of the system process, when the system scale (customer arrival rate) grows to infinity. The dynamic system, describing the behavior of fluid limit trajectories has a very complex structure – it is a switched linear system, which in addition has a reflecting boundary. We derived some sufficient local stability conditions, using the machinery of switched linear systems and common quadratic Lyapunov functions. Our simulation and numerical experiments show good overall performance of the feedback scheme, when the local stability conditions hold. They also suggest that, for our model, the local stability is in fact sufficient for the global stability of fluid limits. Verifying these conjectures, as well as expanding the sufficient local stability conditions, is an interesting subject for future research.

Further generalizations of the agent invitation model are also of interest from both theoretical and practical points of view.

Appendix: Some results for switched linear systems and CQLF

The following facts are used in the proof of our main result (Theorem 2).

Proposition 1 ([16]). *Let $L(\lambda) = \det(A - \lambda I) = 0$ be the characteristic equation of matrix A in $\mathbb{R}^{3 \times 3}$:*

$$L(\lambda) = a_0 \lambda^3 + a_1 \lambda^2 + a_2 \lambda + a_3 = 0 \ , \ a_0 > 0. \tag{66}$$

Matrix A is Hurwitz if and only if a_1, a_2, a_3 are positive and $a_1 a_2 > a_0 a_3$.

Proposition 2 ([10,19]). *The existence of a CQLF for the LTI systems is sufficient for the exponential stability of the switched linear system.*

Proposition 3 ([10,19]). *Let A_1 and A_2 be Hurwitz matrices in $\mathbb{R}^{n \times n}$, and the difference $A_1 - A_2$ has rank one. Then two systems $u'(t) = A_1 u(t)$ and $u'(t) = A_2 u(t)$ have a CQLF if and only if the matrix product $A_1 A_2$ has no negative real eigenvalues.*

Proposition 4 ([18]). *If A_1^{-1} is non-singular, the product $A_1 A_2$ has no negative eigenvalues if and only if $A_1^{-1} + \tau A_2$ is non-singular for all $\tau \geq 0$.*

References

1. Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: a multi-disciplinary perspective on operations management research. Production and Operations Management **16**(6), 665–688 (2007)
2. American Telemedicine Association: *Core Operational Guidelines for Telehealth Services Involving Provider-Patient Interactions* (2014). `http://www.americantelemed.org/docs/default-source/standards/core-operational-guidelines-for-telehealth-services.pdf?sfvrsn=6`
3. Bengtson, S.: Generating better results with crowdsourcing: Leverage a network of high-quality professionals for customer service (2014). URL `http://www.arise.com/`. *White paper*
4. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence. Wiley (1986)
5. Formisano, P.: Flexibility for changing business needs: Improve customer service and drive more revenue with a virtual crowdsourcing solution (2014). URL `http://www.arise.com/`. *White paper*
6. Garnet, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manufacturing and Service Operations Management **4**(3), 208–227 (2002). URL `http://EconPapers.repec.org/RePEc:inm:ormsom:v:4:y:2002:i:3:p:208-227`
7. Gurvich, I., Ward, A.: On the dynamic control of matching queues. Stochastic Systems **4**(2), 479–523 (2014)
8. Hespanha, J.P.: Uniform stability of switched linear systems: extensions of LaSalle's invariance principle. IEEE Transactions on Automatic Control **49**(4), 470–482 (2004)
9. Kashyap, B.R.K.: The double-ended queue with bulk service and limited waiting space. Operations Research **14**(5), 822–834 (1966)
10. Lin, H., Antsaklis, P.J.: Stability and stabilizability of switched linear systems: A survey of recent results. IEEE Transactions on Automatic Control **54**(2), 308–322 (2009)
11. Liu, X., Gong, Q., Kulkarni, V.G.: Diffusion models for doubly-ended queues with renewal arrival processes (2014). Forthcoming in *Stochastic Systems*. DOI: 10.1214/13-SSY113

12. McGee-Smith, S.: Why companies are choosing to deploy the LiveOps cloud-based contact center (2010). `http://www.liveops.com/sites/default/files/uploads/lo_wp_mcgee-smith_analytics.pdf`

13. Nguyen, L.M., Stolyar, A.L.: A service system with randomly behaving on-demand agents. SIGMETRICS Perform. Eval. Rev. **44**(1), 365–366 (2016). DOI 10.1145/2964791.2901484. URL `http://doi.acm.org/10.1145/2964791.2901484`

14. Pang, G., Stolyar, A.L.: A service system with on-demand agent invitations. Queueing Syst. Theory Appl. **82**(3-4), 259–283 (2016). DOI 10.1007/s11134-015-9464-8. URL `http://dx.doi.org/10.1007/s11134-015-9464-8`

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

16. Pontryagin, L.: Ordinary Differential Equations. Adiwes international series in mathematics. Addison-Wesley (1962). URL `https://books.google.com/books?id=SQtKoAEACAAJ`

17. Rudin, W.: Principles of Mathematical Analysis. McGraw-Hill, New York (1976). URL `http://opac.inria.fr/record=b1098274`

18. Shorten, R., Mason, O., O'Cairbre, F., Curran, P.: A unifying framework for the siso circle criterion and other quadratic stabilty criteria. International Journal of Control **77**(1), 1–9 (2004)

19. Shorten, R., Wirth, F., Mason, O., Wulff, K., King, C.: Stability criteria for switched and hybrid systems. SIAM Review **49**(4), 545–592 (2007)

20. Stolyar, A., Reiman, M., Korolev, N., Mezhibovsky, V., Ristock, H.: Pacing in knowledge worker engagement (2010). *United States Patent Application 20100266116-A1*

21. Zeltyn, S., Mandelbaum, A.: Call centers with impatient customers: Many-server asymptotics of the m/m/n + g queue. Queueing Systems **51**(3), 361–402 (2005). DOI 10.1007/s11134-005-3699-8. URL `http://dx.doi.org/10.1007/s11134-005-3699-8`