# A Service System with Randomly Behaving On-demand Agents

Lam M. Nguyen
ISE Department
Lehigh University
Bethlehem, PA 18015
lmn214@lehigh.edu

Alexander L. Stolyar
ISE Department
Lehigh University
Bethlehem, PA 18015
stolyar@lehigh.edu

## ABSTRACT

We consider a service system where agents (or, servers) are invited on-demand. Customers arrive as a Poisson process and join a customer queue. Customer service times are i.i.d. exponential. Agents' behavior is random in two respects. First, they can be invited into the system exogenously, and join the agent queue after a random time. Second, with some probability they rejoin the agent queue after a service completion, and otherwise leave the system. The objective is to design a real-time adaptive agent invitation scheme that keeps both customer and agent queues/waiting-times small. We study an adaptive scheme, which controls the number of pending agent invitations, based on queue-state feedback.

We study 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 derive sufficient conditions for the local stability of fluid limits at the desired equilibrium point (with zero queues). We conjecture that, for our model, 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 the local stability conditions do hold, simulations show good overall performance of the scheme.

## 1. INTRODUCTION

This model is a generalization of that in [9, 7]. It was originally motivated (see [9]) by applications to call/contact centers [1], where what we call agents are "special agents", or "knowledge workers," whose time is expensive, so that it is inefficient to have them working fixed shifts, with inevitable periods of idle time due to random fluctuations in customer demand. It is much more reasonable to invite them on-demand in real time; however, designing an efficient agent invitation strategy is non-trivial due to randomness in agent behavior. Besides efficiency (in terms of minimizing customer and agent waiting times), another highly desirable feature of the invitation scheme is simplicity and robustness.

The model we consider is generic and has other applications, or potential applications. One example is telemedicine [2], in which case "agents" are doctors, invited on-demand to serve patients remotely. Another example is crowdsourcing-based customer service [3]. Also note that 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" [5] and matching systems [4]; although in such models arrivals of all types into the system are typically exogenous, as opposed to being controlled.

We study a feedback-based adaptive scheme of [9, 7]. We consider the system in the asymptotic regime where the customer arrival rate becomes large while the distributions of an agent response times and a service time are fixed. We show convergence of the fluid-scaled process to the fluid limit (Theorem 1). The fluid limit trajectories have very complicated behavior – there are two domains where they follow different ODEs, and a "reflecting" boundary. This poses big challenges for proving *global stability* of the fluid limits, understood as the convergence of their trajectories to the equilibrium point, at which the queues are zero. Given that, the focus of this paper and our main results concern the system *local stability* at the equilibrium point, understood as the stability of the dynamic system which describes fluid limit trajectories away from the boundary. We use the machinery of switched linear systems and common quadratic Lyapunov functions [8] to obtain our **main results** (Theorem 2), providing sufficient local stability conditions. The details of our model, results, proofs, conjectures and numerical/simulation experiments are in [6].

## 2. MODEL AND ALGORITHM

Customers arrive according to a Poisson process, and join the 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. Once being invited, an agent joins the agent queue after a random, exponentially distributed, time with mean $1/\beta$. The customer and agent queues cannot be positive simultaneously: the head-of-the-line customer and agents 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 agent rejoins the agent queue with probability $\alpha \in [0, 1)$. Let $X(t)$ be the number of pending agents that have been invited but have not decided to accept or decline the invitations at time $t$. We

*SIGMETRICS '16 June 14-18, 2016, Antibes Juan-Les-Pins, France*

ACM ISBN 978-1-4503-4266-7/16/06.
DOI: http://dx.doi.org/10.1145/2896377.2901484

define $Y(t) = Q_a(t) - Q_c(t)$ as the difference of the number of customers in the agent queue and in the customer queue at time $t$. Let $Z(t)$ be the number of customers (or agents) in service at time $t$. The system state can be described by three variables $X$, $Y$, $Z$.

The feedback scheme in [9] maintains a "target" $X_{target}(t)$ for the number of invited 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 $\Delta t$ 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, $X_{target}(t) \geq 0$; i.e. if an update of $X_{target}(t)$ makes it negative, its value is immediately reset to zero. To simplify our theoretical analysis, we consider a "stylized" version, 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. Formally, the stylized scheme is defined as follows. There are four 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 $[t, t+dt]$: (i) a customer arrival, (ii) an agent acceptance, (iii) an additional event, and (iv) a service completion with probabilities $\Lambda dt$, $\beta X(t) dt$, $\epsilon |Y(t)| dt$, and $\mu Z(t) dt$, respectively. The changes at these event times are described as follows: (i) Upon a customer arrival, if $Y(t) > 0$, $Z(t)$ changes by $\Delta Z(t) = 1$; and if $Y(t) \leq 0$, $\Delta Z(t) = 0$. $\Delta Y(t) = -1$ and $\Delta X(t) = \gamma$. (ii) Upon the acceptance of an invitation, if $Y(t) < 0$, $\Delta Z(t) = 1$; and if $Y(t) \geq 0$, $\Delta Z(t) = 0$. $\Delta Y(t) = 1$ and $\Delta X(t) = -(\gamma \wedge X(t))$. (iii) Upon the third type of event, if $X(t) \geq 1$, $\Delta X(t) = -\text{sgn}(Y(t))$; and if $X(t) = 0$, $\Delta X(t) = 1$ if $Y(t) < 0$ and $\Delta X(t) = 0$ if $Y(t) \geq 0$. (iv) Upon the service completion, (a) with probability $\alpha$, if $Y(t) < 0$, $\Delta Z(t) = 0$; and if $Y(t) \geq 0$, $\Delta Z(t) = -1$; $\Delta Y(t) = 1$ and $\Delta X(t) = -(\gamma \wedge X(t))$. (b) With probability $(1 - \alpha)$, $\Delta Z(t) = -1$.

## 3. 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 $\lambda r$, while the parameters $\alpha$, $\beta$, $\mu$, $\epsilon$, $\gamma$ are constant. The corresponding process is $(X^r, Y^r, Z^r)$, where $X^r = (X^r(t), t \geq 0)$, $Y^r = (Y^r(t), t \geq 0)$ and $Z^r = (Z^r(t), t \geq 0)$. Let $W = |Y| + 2Z$ be the total number of customers and agents in the system. We are using a new process $(X, Y, W)$, which is more convenient for the analysis. We define new fluid-scaled processes with centering $(\bar{X}^r, \bar{Y}^r, \bar{W}^r) = r^{-1}\left(X^r - \lambda r(1-\alpha)/\beta, Y^r, W^r - 2\lambda r/\mu\right)$.

Theorem 1. *Consider a sequence of processes $(\bar{X}^r, \bar{Y}^r, \bar{W}^r)$, $r \to \infty$, with deterministic initial states such that $(\bar{X}^r(0), \bar{Y}^r(0), \bar{W}^r(0)) \to (x(0), y(0), w(0))$ for some fixed $(x(0), y(0), w(0)) \in \mathbb{R}^3$, $x(0) \geq -\lambda(1-\alpha)/\beta$. 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*

$$(\bar{X}^r, \bar{Y}^r, \bar{W}^r) \to (x, y, w) \quad u.o.c. \quad as \quad r \to \infty \qquad (1)$$

*where $(x, y, w) = [(x(t),\ t \geq 0), (y(t),\ t \geq 0), (w(t),\ t \geq 0)]$ is a locally Lipschitz trajectory such that at any regular point $t \geq 0$*

$$\begin{cases} x' = \begin{cases} -\gamma y' - \epsilon y, & \text{if } x > -\frac{\lambda(1-\alpha)}{\beta} \\ [-\gamma y' - \epsilon y] \vee 0, & \text{if } x = -\frac{\lambda(1-\alpha)}{\beta} \end{cases} \\ y' = \beta x + \frac{1}{2}\alpha\mu(w - |y|) \\ w' = \beta x + \frac{1}{2}(\alpha - 2)\mu(w - |y|). \end{cases} \qquad (2)$$

Consider a dynamic system

$$\begin{cases} x' = -\gamma y'(t) - \epsilon y \\ y' = \beta x + \frac{1}{2}\alpha\mu(w - |y|) \\ w' = \beta x + \frac{1}{2}(\alpha - 2)\mu(w - |y|), \end{cases} \qquad (3)$$

which is (2) "away from boundary," i.e. when $x > -\frac{\lambda(1-\alpha)}{\beta}$.

Theorem 2. *For any set of positive $\beta$, $\mu$, and $\alpha \in (0,1)$, there exist values of $\gamma > 0$ and $\epsilon > 0$ satisfying either*

$$\begin{cases} \frac{\beta\gamma^2}{4} < \epsilon < \frac{\beta\gamma^2}{2} \\ \epsilon > \frac{\beta\gamma^2}{2} - \left(\frac{\alpha\gamma\mu}{2} - \frac{(1-\alpha)\mu^2}{2\beta}\right) \\ \gamma > \frac{(1-\alpha)\mu}{\alpha\beta} \end{cases} \quad or \quad \begin{cases} \epsilon < \frac{\beta\gamma^2}{2} - \frac{\alpha\gamma\mu}{2} \\ \gamma > \frac{\alpha\mu}{\beta}. \end{cases} \qquad (4)$$

*For the parameters, satisfying either the left or right condition of (4), common quadratic Lyapunov function of the system (3) exists, and the system (3) is exponentially stable.*

Note that the right condition of (4) is more robust and is easier to achieve in practice. Indeed, for any given $\epsilon > 0$, it holds for all sufficiently large $\gamma$.

## 4. REFERENCES

[1] Z. Aksin, M. Armony, and V. Mehrotra. 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.

[3] P. Formisano. Flexibility for changing business needs: Improve customer service and drive more revenue with a virtual crowdsourcing solution. *White paper*, 2014.

[4] I. Gurvich and A. Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4(2):479–523, 2014.

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

[6] L. Nguyen and A. Stolyar. A service system with randomly behaving on-demand agents, 2016. http://arxiv.org/pdf/1603.03413v1.pdf.

[7] G. Pang and A. Stolyar. A service system with on-demand agent invitations. *Queueing Systems*, 2015. http://arxiv.org/pdf/1409.7380v2.pdf.

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

[9] A. Stolyar, M. Reiman, N. Korolev, V. Mezhibovsky, and H. Ristock. Pacing in knowledge worker engagement, 2010. *United States Patent Application 20100266116-A1.*