A service system with on-demand agent invitations

Guodong Pang · Alexander L. Stolyar

Ajanthan Mathialagan

Fasil Cheema

Marisa Signorile

1 Project Summary

Our project is concerned with a service system with servers/agents that are invited on demand. The system has two components: the difference between the queues of agents and customers and the number of pending invitations for agents. The model establishes that in the asymptotic regime, that is when the customer arrival rate goes to infinity and the agents' response rate is fixed, both customer and agent waiting times vanish when the system reaches a steady state. In this model the service time itself is not considered, and the customer arrival is modelled via a Poisson arrival process, and the agent responses is an i.i.d exponential distribution. The two variables: the difference between the agent/customer queues and the number of pending invitations are modelled via a CTMC (Continuous Time Markov Chain). Due to the difficulty of modelling the process exactly the analysis is done in the asymptotic regime mentioned earlier, which is known as the many-server asymptotic regime. The queue demonstrated by the call centre in the paper is an M/G/k queue. This is because customer arrival rate is modulated by a poisson process and service times of agents have an exponential distrubution, as well, the queue length in the system has k on demand service agents.

The paper shows that when the customer arrival rate becomes constant the model converges to a fluid limit and uniform global stability of fluid limits. We will show the proof of this first theorem which deals with properties of fluid limits by considering fluid models, defined as locally Lipschitz continuous trajectories (y,x) satisfying the properties in Theorem 1. In other words, if a fluid limit exists (which we prove in this report) it is it is necessarily a fluid model. Then we will prove theorem 2 in the paper showing that the model exhibits process stochastic stability and the limit interchange property. In addition, we will prove that on the diffusion scale, which is represented by r 1/2 scale, where r is a scaling parameter, convergence goes to the diffusion limit process (theorem 3). In proving theorem 3 in the paper, what is particularly interesting is the existence of an 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. What is obtained is the claimed distribution convergence of the limit diffusion process (Y^r, X^r) which follow directly from linear SDEs. By being able to model this agent/ customer queue by a linear SDE, one turns a complex behaviour asymptotic queue, into a model that can be used to predict customer arrival rate approaching infinity. Finally, for the stochastic model tightness and limit-interchange results are proven for the model.

The applications of Brownian motion in queuing theory is very useful for analyzing call centres and internet browsing queues. Brownian motion models are important when simulating and modelling applications of call centers and queues because Brownian motions provide the simplest mechanism in representing stochastic and dynamic phenomena. More specifically, Brownian motions are useful in modelling a service system with on-demand agent invitations be-

cause the Standard Brownian motion is null recurrent, that is, $E(\tau_i) = \infty$ (Hopfner,1990). The limiting regime under the diffusion scaling in this paper is modelled using Brownian motion and leads to a diffusion approximation of the original queue model. Evidently, Standard Brownian motions are employed to demonstrate that both customer and agent waiting times approach 0 as the system scale approaches infinity. The research in this paper is of great importance in call centres, while these special agents with special expertise/knowledge are very valuable, they are very expensive. As a result of this, it is not feasible to have these special agents constantly available. Rather, the model of on-demand agents proposed in this paper assure that both knowledge workers that accepted an invitation should not wait for an extended (expensive) time before they actually start processing a call and 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, keeping a target level of k invited special agents which changes based on current system state via a continuous-time Markov chain (CTMC), the scheme allows us to stabilize the system and keep waiting times of customers and agents low which is beneficial for customers and businesses.

The simulation conducted demonstrates how the unique fluid limit trajectory provides a good approximation of the system dynamics. What is also shown is the behavior and performance of feedback agent invitation schemes, as well as accuracy of the approximations given by the theoretical results in theorems 1-4 from the paper. A simulation was carried out to compare fluid approximations in one sample path with the initial conditions $(Y(0), X(0)) = (0, 0)$, $(Y(0), X(0)) = (1000, 0)$, $(Y(0), X(0)) = (0, 2000)$, $(Y(0), X(0)) = (-1000, 2000)$ which are analogous to figure 2 a,b,c,d in the paper. We have observed that the feedback scheme brings the system very close to its desired operating point, and fluid limit provides a very good approximation of the system trajectory. A noteable conclusion from the study of service systems with on-demand agent invitations is that they demonstrate the desirable performance, that is, both customer and agent waiting delays vanish as the system scale increases to infinity.

2 Proofs of Theorems

Theorem 1

Consider a sequence of processes $(Y^r, X^r), r \to \infty$, with deterministic initial states such that $(Y^r(0), X^(0)) \to (y(0), x(0))$ for some fixed $(y(0), x(0)) \in R^2, x(0) \geq \lambda/\beta$. 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,

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

where

$$x(t) \geq \lambda/\beta t \geq 0$$

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 $x(t) > \lambda/\beta$,

$$y'(t) = \beta x(t),$$
$$x'(t) = \gamma\beta x(t) - \varepsilon y(t),$$

and if x(t) = λ/β,

$$y'(t) = -\lambda$$
$$x'(t) = [\gamma\lambda - \varepsilon y(t)] \vee 0$$

The unique limit trajectory (y, x) will be called a fluid limit starting from (y(0), x(0)).

Proof of Theorem 1

Given the initial state $(Y^r(0), X^r(0))$ in theorem 1, the processes will be proved for all r, that is, for (Y^r, X^r):

$Y^r(t) = Y^r(0) + N_2(\beta \int_0^t X^r(s)ds) - N_1(\lambda rt)$

$X^r(t) = Z^r(t) - (\min_{0\leq s\leq t} Z^r(t)) \vee 0$, since $X^r(t)$ cannot become negative.

$Z^r(t) = X^r(0)+\gamma N_1(\lambda rt)-\gamma N_2(\beta \int_0^t X^r(s)ds)+N_3(\epsilon \int_0^t (Y^r(s))^- ds)+N_4(\epsilon \int_0^t (Y^r(s))^+ ds)$

where $N_i(\cdot)$ are independent unit rate Poisson processes for i=1,..,4.

The functional law of large numbers holds for each Poisson process $N_i(\cdot)$:

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

Define a fluid-scaled processes with centering as

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

Let c be a constant, where $c \geq \|y(0), x(0)\|$. Define (Y_c^r, X_c^r) as a modified fluid scaled process following the same path as (Y^r, X^r) until the first time that $\|(Y^r(t), X^r(t)\| \geq c$. Denote this time as τ_c^r, then, at this time, the process halts at the value $(Y^r(\tau_c^r), X^r(\tau_c^r))$.

Next, we need to prove convergence for the fluid scaled process:
(i) We must show the convergence of (Y_c^r, X_c^r) to a limit trajectory that behaves

like a fluid model as long as the state norm is away from c, constructed above. (ii) Choose c large enough so that the limit trajectory never reaches norm level c, proving it is a unique fluid model. This will imply that on any finite time interval, w.p.1, for all large r ,(Y_c^r, X_c^r) coincides with (Y^r, X^r), where (Y^r, X^r) converges to the fluid model.

For the modified fluid scaled process (Y_c^r, X_c^r), the associated counting process for upward and downward jumps for $t \leq \tau_c^r$ are:

$Y_c^r \uparrow (t) = r^{-1} N_2(r\beta \int_0^t [X_c^r(s) + \lambda/\beta] ds)$
$Y_c^r \downarrow (t) = r^{-1} N_1(\lambda r t)$
$X_c^r \uparrow (t) = r^{-1}\gamma N_1(\lambda r t) + r^{-1} N_3(r\epsilon \int_0^t (Y_c^r(s))^- ds)$
$X_c^r \downarrow (t) = r^{-1}\gamma N_2(r\beta \int_0^t [X_c^r(s) + \lambda/\beta] ds) + r^{-1} N_4(r\epsilon \int_0^t (Y_c^r(s))^+ ds)$

These counting processes are halted at time τ_c^r. For $0 \leq t \leq \tau_c^r$, the original and modified processes (Y^r, X^r) and (Y_c^r, X_c^r) coincide. Thus,

$Y_c^r(t) = Y^r(0) + Y_c^r \uparrow (t) - Y_c^r \downarrow (t)$
$X_c^r(t) = Z_c^r(t) + (-\lambda/\beta - min_{0 \leq s \leq t} Z_c^r(s)) \vee 0$
$Z_c^r(t) = X^r(0) + X_c^r \uparrow (t) - X_c^r \downarrow (t)$

With the functional strong law of large numbers and the fact that the processes Y_c^r and X_c^r are uniformly bounded by construction, we see that, w.p.1. for any subsequence of r , there exists a further sub- sequence along which the set of trajectories $(Y_c^r \uparrow, Y_c^r \downarrow, X_c^r \uparrow, X_c^r \downarrow)$ converges u.o.c. to a set of non-decreasing Lipschitz continuous functions $(y_c^r \uparrow, y_c^r \downarrow, x_c^r \uparrow, x_c^r \downarrow)$. Taking the limit:
$y_c^r \uparrow = \beta \int_0^t (x_c(s) + \lambda/\beta) ds)$
$y_c^r \downarrow = \lambda t$
$x_c^r \uparrow = \gamma\lambda t + \int_0^t y_c^-(s) ds$
$x_c^r \downarrow = \gamma\beta \int_0^t (x_c(s) + \lambda/\beta) ds + \epsilon \int_0^t y_m^+(s) ds$

Along the chosen subsequence, u.o.c . convergence of $(Y^r, X^r$ to the fluid model (y ,x) holds. This means that w.p.1 the u.o.c.convergence of (Y^r, X^r) to (y,x) holds for the original sequence. Thus, theorem 1 is proved.

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 (Y^r, X^r) converges to the Dirac measure concentrated at (0, 0).

Proof of Theorem 2

We use lemma 5 in the paper as an assumption for this proof. That is, 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,
$(y(t), x(t)) \to (0, 0)$,t $\to \infty$ and $max_{t \geq 0} \|(y(t), x(t))\|$ is bounded.

Thus, we need to prove that For all sufficiently large r, the process (X^r, Y^r) is stable, with a unique stationary distribution and the sequence of stationary distributions of (X^r, Y^r) is tight.

Define s(t) = (Y(t), X(t)) as the random process. Next, consider the embedded Markov chain with fixed constants $\delta > 0$ and $\tau_{max} > 0$. For the process starting from a given state s = s(0), consider the random stopping time τ_δ(s), which is the first time t when $|\|s(t)\| - \|s\|| \geq \delta$; we then define the stopping time τ (s)= τ_δ(s) τ_{max}. Define a sequence of stopping times $\tau^{(k)}$, k =1,2,... by
$\tau(1) = \tau$(s(0)), ...
τ (k) = $[\tau^{k-1} + \tau^{max}] \wedge \inf(t > \tau^{k-1} : | \|s(t)\|_* \ \|s(\tau^{k-1})\|_* | \geq \delta)$,
for k = 2, 3, ...

Consider the embedded discrete-time Markov chain s(k), k = 0, 1, ..., using τ(k) as sampling times. Specifically, if s(t), $t \geq 0$, is the original continuous-time Markov process, then:
$\hat{s}(0) = s(0), \hat{s}(k) = s(\tau(k))$, k = 1,2,....

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

$$\mathrm{E}[\phi^2(\hat{s}(1)) - \phi^2(\hat{s}(0)) \mid \hat{s}(0)] \leq -C_1\phi(\hat{s}(0)) + C_2.$$

For some constant $\delta_7 > 0$, for any sequence r $\to \infty$ and corresponding $\hat{s}(0) = \hat{s^r}(0)$ such that $\|s^r(0)\|\hat{*} \uparrow \infty$, we have

$$\mathbf{P}[\phi(s^r(1)) - \phi(s^r(0)) \leq -\delta_7] \to 1$$

It suffices to consider a sequence such that the convergence $\frac{1}{\|s^r(0)\|_*} s^r(0) \to s$ holds, for some vector s with $\|s\|_* = 1$.
We will study the behavior of the continuous-time process s(t), with initial state $s(0) = \hat{s}(0)$, on the interval [0, τ(s(0))].

For any vector s = (y,x) we denote $s\prime = (y\prime, x\prime)$, where $y\prime = \beta$ x, and $x\prime = -\varepsilon y - \gamma\beta x$ Similarly, let $\|s\|_*$ denote (d/dt)$\|s(t)\|_*$ when s(t) = s. Suppose that s=(y,x) and $x > 0$, then, (the sequence of processes can be constructed on a common probability space, such that) w.p.1, u.o.c:

$$s(t/\|s(0)\|_*) - s(0) \to s\prime t \text{ and } \|s(t/\|s(0)\|_*)\|_* - \|s(0)\|* \to \|s\|\prime_* t.$$

We see that $\tau(\mathrm{s}(0)) = \tau_\delta\ (\mathrm{s}(0)) \to 0$, and therefore,

$$\mathbf{P}[\phi(s^r(1)) - \phi(s^r(0)) \le -\delta_7] \to 1 \text{ holds with } \delta_7 = \delta.$$

Consider the sub-case when $[x(0) - (-\lambda/\beta)]/ \mid x\prime \mid \to \infty$;

We check that $s(t/\|s(0)\|_*) - s(0) \to s\prime t$ and $\|s(t/\|s(0)\|_*)\|_* - \|s(0)\|_* \to \|s\|\prime_* t$ 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,$\|s(t)\|_*$, decreases by δ before time τ_{hit}, and $\mathbf{P}[\phi(s^r(1)) - \phi(s^r(0)) \le -\delta_7] \to 1$ follows.

Finally, consider the sub-case when [x (0) -$(-\lambda/\beta)$]/ $\mid x\prime \mid \to c\epsilon[0, \infty)$.

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 $s(t/\|s(0)\|_*) - s(0) \to s\prime t$ and $\|s(t/\|s(0)\|_*) - \|s(0)\|_* \to \|s\|_* t$. From time $\tau_{hit}\|s(0)\|_*$ to infinity, the process continues in actual time, without slowing down. W.p.1. in the limit we obtain the trajectory which satisfies $s(t/\|s(0)\|_*)s(0) \to s\prime t$ and $\|s(t/\|s(0)\|_*)\| - \|s(0)\|_* \to \|s\|_* t$ in the interval [0, c], and then in the interval [c,∞) we have x(t) = $-\lambda/\beta$ and y $\prime(t)$ =-λ. In both intervals, the limit trajectory is such that the norm $\|s(t)\|_*$ is decreasing at least at some positive rate.

From $\mathrm{E}[\phi^2(\hat{s(1)}) - \phi^2(\hat{s(0)}) \mid \hat{s(0)}] \le -C_1\phi(\hat{s(0)}) + C_2$ 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, $\mathrm{E}\phi(\hat{s(\infty)} \le C_2/C_1$.

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

Now we use the relation between stationary distributions of the original continuous-time process and the sampled chain:
$\mathbf{P}[s(\infty)\epsilon A] = \mathbf{E}[\mathrm{E}\ [\int_0^{\tau(s(0)} I[s(t)\epsilon A dt \mid s(0) = \hat{s(\infty)}]]]/[\mathrm{E}[\tau(\hat{s(\infty)})]$
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:

$\mathbf{E}\|s(\infty)\|_* \leq \frac{\mathbf{E}\|\hat{s}\|_* + 2\delta]\tau_{max}}{C_4/2} \geq C_5\mathbf{E}\|s(\infty)\|_* + C_6 \leq C_7$.
Therefore, the uniform bound on the expected norm in steady-state implies the tightness of stationary distributions. Theorem 2 is thus, proved.

Theorem 3

Suppose there exists a sequence of deterministic initial states, 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 fixed in R^2. Then,

$$(\hat{Y}^r, \hat{X}^r)(\hat{Y}, \hat{X})$$

Where $(\hat{Y}, \hat{X})$ are unique solutions to the stochastic differential equations.

Proof of Theorem 3

We have $(\hat{Y}^r(0), \hat{X}^r(0)) \to (\hat{Y}(0), \hat{X}(0))$. In addition we know that (Y^r, X^r) are unique solutions to the fluid scaled process such that $(Y^r(0), X^r(0)) \to (0, 0)$. Applying Theorem 1 we have that, $(Y^r(t), X^r(t))$ converges uniformly on compact to the fluid limit which is $((y(t), x(t)) = (0, 0)$ as $r \to \infty$ Hence, with the probability of 1.

$$Y^r \to 0, X^r \to 0$$

gives us that,

$$\int_0^t Y^r(s)ds \to 0$$

$$\int_0^t X^r(s)ds \to 0$$

which are uniformly on compact. This implies that, with the probability of 1 on a finite time set $[0, T]$, it is not hard to reach The boundary of $\hat{X}^r$ which is sufficiently large r For instance, $\hat{X}^r > -\lambda\frac{\sqrt{r}}{\beta}$ Using the representations of $Y^r(t)$ and $X^r(t)$ in theorem 1 and We have from theorem 1 that,

$$Y^r(t) = Y^r(0) + N_2(\beta \int_0^t X^r(s)ds) - N_1(\lambda rt)$$

$$X^r(t) = Z^r + (-\min_{0\leq s\leq t} Z^r(s))\ 0$$

$$Z^r(t) = X^r(0) + \gamma N_1(\lambda rt) - \gamma N_2(\beta \int_0^t X^r(s)ds) + N_3(\epsilon \int_0^t (Y^r(s))^- ds) - N_4(\epsilon \int_0^t (Y^r(s))^+ ds)$$

applying aspects of Lemma 9 which states. A unit rate Poisson Process $\prod(t) : t \geq 0$ can be realized as some probability space as a standard Brownian motion $\beta(t) : t \geq 0$, such that there exists a random positive variable ϵ which

represents a finite moment generating function in a neighbourhood of the origin. In addition holding that $Z^r = X^r$ in the representations from theorem 1. We finally get,

$$\hat{Y}^r(t) = Y^r(0) + \beta \int_0^t X^r(s)ds + \beta_2(t) - \beta_1(t) + \delta_1^r(t)$$

$$\hat{X}^r(t) = \hat{X}^r(0) - \gamma \int_0^t \beta \hat{X}^r(s)ds - \epsilon \int_0^t \hat{Y}^r(s)ds + \gamma\beta_1(t) + \gamma\beta_2(t) + \delta_2^r(t)$$

where $\delta_i^r(t)$ is a constant for $i = 1, 2$ By letting $r \to \infty$ and applying Lemma 10 which states the mapping of a certain number ψ is continuous in the topology of convergence uniformly on compact. We obtain a probability of 1. Thus the convergence which is uniformly on compact of $(\hat{Y}^r, \hat{X}^r)$ to a limiting distribution process $(\hat{Y}, \hat{X})$. Thus, satisfies the convergence condition of theorem 3.

Theorem 4

The sequence of stationary distributions of 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 $(\hat{(}Y), \hat{(}X))$

Proof of Theorem 4

This proof of the tightness of stationary distributions on diffusion scale—can be obtained by the following steps. Firstly we need to show that 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 $r^1(Y^r, X^r\frac{\lambda r}{\beta})$ are tight and asymptotically concentrate at (0,0). This is shown in theorem 2 such that for all sufficiently large r, the system is stable. The sequence of stationary distributions of the fluid scaled process $(\hat{Y}^r, \hat{X}^r)$ converges to the Dirac measure concentrated at (0,0).

The second step is by showing that $r^{\frac{1}{2}+k}$ - scales tightness, for any k in $(0, \frac{1}{2})$. This is basically saying that the tightness of stationary distribution of $r^{\frac{1}{2}-k}(Y^r, X^r\frac{\lambda r}{\beta})$ This can be shown by adapting Tightness of invariant distributions of a large-scale flexible service system under a priority discipline, which is when considering the scaling limit of the system as the arrival rate of customers and number of servers in each pool tend to infinity in proportion to a scaling parameter r. While having the overall system load remains strictly sub critical. Having to Index the systems by parameter r, we show that the system under LAP discipline is stochastically stable for all sufficiently large r and that the

family of the invariant distributions is tight on scales $r^{-\frac{1}{2}+\epsilon}$ for all $\epsilon > 0$ (More precisely, the sequence of invariant distributions, centered at the equilibrium point and scaled down by $r^{-\frac{1}{2}+\epsilon}$ is tight.

Lastly we need to show the diffusion-scale $(r^{\frac{1}{2}})$-scale tightness is used namely $(r^{\frac{1}{2}+k}$-scale tightness at the starting point and follows that the Diffusion scale tightness of invariant distributions of a large-scale flexible service system. For instance lets a consider a large-scale service system with multiple customer classes and multiple server pools, with the mean service time depending both on the customer class and server pool. Where the allowed activities of routing choices form a tree. We then prove tightness of diffusion-scaled (centered at the equilibrium point and scaled down by $r^{1/2}$ invariant distributions (Similar to the proof in theorem 2). As this happens, we need obtain a limit interchange result which the limit of diffusion-scaled invariant distributions is equal to the invariant distribution of the limiting diffusion process.

References

Höpfner, Reinhard. "Null Recurrent Birth-and-Death Processes, Limits of Certain Martingales, and Local Asymptotic Mixed Normality." Scandinavian Journal of Statistics, vol. 17, no. 3, 1990, pp. 201–215. JSTOR, www.jstor.org/stable/4616169.

Pang, Guodong and Alexander L. Stolyar. "A service system with on-demand agent invitations." Queueing Systems 82 (2016): 259-283.