Probability in the Engineering and Informational Sciences (2025), **39:**3 344–369
doi:10.1017/S0269964824000251

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The number of overlapping customers in Erlang-A queues: an asymptotic approach

Jamol Pender[1], Young Myoung Ko[2] and Jin Xu[3]

[1]School of Operations Research and Information Engineering, Cornell University, Ithaca, New York, USA
[2]Department of Industrial and Management Engineering, Pohang University of Science Technology, Pohang-si, Gyeongsangbuk-do, Korea
[3]School of Management, Huazhong University of Science and Technology, Wuhan, Hubei, China
Corresponding author: Jamol Pender; Email: jjp274@cornell.edu

Keywords: abandonment; Covid-19; epidemics; Markov processes; queueing

Abstract
In this paper, we investigate the number of customers that overlap or coincide with a virtual customer in an Erlang-A queue. Our analysis starts with the fluid and diffusion limit differential equations to obtain the mean and variance of the queue length. We then develop precise approximations for waiting times using fluid limits and the polygamma function. Building on this, we introduce a novel approximation scheme to calculate the mean and variance of the number of overlapping customers. This method facilitates the assessment of transient overlap risks in complex service systems, offering a useful tool for service providers to mitigate significant overlaps during pandemic seasons.

1. Introduction

Gathering during pandemic seasons can be risky. As people congregate in stores to buy essential goods such as water and non-perishable items, the risk of spreading the virus increases. Many service facilities have implemented protective measures, including transparent barriers, air filtration systems, and mandatory mask-wearing for all customers. Additionally, various social and physical distancing protocols have been adopted to limit close interactions between individuals, as discussed by Bove and Benoit [4]. Although these measures can reduce the risk of infection to some extent, customers still inevitably interact with each other when present in the same area. In certain places where maintaining distance is not always feasible, such as in workplaces or crowded retail environments, understanding the extent to which customers interact with one another is important.

Most previous studies on COVID-19 have used deterministic compartmentalized models to estimate infection rates and spread dynamics (e.g., Kaplan [22], Nguemdjo *et al*. [33], and Dandekar *et al*. [7]). However, in service systems where customers' arrival and departure processes are stochastic, it is crucial to incorporate the stochastic effects in modeling the infection risk (Kang *et al*. [21], Forien *et al*. [17], Drakopoulos *et al*. [13] , and Palomo *et al*. [37]).

In this work, we aim to characterize customer interactions in service systems, where arrivals, service, and abandonment behaviors are stochastic. The number of overlaps, defined as the count of people a customer interacts with during their time in the service system, is a key metric for assessing close contact. This metric is particularly important in the context of contact tracing (see World Health Organization [47]). A higher overlap count increases the likelihood of a customer being exposed to an infected individual. Conversely, when an infected customer is present, this metric also indicates how many others are at risk of exposure.

Recent research has shed light on calculating customers' overlaps in queueing systems. Most of the existing studies evaluate customers' overlaps in the stationary queues. For instance, Kang

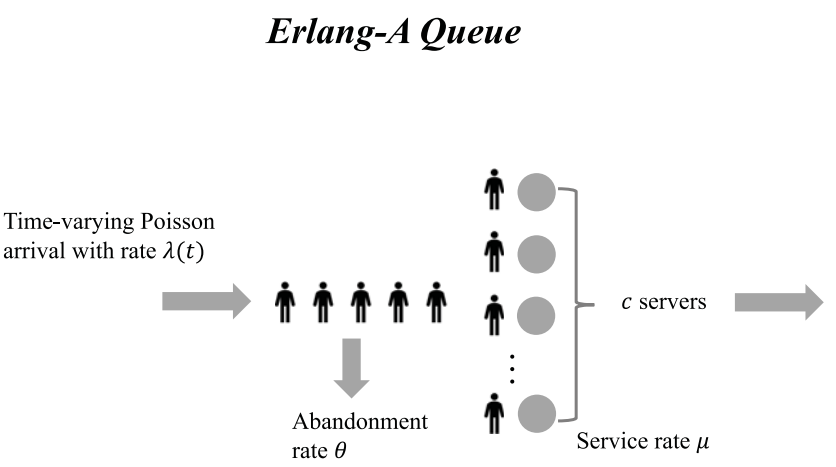

Figure 1. *A demonstrative graph of the $M_t/M/c+M$ queueing system. A virtual customer who arrives at this particular time will overlap with* $c+5$ *customers immediately and will overlap with other new arrivals during her sojourn time in the queue.*

et al. [21] demonstrate how overlaps can be used to calculate a new R_0 value for analyzing infection rates in stationary $M/M/c$ queues. Xu *et al*. [48] derive the number of overlaps in the $M/M/c$ systems with different queueing topology. Meanwhile, Palomo and Pender [35] prove that the tail of the distribution of the overlap time between any pair of customers is exponential for the $M/M/1$ queue and depends explicitly on the service distribution, where an overlap time is defined as the duration that both customers are in the queue together. Palomo and Pender [38] extend the analysis to the case of batch arrivals, highlighting the practical applications of this analysis for transportation systems like trains and buses. Boxma and Pender [5] consider the overlap times in $G/G/1$ queues. These studies mainly concentrate on steady-state overlap distributions, neglecting the transient dynamics of overlaps.

There are very few studies considering the transient behavior of overlaps. Ko and Xu [23] provide an approximation scheme of the overlap times in a time-varying multi-server queue. Palomo and Pender [36] investigate the number of overlaps in an infinite server queue, where the number of overlaps is equivalent to the queue length upon arrival plus the number of additional arrivals during service, as there is no waiting. However, these studies do not capture the realistic scenario where the service provider only has a finite number of servers and customers can thus be impatient due to long waits. Understanding the transient behavior of overlaps in such a finite-server system where customers may abandon the queue is crucial for assessing the risk of joining realistic service systems during peak times.

Our work aims to address these research gaps in the literature by investigating the transient distribution of the number of overlaps in $M_t/M/c+M$ systems. In this system, customers arrive according to a time-varying Poisson process with a rate $\lambda(t)$, and service at each of the c servers follows an exponential distribution with a rate μ. Customers will abandon the queue if their waiting time exceeds their patience threshold, which is exponentially distributed with rate θ. A demonstrative graph for the $M_t/M/c+M$ system is provided in Figure 1. This investigation into the transient behavior of overlaps is challenging because the analytic frameworks used for steady-state analysis in previous studies by the authors in Refs. [5, 48] will not apply. Additionally, the number of overlaps derived for an infinite server system, as discussed by Palomo and Pender [36] can only serve as a lower bound for any finite server system that includes customer abandonment.

1.1. Main contributions of the paper

The main contributions of this work can be summarized as follows:

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

- To address the challenges identified, we derived exact expressions for the Erlang-A fluid and diffusion differential equations, which allowed us to establish fluid and diffusion limits for the queue length process. Using the digamma and trigamma functions, we then developed new approximations for the mean and variance of waiting times in the Erlang-A queue. Building on these results, we introduced new approximations for the mean and variance of the number of overlaps, which can potentially be adapted to other time-varying queueing systems.
- Our paper presents a new analysis that provides critical insights into finite server systems with customer abandonment. Specifically, our analysis enables us to determine the number of overlaps that occur accurately and sheds light on the distribution of potential overlaps, allowing us to establish prediction intervals. This analysis is practically applicable in preventing large overlaps and serves as a valuable tool for designing service systems by adjusting arrival rates, service distributions, and server numbers. Our work significantly enhances the understanding of queueing systems and can guide decision-making across various service system applications.

1.2. Organization of the paper

The remainder of this paper is structured as follows. Section 2 begins by introducing the Erlang-A queueing model and deriving exact expressions for the fluid and diffusion variance differential equations. In Section 3, we present new approximations for the mean and variance of waiting times in the Erlang-A queue. Building on these, we introduce novel approximations for the mean and variance of the number of overlaps for a virtual customer. The effectiveness of these approximations is then validated through simulation experiments in Section 4. Lastly, Section 5 summarizes our findings and outlines promising directions for future research in this field.

2. The Erlang-A queueing model

The Erlang-A model incorporates customer abandonment, which is a very important feature of real-world service systems. In particular, Mandelbaum *et al*. [30] show that the queue length process for an $M_t/M/c+M$ queueing system $Q \equiv \{Q(t)|t \geq 0\}$ is represented by the following stochastic process:

$$Q(t) = Q(0) + \Pi_1\left(\int_0^t \lambda(s)ds\right) - \Pi_2\left(\int_0^t \mu \cdot (Q(s) \wedge c)ds\right) - \Pi_3\left(\int_0^t \theta \cdot (Q(s) - c)^+ ds\right), \tag{1}$$

where $\Pi_i \equiv \{\Pi_i(t)|t \geq 0\}$ for $i = 1, 2, 3$ are i.i.d. standard (rate 1) Poisson processes, $(x \wedge y) = \min\{x, y\}$, and $x^+ = \max\{x, 0\}$. In view of Mandelbaum *et al*. [30], the sample path of the queue length process of the Erlang-A queue described in Eq. (1) can be uniquely decomposed into three independent unit rate Poisson processes. These processes include the arrival process with rate $\lambda(t)$, the service departure process with rate μ, and the abandonment process with rate θ. The general idea of Eq. (1) is that the queue length at any given time is the initial queue length at time 0, plus the number of arrivals over the period $[0, t]$, minus the number of departures due to service completion or abandonment.

The Erlang-A queueing model and its variants are extensively studied in the queueing literature, see for example Zeltyn and Mandelbaum [49], Whitt [46], Mandelbaum and Zeltyn [31], Gurvich *et al*. [18], Engblom and Pender [16], Pender [41], Niyirora and Pender [34], Aktekin and Ekin [1], Braverman *et al*. [6], Pender [42], Pender and Massey [43], Bitton *et al*. [3], Azriel *et al*. [2], and van Leeuwaarden *et al*. [45]. The Erlang-A queue is widely investigated because three classic queueing models are special cases of it. The $M/M/\infty$, $M/M/c/c$, and $M/M/c/\infty$ queues are all special cases of the Erlang-A queue. The $M/M/\infty$ is obtained from the Erlang-A queue in two ways. The first way is to set the number of servers to infinity, that is, $c = \infty$. The second way is to make $\mu = \theta$. The $M/M/c/c$ is obtained from the Erlang-A queue by letting θ get large, that is, $\theta \to \infty$. This blocking phenomenon was first observed

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

in Hampshire *et al*. [20]. Finally, the $M/M/c/\infty$ is obtained by letting $\theta = 0$. This is obvious as the abandonment process is shut off.

In Halfin and Whitt [19], an important insight emerged regarding multi-server queueing systems, emphasizing the ability to scale up both the arrival rate and the number of servers simultaneously. This scaling approach, known as the *Halfin–Whitt* scaling, has become pivotal in modeling call centers within queueing literature, as illustrated in works such as Pang *et al*. [39]. Given that the $M(t)/M/c+M$ queueing process is a specialized instance of a single-node *Markovian service network*, we can extend this concept to construct a *uniformly accelerated* queueing process. In this accelerated scenario, both the new arrival rate $\eta \cdot \lambda(t)$ and the new number of servers $\eta \cdot c$ are proportionally scaled by a common factor $\eta > 0$. By applying the *Halfin–Whitt* scaling to the Erlang-A model, we derive the following sample path representation for the queue length process as

$$\begin{aligned} Q^{\eta}(t) &= Q^{\eta}(0) + \Pi_1\left(\int_0^t \eta \cdot \lambda(s)ds\right) - \Pi_2\left(\int_0^t \mu \cdot (Q^{\eta}(s) \wedge \eta \cdot c)ds\right) \\ &\quad - \Pi_3\left(\int_0^t \theta \cdot (Q^{\eta}(s) - \eta \cdot c)^+ ds\right) \\ &= Q^{\eta}(0) + \Pi_1\left(\int_0^t \eta \cdot \lambda(s)ds\right) - \Pi_2\left(\int_0^t \eta \cdot \mu \cdot \left(\frac{Q^{\eta}(s)}{\eta} \wedge c\right) ds\right) \\ &\quad - \Pi_3\left(\int_0^t \eta \cdot \theta \cdot \left(\frac{Q^{\eta}(s)}{\eta} - c\right)^+ ds\right). \end{aligned} \tag{2}$$

The derivation of Eq. (2) is similar to Eq. (1) and the only difference is that the arrival process and the number of servers are scaled by a factor of η. Moreover, taking the Halfin–Whitt limit gives us the *fluid* models of Mandelbaum *et al*. [30], that is,

$$\lim_{\eta \to \infty} \frac{1}{\eta} Q^{\eta}(t) = q(t) \text{ a.s.} \tag{3}$$

where the deterministic process $q(t)$, the *fluid mean*, satisfies the following one-dimensional ordinary differential equation (ODE),

$$\dot{q}(t) = \lambda(t) - \mu \cdot (q(t) \wedge c) - \theta \cdot (q(t) - c)^+. \tag{4}$$

Moreover, the diffusion limit converges to a diffusion process, that is,

$$\lim_{\eta \to \infty} \sqrt{\eta}\left(\frac{1}{\eta} Q^{\eta}(t) - q(t)\right) \Rightarrow \tilde{Q}(t),$$

and the variance of the diffusion is given by the following ordinary differential equation (the detailed derivation can be found in Theorem 5.2 of Mandelbaum *et al*. [30]),

$$\begin{aligned} \dot{\text{Var}}\left[\tilde{Q}(t)\right] &= \lambda(t) + \mu \cdot (q(t) \wedge c) + \theta \cdot (q(t) - c)^+ \\ &\quad - 2 \cdot \text{Var}\left[\tilde{Q}(t)\right] \cdot (\mu \cdot \{q(t) < c\} + \theta \cdot \{q(t) \geq c\}). \end{aligned} \tag{5}$$

In the following sections, we provide an extensive analysis of the differential equations derived from the fluid and diffusion limits. Although these equations are well-documented in the literature, this paper offers the first detailed analysis of the fluid and diffusion differential equation dynamics in the constant arrival rate case. We begin with the analysis of the differential equations of the fluid limits.

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

2.1. Fluid analysis

In this section, we present results regarding the fluid limit dynamics under a constant arrival rate. Before delving into our main findings related to the fluid analysis, we first provide a standard result for linear differential equations.

Lemma 2.1. *Let $q(t)$ be the solution to the following differential equation*

$$\dot{q} = \lambda(t) - \mu(t)q(t)$$

where $q(0) = q_0$. Then the solution for any value of t is given by

$$q(t) = q_0 \exp\left\{-\int_0^t \mu(s)ds\right\} + \left(\exp\left\{-\int_0^t \mu(s)ds\right\} \cdot \left(\int_0^t \lambda(s) \exp\left\{\int_0^s \mu(r)dr\right\} ds\right)\right).$$

Proof. This follows from standard results on ordinary differential equations by varying parameters (see Tenenbaum and Pollard [44]). □

Now that we characterize the dynamics of $q(t)$ with time-varying arrival rate $\lambda(t)$ and service rate $\mu(t)$, our next result provides the explicit solution to the fluid mean $q(t)$ in the steady state, that is, when the arrival and service rates are both constant.

Proposition 2.2. *When $\lambda(t)$ and $\mu(t)$ are constants, the solution $q(t)$ to Eq. (4) is given as follows:*

$$q(t) = \begin{cases} \frac{\lambda-\mu c+\theta c}{\theta} + \left(q(0) - \frac{\lambda-\mu c+\theta c}{\theta}\right) e^{-\theta t} & \text{if } q(0) > c, \lambda > \mu c \\ \frac{\lambda-\mu c+\theta c}{\theta} + \left(q(0) - \frac{\lambda-\mu c+\theta c}{\theta}\right) e^{-\theta t}, & \text{if } q(0) > c, \lambda \le \mu c, t \le t_1^* \\ \frac{\lambda}{\mu} + \left(c - \frac{\lambda}{\mu}\right) e^{-\mu(t-t^*)}, & \text{if } q(0) > c, \lambda \le \mu c, t > t_1^* \\ \frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right) e^{-\mu t} & \text{if } q(0) \le c, \lambda \le \mu c \\ \frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right) e^{-\mu t}, & \text{if } q(0) \le c, \lambda > \mu c, t \le t_2^* \\ \frac{\lambda-\mu c+\theta c}{\theta} + \left(\frac{-\lambda+\mu c}{\theta}\right) e^{-\theta(t-t^*)}, & \text{if } q(0) \le c, \lambda > \mu c, t > t_2^* \end{cases} \tag{6}$$

where $t_1^* = \frac{\log\left(\frac{\theta q(0)-\lambda+\mu c-\theta c}{\mu c-\lambda}\right)}{\theta}$ *and* $t_2^* = \frac{\log\left(\frac{q(0)-\frac{\lambda}{\mu}}{c-\frac{\lambda}{\mu}}\right)}{\mu}$.

Proof. When $q(0) > c$ and $\lambda > \mu c$, we apply Lemma 2.1 to the differential equation given in Eq. (4). When $q(0) > c$ and $\lambda \le \mu c$, from time 0 to time t_1^* (which is the time that the differential equation hits the value c), we can apply Lemma 2.1 to the differential equation to obtain the solution. After t_1^*, we know the solution is larger than c so that it follows a new differential equation and the solution is also given by Lemma 2.1. The proofs of the other cases are similar, so we omit the details here. This completes the proof. □

Now that we have completely described the dynamics of the fluid queue length as a function of time in the previous proposition, hence we can easily obtain the next corollary following Proposition 2.2.

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

Corollary 2.3. *Suppose that $\lambda(t)$ and $\mu(t)$ are constants, and $q(\infty) = \lim_{t\to\infty} q(t)$. Then we have*

$$q(\infty) = \begin{cases} \frac{\lambda-\mu c+\theta c}{\theta} & \text{if } \lambda > \mu c \\ \frac{\lambda}{\mu} & \text{if } \lambda \le \mu c. \end{cases}$$

Proof. *The results follow from Proposition 2.2 by letting $t \to \infty$.* □

We then turn our attention to analyzing the diffusion variance differential equations. Analyzing the fluid equations first is essential since the diffusion variance differential equations depend on the fluid dynamics in an explicit way.

2.2. Diffusion variance analysis

In this section, we provide an analysis of the diffusion variance differential equations given in Eq. (5). We start with explicit solutions to the diffusion variance equations.

Proposition 2.4. *When $\lambda(t)$ and $\mu(t)$ are constants, the solution $v(t)$ to Eq. (5), is given as follows:*

$$v(t) = \begin{cases} \left(v(0) - q(0) - \frac{\mu c-\theta c}{\theta}\right) e^{-2\theta t} + \frac{\lambda}{\theta} + \left(q(0) - \frac{\lambda-\mu c+\theta c}{\theta}\right) e^{-\theta t}, & \text{if } q(0) > c, \lambda > \mu c \\ \left(v(0) - q(0) - \frac{\mu c-\theta c}{\theta}\right) e^{-2\theta t} + \frac{\lambda}{\theta} + \left(q(0) - \frac{\lambda-\mu c+\theta c}{\theta}\right) e^{-\theta t}, & \text{if } q(0) > c, \lambda \le \mu c, t \le t_1^* \\ (v(t_1^*) - c)e^{-2\mu(t-t_1^*)} + \frac{\lambda}{\mu} + \left(c - \frac{\lambda}{\mu}\right) e^{-\mu(t-t_1^*)}, & \text{if } q(0) > c, \lambda \le \mu c, t > t_1^* \\ (v(0) - q(0))\, e^{-2\mu t} + \left(1 - e^{-\mu t}\right) \cdot \frac{\lambda}{\mu} + q(0)e^{-\mu t}, & \text{if } q(0) \le c, \lambda \le \mu c \\ (v(0) - q(0))\, e^{-2\mu t} + \left(1 - e^{-\mu t}\right) \cdot \frac{\lambda}{\mu} + q(0)e^{-\mu t}, & \text{if } q(0) \le c, \lambda > \mu c, t \le t_2^* \\ (v(t_2^*) - c)e^{-2\theta(t-t_2^*)} + \frac{\lambda}{\theta} + \left(\frac{-\lambda+\mu c}{\theta}\right) e^{-\theta(t-t_2^*)}, & \text{if } q(0) \le c, \lambda > \mu c, t > t_2^* \end{cases} \tag{7}$$

where

$$t_1^* = \frac{\log\left(\frac{\theta q(0)-\lambda+\mu c-\theta c}{\mu c-\lambda}\right)}{\theta} \quad \text{and} \quad t_2^* = \frac{\log\left(\frac{q(0)-\frac{\lambda}{\mu}}{c-\frac{\lambda}{\mu}}\right)}{\mu}.$$

Proof. *See Appendix A.* □

Now that we have provided an analysis of the transient dynamics of the diffusion variance differential equations, we analyze the steady-state behavior of the diffusion variance differential equation in the following result.

Corollary 2.5. *Let $v(t)$ be the solution to Eq. (5), then in steady state, we have*

$$v(\infty) = \begin{cases} \frac{\lambda}{\mu}, \text{if } \lambda \le \mu c \\ \frac{\lambda}{\theta}, \text{if } \lambda > \mu c. \end{cases}$$

Proof. *The result follows directly from Proposition 2.4 by letting $t \to \infty$ in Eq. (7).* □

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

What is apparent from the steady-state variance is that, it is the same as the steady-state mean queue length when $\lambda \leq \mu c$. This is because, in this regime, the system behaves like an infinite server queue with a Poisson arrival process. Furthermore, when the arrival rate λ is larger than the maximum service rate μc, the steady-state variance becomes λ/θ. This observation is intriguing because when $\theta < \mu$, the mean is smaller than the variance, indicating over-dispersion. When $\theta \geq \mu$, the steady-state variance is smaller than the steady-state mean queue length, indicating under-dispersion. For a more detailed discussion on the relationships between mean and variance in the Erlang-A queue, see Daw and Pender [11].

Now that we have a good understanding of the fluid mean and diffusion variance of the Erlang-A queue, we will show how to use these results in the context of studying overlaps in the Erlang-A queue.

3. The number of overlaps

In this section, we introduce the virtual overlap process, which counts the number of customers that the virtual customer (a hypothetical customer arriving at time t) will overlap with during their time in the queue. This process is crucial from an epidemiological perspective, as it represents the number of individuals who would need to be contact traced for potential exposure if the virtual customer were infectious. For further details, see Kang *et al*. [21], Palomo *et al*. [37], and Xu *et al*. [48].

Similar to the infinite server setting, the virtual customer will overlap with the customers already present in the queue and the customers that arrive during the virtual customer's service time. However, unlike the infinite server setting, the virtual customer—who we assume does not abandon the queue—must also overlap with customers who arrive during their wait for service. Thus, the total number of overlaps for the virtual customer equals the number of customers present upon their arrival plus those they encounter during their sojourn time in the queue. This number of overlaps can be expressed in terms of the queue length $Q(t)$, the virtual waiting time $W(t)$, and the service time of the virtual customer $\mathcal{S}$, that is,

$$O(t) = \underbrace{N\left(t+\mathcal{S}+W(t)\right)-N(t)}_{\text{Arrivals During Service and Wait}} + \underbrace{Q(t)}_{\text{Queue Upon Arrival}}, \tag{8}$$

where $N(t)$ denotes the number of arrivals until time t. Using the above representation, we can compute the mean number of overlaps at time t by taking the expectation of the overlap process. Thus, the mean number of overlaps can be written as

$$\begin{aligned}\mathbb{E}\left[O(t)\right] &= \mathbb{E}\left[N\left(t+\mathcal{S}+W(t)\right)-N(t)+Q(t)\right]\\ &= \mathbb{E}\left[N\left(t+\mathcal{S}+W(t)\right)-N(t)\right]+\mathbb{E}\left[Q(t)\right]\\ &= \lambda\mathbb{E}\left[\mathcal{S}\right]+\lambda\mathbb{E}\left[W(t)\right]+\mathbb{E}\left[Q(t)\right].\end{aligned} \tag{9}$$

The last equality of Eq. (9) follows from the fact that the arrival process is Poisson. Unfortunately, the transient mean queue length and the transient mean wait time are not known in closed form for the Erlang-A queue, except in the case, where $\mu = \theta$. This is a major difference between the Erlang-A and the infinite server queue. In the infinite server queue, the mean wait time is zero and the mean queue can be written as an explicit integral with respect to the service time distribution, see, for example, Eick *et al*. [14, 15]. Consequently, we will use limit theory to approximate the mean number of overlaps.

3.1. The transient mean number of overlaps

In this section, we show how to approximate the transient mean number of overlaps using asymptotic analysis. We first leverage the results of Mandelbaum *et al*. [30], which prove almost sure limit theorems for the queue length process in the Halfin–Whitt regime. However, we also need results for the virtual

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

waiting time in order to fully analyze the overlap process. To do this, we will exploit a recent result by Massey and Pender [32], which proves the following theorem.

Theorem 3.1. *Let $W^\eta(t)$ be the virtual wait time of a customer at time t who is not going to abandon, in the scaled process. Then we have*

$$\lim_{\eta\to\infty} W^\eta(t) \quad \stackrel{a.s}{=} \quad w(t),$$

where w(t) satisfies the following equation

$$w(t) \quad = \quad \frac{1}{\theta}\log\left(1+\frac{\theta\cdot(q(t)-c)^+}{\mu c}\right). \tag{10}$$

Moreover, as $t\to\infty$ we have that when $\lambda > \mu c$

$$\lim_{t\to\infty} w(t) \quad = \quad \frac{1}{\theta}\log\left(\frac{\lambda}{\mu c}\right).$$

Proof. See Theorem 6 of Massey and Pender [32]. □

This result shows that the limiting virtual waiting time is a function of the fluid queue length function $q(t)$. When the fluid queue length is less than the number of servers c, then the virtual waiting time is equal to zero. Moreover, when the fluid queue length is greater than the number of servers c, then the virtual waiting time is positive. What is more important is that the fluid limit for the virtual waiting time yields a deterministic function of time. Thus, we are able to approximate the virtual waiting time with a non-random function of time. Now that we have a limiting expression for the virtual waiting time, we can define the scaled overlap process as follows:

$$O^\eta(t) = N^\eta\left(t+\mathcal{S}+W^\eta(t)\right) - N^\eta(t) + Q^\eta(t). \tag{11}$$

We will use the scaled overlap process to prove our main result of the paper, which gives us a deterministic approximation for the transient mean number of overlaps in the Erlang-A queue.

Theorem 3.2. *Let $O^\eta(t)$ be the scaled number of people that a customer who arrives at time t will overlap with. Then, we have*

$$\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[O^\eta(t)\right] \quad = \quad \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(1+\frac{\theta\cdot(q(t)-c)^+}{\mu c}\right)+q(t),$$

where q(t) is given in Proposition 2.2. Moreover, when $t\to\infty$ we have that

$$\lim_{t\to\infty}\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[O^\eta(t)\right] \quad = \quad \begin{cases}\frac{2\lambda}{\mu}, \text{if } \lambda\le\mu c\\ \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(\frac{\lambda}{\mu c}\right)+c+\frac{\lambda-\mu c}{\theta}, \text{if } \lambda>\mu c.\end{cases} \tag{12}$$

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

Proof. Based on Eqs. (3), (9), (10), and (11), we have

$$\begin{aligned}\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[O^{\eta}(t)\right] &= \lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[\left(N^{\eta}\left(t+\mathcal{S}+W^{\eta}(t)\right)-N^{\eta}(t)\right)\right]+\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[Q^{\eta}(t)\right]\\ &= \lim_{\eta\to\infty}\frac{1}{\eta}\left(\eta\lambda\mathbb{E}\left[\mathcal{S}\right]+\eta\lambda\mathbb{E}\left[W^{\eta}(t)\right]\right)+q(t)\\ &= \frac{\lambda}{\mu}+\lambda w(t)+q(t)\\ &= \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(1+\frac{\theta\cdot(q(t)-c)^{+}}{\mu c}\right)+q(t).\end{aligned}$$

In particular, when we look at steady state, we have that when $\lambda \le \mu c$,

$$\begin{aligned}\lim_{t\to\infty}\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[O^{\eta}(t)\right] &= \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(1+\frac{\theta\cdot(q(\infty)-c)^{+}}{\mu c}\right)+q(\infty)\\ &= \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(1+\frac{\theta\cdot\left(\frac{\lambda}{\mu}-c\right)^{+}}{\mu c}\right)+\frac{\lambda}{\mu}\\ &= \frac{2\lambda}{\mu}.\end{aligned}$$

When $\lambda > \mu c$, we have

$$\begin{aligned}\lim_{t\to\infty}\lim_{\eta\to\infty}\frac{1}{\eta}\mathbb{E}\left[O^{\eta}(t)\right] &= \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(1+\frac{\theta\cdot(q(\infty)-c)^{+}}{\mu c}\right)+q(\infty)\\ &= \frac{\lambda}{\mu}+\frac{\lambda}{\theta}\log\left(\frac{\lambda}{\mu c}\right)+c+\frac{\lambda-\mu c}{\theta}.\end{aligned}$$

This completes the proof. □

It is important to note that the expectation for Theorem 3.2 is necessary. Without the expectation the overlap process would explicitly depend on $\mathcal{S}$ and would be random. It is also worth noting that in the case $\lambda \le \mu c$, the steady-state mean number of overlaps is equal to $\frac{2\lambda}{\mu}$. This is important from a probability perspective since it is also equal to the steady-state mean number of overlaps in the infinite server queue as well. One should also note that the result does not depend on the staffing level c since it behaves like an infinite server in the steady-state setting.

As we stated earlier, the Erlang-A model is a generalization of multiple queueing systems. We now discuss several special cases of the Erlang-A model in the following based on our results in Theorem 3.2.

Remark 3.3. (M/M/∞ system): When $c \to \infty$, all the customers will enter the service directly upon arrival. The system with constant arrival and service rates becomes the M/M/∞ system. We find from Eq. (12) that the number of overlapped customers will converge to $\frac{2\lambda}{\mu}$. This result matches our analysis in Xu *et al.* [48]. It is important to distinguish between letting $c \to \infty$ and letting $\theta \to \mu$. Notably, as c approaches infinity, the waiting time converges to zero. Conversely, when θ tends toward μ, the waiting time does not approach zero. Consequently, although the queue lengths are identical and share the same sample path structure, the expected number of overlaps differs between the systems. This observation of the two different systems emphasizes that the number of overlaps is dependent on the customer's experience instead of only the queue length process. In particular, the customers in the case where

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

$c \to \infty$ all have the same experience in the system, however, in the case where $\theta \to \mu$, abandoning customers have a different overlap experience than those who wait in the queue and then get served. Finally, it is also worth noting that the diffusion limits of the two systems are different, see for example page 167 of Mandelbaum *et al*. [30]. This further highlights the difference in the number of overlaps experienced by customers in the two systems.

Remark 3.4. (Erlang-B system): When letting the abandonment rate $\theta \to \infty$, we have the Erlang-B system. According to Theorem 3.2, the expected number of people that the virtual customer overlaps in the steady state is $\frac{2\lambda}{\mu}$ when $\lambda \leq \mu c$, and is $\frac{\lambda}{\mu} + c$ when $\lambda > \mu c$. The reason is that no customer will wait in the queue, so the virtual customer will always enter the service directly. According to Corollary 2.3, the virtual customer will overlap with $\frac{\lambda}{\mu}$ customers in the system when $\lambda \leq \mu c$, and c customers when $\lambda > \mu c$. Moreover, during the service time, the virtual customer is expected to overlap with $\frac{\lambda}{\mu}$ newly arrived customers. Note that although the abandonment rate is infinity, we still assume that every customer will arrive at the system first before abandoning, which explains why the virtual customer will overlap with new arrivals.

Remark 3.5. (Erlang-C system): When shutting off the abandonment process, we have the Erlang-C system. Based on Eq. (12), the number of overlapped customers in the fluid limit becomes $2\frac{\lambda}{\mu}$ for $\lambda \leq \mu c$. The reason is that the fluid limit of the virtual waiting time becomes 0 in the steady state, according to Theorem 3.1. In the fluid limit, the virtual customer will overlap with $\frac{\lambda}{\mu}$ customers upon arrival, and other $\frac{\lambda}{\mu}$ customers during service.

The transient mean queue length and wait time are not known in closed form. However, if we condition on the queue length, the waiting time distribution is known in closed form for the Erlang-A model. By conditioning on the number of customers ahead of you given that you are waiting, it is easily seen that the waiting time has a hypoexponential distribution, that is,

$$W_k = \sum_{j=0}^{k} Y_j$$

where $Y_j \sim \text{Exp}(\mu c + \theta \cdot j)$ and k is the number of customers that are ahead of the customer upon arrival. Note that j is allowed to be zero when the queue length is identical to the number of servers since in this case, the customer needs to wait an exponential amount of time with rate μc. With this hypoexponential representation, we can compute the conditional mean and variance of the waiting time. The conditional mean waiting time is

$$\mathbb{E}[W_k] = \sum_{j=0}^{k} \frac{1}{\mu c + \theta \cdot j} = \frac{1}{\theta} \sum_{j=0}^{k} \frac{1}{\frac{\mu c}{\theta} + j} = \frac{1}{\theta} \left(\psi\left(\frac{\mu c}{\theta} + k + 1 \right) - \psi\left(\frac{\mu c}{\theta} \right) \right) \tag{13}$$

where $\psi(x)$ is the digamma function that satisfies $\psi(z+1) - \psi(z) = \frac{1}{z}$. Moreover, since Y_i are independent exponential variables, the variance of W_k can be given as

$$\text{Var}[W_k] = \sum_{j=0}^{k} \frac{1}{(\mu c + \theta \cdot j)^2} = \frac{1}{\theta^2} \left(\psi^{(1)}\left(\frac{\mu c}{\theta} \right) - \psi^{(1)}\left(\frac{\mu c}{\theta} + k + 1 \right) \right) \tag{14}$$

where $\psi^{(1)}(x)$ is the trigamma function that satisfies $\psi^{(1)}(z+1) - \psi^{(1)}(z) = -\frac{1}{z^2}$. One should note that both the digamma and the trigamma functions are special cases of the Hurwitz-Riemann zeta function defined as

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

$$\zeta(s,\alpha) = \sum_{j=0}^{\infty} \frac{1}{(j+\alpha)^s}.$$

Now that we have explicit expressions for the mean and variance of the waiting time given that there are k customers in front of the current arrival, we should be able to leverage the fluid limits for the queue length process to approximate the mean and variance of the waiting time at any time t.

For a continuously differentiable function $f(x)$, we have from a first-order Taylor expansion around the mean $\mathbb{E}[Q(t)]$ that

$$f(Q(t)) \approx f(\mathbb{E}[Q(t)]) + f'(\mathbb{E}[Q(t)]) \cdot (Q(t) - \mathbb{E}[Q(t)]). \tag{15}$$

Thus, taking the expectation on both sides of Eq. (15), the mean of the function can be approximated by

$$\mathbb{E}[f(Q(t))] \approx f(\mathbb{E}[Q(t)]) \approx f(q(t)). \tag{16}$$

Similarly, when taking the variance on both sides of Eq. (15), we can approximate the variance of the function by

$$\mathrm{Var}[f(Q(t))] \approx f'(\mathbb{E}[Q(t)])^2 \cdot \mathrm{Var}[Q(t)] \approx f'(q(t))^2 \cdot v(t). \tag{17}$$

Finally, it is also important to compute the following approximation for the covariance as well since we will need it later for computing the variance of the number of overlaps in the Erlang-A queue.

$$\begin{aligned} \mathrm{Cov}[f(Q(t)), Q(t)] &\approx \mathrm{Cov}[f(\mathbb{E}[Q(t)]) + f'(\mathbb{E}[Q(t)])(Q(t) - \mathbb{E}[Q(t)]), Q(t)] \\ &= f'(\mathbb{E}[Q(t)]) \cdot \mathrm{Var}[Q(t)] \\ &\approx f'(q(t)) \cdot v(t). \end{aligned} \tag{18}$$

It is worth mentioning that the covariance result is well-known in physics as the linear noise approximation. We will show in the sequel how to use these approximations for estimating the mean and variance of the waiting time for the Erlang-A queue. We will also find that these approximations of the waiting time are essential for computing the mean and variance of the number of overlaps in the Erlang-A queue as well.

Combining Eqs. (13) and (16), we obtain the following expression as an approximation for the mean waiting time

$$\begin{aligned} \mathbb{E}[W(t)] &= \mathbb{E}\left[\frac{1}{\theta}\left(\psi\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right) - \psi\left(\frac{\mu c}{\theta}\right)\right)\right] \\ &= \frac{1}{\theta}\mathbb{E}\left[\left(\psi\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right)\right)\right] - \frac{1}{\theta}\psi\left(\frac{\mu c}{\theta}\right) \\ &\approx \frac{1}{\theta}\left(\psi\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right) - \psi\left(\frac{\mu c}{\theta}\right)\right). \end{aligned} \tag{19}$$

Moreover, the first-order Taylor expansion also yields the following approximation for the variance of the waiting time

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

$$\begin{aligned}
\mathrm{Var}\left[W(t)\right] &= \mathbb{E}\left[\mathrm{Var}\left[W(t) \mid Q(t)\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[W(t) \mid Q(t)\right]\right] \\
&= \mathbb{E}\left[\frac{1}{\theta^2}\left(\psi^{(1)}\left(\frac{\mu c}{\theta}\right) - \psi^{(1)}\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right)\right)\right] \quad \text{by Equations (13) and (14)} \\
&+ \mathrm{Var}\left[\frac{1}{\theta}\left(\psi\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right) - \psi\left(\frac{\mu c}{\theta}\right)\right)\right] \\
&= \frac{1}{\theta^2}\left(\psi^{(1)}\left(\frac{\mu c}{\theta}\right) - \mathbb{E}\left[\psi^{(1)}\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right)\right]\right) \\
&+ \frac{1}{\theta^2}\mathrm{Var}\left[\psi\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right)\right] \\
&\approx \frac{1}{\theta^2}\left(\psi^{(1)}\left(\frac{\mu c}{\theta}\right) - \psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right)\right) \\
&+ \frac{1}{\theta^2}\cdot\psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right)^2 \cdot \{q(t) > c\} \cdot v(t) \qquad (20)
\end{aligned}$$

The last approximation of the above equation follows from Eqs. (16) and (17). Note that in the above approximation, we have replaced the value k in the conditional mean and variance formulas in Eqs. (13) and (14) with the fluid limit queue length at time t.

3.2. The transient variance of the number of overlaps

In addition to the mean, we are also interested in approximating the variance of the number of overlaps. With the variance, we are able to understand the variation around our approximations of the mean. This implies that we can construct prediction intervals for the number of overlaps one might expect at any time t. Using (8), note that the overlap process satisfies the following equation

$$O(t) = N\left(t + \mathcal{S} + W(t)\right) - N(t) + Q(t).$$

Thus, the variance of the number of overlaps is given in the following lemma.

Lemma 3.6. *The variance of the number of overlaps is equivalent to the following expression*

$$\begin{aligned}
\mathrm{Var}[O(t)] &= \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \lambda\mathbb{E}\left[W(t)\right] + \lambda^2\mathrm{Var}[W(t)] + \mathrm{Var}\left[Q(t)\right] \\
&+ 2\lambda\mathrm{Cov}\left[W(t), Q(t)\right].
\end{aligned}$$

Proof. See Appendix B. □

The exact variance of both $W(t)$ and $Q(t)$ and their covariance are unknown. Fortunately, because of the fluid and diffusion limits, we have approximations for the mean and variance of the queue length and waiting times. Thus, it remains for us to derive a transient approximation for the covariance between the waiting time and the queue length at time t. In order to compute the covariance of the two processes, we use a conditioning argument based on the queue length, which is natural given the conditional mean waiting time formula of Eq. (13). We outline this argument below.

Lemma 3.7. *The covariance of the waiting time process and the queue length process has the following expression in terms of the queue length process*

$$\mathrm{Cov}\left[W(t), Q(t)\right] = \mathrm{Cov}\left[\psi\left(\frac{\mu c}{\theta} + (Q(t)-c)^+\right), Q(t)\right] \qquad (21)$$

Proof. See Appendix C. □

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

Based on Lemma 3.7, we exploit the Taylor expansion for the covariance as Eq. (18), we have that

$$\text{Cov}\left[W(t), Q(t)\right] \approx v(t) \cdot \psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right) \cdot \{q(t) > c\}, \tag{22}$$

which provides an approximation for the covariance of the waiting time and the queue length process.

Based on Lemma 3.6, we now apply Eq. (19) digamma to approximate the expectation of waiting time, Eq. (20) to approximate the variance of waiting time, Eq. (22) to approximate the covariance between waiting time and queue length, and the diffusion variance of the solution to Eq. (5) to approximate the variance of queue length. We then obtain the approximation of the transient variance of the number of overlaps as follows:

$$\begin{aligned}
\text{Var}[O(t)] &= \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \lambda \mathbb{E}\left[W(t)\right] + \lambda^2 \text{Var}[W(t)] + \text{Var}\left[Q(t)\right] + 2\lambda \text{Cov}\left[W(t), Q(t)\right] \\
&\approx \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \frac{\lambda}{\theta}\left(\psi\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right) - \psi\left(\frac{\mu c}{\theta}\right)\right) + v(t) \\
&+ \frac{\lambda^2}{\theta^2}\left(\psi^{(1)}\left(\frac{\mu c}{\theta}\right) - \psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right)\right) \\
&+ \frac{\lambda^2}{\theta^2} \cdot \psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right)^2 \cdot \{q(t) > c\} \cdot v(t) \\
&+ 2\lambda v(t) \cdot \psi^{(1)}\left(\frac{\mu c}{\theta} + (q(t)-c)^+\right) \cdot \{q(t) > c\},
\end{aligned} \tag{23}$$

where $v(t)$ is given in Proposition 2.4. We can approximate the steady-state variance by setting $t \to \infty$. Thus, when $\lambda > \mu c$

$$\begin{aligned}
\lim_{t\to\infty} \text{Var}[O(t)] &\approx \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \frac{\lambda}{\theta}\left(\psi\left(\frac{\lambda}{\theta}\right) - \psi\left(\frac{\mu c}{\theta}\right)\right) + \frac{\lambda}{\theta} + \frac{\lambda^3}{\theta^3}\psi^{(1)}\left(\frac{\lambda}{\theta}\right)^2 \\
&+ \frac{\lambda^2}{\theta^2}\left(\psi^{(1)}\left(\frac{\mu c}{\theta}\right) - \psi^{(1)}\left(\frac{\lambda}{\theta}\right)\right) + 2\frac{\lambda^2}{\theta} \cdot \psi^{(1)}\left(\frac{\lambda}{\theta}\right)
\end{aligned}$$

and when $\lambda < \mu c$

$$\lim_{t\to\infty} \text{Var}[O(t)] \approx \frac{2\lambda}{\mu} + \frac{\lambda^2}{\mu^2}.$$

It is important to note that the last equation is precisely the same variance of the infinite server queue setting. Thus, when the queue is underloaded, the number of overlaps behaves similarly to an infinite server queue.

4. Numerical experiments

To understand how our approximations for the mean queue length and the mean number of overlapping customers perform, we present eight different numerical examples below. Before we provide the examples, we list the parameter values for each of the examples in Table 1. In all simulation examples, we simulate the sample paths 10,000 times to produce each curve.

In Figure 2, we present a plot of the simulated mean queue length, which has been approximated by employing the fluid limit as expressed in Eq. (4), denoted as "analytical" in the figures. Our findings indicate that across various parameter settings outlined in Table 1, the fluid approximation consistently

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

Table 1. *Parameters for examples.*

Parameters	λ	μ	θ	C	$Q(0)$
Figure (a)	10	1	0.5	30	10
Figure (b)	10	1	0.5	30	50
Figure (c)	10	1	2	30	10
Figure (d)	10	1	2	30	50
Figure (e)	40	1	0.5	30	10
Figure (f)	40	1	0.5	30	50
Figure (g)	40	1	2	30	10
Figure (h)	40	1	2	30	50

and accurately estimates the mean dynamics when juxtaposed with the simulated values. This observation underscores the reliability and effectiveness of the fluid approximation in accurately characterizing the behavior of the mean queue length.

In Figure 3, we plot the standard deviation of the queue length for the parameters outlined in Table 1. We observe that the approximation provided by the diffusion variance, as given in Eq. (5), consistently performs well at approximating the corresponding simulated values, for all of the parameter values. Thus, this illustrates the accuracy of the diffusion variance in effectively approximating the dynamic behavior of the queue length standard deviation. Moreover, with good approximations for the standard deviation, it becomes feasible to construct prediction intervals for the queue length. Such prediction intervals serve as valuable tools for assessing the range within which the actual queue length is likely to fall, providing a measure of confidence in our approximations.

In Figure 4, we plot the simulated mean virtual waiting time with two different approximations. The first approximation, denoted as "analytical 1" in the figures, is derived from the fluid limit, as outlined in Theorem 3.1 and is given in Eq. (10). The second approximation, denoted as "analytical 2" in the figure, is given by Eq. (19), which depends on the digamma function and the fluid queue length. We observe that across all parameter settings provided in Table 1, the approximation employing the digamma function consistently outperforms the fluid-based version in estimating the virtual waiting time.

In Figure 5, we plot the simulated standard deviation of the virtual waiting time, utilizing an approximation derived from the trigamma function as given in Eq. (20). The derivation of this approximation involved a Taylor expansion of the variance and the utilization of conditional mean and variance formulas for the wait time, as outlined in Eqs. (13) and (14). Our observations reveal that regardless of the parameter settings examined in Table 1, the trigamma function consistently demonstrates excellent performance in approximating the standard deviation of the virtual waiting time. This consistency and accuracy highlight the reliability and effectiveness of the trigamma-based approximation method.

Moving on to Figure 6, we explore the mean number of overlapping customers as a function of time. It is noteworthy that the approximations presented in Theorem 3.2, denoted as "analytical" in the figures, show remarkable accuracy across all parameter values considered. The high level of precision maintained by these approximations reinforces their reliability and usefulness in practical settings.

Figure 7 plots the standard deviation of the number of overlapping customers over time for the aforementioned eight examples in Table 1. Our investigation reveals that the two variance approximations presented in Eq. (5) ("analytical 1" in the figures) and Eq. (23) ("analytical 2" in the figures) both exhibit strong performance across all parameter values. The approximation based on the fluid and diffusion limits, as outlined in Eqs. (4) and (5), consistently outperforms the approximation given by Eq. (23). Consequently, with these robust approximations at our disposal, it becomes feasible to construct reliable prediction intervals for determining the customers that a virtual customer may overlap with.

By extending and refining our analysis using advanced approximation techniques derived from trigamma functions and fluid and diffusion limits, we have established a solid foundation for predicting and understanding various aspects of the virtual waiting time and customer overlap dynamics.

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

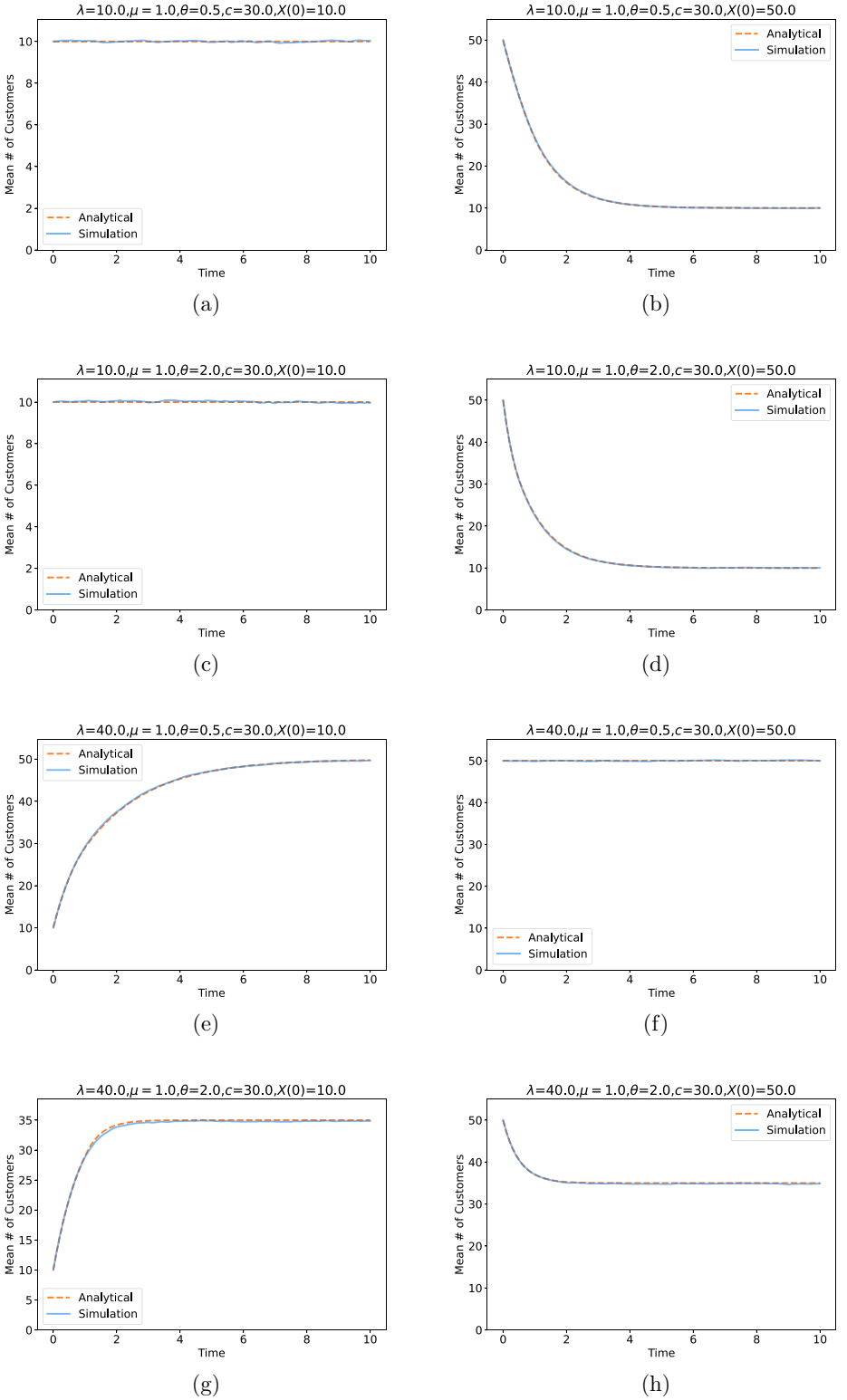

Figure 2. *Fluid mean number in system vs. simulation.*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

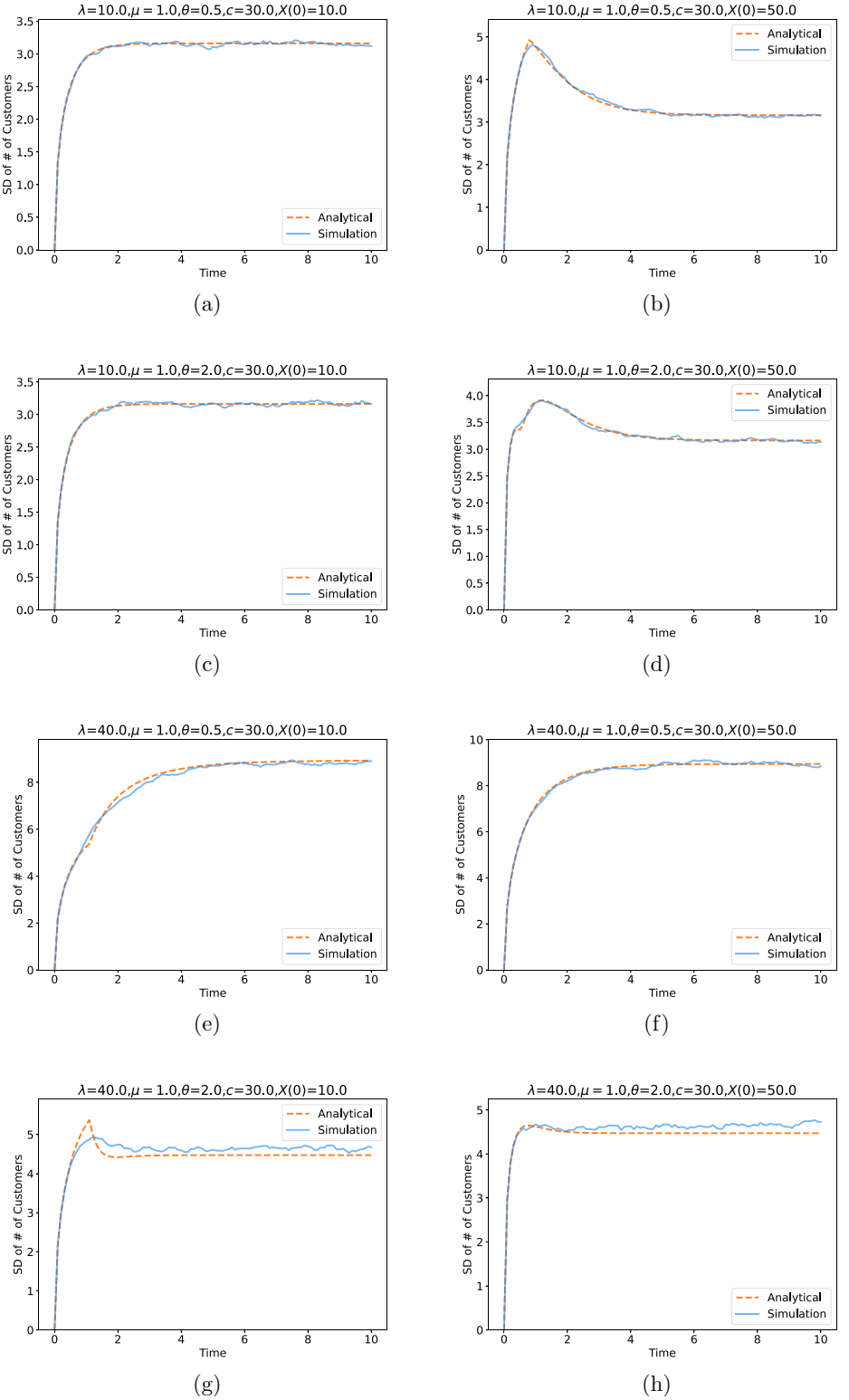

Figure 3. *Standard deviation number of customers (analytical vs. simulation).*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

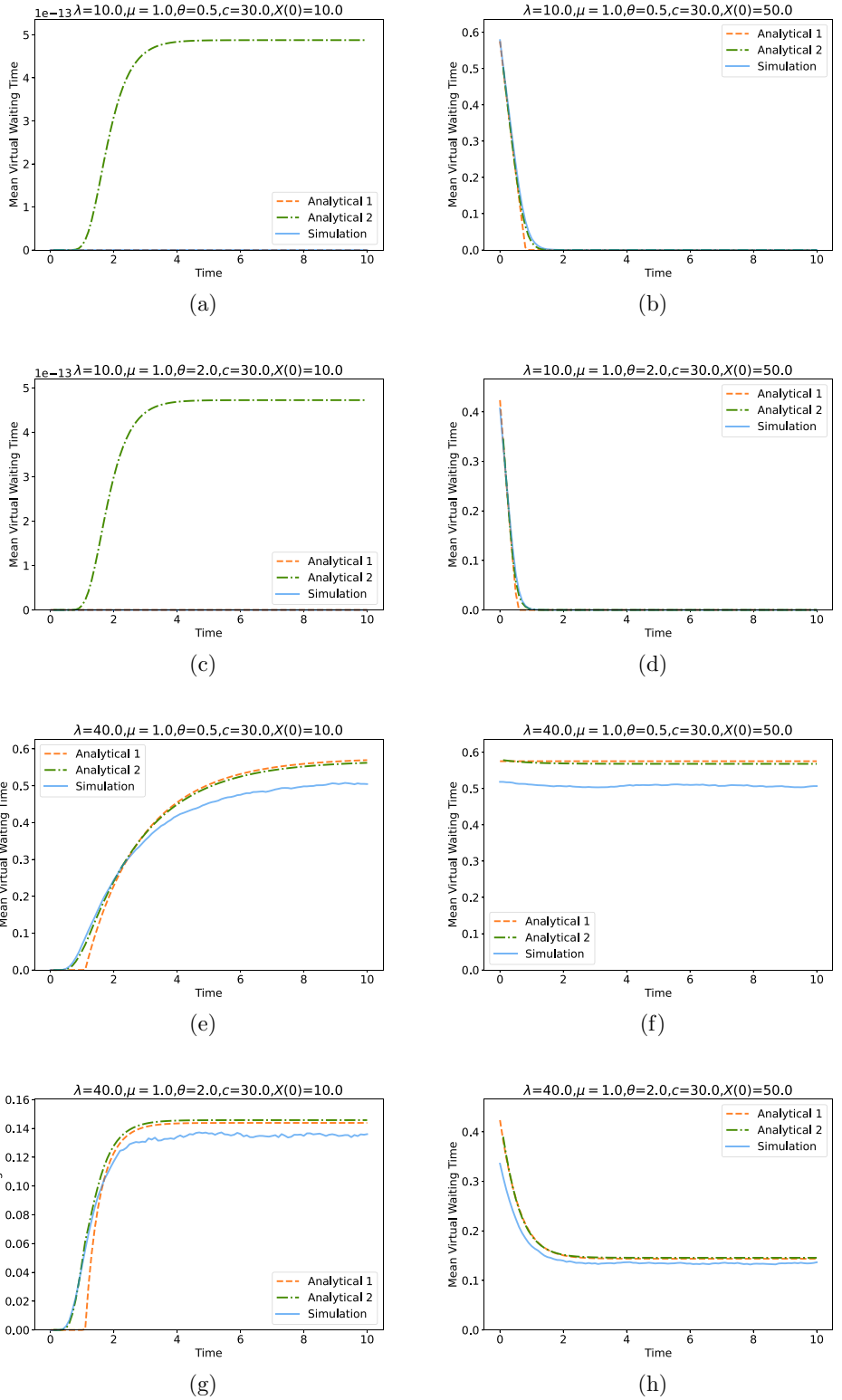

Figure 4. *Mean virtual waiting time (analytical vs. simulation).*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

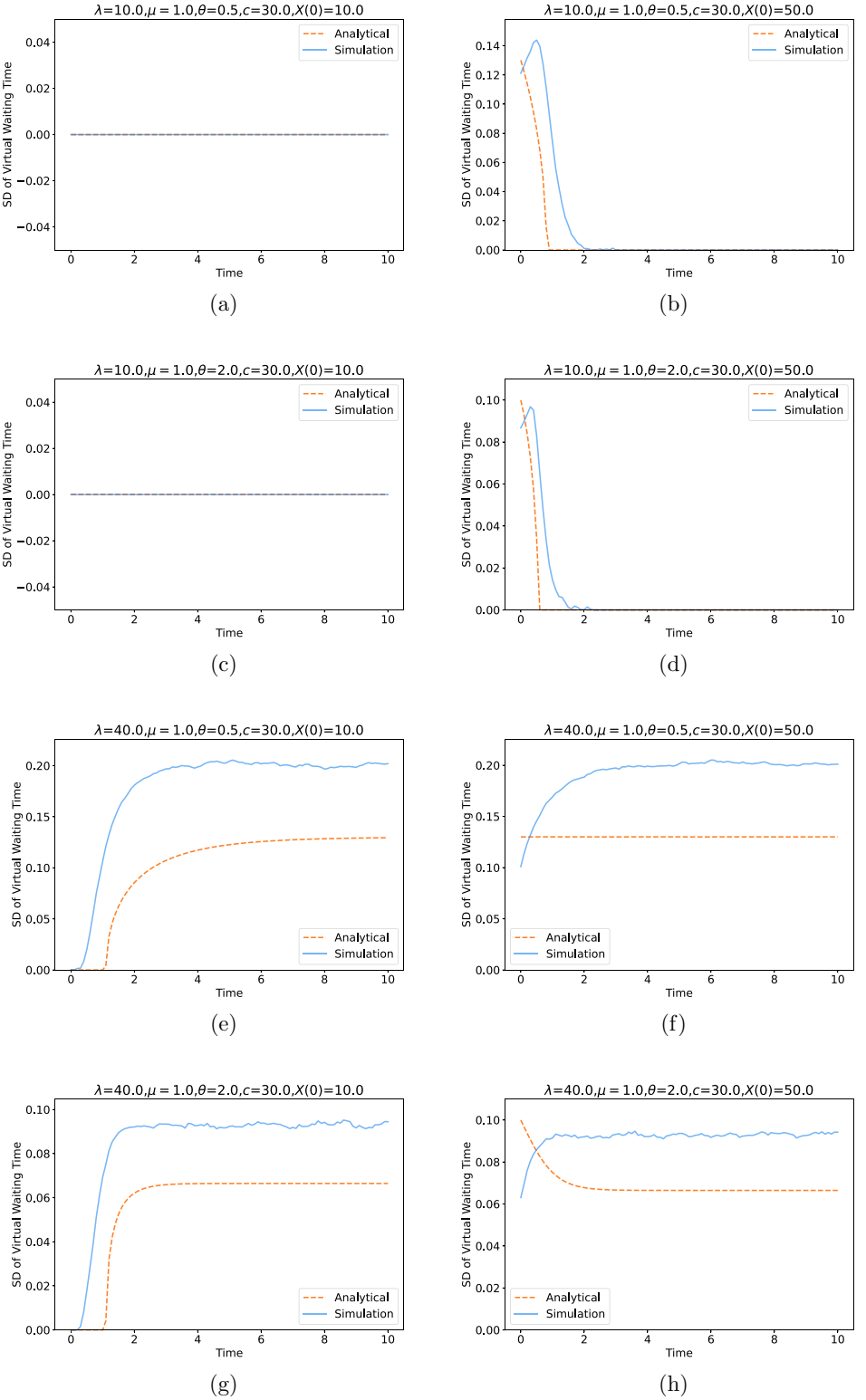

Figure 5. *Standard deviation of virtual waiting time (analytical vs. simulation).*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

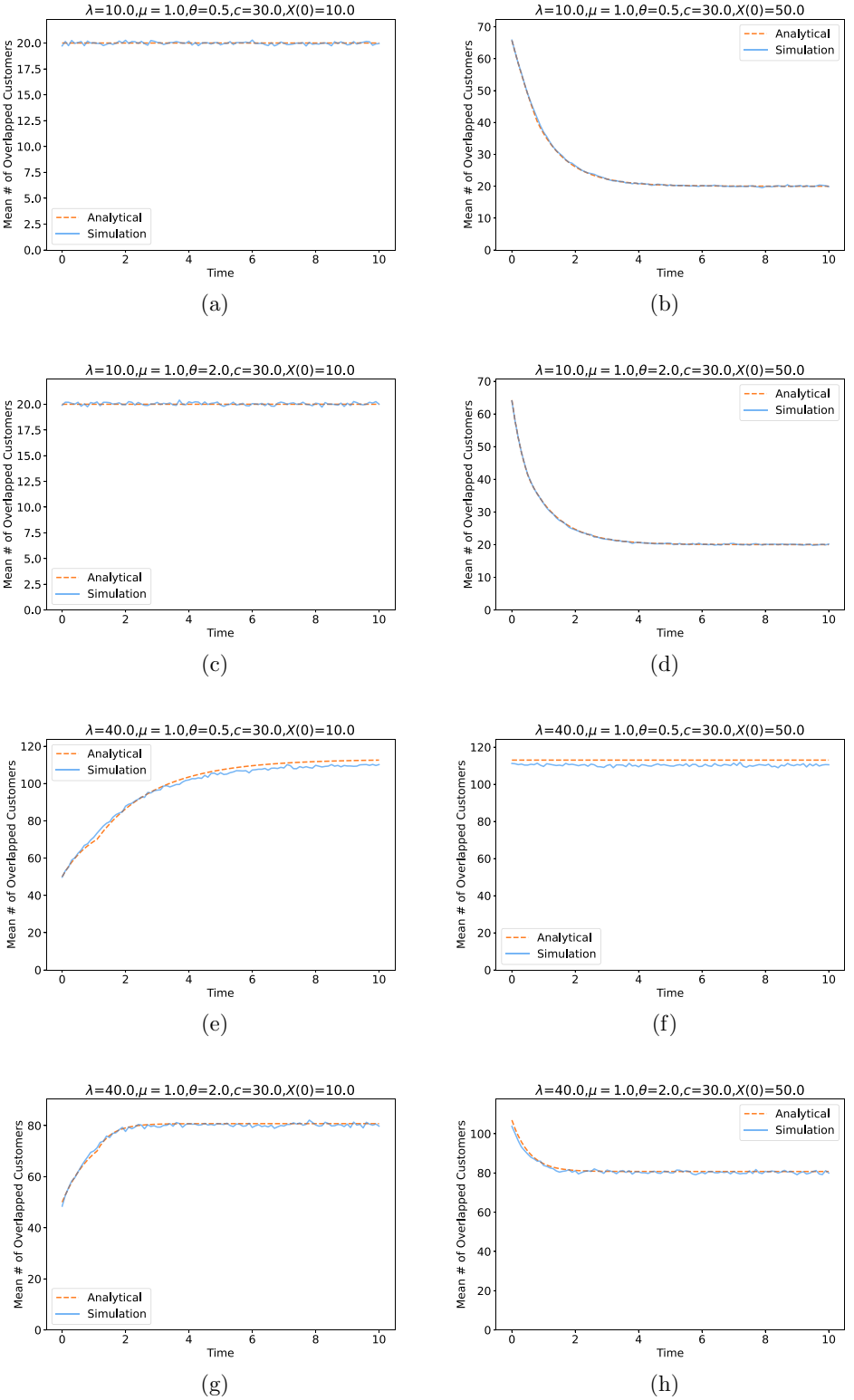

Figure 6. *Mean number of overlapping customers (analytical vs. simulation).*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

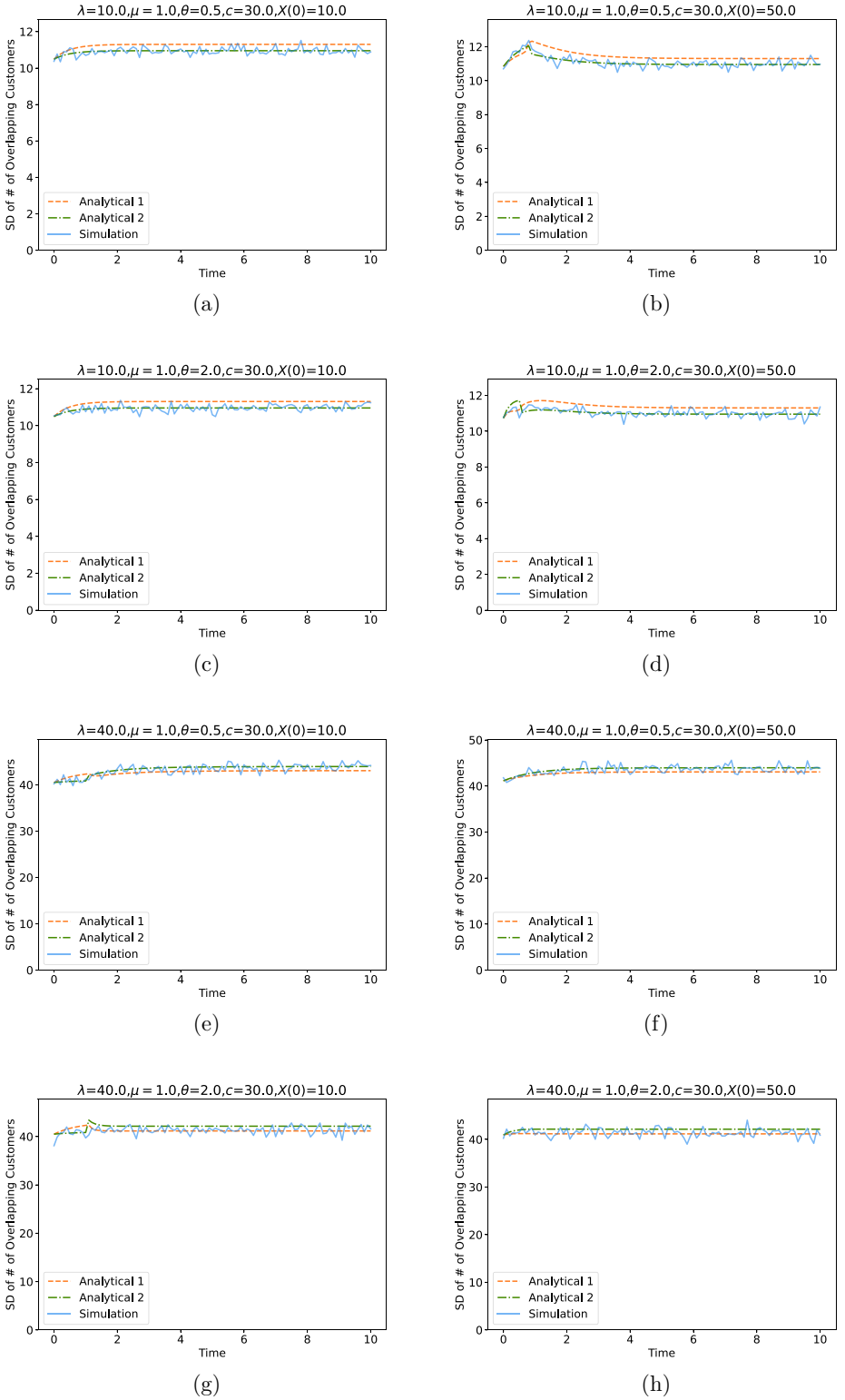

Figure 7. *Standard deviation of number of overlapping customers (analytical vs. simulation).*

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

These findings have significant implications for enhancing the efficiency and effectiveness of virtual customer service systems in a number of practical applications.

5. Conclusion

In this paper, we present a novel analysis of the mean and variance of the number of overlaps in the Erlang-A queue. Our contribution extends the current literature by considering both abandonment and a finite number of servers, thus providing a more realistic model. To achieve this, we employ a methodology based on the fluid and diffusion differential equations introduced by Mandelbaum *et al*. [29, 30] and Massey and Pender [32]. Specifically, we derive exact expressions for these equations using the theory of linear differential equations. Moreover, we utilize these exact expressions to approximate the number of overlaps for a virtual customer that will not be abandoned. Our results show that our fluid and diffusion-based approximations offer reliable estimates of the mean and variance of the number of overlapping customers in the Erlang-A queue.

As a side result, we also present new approximations for the mean and variance of the waiting time in the Erlang-A queue. Notably, our approximations are functions of the digamma and trigamma functions, respectively. These approximations offer a significant improvement over existing results and can be used to enhance the performance of queueing systems in practical applications. Overall, our work contributes to a better understanding of the behavior of the Erlang-A queue and provides useful tools for its analysis in real-world scenarios.

This work opens up several potential avenues for future research that could be valuable to pursue. First, we could explore a more general queueing model with abandonment, such as the $G/G/C + G$ queue. Although some limit theorems exist for this model Liu and Whitt [26, 27], the analysis of the virtual waiting time and its relationship with the queue length is currently unavailable. A thorough investigation of this relationship would provide insights into how the generality of the arrival, service, and abandonment processes could impact the number of overlaps.

Additionally, we could consider other types of queueing models, such as multidimensional network queueing models like those of Liu and Whitt [25, 28] and Pender and Massey [43], batch queueing models like those explored in Pang and Whitt [40], Daw and Pender [12], and Daw *et al*. [9], and even models with self-exciting arrivals like those in Koops *et al*. [24], Daw and Pender [10], and Daw *et al*. [8]. These extensions would offer further opportunities to investigate the impact of various system parameters on the number of overlaps and the waiting time and they represent promising avenues for future research that we intend to pursue.

Acknowledgements. Jamol Pender is supported by the National Science Foundation DMS Award # 2,206,286. Young Myoung Ko is supported in part by the National Research Foundation of Korea (NRF) grants (No. 2021R1A2C1094699 and 2021R1A4A1031019) funded by the Korea government (Ministry of Science and ICT, MSIT). Jin Xu is supported in part by the National Natural Science Foundation of China under Grant 72,301,113.

References

[1] Aktekin, T. & Ekin, T. (2016). Stochastic call center staffing with uncertain arrival, service and abandonment rates: A Bayesian perspective. *Naval Research Logistics (NRL)* 63(6): 460–478.
[2] Azriel, D., Feigin, P.D. & Mandelbaum, A. (2019). Erlang-s: a data-based model of servers in queueing networks. *Management Science* 65(10): 4607–4635.
[3] Bitton, S., Cohen, I. & Cohen, M. (2019). Joint repair sourcing and stocking policies for repairables using Erlang-A and Erlang-B queueing models. *IISE Transactions* 51(10): 1151–1166.
[4] Bove, L.L. & Benoit, S. (2020). Restrict, clean and protect: signaling consumer safety during the pandemic and beyond. *Journal of Service Management* 31(6): 1185–1202.
[5] Boxma, O. & Pender, J. (2024). Overlap times in the g/g/1 queue via laplace transforms.
[6] Braverman, A., Dai, J.G. & Feng, J. (2017). Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stochastic Systems* 6(2): 301–366.
[7] Dandekar, R., Henderson, S.G., Jansen, H.M., McDonald, J., Moka, S., Nazarathy, Y., Rackauckas, C., Taylor, P.G. & Vuorinen, A. (2021). Safe blues: the case for virtual safe virus spread in the long-term fight against epidemics. *Patterns* 2(3): 100220.

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

[8] Daw, A., Castellanos, A., Yom-Tov, G.B., Pender, J. & Gruendlinger, L. (2020a). The co-production of service: modeling service times in contact centers using Hawkes processes. *Management Science*. preprint arXiv:2004.07861.

[9] Daw, A., Fralix, B. & Pender, J. (2020b). Non-stationary queues with batch arrivals. preprint arXiv:2008.00625.

[10] Daw, A. & Pender, J. (2018). Queues driven by Hawkes processes. *Stochastic Systems* 8(3): 192–229.

[11] Daw, A. & Pender, J. (2019b). New perspectives on the Erlang-A queue. *Advances in Applied Probability* 51(1): 268–299.

[12] Daw, A. & Pender, J. (2019a). On the distributions of infinite server queues with batch arrivals. *Queueing Systems* 91(3): 367–401.

[13] Drakopoulos, K., Ozdaglar, A. & Tsitsiklis, J.N. (2017). When is a network epidemic hard to eliminate? *Mathematics of Operations Research* 42(1): 1–14.

[14] Eick, S.G., Massey, W.A. & Whitt, W. (1993a). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39(2): 241–252.

[15] Eick, S.G., Massey, W.A. & Whitt, W. (1993b). The physics of the $M_t/G/\infty$ queue. *Operations Research* 41(4): 731–742.

[16] Engblom, S. & Pender, J. (2014). Approximations for the moments of nonstationary and state dependent birth-death queues. preprint arXiv:1406.6164.

[17] Forien, R., Pang, G. & Pardoux, E. (2020). Epidemic models with varying infectiosity. preprint arXiv:2006.15377.

[18] Gurvich, I., Huang, J. & Mandelbaum, A. (2014). Excursion-based universal approximations for the Erlang-A queue in steady-state. *Mathematics of Operations Research* 39(2): 325–373.

[19] Halfin, S. & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations research* 29(3): 567–588.

[20] Hampshire, R.C., Jennings, O.B. & Massey, W.A. (2009). A time-varying call center design via Lagrangian mechanics. *Probability in the Engineering and Informational Sciences* 23(2): 231–259.

[21] Kang, K., Doroudi, S., Delasay, M. & Wickeham, A. (2023). A queueing-theoretic framework for evaluating transmission risks in service facilities during a pandemic. *Production and Operations Management* 32(5): 1453–1470.

[22] Kaplan, E.H. (2020). OM forum—COVID-19 scratch models to support local decisions. *Manufacturing & Service Operations Management* 22(4): 645–655.

[23] Young Myoung, K. & Jin, X. (2022). Overlapping time of a virtual customer in time-varying many-server queues. preprint arXiv:2211.03962.

[24] Koops, D.T., Saxena, M., Boxma, O.J. & Mandjes, M. (2018). Infinite-server queues with Hawkes input. *Journal of Applied Probability* 55(3): 920–943.

[25] Liu, Y. & Whitt, W. (2011). A network of time-varying many-server fluid queues with customer abandonment. *Operations research* 59(4): 835–846.

[26] Liu, Y. & Whitt, W. (2012). The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4): 405–444.

[27] Liu, Y. & Whitt, W. (2014a). Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability* 24(1): 378–421.

[28] Liu, Y. & Whitt, W. (2014b). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28(4): 419–449.

[29] Mandelbaum, A., Massey, W.A., Reiman, M.I. & Stolyar, A.L. (1999). Waiting time asymptotics for time varying multi-server queues with abandonment and retrials. In Proceedings of the Annual Allerton Conference on Communication Control and Computing, pp. 1095–1104. Vol. 37, The University, Citeseer.

[30] Mandelbaum, A., Massey, W.A. & Reiman, M.I.I. (1998). Strong approximations for Markovian service networks. *Queueing Systems* 30(1): 149–201.

[31] Mandelbaum, A. & Zeltyn, S. (2007). *Service engineering in action: the Palm/Erlang-A queue, with applications to call centers*. In *Advances in Services innovations*, Berlin, Heidelberg: Springer, pp.17–45.

[32] Massey, W.A. & Pender, J. (2018). Dynamic rate Erlang-A queues. *Queueing Systems* 89(1): 127–164.

[33] Nguemdjo, U., Meno, F., Dongfack, A. & Ventelou, B. (2020). Simulating the progression of the COVID-19 disease in Cameroon using SIR models. *PloS one* 15(8): e0237832.

[34] Niyirora, J. & Pender, J. (2016). Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)* 63(8): 615–630.

[35] Palomo, S. & Pender, J. (2021). Measuring the overlap with other customers in the single server queue. Submitted to the Proceedings of the 2021 Winter Simulation Conference. Phoenix: Institute of Electrical and Electronics Engineers, Inc, In KH Bae, B Feng, S Kim, S Lazarova-Molnar, Z Zheng, T Roeder, R Thiesing, editors, Piscataway, New Jersey.

[36] Palomo, S. & Pender, J. (2023a). Overlap times in the infinite server queue. *Probability in the Engineering and Informational Sciences* 1–7.

[37] Palomo, S., Pender, J.J., Massey, W.A. & Hampshire, R.C. (2023). Flattening the curve: Insights from queueing theory. *Plos one* 18(6): e0286501.

[38] Palomo, S.D. & Pender, J. (2023b). Overlap Times in the $GI^B/GI/\infty$ Queue. *arXiv preprint arXiv:2302.07410*.

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

[40] Pang, G. & Whitt, W. (2012). Infinite-server queues with batch arrivals and dependent service times. *Probability in the Engineering and Informational Sciences* 26(2): 197–220.

[41] Pender, J. (2014). Gram Charlier expansion for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics* 74(4): 1238–1265.
[42] Pender, J. (2017). Sampling the functional Kolmogorov forward equations for nonstationary queueing networks. *INFORMS Journal on Computing* 29(1): 1–17.
[43] Pender, J. & Massey, W.A. (2017). Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Informational Sciences* 31(1): 1–42.
[44] Tenenbaum, M. & Pollard, H. (1985). *Ordinary Differential equations: an Elementary Textbook for Students of mathematics, engineering, and the sciences*, Massachusetts: Courier Corporation.
[45] van Leeuwaarden, J.S.H., Mathijsen, B.W.J. & Zwart, B. (2019). Economies-of-scale in many-server queueing systems: tutorial and partial review of the QED Halfin-Whitt heavy-traffic regime. *SIAM Review* 61(3): 403–440.
[46] Whitt, W. (2006). Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations research* 54(2): 247–260.
[47] World Health Organization. Covid-19: physical distancing. https://www.who.int/westernpacific/emergencies/covid-19/information/physical-distancing, 2022.
[48] Jin, X., Young Myoung, K., Kong, M. & Pender, J. (2023). Queueing management for reducing the overlaps of customers in service systems. *Available at SSRN 4384706*.
[49] Zeltyn, S. & Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/N+ G queue. *Queueing Systems* 51(3): 361–402.

Appendix A. Proof of Proposition 2.4

Proof. We prove the results in Eq. (7) by discussing the following four cases.
CASE 1: For the first case with $q(0) > c$ and $\lambda > \mu c$, we have

$$\dot{v} = \lambda + \mu \cdot c + \theta \cdot (q(t) - c) - 2 \cdot \theta \cdot v(t).$$

Using the theory of linear ODEs, this implies that

$$\begin{aligned}
v(t) &= v(0)e^{-2\theta t} + e^{-2\theta t} \int_0^t e^{2\theta s} \left(\lambda + \mu c - \theta c + \theta q(s)\right) ds \\
&= v(0)e^{-2\theta t} + (1 - e^{-2\theta t}) \cdot \left(\frac{\lambda + \mu c - \theta c}{2\theta}\right) + \theta e^{-2\theta t} \int_0^t e^{2\theta s} q(s) ds \\
&= v(0)e^{-2\theta t} + (1 - e^{-2\theta t}) \cdot \left(\frac{\lambda + \mu c - \theta c}{2\theta}\right) \\
&+ \theta e^{-2\theta t} \int_0^t e^{2\theta s} \left(\frac{\lambda - \mu c + \theta c}{\theta} + e^{-\theta s} \left(q(0) - \frac{\lambda - \mu c + \theta c}{\theta}\right)\right) ds \\
&= v(0)e^{-2\theta t} + (1 - e^{-2\theta t}) \cdot \left(\frac{\lambda + \mu c - \theta c}{2\theta}\right) \\
&+ \left(\frac{\lambda - \mu c + \theta c}{2\theta}(1 - e^{-2\theta t}) + \left(q(0) - \frac{\lambda - \mu c + \theta c}{\theta}\right) \cdot \left(e^{-\theta t} - e^{-2\theta t}\right)\right) \\
&= \left(v(0) - q(0) + \frac{-\mu c + \theta c}{\theta}\right) e^{-2\theta t} + \frac{\lambda}{\theta} + \left(q(0) - \frac{\lambda - \mu c + \theta c}{\theta}\right) e^{-\theta t}.
\end{aligned}$$

We thus showed the first item of Eq. (7).
CASE 2: For the second case where $q(0) > c$ and $\lambda < \mu c$, we have

$$\dot{v} = \begin{cases} \lambda + \mu c + \theta q(t) - \theta c - 2\theta v(t), & \text{if } t \le t^* \\ \lambda + \mu q(t) - 2\mu v(t), & \text{if } t > t^* \end{cases}$$

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

where t^* is equal to

$$t_1^* = \frac{\log\left(\frac{\theta q(0)+\lambda-\mu c+\theta c}{\mu c-\lambda}\right)}{\theta}.$$

Using the theory of linear ODEs, this implies for $t \le t_1^*$ that

$$v(t) = \left(v(0) - q(0) + \frac{-\mu c + \theta c}{\theta}\right) e^{-2\theta t} + \frac{\lambda}{\theta} + \left(q(0) - \frac{\lambda - \mu c + \theta c}{\theta}\right) e^{-\theta t}.$$

Lastly for $t > t_1^*$ we have that

$$v(t) \quad = \quad (v(t_1^*) - c)e^{-2\mu(t-t_1^*)} + \frac{\lambda}{\mu} + \left(c - \frac{\lambda}{\mu}\right) e^{-\mu(t-t_1^*)}.$$

We hence proved the second and third items in Eq. (7).

CASE 3: For the third case where $q(0) \le c$ and $\lambda < \mu c$, we have

$$\dot{v} \quad = \quad \lambda + \mu \cdot q(t) - 2 \cdot \mu \cdot v(t).$$

Using the theory of linear ODEs, this implies that

$$\begin{aligned}
v(t) &= v(0)e^{-2\mu t} + e^{-2\mu t} \int_0^t e^{2\mu s} \left(\lambda + \mu q(s)\right) ds \\
&= v(0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right) \cdot \frac{\lambda}{2\mu} + \mu e^{-2\mu t} \int_0^t e^{2\mu s} q(s) ds \\
&= v(0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right) \cdot \frac{\lambda}{2\mu} + \mu e^{-2\mu t} \int_0^t e^{2\mu s} \left(\frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right) e^{-\mu s}\right) ds \\
&= v(0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right) \cdot \frac{\lambda}{\mu} + \mu e^{-2\mu t} \int_0^t \left(q(0) - \frac{\lambda}{\mu}\right) e^{\mu s} ds \\
&= v(0)e^{-2\mu t} + \left(1 - e^{-2\mu t}\right) \cdot \frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right) \left(e^{-\mu t} - e^{-2\mu t}\right) \\
&= \left(v(0) - q(0)\right) e^{-2\mu t} + \left(1 - e^{-\mu t}\right) \cdot \frac{\lambda}{\mu} + q(0)e^{-\mu t}.
\end{aligned}$$

We hence have the fourth item of Eq. (7).

CASE 4: For the fourth case where $\lambda > \mu c$ and $q(0) \le c$, we have

$$\dot{v} \quad = \quad \begin{cases} \lambda + \mu q(t) - 2\mu v(t), & \text{if } t \le t_2^* \\ \lambda + \mu c + \theta q(t) - \theta c - 2\theta v(t), & \text{if } t > t_2^* \end{cases}$$

where t^* is equal to

$$t_2^* = \frac{\log\left(\frac{q(0)-\frac{\lambda}{\mu}}{c-\frac{\lambda}{\mu}}\right)}{\mu}.$$

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

Using the theory of linear ODEs, this implies for $t \le t_2^*$ that

$$\begin{aligned} v(t) &= v(0)e^{-2\mu t} + e^{-2\mu t}\int_0^t e^{2\mu s}\left(\lambda + \mu q(s)\right) ds \\ &= v(0)e^{-2\mu t} + \frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right) \\ &+ \mu e^{-2\mu t}\int_0^t e^{2\mu s}\left(\frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right)e^{-\mu s}\right) ds \\ &= v(0)e^{-2\mu t} + \left(\frac{\lambda}{2\mu} + \frac{\lambda}{2\mu}\right)\left(1 - e^{-2\mu t}\right) + \mu e^{-2\mu t}\int_0^t e^{\mu s}\left(q(0) - \frac{\lambda}{\mu}\right) ds \\ &= v(0)e^{-2\mu t} + \left(\frac{\lambda}{2\mu} + \frac{\lambda}{2\mu}\right)\left(1 - e^{-2\mu t}\right) + \left(q(0) - \frac{\lambda}{\mu}\right)\left(e^{-\mu t} - e^{-2\mu t}\right) \\ &= (v(0) - q(0))e^{-2\mu t} + \frac{\lambda}{\mu} + \left(q(0) - \frac{\lambda}{\mu}\right)e^{-\mu t}. \end{aligned}$$

Lastly for $t > t_2^*$ we have that

$$v(t) = (v(t_2^*) - c)e^{-2\theta(t - t_2^*)} + \frac{\lambda}{\theta} + \left(\frac{-\lambda + \mu c}{\theta}\right)e^{-\theta(t - t_2^*)}.$$

We thus proved the fifth and sixth items of Eq. (7). □

Appendix B. Proof of Lemma 3.6

Proof. To compute the variance of the queue length $Q(t)$, we have

$$\begin{aligned} \mathrm{Var}[O(t)] &= \mathrm{Var}\left[N\left(t + \mathcal{S} + W(t)\right) - N(t) + Q(t)\right] \\ &= \mathrm{Var}\left[N\left(t + \mathcal{S}\right) - N(t) + N\left(t + \mathcal{S} + W(t)\right) - N\left(t + \mathcal{S}\right) + Q(t)\right] \\ &= \mathrm{Var}\left[N\left(t + \mathcal{S}\right) - N(t)\right] + \mathrm{Var}\left[N\left(t + \mathcal{S} + W(t)\right) - N\left(t + \mathcal{S}\right) + Q(t)\right] \\ &= \mathrm{Var}\left[N\left(t + \mathcal{S}\right) - N(t)\right] + \mathrm{Var}\left[N\left(W(t)\right) + Q(t)\right] \quad \text{by stationary Poisson} \\ &= \mathrm{Var}\left[N\left(t + \mathcal{S}\right) - N(t)\right] + \mathrm{Var}\left[N\left(W(t)\right)\right] + \mathrm{Var}\left[Q(t)\right] \\ &+ 2\mathrm{Cov}\left[N\left(W(t)\right), Q(t)\right] \\ &= \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} + \lambda\mathbb{E}\left[W(t)\right] + \lambda^2\mathrm{Var}[W(t)] + \mathrm{Var}\left[Q(t)\right] \\ &+ 2\lambda\mathrm{Cov}\left[W(t), Q(t)\right], \end{aligned}$$

where the last equality follows from the facts that

$$\begin{aligned} \mathrm{Var}\left[N(t + \mathcal{S}) - N(t)\right] &= \mathrm{Var}\left[N(\mathcal{S})\right] \\ &= \mathbb{E}[\mathrm{Var}(N(\mathcal{S})|\mathcal{S})] + \mathrm{Var}\Big(\mathbb{E}[(N(\mathcal{S})|\mathcal{S})]\Big) \\ &= \lambda\mathbb{E}[\mathcal{S}] + \mathrm{Var}[\lambda\mathcal{S}] \\ &= \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu^2} \end{aligned}$$

https://doi.org/10.1017/S0269964824000251 Published online by Cambridge University Press

and

$$\begin{aligned}\mathrm{Var}\left[N(W(t))\right] &= \mathbb{E}\left[\mathrm{Var}(N(W(t))|W(t))\right] + \mathrm{Var}\Big(\mathbb{E}[(N(W(t))|W(t))]\Big)\\ &= \lambda\mathbb{E}[W(t)] + \lambda^2\mathrm{Var}[W(t)].\end{aligned}$$

Hence proved. □

Appendix C. Proof of Lemma 3.7

Proof. To compute the covariance of waiting time and queue length, we have

$$\begin{aligned}\mathrm{Cov}\left[W(t), Q(t)\right] &= \mathbb{E}\left[W(t)\cdot Q(t)\right] - \mathbb{E}\left[W(t)\right]\cdot\mathbb{E}\left[Q(t)\right]\\ &= \sum_{k=0}^{\infty} k\cdot\mathbb{E}\left[W(t)|Q(t)=k\right]\cdot\mathbb{P}\left(Q(t)=k\right) - \mathbb{E}\left[W(t)\right]\cdot\mathbb{E}\left[Q(t)\right]\\ &= \sum_{k=c}^{\infty} k\cdot\left(\sum_{j=0}^{k-c}\frac{1}{\mu c+\theta\cdot j}\right)\cdot\mathbb{P}\left(Q(t)=k\right) - \mathbb{E}\left[W(t)\right]\cdot\mathbb{E}\left[Q(t)\right]\\ &= \sum_{k=c}^{\infty} k\cdot\left(\psi\left(\frac{\mu c}{\theta}+k-c\right)-\psi\left(\frac{\mu c}{\theta}\right)\right)\cdot\mathbb{P}\left(Q(t)=k\right) - \mathbb{E}\left[W(t)\right]\cdot\mathbb{E}\left[Q(t)\right]\\ &= \mathbb{E}\left[Q(t)\left(\psi\left(\frac{\mu c}{\theta}+(Q(t)-c)^+\right)-\psi\left(\frac{\mu c}{\theta}\right)\right)\right] - \mathbb{E}\left[W(t)\right]\cdot\mathbb{E}\left[Q(t)\right]\\ &= \mathrm{Cov}\left[\psi\left(\frac{\mu c}{\theta}+(Q(t)-c)^+\right)-\psi\left(\frac{\mu c}{\theta}\right), Q(t)\right]\\ &= \mathrm{Cov}\left[\psi\left(\frac{\mu c}{\theta}+(Q(t)-c)^+\right), Q(t)\right].\end{aligned}$$

We hence proved the lemma. □

Cite this article: Pender J., Myoung Ko Y. and Xu J. (2025). The number of overlapping customers in Erlang-A queues: an asymptotic approach. *Probability in the Engineering and Informational Sciences* 39(3): 344–369. https://doi.org/10.1017/S0269964824000251