Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

SIMULATING CUSTOMER WAIT-TIME METRICS IN NONSTATIONARY QUEUES: A QUEUE-BASED CONDITIONAL ESTIMATOR FOR VARIANCE REDUCTION

Yunan Liu[1], and Ling Zhang[2]

[1]Amazon, Supply Chain Optimization Technologies, New York City, New York, NY, USA
[2]Amazon, Worldwide Last Mile Science, Bellevue, Washington, WA, USA

ABSTRACT

We introduce a new simulation method to estimate customer-averaged service metrics in nonstationary queueing systems with time-varying arrivals and staffing. Practical delay-based metrics—such as the fraction of customers waiting under a threshold or the average waiting time—are difficult to estimate using standard discrete-event simulation (DES) due to high implementation and variance complexity. We propose a queue-based conditional estimator that models system performance based on the time-dependent queue-length process. By conditioning on this process, our approach significantly reduces variance and simplifies simulation. We further enhance accuracy by incorporating many-server heavy-traffic approximations. Numerical experiments confirm that our estimator offers substantial computational gains over DES while maintaining accuracy. This method provides a scalable solution for evaluating service-level metrics in complex, time-varying environments.

1 INTRODUCTION

Customer waiting times are among the most commonly used performance metrics for evaluating service quality in queueing systems. Beyond the mean waiting time, which reflects the average customer experience, several delay-based *service-level* (SL) metrics are widely used in practice. One such metric is the *probability of delay* (PoD), which represents the likelihood that an arriving customer cannot immediately enter service and must wait. Another important metric is the *probability of abandonment* (PoA), defined as the probability that a customer leaves the queue after exceeding their patience threshold. PoA is particularly relevant in service systems like call centers (Liu and Whitt 2012c). The most prevalent SL metric is the *tail probability of delay* (TPoD), that is the probability that a customer's waiting time exceeds a designated delay target $w > 0$. TPoD is a crucial performance target in various practical settings, particularly in emergency departments. For instance, the Canadian Triage and Acuity Scale (Murray 2003) sets TPoD targets for different patient severity levels, with class-dependent thresholds w_i ranging from 15 to 120 minutes. Similarly, many call centers adopt TPoD targets, such as aiming to answer 80% of calls within 20 seconds (Preece et al. 2018), which translates to a TPoD target of $0.2 = 1 - 80\%$ with $w = 20$ seconds. See Liu (2018), Liu et al. (2022) for additional discussions on TPoD.

The prediction of wait-time SL metrics is relatively less challenging in stationary queueing models with constant demand rate and staffing level, because customers have statistically similar waiting time experiences. However, in real-world queueing systems, the demand function often exhibits significant variability in time (Green et al. 2007). In order to alleviate the performance fluctuations in time, the staffing level ought to be time-inhomogeneous to cope with the nonstationary demand. This gives rise to increased complexity of the waiting-time analysis. In this paper, we propose new simulation methods to estimate the customer-averaged wait-time metrics in nonstationary queueing systems.

1.1 Customer-averaged wait-time SL metrics

Prior research in the nonstationary queueing literature has primarily focused on predicting pointwise SL metrics, such as the expected waiting time $\mathbb{E}[W_t]$ and the tail probability $\mathbb{P}(W_t > \tau)$ at each time t (Aras et al. 2018; Liu 2018; Liu et al. 2022; Liu and Whitt 2012a; Liu and Whitt 2014c; Garyfallos et al. 2024). However, controlling these pointwise SL metrics can be either infeasible or prohibitively costly in practical service systems. For instance, enforcing a stringent requirement like $\mathbb{E}[W_t] \leq w$ for all t over the course of a day demands highly precise control of the staffing function. In contrast, a more practical and flexible approach is to impose a customer-averaged SL constraint, e.g., requiring that the average waiting time of all customers throughout the day does not exceed w. This relaxation accommodates natural fluctuations in system performance over time and reduces the operational burden of strict pointwise control. These metrics are widely used in real-world service systems, making them more meaningful for industrial applications (Preece et al. 2018).

The present paper emphasizes *customer-averaged* SL metrics. Examples include the average waiting time of all customers over a day or the daily fraction of customers who wait longer than a given threshold. These metrics aggregate performance over time and across customers, providing a holistic view of system behavior. In contrast, pointwise SL metrics, those evaluated at specific time points, are typically easier to analyze, often leveraging approximation formulas from fluid or diffusion limits in heavy-traffic regimes. However, customer-averaged SL metrics are significantly more challenging to study, particularly in nonstationary queues, where time-varying dynamics prevent the use of steady-state performance measures that are often available in stationary settings.

Consider a finite time interval $[0,T]$. Given a arrival rate $\lambda \equiv \{\lambda(t), t \in [0,T]\}$ and a staffing plan (i.e., number of servers) $\mathbf{n} \equiv \{n(t), t \in [0,T]\}$, our goal is to estimate the customer-averaged SL metric over a finite time interval $[0,T]$ in the following form:

$$\beta_u \equiv \beta_u(T, \mathbf{n}, \lambda) \equiv \mathbb{E}\left[\frac{1}{N(T)} \sum_{i=1}^{N(T)} u(W_i)\right], \tag{1}$$

where $N(T)$ is the total number of customer arrivals in $[0,T]$, W_i is the waiting time of the i^{th} customer, and $u(\cdot)$ is a utility function that maps W_i to a designated SL metric. We refer to the general form defined in (1) as the *customer-averaged service experience* (CASE). It is straightforward to realized that the generality of $u(\cdot)$ enables CASE to cover all above-mentioned wait-time SL metrics. For example, CASE defined in (1) reduces to (i) the average waiting time with $u(x) = x$, (ii) PoD with $u(x) = \mathbf{1}_{\{x>0\}}$, and (iii) TPoD with $u(x) = \mathbf{1}_{\{x>w\}}$ for $w > 0$, where $\mathbf{1}_A$ is the indicator for event A.

In queueing systems with customer abandonment (e.g., call centers and healthcare), the W_i in (1) can represent different definitions of waiting time, depending on the context and intended use case. The most commonly studied version in queueing theory is the *potential waiting time* (PWT), the offered waiting time assuming a customer remains indefinitely patient. PWT is independent of a customer's abandonment behavior and serves as a key metric for assessing system workload at a given time. In contrast, practical service systems more frequently use the *actual waiting time* (AWT), which measures the total time customers spend in the queue, regardless of whether they abandon or receive service. AWT is more directly observable from real-world data. For a detailed discussion on different waiting time definitions, see Liu et al. (2022). The generality of our approach here allows us to treat CASE driven by either PWT or AWT.

1.2 Crude Monte-Carlo Estimator

Given the complex dynamics and intractable nature of nonstationary queues, the most conceptually straightforward approach to compute CASE in (1) may be Monte-Carlo (MC) simulation. We envision an MC algorithm should follow the framework specified in Algorithm 1.

To distinguish Algorithm 1 from the other algorithms introduced later, we refer to the MC estimator in Algorithm 1 as the crude MC (CMC) estimator. Despite the apparent simplicity of Algorithm 1, the simulation of the CMC estimator involves subtleties that make it more complex than it may first appear. First, designing and implementing a nonstationary queueing system is already significantly more challenging than its stationary counterpart, as it requires tracking the transient evolution of system dynamics. Second,

Algorithm 1 Crude Monte Carlo Method.

```
for $j = 1$ to $n$ do
    Generate all customer waiting times $W_1^j, W_2^j, \dots, W_{N^j(T)}^j$ using a discrete-event simulator in $[0,T]$
    Compute $\theta^j \leftarrow \frac{1}{N^j(T)} \sum_{i=1}^{N^j(T)} f(W_i^j)$
end for
Output the sample mean $\bar{\Theta}(n) \leftarrow \frac{1}{n} \sum_{j=1}^{n} \theta^j$
```

Note: Superscript "j" indicates samples on the j^{th} MC trial.

simulating customer-level waiting times requires generating high-granularity events, making it significantly more complex and challenging than simulating system-level queue length. For instance, systematically modeling the PWT as a function of time necessitates periodically generating virtual customer arrivals to observe and experience the waiting process. A virtual customer is able to record the PWT and then dismissed only when it is about to "enter" service (Liu and Whitt 2012c).

Besides its high implementation complexity, Algorithm 1 may require a large number of MC trials to achieve a reasonably tight confidence interval. This is because simulating individual customer experiences in a bottom-up fashion involves generating a large number of random variables, which collectively contribute to the system's overall stochastic variability and inflate the variance of the CMC estimator. In line with general principles in simulation, the more random variables a model generates, the greater the inherent stochastic fluctuations, resulting in higher variance.

The primary focus of this paper is to develop novel estimators for CASE, as defined in (1), for nonstationary queues, that are more efficient than the CMC estimator in two key aspects: (1) they should be easier to simulate, requiring much less implementation effort, and (2) they leverage *conditioning* techniques to achieve significant variance reduction. To apply the conditioning technique, our point of departure is to separately model and treat (i) the randomness in the queue size "observed" by a customer upon her arrival and (ii) that in this customer's (future) waiting-time experience assuming a given number of existing customers yet to be processed before this customer enters service. For a given queue-length process, we derive an expression for CASE in (1) as a function of the queue-length process. Next, we take some extra steps to characterize the transient trajectory of the queue length. Following Liu et al. (2016), we leverage the heavy-traffic limit to approximate the queue length process by truncated Gaussian approximation.

1.3 Contributions and organization

We summarize our contributions below:

- We contribute to the performance analysis of nonstationary queueing systems by developing efficient and effective estimators for customer-averaged wait-time service-level (SL) metrics over a finite time horizon. Unlike CMC, which directly simulate individual customer experiences, our proposed estimator only requires simulating the queue length, which achieves a lower variance and reduced implementation complexity.
- Our variance reduction is achieved by conditioning on the queue-length process, we express CASE defined in (1) as a weighted sum of the time-averaged queue length at different time step. This derivation utilizes the relationship between customer waiting times and the queue length process in nonstationary queues.
- We further refine our estimator using a Gaussian approximation derived from process-level heavy-traffic diffusion limits, leading to additional variance reduction. We conduct numerical experiments to investigate the solution accuracy, computational efficiency, and robustness of our algorithms.

Organization of the paper. In Section 2 we derive our conditioning estimator as a function of the queue length process. In Section 3 we further refine the queue-based estimator using the truncated Gaussian approximation derived from heavy-traffic limit. In Section 4, we conduct numerical experiments and computer simulations to evaluate the performance of our new algorithm. In Section 5 we give concluding remarks and discuss future directions.

2 A QUEUE-BASED CONDITIONAL ESTIMATOR

We work with the $M_t/M/s_t+M$ nonstationary queueing model having a Poisson arrival process with time-varying rate $\lambda(t)$, *independent and identically distributed* (I.I.D.) service times following an exponential distribution with rate $\mu > 0$, a time-varying staffing level $n(t)$, and customer abandonment according to I.I.D. exponential abandonment times with rate $\theta \geq 0$. Customers are served according to the *first-come first-served* (FCFS) discipline. Our model covers both Erlang-A (with $\theta > 0$) and Erlang-C (with $\theta = 0$) models. Let $Q(t)$ and $B(t)$ be the numbers of customers in the waiting queue and in service at time t, and let $X(t) \equiv Q(t) + B(t)$ be the total number of customers in the system (in queue or in service) at time t. In what follows, we develop an estimator for CASE by conditioning on the queue length. We do so in two steps: First, we connect the customer-averaged SL metric in $[0,T]$ to the pointwise queue-based SL by conditioning on the queue length (Section 2.1). Second, we explain how to compute the pointwise SL at a fixed time t given the queue length $Q(t)$ (Section 2.2).

2.1 Q-CASE: A Conditional Estimator on Queue Length

We condition on the queue-length process $\mathbf{Q} \equiv \{Q(t), 0 \leq t \leq T\}$. Aligned with the philosophy of the technique of conditioning, this approach removes the extraneous randomness in CASE beyond that inherent in the queue-length process. To proceed, our key idea here is to properly separate (i) the total number of arrivals in $[0,T]$, (ii) the random distribution of their arrival times, and (iii) the pointwise SL for each customer given her arrival time.

We analyze the random arrival times (and their "observed" queue lengths) by looking into all potential *scenarios*, the k^{th} of which represents the event $\{N(T) = k\}$, meaning *there are in total k arrivals in* $[0,T]$. Next, we properly treat the arrival times of the k customers which, along with their "observed" queue length, provides an estimator of the SL. The most important step is to properly aggregate the SL experienced by all future customers in $[0,T]$ using weights designed based on both *spatial* and *temporal* attributes of the problem.

Conditional arrival times. It is well known that, given there are in total $N(T) = k$ arrivals during $[0,T]$, the k arrival times $0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < \tau_k < T$ follow the distribution of the *order statistics* of k I.I.D. random variables each follows the *probability density function* (PDF)

$$f(t) = \frac{\lambda(t)}{\int_0^T \lambda(s)ds}, \qquad t \in [0,T]. \tag{2}$$

Now, we consider a discrete-time framework where the interval $[0,T]$ is evenly split to a finite number m consecutive time steps: $0 \equiv t_0 < t_1 < \cdots < t_{m-1} < t_m \equiv T$, with $t_i \equiv i\Delta T$, $\Delta T \equiv T/m$. All arrival times $\tau_1, \ldots, \tau_k$ are distributed on the m time grids, which are order statistics of random variables following the discrete-time analog of (2), with a *probability mass function* (PMF)

$$p(t_i) = \frac{\lambda(t_i)}{\sum_{j=1}^m \lambda(t_j)}, \qquad t \in \{t_0, \ldots, t_m\}. \tag{3}$$

Given that $N(T) = k$, define N_i as the total number of arrivals among all k arrivals that occur at time t_i, $i = 0, 1, \ldots, m$. We know that $N_i \sim \text{Bino}(m, p(t_i))$, a binomial distribution having m trials and a success probability $p(t_i)$. Next let $W_j^{(i)}$ be the waiting time of the j^{th} customer among all N_i arrivals at time t_i, $1 \leq j \leq N_i$.

Connecting to the pointwise SL. For a tagged customer who arrives at time t, define

$$\widehat{\beta}(t,q) \equiv \mathbb{E}[u(W(t))|Q(t) = q] \tag{4}$$

as the conditional expected SL specified by waiting time $W(t)$ at t and utility function $u(\cdot)$, given that the tagged customer observes a queue length q (excluding herself). We refer to (4) as the *pointwise* queue-snapshot SL in order to distinguish from CASE as defined in (1). Here the function $\widehat{\beta}$ maps the dynamic queue-length information at t and other static inputs (e.g., arrival rate and staffing level) to the desired wait-time SL metric at t. Our idea is to develop CASE by aggregating the pointwise SL at different time steps over the temporal domain. We discuss the computational procedure of $\widehat{\beta}$ in Section 2.2.

Given $N(T)=k$, the conditional expectation of the discrete-time version of CASE in (1) is

$$\begin{aligned}\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{N_i}u(W_j^{(i)})\middle| N(T)=k\right] &= \sum_{l=0}^{k}\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{l}u(W_j^{(i)})\middle| N(T)=k,N_i=l\right]\mathbb{P}(N_i=l|N(T)=k)\\ &= \frac{1}{k}\sum_{l=0}^{k}\sum_{i=1}^{m}\sum_{j=1}^{l}\mathbb{E}\left[u(W_j^{(i)})\middle| N(T)=k,N_i=l\right]\binom{k}{l}p(t_i)^l(1-p(t_i))^{k-l}\\ &= \frac{1}{k}\sum_{i=1}^{m}\sum_{l=0}^{k}l\mathbb{E}\left[\widehat{\beta}(t_i,Q_k(t_i)\right]\binom{k}{l}p(t_i)^l(1-p(t_i))^{k-l} = \sum_{i=1}^{m}p(t_i)\mathbb{E}\left[\widehat{\beta}(t_i,Q_k(t_i)\right],\end{aligned}$$

where $Q_k(t_i)$ is the queue length at t_i given $N(T)=k$. Unconditioning on $N(T)$, we write the discrete-time CASE (with the discrete time step Δ included in the superscript) as

$$\beta_u^{\Delta} = \sum_{k}\sum_{i=0}^{m}p(t_i)\mathbb{E}\left[\widehat{\beta}(t_i,Q_k(t_i))\right]\cdot\mathbb{P}(N(T)=k) = \mathbb{E}\left[\sum_{i=0}^{m}p(t_i)\widehat{\beta}(t_i,Q(t_i))\right] \tag{5}$$

$$\to \mathbb{E}\left[\int_0^T f(t)\widehat{\beta}(t,Q(t))dt\right] = \beta_u, \qquad \text{as } \Delta\to 0. \tag{6}$$

Remark 1 (Individual vs. population, ensemble average vs. time average) Formula (6) reveals some good insights: First, for each time point t_i, the snapshot term $\widehat{\beta}(t,Q(t))$ quantifies a pointwise *individual* SL experience, and this experience is estimated here in the ensemble sense with repeated MC trials. Next, the "weight" $f(t)$ (and $p(t_i)$ in (5)) describes how frequent a customer arrival should occur at time t (or equivalently how many customers will arrive at time t_i over many "days"), so that it takes into account the *population* level effect. Finally, the weighted sum gives the *time-averaged* SL outcome over the entire $[0,T]$ period.

It is straightforward to see that (6) can be further simplifies to $\mathbb{E}\left[\mathbb{E}\left[\widehat{\beta}(\tau,Q(\tau))|\tau\right]\right]=\mathbb{E}\left[\widehat{\beta}(\tau,Q(\tau))\right]$, where τ is a random arrival time following the PDF in (2). This seems to suggest that the most conceptually simple queue-based estimator is $\widehat{\beta}(\tau,Q(\tau))$, which requires sampling an independent arrival time τ and the queue length $Q(\tau)$ at τ, making it easy to implement. In terms of variance reduction, the estimator $\widehat{\beta}(\tau,Q(\tau))$ conditions on the queue length to eliminate the need to sample waiting times for all customers, thereby removing additional sources of stochasticity beyond what is already captured by the queue length. However, collapsing the customer heterogeneity within their arrival times into a single representative arrival time can significantly amplify overall variability. Therefore, we propose the estimator in (6), which can be interpreted as $\mathbb{E}\left[\widehat{\beta}(\tau,Q(\tau))|\tau\right]$, the conditional estimator of $\widehat{\beta}(\tau,Q(\tau))$ given τ. We call it the queue-conditioning CASE (Q-CASE), with (5) and (6) representing its discrete-time and continuous-time versions, respectively. See Algorithm 2 below for the formal statement of the algorithm.

2.2 Pointwise SL conditional on queue length

We now describe how to compute the queue-conditioning pointwise SL $\widehat{\beta}(t,q)$ for a given queue length $q\geq 0$. We hereby focus on PWT, but all analysis easily generalizes to AWT; see for example Section 2.1

Algorithm 2 Q-CASE

```
for $j = 1$ to $n$ do
    Generate an independent sample path of $Q^{(j)}(t_i)$ for all $t_1, \ldots, t_m$
    Compute $\widehat{\beta}(t_i, Q^{(j)}(t_i))$ for all $t_1, \ldots, t_m$
    Compute the sum $\theta^j \leftarrow \sum_{i=1}^{m} p(t_i)\widehat{\beta}(t_i, Q^{(j)}(t_i))$
end for
Output the sample mean $\bar{\Theta}(n) \leftarrow \frac{1}{n}\sum_{j=1}^{n} \theta^j$
```

Note: Superscript "j" indicates samples on the j^{th} MC trial.

in Konrad and Liu (2023b). In addition, for the ease of explanation, below we assume a constant staffing function n. The full treatment of $n(t)$ is given in Konrad and Liu (2023b).

We view the waiting time at t as a phase-type (PH) distribution associated with a *continuous-time Markov chain* (CTMC) $\{Y_q(s) : s \geq t\}$ which models the future *queueing position* of a customer who arrives at t. We refer to this customer as the "tagged" customer. Because the system is operated according to FCFS, the tagged customer's waiting time is independent with any future arrivals after t, so that Y_q is a pure-death process; see Figure 1 for the transition rate diagram. Suppose at time t, the tagged customer arrives, seeing that there are q existing customers waiting in the queue. According to the definition of PWT, the tagged customer is infinitely patient, who remains in the waiting line until $q+1$ "departures" occur (either due to service completions or customer abandonments). Formally, the waiting time conditioning on the queue length q, denoted by $W(n,q)$, is a PH random variable associated with the absorbing CTMC Y_q having a state space $\mathscr{S} \equiv \{n-1, n, \ldots, n+q-1, n+q\}$, where state $n+q$ is the initial transient state and state $n-1$ is the only absorbing state. To see this, we can write $W(n,q)$ as a sum of $q+1$ independent random variables:

$$W(n,q) \overset{\text{d}}{=} D_q + D_{q-1} + \cdots + D_1 + D_0, \tag{7}$$

where D_i is an exponential random variable with rate $i\theta + n\mu$, and "$\overset{\text{d}}{=}$" means "equal in distribution". Here D_i is the i^{th} inter-departure time of all required $q+1$ departures.

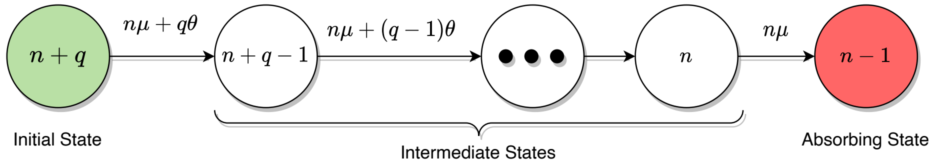

Figure 1: Transition rate diagram.

The transition rate matrix for Y_q is given below as

$$\mathbf{Q} = \begin{array}{r} n-1 \\ n \\ \vdots \\ n+q \end{array} \left[\begin{array}{c|ccccc} 0 & 0 & \cdots & 0 & & 0 \\ \hline n\mu & -n\mu & 0 & \cdots & & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & n\mu + q\theta & & -n\mu - q\theta \end{array}\right] = \left[\begin{array}{c|c} 0 & \mathbf{0} \\ \hline \mathbf{S}^0 & \mathbf{S} \end{array}\right], \tag{8}$$

where $\mathbf{S}$ is a $(q+1) \times (q+1)$ matrix, and $S_{i,i-1} = n\mu + (i-n)\theta$ for $i = n, \cdots, q+n$, $S_{i,j} = 0$ for $j \notin \{i, i-1\}$, $i = n+1, \cdots, q+n$. The *cumulative distribution function* (CDF) of $W(n,q)$ is given by

$$F(w; n, q) \equiv \mathbb{P}(W(n,q) \leq w) = 1 - \alpha e^{\mathbf{S}w}\mathbf{e}, \tag{9}$$

where $\mathbf{e}$ is a column vector of 1's, row vector $\alpha = [0, \ldots, 0, 1]$ (because the initial state is $n+q$), and $e^{\mathbf{X}} = \sum_{k=0}^{\infty}(1/k!)\mathbf{X}^k$ the matrix exponential. See Latouche and Ramaswami (1999) for detailed review of

PH distributions. Using the CDF in (9), we can easily compute

$$\widehat{\beta}(t,q) = \mathbb{E}[u(W_j)|Q(t)=q] = \int_w u(w)\,dF(w;n,q). \tag{10}$$

For example, to compute the TPoD at w, we have

$$\mathbb{E}[\mathbf{1}_{\{W_j>w\}})|Q(t)=q] = 1-F(w;n,q) = \alpha e^{Sw}\mathbf{e}.$$

See Konrad and Liu (2023b) for the generalized treatment of a time-varying $n(t)$.

Remark 2 (Inverse Laplace transformation) As an alternative to the above PH method for computing the distribution of $W(n,q)$, the decomposition in (7) enables us to derive its Laplace transform (LT) directly, which has the following product form:

$$\mathscr{L}_{W(n,q)}(s) = \prod_{k=0}^{q}\mathscr{L}_{D_k}(s) = \prod_{k=0}^{q}\frac{n\mu+k\theta}{n\mu+k\theta+s}. \tag{11}$$

To recover the original PDF of $W(n,q)$ we resort to the numerical inversion of LT developed by Abate and Whitt (1995).

3 G-CASE: A GAUSSIAN APPROXIMATION FOR Q-CASE

In this section, we consider a heuristic approach leveraging the heavy-traffic approximation for the queue length process - the fluid and diffusion models.

3.1 Heavy-Traffic Fluid Limit

In a nonstationary queueing system having time-varying arrival rate and staffing level, the challenges in establishing analytic solutions are due to two main factors: (i) *temporal variability* (e.g., the system's states at different time points during a day, such as peak and off-peak hours) and (ii) *stochastic variability* or ensemble variability (e.g., the variability across different days but at the same time during a day). As the system scale increases, the former begins to dominate the latter (Liu and Whitt 2012a). Fluid models, arising from the large-scale limits of their associated stochastic queueing models with the scale approaching infinity, have been proven effective in the performance analysis of multiserver queueing systems (Liu and Whitt 2012a). Fluid models focus on characterizing the temporal variability while omitting the stochastic variability; they are determinisitic models described by a system of differential equations so they are much more tractable than the corresponding stochastic root models. Also see Liu and Whitt (2011), Liu and Whitt (2012a), Liu and Whitt (2012b), Liu and Whitt (2014a), Lee et al. (2021).

The better explain the fluid functions, we first describe the queueing equations for their stochastic counterparts. The Markovian structure enables us to use the so-called *random time change of unit-rate Poisson* approach (e.g., Section 2.1 in Pang et al. (2007)) to construct the above-mentioned performance functions. Specifically, let $N^{\lambda}(\cdot)$, $N^{s}(\cdot)$ and $N^{a}(\cdot)$ be three I.I.D. unit-rate Poisson processes. We have

$$\begin{aligned} X(t) &= X(0) + N^{\lambda}\left(\int_0^t \lambda(s)\mathrm{d}s\right) - N^{s}\left(\int_0^t \mu B(s)\mathrm{d}s\right) - N^{a}\left(\int_0^t \theta Q(s)\mathrm{d}s\right) \\ &= X(0) + N^{\lambda}\left(\int_0^t \lambda(s)\mathrm{d}s\right) - N^{s}\left(\int_0^t \mu X(s)\wedge n(s)\mathrm{d}s\right) - N^{a}\left(\int_0^t \theta(X(s)-n(s))^{+}\mathrm{d}s\right). \end{aligned} \tag{12}$$

As we can see from the above formula, the system's randomness is attributed to three separate random sources: the arrival process (modeled by N^{λ}), the service process (modeled by N^{s}) and the abandonment process (modeled by N^{a}).

We use lower-case letters $q(t)$, $b(t)$ and $x(t)$ to denote the fluid limits of their stochastic counterparts. The dynamics of the total fluid content $x(t)$ is described by the following *ordinary differential equation* (ODE) having a piecewise drift:

$$x'(t) = \lambda(t) - \mu \cdot b(t) - \theta \cdot q(t), \quad \text{where} \;\; q(t) \equiv (x(t) - n(t))^{+} \quad \text{and} \quad b(t) \equiv x(t) \wedge n(t) \tag{13}$$

are the fluid contents in queue and in service at t, $x^{+} \equiv \max(x, 0)$, and $x \wedge y \equiv \min(x, y)$.

To formalize the MSHT framework, we define a sequence of $M_t/M/s_t + M$ models indexed by $\eta = 1, 2, 3, \ldots$, where the η^{th} model has an η-scaled arrival rate $\lambda_\eta(t) = \eta\lambda(t)$ and service capacity $n_\eta(t) = \eta n(t)$, but unscaled service rate μ and abandonment rate θ. Let $X_\eta(t)$, $Q_\eta(t)$ and $B_\eta(t)$ be the total queue length, waiting queue length and number of busy servers at t in the η^{th} model. We expect the following *functional weak law of large number* (FWLLN)

$$X_\eta(t)/\eta \Rightarrow x(t) \qquad \text{as} \quad \eta \to \infty, \tag{14}$$

where $x(t)$ is the fluid limit specified in (13). A straightforward comparison of (13) and (12) reveals that the fluid model is to treat the original stochastic model as if the random fluctuation may be omitted (with all three Poisson shells N^{λ}, N^{s} and N^{a} removed) while still preserving the system's trend in the temporal domain. See Liu and Whitt (2012b), Liu and Whitt (2014b), Whitt (2006) for detailed theoretical framework of FWLLN.

3.2 Gaussian Approximation for the Queue Length

Building on the fluid limit specified in (13), we further expect that the following *functional central limit theorem* (FCLT) holds:

$$(X_\eta(t) - \eta x(t))/\sqrt{\eta} \Rightarrow \widehat{X}(t) \qquad \text{as} \quad \eta \to \infty, \tag{15}$$

where the diffusion limit $\widehat{X}(t)$ is a zero-mean Gaussian process. Before we carefully investigate the dynamics of $\widehat{X}(t)$, we first explain how the Gaussian approximation works. According to (15), for a large system scale η, we have the approximation

$$X_\eta(t) \approx \eta x(t) + \sqrt{\eta}\widehat{X}(t) \overset{\mathrm{d}}{=} \mathscr{N}\left(\eta x(t), \eta\sigma^2(t)\right) \qquad \text{for any time } t, \tag{16}$$

where $\mathscr{N}(a,b)$ is a Gaussian random variable with mean a and variance b, and $\sigma^2(t) \equiv \mathrm{Var}(\widehat{X}(t))$.

Next, following the truncated Gaussian approximation in Liu et al. (2016), we write

$$Q_\eta(t) = (X_\eta(t) - n_\eta(t))^{+} \overset{\mathrm{d}}{\approx} \mathscr{N}\left(\eta x(t) - n_\eta(t), \eta\sigma^2(t)\right)^{+} \overset{\mathrm{d}}{=} (\eta x(t) - n_\eta(t) + \sqrt{\eta}\sigma(t)\mathscr{Z})^{+} \equiv \widehat{Q}_\eta(t), \tag{17}$$

$$B_\eta(t) = X_\eta(t) \wedge n_\eta(t) \overset{\mathrm{d}}{\approx} \mathscr{N}\left(\eta x(t), \eta\sigma^2(t)\right) \wedge n_\eta(t) \overset{\mathrm{d}}{=} (\eta x(t) + \sqrt{\eta}\sigma(t)\mathscr{Z}) \wedge n_\eta(t) \equiv \widehat{B}_\eta(t), \tag{18}$$

where $\mathscr{Z}$ is a standard Gaussian random variable.

In the Q-CASE estimator defined in (6), replacing $Q(t)$ by $\widehat{Q}_\eta(t)$ specified in (17) yileds the Gaussian approximation version of Q-CASE:

$$\beta^{G} \equiv \mathbb{E}\left[\int_0^T f(t)\widehat{\beta}\left(t, (x(t) - n(t) + \sigma(t)\mathscr{Z})^{+}\right) dt\right]. \tag{19}$$

Here in (19) we slightly abuse our notation by using $x(t)$ in place of $\eta x(t)$ as the scale-free fluid process and $\sigma(t)$ in place of $\sqrt{\eta}\sigma(t)$ as the scale-free standard deviation process.

The final missing piece in (19) is the variance $\sigma^2(t)$. Our $M_t/M/s_t+M$ model is a special case of results in the $G_t/M/s_t+G$ model in (Liu and Whitt 2014b) and the Markovian queueing network in Mandelbaum et al. (1998). For this step, we mainly draw from Mandelbaum et al. (1998). The diffusion limit of the total queue length in our $M_t/M/s_t+M$ model solves the following *stochastic differential equation* (SDE)

$$\begin{aligned} d\widehat{X}(t) &= [\mu\mathbf{1}(x(t)\le n(t))+\theta\mathbf{1}(x(t)>n(t))]\widehat{X}(t)^- - [\mu\mathbf{1}(x(t)<n(t))+\theta\mathbf{1}(x(t)\ge n(t))]\widehat{X}(t)^+ \\ &\quad - d\mathscr{B}_\lambda\left(\int_0^t \lambda(s)ds\right) - d\mathscr{B}_a\left(\int_0^t \theta(x(s)-n(s))^+ds\right) - d\mathscr{B}_s\left(\int_0^t \mu(x(s)\wedge n(s))ds\right), \qquad (20)\end{aligned}$$

where $x^- \equiv \max(-x,0)$, and $\mathscr{B}_\lambda$, $\mathscr{B}_s$ and $\mathscr{B}_a$ are three I.I.D. standard *Brownian motions* (BMs).

Next, if we omit the time points at which the fluid model is critically loaded, that is $x(t)=n(t)$ (which rarely occurs in practice), we can simplify (20) to the following *Ornstein-Uhlenbeck* (OU) process having a piecewise drift:

$$\begin{aligned} d\widehat{X}(t) &= -[\mu\mathbf{1}(x(t)<n(t))+\theta\mathbf{1}(x(t)>n(t))]\widehat{X}(t) \\ &\quad - d\mathscr{B}_\lambda\left(\int_0^t \lambda(s)ds\right) - d\mathscr{B}_a\left(\int_0^t \theta(x(s)-n(s))^+ds\right) - d\mathscr{B}_s\left(\int_0^t \mu(x(s)\wedge n(s))ds\right), \qquad (21)\end{aligned}$$

of which the variance $\sigma^2(t)$ solves the following ODE.

$$\frac{d\sigma^2(t)}{dt} = -2\left[\theta\mathbf{1}(x(t)>n(t))+\mu\mathbf{1}(x(t)<n(t))\right]\sigma^2(t)+\lambda(t)+\theta(x(t)-n(t))^+ +\mu(x(t)\wedge n(t)), \qquad (22)$$

with initial condition $\sigma^2(0)=0$.[1]

We are now ready to describe the MC algorithm arising from (19). Note that the main appeal of (19) is its ease of implementation: For each MC trial, instead of generating a full sample path of the queue-length process, it now suffices to generate a single random variable $\mathscr{Z}$. This advantage can further enable the use of additional variance reduction technique: For example, because the pointwise SL function $\widehat{\beta}(t,q)$ is monotonically increasing in q, we can apply the technique of *antithetic variables*, that is

$$\beta_A^G \equiv \mathbb{E}\left[\int_0^T f(t)\left(\frac{\widehat{\beta}\left(t,(x(t)-n(t)+\sigma(t)\mathscr{Z})^+\right)+\widehat{\beta}\left(t,(x(t)-n(t)+\sigma(t)(-\mathscr{Z}))^+\right)}{2}\right)dt\right]. \qquad (23)$$

Below we formally state the Gaussian-antithetic version of Q-CASE in Algorithm 3. We call it G-CASE.

Algorithm 3 G-CASE

```
for $j=1$ to $n$ do
    Generate $Z_j \sim \mathcal{N}(0,1)$;
    Compute $x(t)$ and $\sigma^2(t)$ using (13) and (22).
    Compute $\widehat{\beta}\left(t_i,(x(t_i)-n(t_i)+\sigma(t_i)Z_j)^+\right)$ and $\widehat{\beta}\left(t_i,(x(t_i)-n(t_i)+\sigma(t_i)(-Z_j))^+\right)$ for all $t_1,\dots,t_m$;

    Compute the sum $\theta^j \leftarrow \sum_{i=1}^m p(t_i)\frac{\widehat{\beta}\left(t_i,(x(t_i)-n(t_i)+\sigma(t_i)Z_j)^+\right)+\widehat{\beta}\left(t_i,(x(t_i)-n(t_i)+\sigma(t_i)(-Z_j))^+\right)}{2}$
end for
Output the sample mean $\bar{\Theta}(n) \leftarrow \frac{1}{n}\sum_{j=1}^n \theta^j$
```

Note: Superscript "j" indicates samples on the j^{th} MC trial.

[1] The initial queue value is the observation of the system state at the beginning of the interval $[0,T]$, which is always deterministic. Hence, the variance of the initial queue size is 0.

4 NUMERICAL EXPERIMENTS

To evaluate the accuracy and effectiveness of our new estimators, we conduct a numerical studies which benchmark Q-CASE and G-CASE to CMC. Following Liu and Whitt (2012a), we first consider an $M_t/M/s_t+M$ example having a sinusoidal arrival rate $\lambda(t)=n(1+0.2\cos(t))$, a constant staffing level with n servers, service rate $\mu=1$, and abandonment rate $\theta=0.5$. Let $n=1000$. We conduct 1000 MC trials for all estimators. We focus on two predominant SL metrics in form of CASE in (1): (i) customer-averaged waiting time in $[0,t]$ (with $u(x)=x$), and (ii) fraction of customers in $[0,t]$ whose waiting times are $\le w$ (with $u(x)=\mathbf{1}_{\{x\le w\}}$, $w=0.5$). See Table 1 for CIs of all estimators at four representative time points. Table 1 shows that the use of antithetic variables in G-CASE is able to achieve additional variance reduction beyond Q-CASE.

In terms of end-to-end simulation time, CMC requires around 1 minute and 40 seconds to complete. In contrast, G-CASE achieves comparable accuracy with runtime of just 0.3 seconds. This efficiency is enabled by a preprocessing step where all unique integer-valued queue lengths are identified and their corresponding $\hat{\beta}(t,q)$ values are precomputed. This is not surprising because each trial of G-CASE requires only the generation of a single Gaussian random variable, while that of CMC requires a complete DES simulation of detailed waiting time data for all customers.

Table 1: CMC vs. Q/G-CASE: confidence intervals for (i) customer-averaged waiting time, and (ii) fraction of customers with waiting times below a target, in interval $[0,t]$.

	Time	$t=8$	$t=10$	$t=16$	$t=20$
Customer-averaged waiting time	CMC	0.044 ± 3e-03	0.045 ± 3e-03	0.056 ± 2e-03	0.058 ± 2e-03
	Q-CASE	0.045 ± 1e-03	0.046 ± 1e-03	0.057 ± 1e-03	0.059 ± 1e-03
	G-CASE	0.043 ± 8e-06	0.044 ± 4e-05	0.055 ± 4e-05	0.057 ± 4e-05
Fraction of waiting times below a target	CMC	0.223 ± 2e-02	0.221 ± 2e-02	0.275 ± 1e-02	0.284 ± 1e-02
	Q-CASE	0.228 ± 5e-03	0.23 ± 6e-03	0.286 ± 5e-03	0.293 ± 5e-03
	G-CASE	0.221 ± 2e-03	0.224 ± 1e-03	0.282 ± 2e-03	0.288 ± 2e-03

Supplementing the above example with a theoretical arrival rate, we next consider a more realistic arrival rate that resembles the demand pattern in real-world large-scale call centers (representing two demand surges throughout the day). In Figure 2, we plot all three estimators for (i) and (ii) within different interval $[0,t]$, for $t\in[0,T]$, $T=24$. See the two panels at the bottom of Figure 2. Across both metrics, both Q-CASE and G-CASE closely agree with CMC, confirming their accuracy. On the other hand, both estimators achieve significantly smaller *confidence intervals* (CIs) as demonstrated in Table 2.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose new methodologies for efficiently estimating customer-averaged service level metrics in nonstationary queueing systems, with a particular focus on developing computationally efficient estimators that outperform traditional Monte Carlo approaches. By leveraging conditioning techniques on

Table 2: CMC vs. Q/G-CASE: confidence intervals for the example in Figure 2, $n=400$ servers, mean service time is 10 minutes, and mean abandonment time is 20 minutes.

	Time	$t=10$	$t=12$	$t=16$	$t=20$
Customer-averaged waiting time	CMC	0.031 ± 1e-03	0.052 ± 1e-03	0.056 ± 8e-04	0.049 ± 7e-04
	Q-CASE	0.031 ± 9e-04	0.051 ± 8e-04	0.053 ± 6e-04	0.048 ± 6e-04
	G-CASE	0.032 ± 6e-06	0.051 ± 3e-05	0.053 ± 6e-05	0.048 ± 6e-05
Fraction of waiting times below a target	CMC	0.041 ± 5e-03	0.162 ± 9e-03	0.103 ± 6e-03	0.08 ± 5e-03
	Q-CASE	0.034 ± 9e-03	0.168 ± 2e-02	0.107 ± 1e-02	0.082 ± 8e-03
	G-CASE	0.036 ± 3e-03	0.165 ± 3e-03	0.106 ± 4e-03	0.081 ± 3e-03

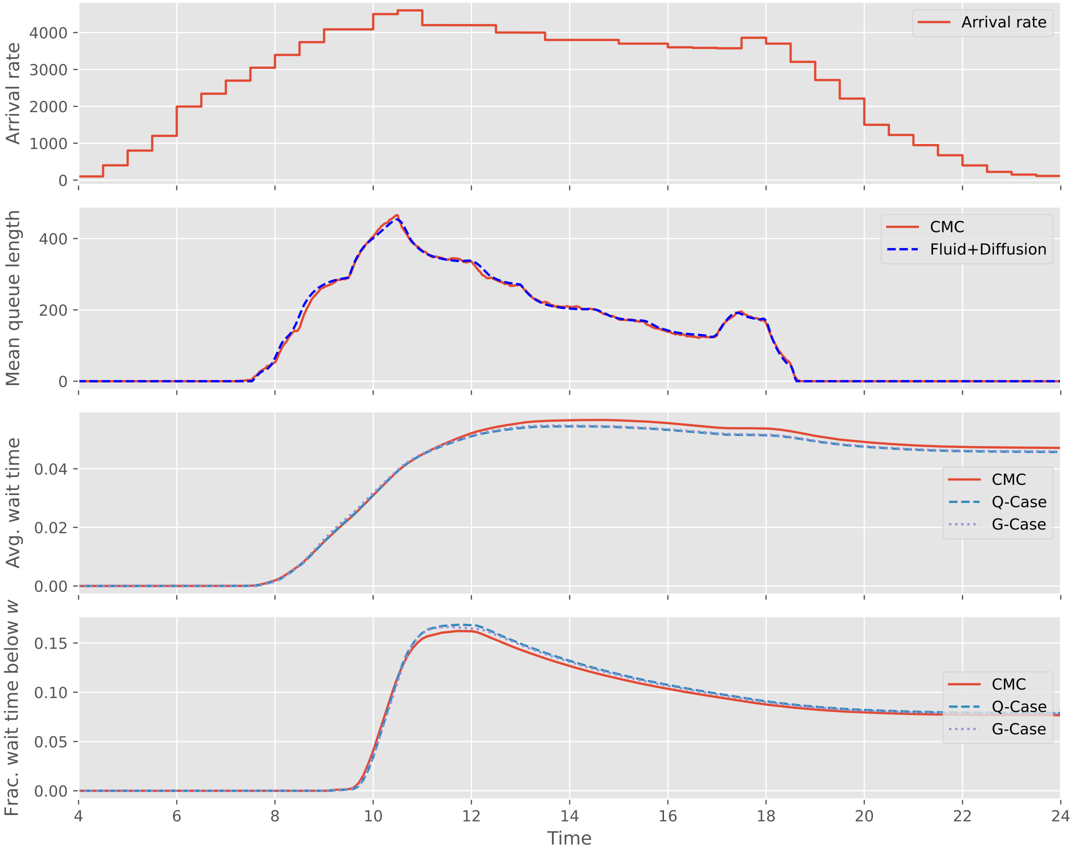

Figure 2: CMC vs. Q/G-CASE: (1) average wait time and (2) fraction of wait times below target w, in $[0,t]$, for $t \in [0,T]$, $T = 24$, $n = 400$ servers, mean service time is 10 minutes, and mean abandonment time is 20 minutes.

the queue length and its heavy-traffic approximations, the proposed new estimators significantly reduce variance and end-to-end simulation time while maintaining high accuracy in predicting customer wait-time metrics. Through extensive numerical experiments, we demonstrate the effectiveness of these methods.

There are several important venues for future research. Although the present paper focuses on the time-varying Erlang-A model, it is possible to extend our results to non-Erlang models. For example, the development of the Gaussian-based estimator can leverage useful heavy-traffic results of non-Markovian queues in Aras et al. (2018), Liu and Whitt (2012a). Another future direction is to extend our methods to multi-class queueing systems, which requires the estimator to be aware of the scheduling and routing policies (Liu et al. 2022). Finally, a practical future work is to turn the performance prediction result to a decision-support tool; we can design an efficient simulation optimization framework to determine the optimal operational decisions (e.g., staffing) subject to SL constraints (Konrad and Liu 2023a).

REFERENCES

Abate, J., and W. Whitt. 1995. “Numerical Inversion of Laplace Transforms of Probability Distributions”. *ORSA Journal on computing* 7(1):36–43.

Aras, K., X. Chen, and Y. Liu. 2018. “Many-Server Gaussian Limits for Non-Markovian Queues with Customer Abandonment”. *Queueing Systems* 89(1):81–125.

Garyfallos, S., Y. Liu, P. Barlet-Ros, and A. Cabellos-Aparicio. 2024. “Service Level Prediction in Non-Markovian Nonstationary Queues: A Simulation-Based Deep Learning Approach”. In *2024 Winter Simulation Conference (WSC)*, 2655–2666 https://doi.org/10.1109/WSC63780.2024.10838828.

Green, L. V., P. J. Kolesar, and W. Whitt. 2007. “Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System”. *Production and Operations Management* 16(1):13–39.

Konrad, K., and Y. Liu. 2023a. "Achieving Stable Service-Level Targets in Time-Varying Service Systems. A Simulation-based Offline Learning Staffing Algorithm". In *2023 Winter Simulation Conference (WSC)*, 327–338 https://doi.org/10.1109/WSC60868.2023.10408273.

Konrad, K., and Y. Liu. 2023b. "Real-Time Estimations for the Waiting-Time Distribution in Time-Varying Queues". In *2023 Winter Simulation Conference (WSC)*, 315–326 https://doi.org/10.1109/WSC60868.2023.10407204.

Latouche, G., and V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM.

Lee, C., X. Liu, Y. Liu, and L. Zhang. 2021. "Optimal Control of a Time-Varying Double-Ended Production Queueing Model". *Stochastic Systems* 11:140–173.

Liu, Y. 2018. "Staffing to Stabilize the Tail Probability of Delay in Service Systems with Time-Varying Demand". *Operations Research* 66(2):514–534.

Liu, Y., X. Sun, and K. Hovey. 2022. "Scheduling to Differentiate Service in a Multiclass Service System". *Operations Research* 670:527–544.

Liu, Y., and W. Whitt. 2011. "A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment". *Operations Research* 59:835–846.

Liu, Y., and W. Whitt. 2012a. "The $G_t/GI/s_t+GI$ Many-Server Fluid Queue". *Queueing Systems* 71:405–444.

Liu, Y., and W. Whitt. 2012b. "A Many-Server Fluid Limit for the $G_t/GI/s_t+GI$ Queueing Model experiencing Periods of Overloading". *Operations Research Letters* 40:307–312.

Liu, Y., and W. Whitt. 2012c. "Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals". *Operations Research* 60(6):1551–1564.

Liu, Y., and W. Whitt. 2014a. "Algorithms for Time-Varying Networks of Many-Server Fluid Queues". *INFORMS Journal on Computing* 26:59–73.

Liu, Y., and W. Whitt. 2014b. "Many-Server Heavy-Traffic Limits for Queues with Time-Varying Parameters". *Annals of Applied Probability* 24:378–421.

Liu, Y., and W. Whitt. 2014c. "Stabilizing Performance In Networks Of Queues With Time-Varying Arrival Rates". *Probability in the Engineering and Informational Sciences* 28(4):419–449.

Liu, Y., W. Whitt, and Y. Yu. 2016. "Approximations for Heavily Loaded $G/GI/n+GI$ Queues". *Naval Research Logistics* 63:187–217.

Mandelbaum, A., W. A. Massey, and M. I. Reiman. 1998. "Strong Approximations for for Markovian service networks". *Queueing Systems* 30:149–201.

Murray, M. J. 2003. "The Canadian Triage and Acuity Scale: A Canadian perspective on emergency department triage". *Emergency Medicine* 15(1):6–10.

Pang, G., R. Talreja, and W. Whitt. 2007. "Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues". *Probability Surveys* 4:193–267.

Preece, D., F. Sherlock, and B. Bischoff. 2018. "What Are the Industry Standards for Call Centre Metrics?". *Call Centre Helper*. https://www.callcentrehelper.com/industry-standards-metrics-125584.htm, accessed April 17th.

Whitt, W. 2006. "Fluid Models for Multiserver Queues with Abandonments". *Operations Research* 54:37–54.

AUTHOR BIOGRAPHIES

YUNAN LIU is a Principal Research Scientist in the Supply Chain Optimization Technology team at Amazon. He is also an Adjunct Professor in the ISE Department of NC State University. He earned his Ph.D. in Operations Research from Columbia University. His research interests include stochastic modeling, simulation, optimal control and online learning, with applications to supply chain and call centers. His work was awarded first place in the INFORMS Junior Faculty Interest Group Paper Competition in 2016. His email address is yunanliu@amazon.com. His website is https://yliu48.github.io/.

LING ZHANG is a Senior Research Scientist in the Last Mile Science team at Amazon. He earned his Ph.D. in Industrial and Systems Engineering from NC State University. His research interests include queueing theory, stochastic modeling, simulation and large language model application for OR practitioners. His email address is lzhangn@amaon.com. His LinkedIn is www.linkedin.com/in/lingzhang7