Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

REAL-TIME ESTIMATIONS FOR THE WAITING-TIME DISTRIBUTION IN TIME-VARYING QUEUES

Kurtis Konrad
Yunan Liu

Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27607, USA

ABSTRACT

Customers' waiting times are the most commonly used performance data to measure the quality of service in service systems such as call centers and healthcare. Unlike stationary queueing models where customers' waiting times are statistically similar, the prediction of waiting times is far less straightforward in time-varying queues having nonstationary demand (i.e., arrival rate) and supply (i.e., number of servers). In this paper, we develop a novel methodology for more accurately computing the wait time distribution in a time-varying queueing system. We design extensive simulation experiments to evaluate our prediction methods. In addition, we discover that the waiting-time prediction is highly sensitive to the work-releasing policy of the staffing plan, i.e., the rule under which the number of servers changes in time.

1 INTRODUCTION

Customers' waiting times are the most commonly used performance data to describe the quality of service in queueing systems. Besides the mean waiting time that reports the average customer experience in the waiting queue, there exist several delay-based *service-level* (SL) metrics in practice. One example is the *probability of delay* (PoD), meaning the probability that an arriving customer cannot immediately enter service (experiencing a positive waiting time). A second example is the *probability of abandonment* (PoA), i.e., the probability that a customer runs out of patience and abandons from the queue (with the offered waiting time exceeding her own patience level). PoA is considered in queueing models with applications to service systems such as call centers (Liu and Whitt 2012b). 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 $\tau > 0$. This is an important objective that is used in many practical settings, one of the most notable is in emergency departments. The Canadian Triage and Acuity Scale (Murray 2003) defines TPoD targets for different patient classes (i.e., severity levels $i = 1, 2, \ldots, 5$) with class-dependent τ_i ranging from 15 to 120 minutes. It is also known that many call centers utilize TPoD targets, aiming to answer 80% of calls within 20 seconds (Preece et al. 2018), which translates into a TPoD target $0.2 = 1 - 80\%$ with $\tau = 20$ seconds. See Liu (2018), Liu et al. (2021) for additional discussions on TPoD.

Unlike the mean waiting time, all three above-mentioned SL metrics utilize the distributional information beyond the mean, hence their computation draws from the distribution function of the waiting time. Specifically, let W be a generic waiting time, and let $F(x) \equiv \mathbb{P}(W \leq x)$ and $F^c(x) \equiv 1 - F(x)$ be its *cumulative distribution function* (CDF) and complementary CDF (CCDF), we can write the mean waiting time, PoD, PoA and TPoD all in the form of F as below:

$$\mathbb{E}[W] = \int_0^\infty F^c(x)dx, \quad \mathbb{P}(W > 0) = F^c(0) \quad \text{and} \quad \mathbb{P}(W > A) = \int F^c(x) f_A(x)dx, \quad \mathbb{P}(W > \tau) = F^c(\tau),$$

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

with A denoting a customer's abandonment time (the time a customer is willing to wait in the queue) and $f_A(x)$ denoting its *probability density function* (PDF). Hence, there is an increasing need for tools to better predict the waiting-time distribution (e.g., the CDF $F(\cdot)$).

The prediction of waiting-time distributions 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 achieve time-stable performance, the staffing level (i.e., number of servers) ought to be time-inhomogeneous to cope with the nonstationary demand. When the staffing level changes in time, the prediction of the waiting time distribution at time t should not only depend on the present system state (e.g., queue length), but also on the (near) future service capacity, that is, the staffing dynamics n_s for $s \geq t$. However, this gives rise to increased complexity of the waiting-time analysis, which is the case we treat in the rest of the paper. Of course, a simple idea is to predict the waiting time distribution at t as if the number of presently staffed servers remains a constant after t. Because this idea ignores the future dynamics of the staffing function, we dubbed this the *myopic* prediction. To emphasize the importance of the future information of the staffing function, we use a simple example to show that the myopic prediction can lead to significantly inaccurate solutions (with large prediction errors).

1.1 A Motivating Example

Consider an $M/M/n_t + M$ queueing model where customers arrive according to a Poisson process with rate $\lambda = 20$, exponential service times with rate $\mu = 2$, exponential abandonment times with rate $\theta = 0.5$, and a time-varying staffing level n_t which fluctuates between 5 and 15 following a nearly sinusoidal pattern (see the bottom panel of Figure 1). Here our goal is to predict the trajectories of the mean and variance of the waiting time process in the interval $[0, T]$ with $T = 12$. First, we conduct Monte-Carlo simulations to estimate the "ground truth" values using 1000 independent replications (the dashed lines in Figure 1. Next, we compute the predicted curves under the myopic prediction (the solid lines) and compare them to their ground truth values. We observe that both the myopically predicted mean (top panel) and variance (middle panel) exhibit significant errors. According to Figure 1, we observe that, as the staffing level decreases (increases), the myopic prediction underestimates (overestimates) the true wait time. This simple example illustrates the need to incorporate time-varying staffing information into the prediction of wait time. To do so, we develop a method called the *delay estimator under time-varying staffing* (DETS). As showed in Figure 1, the predicted values under DETS are nearly identical to the ground truth values for both the mean and variance of the waiting times. The detailed descriptions of the prediction methods (both myopic and DETS) are presented in Sections 2–3.

1.2 Related Literature

While much of the literature has been concerned with the analytical modeling of time-varying queues, there has also been a large focus on making staffing decisions to control these systems. Green and Kolesar (1991) developed the pointwise stationary approximation (PSA) method, which treats a time-varying queue as a sequence of time-indexed stationary models with a constant arrival rate. Jennings et al. (1996) used an infinite-server approximation to determine appropriate staffing levels in a time-varying model without customer abandonment. This approach is often referred to as the modified-offered-load (MOL) approximation, which has been proven effective in stabilizing metrics such as the probability of delay (Jennings et al. 1996; Yom-Tov and Mandelbaum 2014; Feldman et al. 2008), the mean waiting time and probability of abandonment (Liu and Whitt 2012b; Liu and Whitt 2014b; Liu and Whitt 2017), the blocking probability in loss models (Li et al. 2016; Whitt and Zhao 2017), and the TPoD in multiclass queues (Liu 2018; Liu et al. 2021). Also see He et al. (2016) and Sun and Liu (2021) for studies on the impact of the non-Poisson arrivals on the time-varying staffing levels. The analysis and control of time-varying queues also benefit from heavy-traffic theories. See Whitt (1992), Mandelbaum and Pats (1998), Mandelbaum et al. (1998), Garnett et al. (2002), Borst et al. (2004), Liu and Whitt (2011), Liu and Whitt (2012a), Liu and Whitt (2014a), Liu et al. (2021) for treatments of time-varying queues using heavy-traffic approximations.

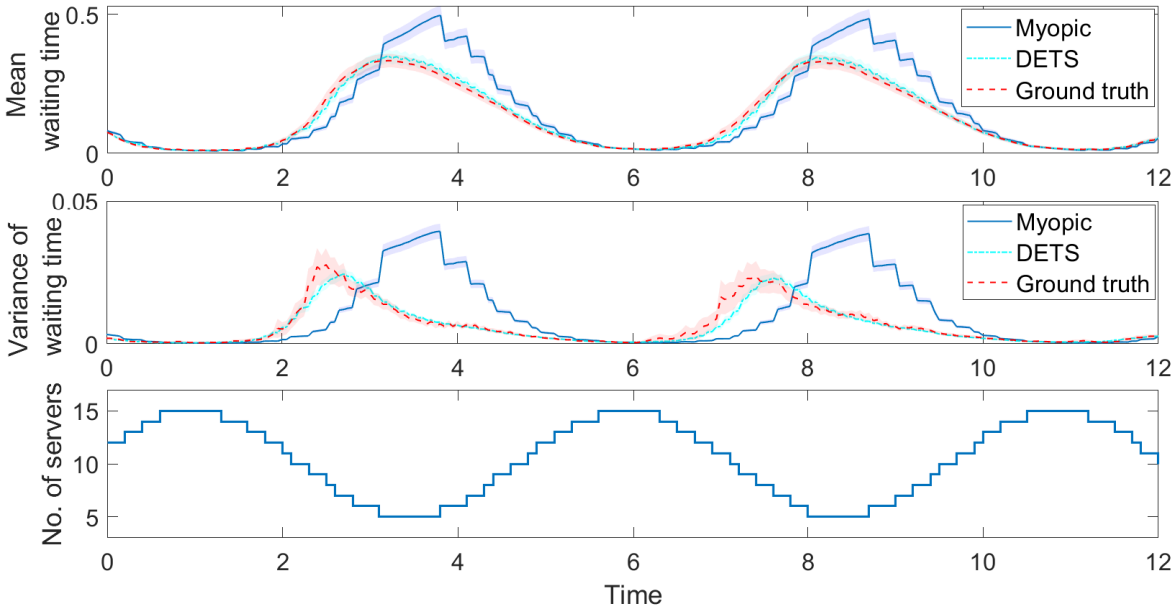

Figure 1: Waiting time prediction under time-varying staffing: myopic vs. DETS.

See Whitt (2018) for a comprehensive review of time-varying queues. Distinct from the above literature, we develop new methods that produce the exact waiting time distribution under time-varying staffing by explicitly modeling the discrete nature of queueing systems. Our work is also somewhat related to the delay-announcement literature because well-predicted waiting-time distributions can enable managers to build effective delay-announcement mechanisms. A recent survey paper on delay announcements (Ibrahim 2018) observed that many challenges exist when the staffing levels fluctuate. Also see Ibrahim and Whitt (2011) for delay announcement estimators in queueing models under various settings.

1.3 Contributions and Organization

We make the following contributions:

- We develop analytic formulas for the real-time prediction of the waiting time distributions in a time-varying queueing model. Our results properly take into account the impact of staffing changes on the waiting time process.
- We also discuss several practical policies for how busy servers should conclude their work, and we investigate the policies' impacts on our results. Ours is the first paper to formalize all three policies, which can provide a basis for future works on queues with time-varying staffing.
- We conduct a comprehensive set of computer simulations to evaluate the effectiveness of our methods and gain insights into the roles of different work-releasing policies.

The remainder of this paper shall proceed as follows. We focus on the simple case where staffing is fixed in Section 2. In Section 3, we treat the general case with time-varying staffing; we also discuss various work-releasing policies and study their impacts on wait times. In Section 4, we conduct simulation studies to examine the effectiveness of our results. Finally, we provide concluding remarks in Section 5.

2 QUEUE-BASED WAITING-TIME PREDICTION UNDER CONSTANT STAFFING

We consider the Markovian $M_t/M/s_t+M$ model having time-varying Poisson arrivals (the M_t), exponentially distributed service times with rate μ (the M), a time-varying staffing function (the s_t), and exponentially

distributed customer abandonment with rate θ (the $+M$). We assume the system is operated under the *first-come first-served* (FCFS) rule. We aim to predict the waiting-time distribution for a newly arrived customer conditional on the number of existing customers in the queue $Q = q$. Note that this new customer can begin receiving service only when all q existing customers have departed the queue. Because a new customer's waiting time depends only on the dynamics of the existing customers, not on any future arrivals (thanks to the FCFS rule), our prediction method does not require any specific assumptions on the arrival process. In other words, our results remain correct for the more general G_t arrivals beyond M_t.

In this section, we describe our waiting-time prediction method by focusing on the simple case: constant staffing. There two commonly used versions for the waiting times are: (i) *potential waiting time* (PWT) - the time that an infinitely patient customer would spend in the queue before entering service, and (ii) *actual waiting time* (AWT) - the total time a customer spends in the waiting line regardless of whether the customer eventually abandons or is served. PWT is a customer-patience independent metric that focuses on describing the system-side congestion or workload; AWT is the mixture of a customer's patience level and the system's workload. Let W be the PWT and A be a generic patience time, AWT is the minimum of W and A. In this paper, we focus on the prediction of the distribution of PWT W. However, at the end of this section, we will also briefly remark how to extend our method to treat AWT.

2.1 Waiting-Time Prediction

Suppose there are $s_t = s$ servers. Let W_Q denote the PWT of a new customer seeing Q existing customers upon arrival. If $Q < s$, then $W_Q = 0$ since that new customer immediately enters service without needing to wait in queue. Hence, we focus on the non-trivial case where all servers are currently occupied. Let $\bar{q} = Q - s$ correspond to the number of existing customers in the queue.

We model W_Q as a *phase-type* (PH) random variable. Specifically, W_Q is the first-passage time until absorption in a *continuous-time Markov chain* (CTMC) which tracks the dynamics of the tagged customer's position in queue. This CTMC has a state space $\{-1, 0, 1, \ldots, \bar{q}\}$, where a state $q = 0, 1, \ldots, \bar{q}$ means that there are currently q customers remaining in the queue who arrived before the tagged customer, and the state -1 (called "Exit Queue") is an absorbing state that indicates that the customer of interest has entered service. We can transition from state q to state $q-1$ with rate $q\theta + s\mu$ as soon as one of the q customers either abandons the queue or enters service. In Figure 2, we present a transition rate diagram for this queueing system. We also give the transition rate matrix $\mathbf{Q}_{\bar{q}}$ for this system in (1).

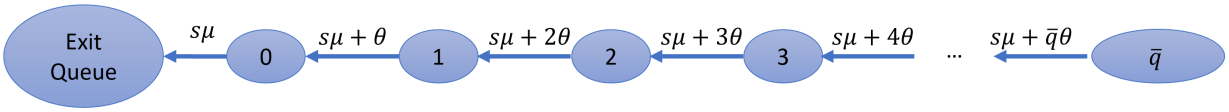

Figure 2: PWT: Transition rate diagram tracking the tagged customer's queue position.

$$\mathbf{Q}_{\bar{q}} = \begin{pmatrix} -s\mu & 0 & 0 & 0 & \cdots & 0 \\ s\mu+\theta & -(s\mu+\theta) & 0 & 0 & \cdots & 0 \\ 0 & s\mu+2\theta & -(s\mu+2\theta) & 0 & \cdots & 0 \\ 0 & 0 & s\mu+3\theta & -(s\mu+3\theta) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & s\mu+\bar{q}\theta & -(s\mu+\bar{q}\theta) \end{pmatrix}. \tag{1}$$

According to Latouche and Ramaswami (1999), the waiting time distribution can then be calculated as

$$F_{\bar{q}}(\tau) \equiv \mathbb{P}(W_{\bar{q}} \leq \tau) = 1 - \beta_0 \left(e^{\mathbf{Q}_{\bar{q}}\tau} \right) \mathbf{e}, \qquad \tau > 0, \tag{2}$$

where β_0 is a row vector of zeros with a one at the end, and $\mathbf{e}$ is a column vector of all ones.

Remark 1 (Treating AWT) To treat AWT, it suffices to modify the transition rate matrix as follows: for every state we will add an additional rate θ transitioning from that state to the "Exit Queue" state, indicating the rate at which the individual customer of interest abandons. See Figure 3 for the transition rate diagram for the AWT. The new AWT transition rate matrix, denoted as $\mathbf{Q}^A_{\bar{q}}$ is given by as $\mathbf{Q}^A_{\bar{q}} = \mathbf{Q}_{\bar{q}} - \theta\mathbf{I}$, with $\mathbf{I}$ being the identity matrix with the proper dimension. To compute AWT's distribution, it suffices to replace $\mathbf{Q}_{\bar{q}}$ by $\mathbf{Q}^A_{\bar{q}}$ in (2).

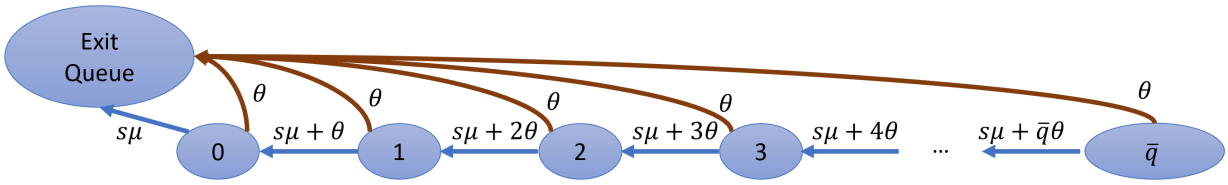

Figure 3: AWT: Transition rate diagram tracking the tagged customer's queue position.

3 WAITING-TIME PREDICTION UNDER TIME-VARYING STAFFING

We now examine how to incorporate time-varying staffing into our waiting-time prediction. We first define three *work-releasing* policies for busy servers upon planned staffing changes; specifically, we explain how to remove a busy server who is scheduled to depart. Next, we set up the general framework for computing the wait times. Finally, we compute the waiting time distribution under all three work-releasing rules.

3.1 Work-Releasing Policies for Busy Servers

When the staffing level (i.e., number of servers) changes over time, we need to clarify the ways to handle staffing changes. An idle server leaves immediately upon its scheduled departure time. On the other hand, when a busy server is scheduled to depart, it is far less straightforward. Below we describe three primary policies: *preemptive* (PE), *exhaustive completion* (EC), and *exhaustive handoff* (EH).

1. **Preemptive (PE).** Under PE, any busy servers who are scheduled to depart will return their customers to the head of the queue. The customer will resume service as soon as the next server becomes available. We assume that our measurement of PWT is only until the time of first service, and does not include any additional wait time that may arise after the service begins. PE has been primarily studied in the literature due to its mathematical simplicity (see Mandelbaum et al. 1998). In practice, PE is often used in priority queueing systems. For example, a multi-skilled agent in a customer contact center may switch to high priority customers (e.g., calls) in the middle of the processing of low priority jobs (e.g., messengers and emails). For mathematical convenience, we assume a customer returned to the queue due to staffing changes can still abandon from the queue with the same rate θ. However, it is not too difficult to treat the case of changed abandonment behavior; for example, the abandonment rate becomes $\theta' \neq \theta$ (a special case $\theta' = 0$ means that customers do not abandon after their services begin).
2. **Exhaustive completion (EC).** Under EC, when servers are due to depart, they complete service on their current customer before departing. EC prevails when service times are relatively short. Practical examples of EC include grocery store check-out lanes and call centers; see Ingolfsson et al. (2007). At the time a server is scheduled to depart, this server, although it continues to serve its present customer, instantaneously turns into an "isolated" server, and it no longer affects the rate at which a waiting customer advances in line.
3. **Exhaustive Handoff (EH).** Similar to EC, a busy server continues serving its present customer at the time of its scheduled departure. The distinction is that, rather than completing the full service, that server can transfer the customer to the next idle server as soon as one becomes available. Both

EC and EH require busy servers to do some overtime (extending their work times beyond their scheduled depart time). However, EC delays a server's departure by nearly a full service duration, while EH demands a much shorter overtime, especially in a large-scale system where any one of the many other servers can be a potential "savior". EH works well when service times are relatively long and the time and effort required to transfer a customer between servers is relatively small. Practical applications of EH include hospitals and repair shops (Li et al. 2016).

3.2 DETS: The General Framework

We use two sequences to quantify a given staffing plan: $\{t_1, t_2, \ldots, t_k\}$ and $\{\Delta s_1, \Delta s_2, \ldots, \Delta s_k\}$, with t_k and Δs_k denoting the time and net amount of the k^{th} staffing change. Hence, the staffing level is a piecewise stair function given by

$$s(t) = s_0 + \sum_i \Delta s_i \mathbf{1}_{\{t_i \le t\}}, \tag{3}$$

where $\mathbf{1}_A = 1$ if A occurs and is 0 otherwise. For simplicity, we assume that at the time of the k^{th} staffing change, only one type of event should occur: busy servers depart or new servers begin working, but not both. We will show how this setting may be generalized at the end of this section.

Instead of the CDF, we hereby compute the CCDF, that is the probability that a customer's waiting time exceeds a wait time target τ, or equivalently, a customer waits in the queue for at least τ time units. Unlike the case of constant staffing (as illustrated in Section 2), where the problem can be solved using a single Markov chain, the main challenge here is that the changes in the staffing level give rise to changes in both the service completion rate as well as the system state. To see this, note that the service completion rate decreases as $s(t)$ decreases, while the queue length decreases as $s(t)$ increases (with customers at the head of the line immediately admitted into service), which also reduces the abandonment rate. In any case, the probability structure of the Markov chain needs to be updated every time the staffing level changes. The time-dependent structure of the Markov chain can be determined in two steps:

- **Inter-changeover:** First, we describe the transition probability between two epochs of staffing changeovers t_k and t_{k+1}, which can be given for transient states as $e^{\mathbf{Q}^{(k)} \cdot (t_{k+1} - t_k)}$, where $\mathbf{Q}^{(k)}$ is the transient transition rate matrix updated right after time t_k and $e^{\mathbf{X}}$ represents the exponential function of a matrix $\mathbf{X}$. The detailed definition of $\mathbf{Q}^{(k)}$ depends on the work-releasing policy under consideration and will be specified later.
- **Upon-changeover:** Next, we quantitatively describe how the system state should be updated at t_k. To do so, we use a transformation matrix $\mathscr{L}^{(k)}$ that maps transient state q to transient state q' in a way that depends on staffing changeover. (For instance, if there are three customers in the queue, and two servers begin service, then our system experiences an instantaneous and deterministic change from state 3 to state 1 because two customers can begin service with the two new servers.) We give the details for $\mathscr{L}^{(k)}$ later, which again depends upon the work-releasing policy.

Given $\left\{ \left(\mathbf{Q}^{(k)}, \mathscr{L}^{(k)}, t_k \right), k = 1, 2, \ldots, K \right\}$, we can now give the distribution function of the wait time.

Theorem 1 (Delay estimation under time-varying staffing (DETS)) The CCDF of the wait time is given by

$$F^c(\tau) = \mathbb{P}(W > \tau) = \beta_0 \prod_{k=1}^{K} \left(e^{\mathbf{Q}^{(k-1)} \cdot (t_k - t_{k-1})} \mathscr{L}^{(k)} \right) e^{\mathbf{Q}^{(K)} \cdot (\tau - t_k)} \mathbf{e}, \qquad \tau > 0, \tag{4}$$

where β_0 is a row vector of zeros except for its last component which is 1, and $\mathbf{e}$ is a column vector of all ones and the specific forms of $\mathbf{Q}^{(k)}$ and $\mathscr{L}^{(k)}$ depend on the work-releasing policy in use. Further, $K \equiv \max\{k \ge 0 | t_k < \tau\}$ is the total number of staffing changes covered by the interval $[0, \tau]$.

In the next subsections, we give the details for $\mathbf{Q}^{(k)}$ and $\mathscr{L}^{(k)}$ under all three policies. We give three versions of DETS, called DETS-PE, DETS-EC and DETS-EH.

3.3 Preemptive: DETS-PE

We first derive $\mathbf{Q}^{(k)}$ under PE. Consider a Markov chain with the state space spanning from 0 to $\bar{q}^{(k)}$, where $\bar{q}^{(k)}$ is the maximum possible number of customers waiting in line in the k^{th} staffing interval. A transition occurs from state q to state $q-1$ if a customer either abandons or finishes service, with rate $s_k\mu + q\theta$. Note that a service completion at state 0 means that the customer of interest enters service, which ends the whole process. The transition diagram and rate matrix are identical to those in Figure 2 and (1), except that s should be repaced by s_k.

We next describe the transformation matrix $\mathscr{L}^{(k)}$ under PE. If $\Delta s_k = 0$, then there is no change, and $q' = q$. If $\Delta s_k > 0$, Δs_k customers enter service immediately so that only the states bigger than Δs_k remain transient. Hence, $q' = q - \Delta s_k$. If $\Delta s_k < 0$, then Δs_k already-in-service customers return to the head of the line, leading to $q' = q + |\Delta s_k|$. Because we assume that any customers who are sent back to the queue maintain their original abandonment rate θ, there is no need to distinguish the customer who have previously entered service from those who haven't. Hence, the matrix $\mathscr{L}^{(k)}$ having size $(\bar{q}^{(k-1)}+1)\times(\bar{q}^{(k)}+1)$ is specified as:

$$\mathscr{L}^{(k)} = \begin{cases} \overline{\mathscr{L}}^{(k)}, & \Delta s_k > 0 \\ \mathbf{I}, & \Delta s_k = 0 \\ \underline{\mathscr{L}}^{(k)}, & \Delta s_k < 0 \end{cases}, \quad \overline{\mathscr{L}}^{(k)}_{ij} = \begin{cases} 1, & if\ j = i - \Delta s_k \\ 0, & \text{otherwise} \end{cases}, \quad \underline{\mathscr{L}}^{(k)}_{ij} = \begin{cases} 1, & if\ j = i + |\Delta s_k| \\ 0, & \text{otherwise} \end{cases}.$$

The maximum queue length through time can be recursively defined as:

$$\bar{q}^{(k)} = \bar{q}^{(k-1)} - \Delta s_k, \qquad \bar{q}^{(k)} = \bar{q}^{(0)} - \sum_{i=1}^{k} \Delta s_i$$

3.4 Exhaustive Completion: DETS-EC

Although the transition rate matrix $\mathbf{Q}^{(k)}$ under EC is identical to that under PE, the distinction lies in the matrix $\mathscr{L}^{(k)}$. Specifically, if $\Delta s_k < 0$, the departing servers will not be available to serve any future customers. However, the customers presently dedicated to them will not return to the queue, so that the size of the waiting queue remains unchanged (i.e., $q' = q$). So we have

$$\mathscr{L}^{(k)} = \begin{cases} \overline{\mathscr{L}}^{(k)}, & \text{if } \Delta s_k > 0 \\ \mathbf{I}, & \text{otherwise} \end{cases} \quad \text{and} \quad \overline{\mathscr{L}}^{(k)}_{ij} = \begin{cases} 1, & \text{if } j = i - \Delta s_k \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Although the size of $\mathscr{L}^{(k)}$ remains $(\bar{q}^{(k-1)}+1)\times(\bar{q}^{(k)}+1)$, the updating procedure of the queue length is different (because the queue length can no longer increase):

$$\bar{q}^{(k)} = \bar{q}^{(k-1)} - (\Delta s_k)^+, \qquad \bar{q}^{(k)} = \bar{q}^{(0)} - \sum_{i=1}^{k} (\Delta s_i)^+. \tag{6}$$

3.5 Exhaustive Handoff: DETS-EH

The treatment of EH is more complex than those of EC and PE, because we have to track the number of servers who are about to depart but haven't yet due to the lack of "helpers" (the first idle servers available to take over their customers). We do so by augmenting the state space of our Markov chain, adding an additional element r to the state definition, where r represents the number of servers who have been

scheduled to depart, but have not left yet. We first determine the transition rates between two staffing changeover epochs. If we have a service completion while $r > 0$, then the Markov chain transitions from (q,r) to $(q,r-1)$, otherwise it goes from $(q,0)$ to $(q-1,0)$. Absorption (exiting the queue and entering service) occurs from state $(0,0)$ only. See Figure 4 for the transition diagram.

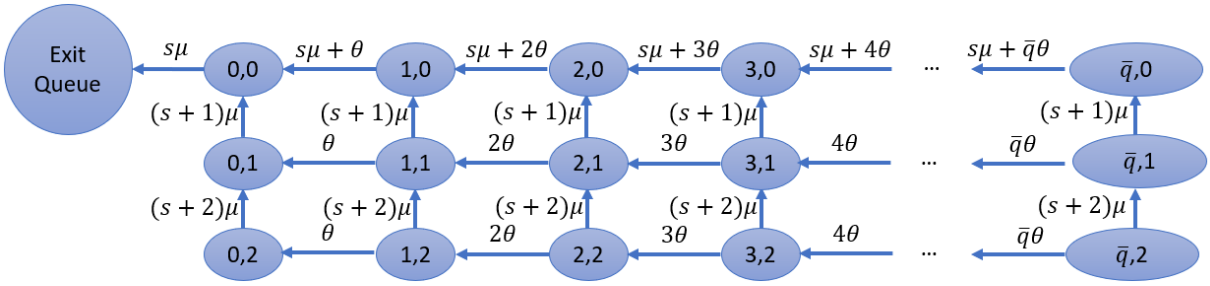

Figure 4: The transition rate diagram for the exhaustive handoff policy.

The transition rate diagram reveals a quasi-birth-and-death structure (Latouche and Ramaswami 1999), where the levels are the number of servers who are "about" to depart. Below we give the transition rate matrix using submatrices $\mathbf{A}^{(k)}$, $\mathbf{B}_r^{(k)}$, and $\mathbf{C}_r^{(k)}$, all of which are square matrices of size $(\bar{q}^{(k)}+1)$:

$$\mathbf{Q}^{(k)} = \left(\begin{array}{c|c|c|c} \mathbf{A}^{(k)} & \mathbf{0} & \mathbf{0} & \cdots \\ \hline \mathbf{B}_1^{(k)} & \mathbf{C}_1^{(k)} & \mathbf{0} & \ddots \\ \hline \mathbf{0} & \mathbf{B}_2^{(k)} & \mathbf{C}_2^{(k)} & \ddots \\ \hline \vdots & \ddots & \ddots & \ddots \end{array}\right), \quad \text{where} \quad \mathbf{C}_r^{(k)} = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ \theta & -\theta & 0 & \ddots \\ 0 & 2\theta & -2\theta & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} - \mathbf{B}_r^{(k)}.$$

describes the intra-level transitions, $\mathbf{A}^{(k)}$ is nearly identical to the matrix in (1) (using s_k instead of s), and $\mathbf{B}_r^{(k)}$ describes the movement from one level to another with $\mathbf{B}_r^{(k)} = (s_k + r)\mu\mathbf{I}$. The overall matrix $\mathbf{Q}^{(k)}$ is a square matrix of size $(\bar{r}^{(k)}+1)\cdot(\bar{q}^{(k)}+1)$, with $\bar{r}^{(k)}$ denoting the maximum number of servers who are scheduled to leave.

We next specify the matrix $\mathscr{L}^{(k)}$ to describe the transformation needed upon a change in staffing. First, if $\Delta s_k = 0$, then state $(q',r') = (q,r)$, and $\mathscr{L}^{(k)} = \mathbf{I}$. Next, if $\Delta s_k < 0$, we transition from state (q,r) to state $(q, r+|\Delta s_k|)$. The case $\Delta s_k > 0$ is somewhat more complex: When servers are added to the service capacity, they first replace those servers with delayed departures (if any) by taking over their unfinished services. If all departing servers have left, they begin serving additional customers from the head of the waiting queue. In Figure 4, we first move vertically up the transition rate diagram, and then we move left until we travel a total of Δs_k states. This corresponds to a move from state (q,r) to a state $(q-y,x)$, where $y = (\Delta s_k - r)^+$ and $x = (r - \Delta s_k)^+$. Specifically, we have

$$\mathscr{L}^{(k)} = \begin{cases} \overline{\mathscr{L}}^{(k)} & ,\text{if} \quad \Delta s_k > 0 \\ \mathbf{I} & ,\text{if} \quad \Delta s_k = 0 \\ \underline{\mathscr{L}}^{(k)} & ,\text{if} \quad \Delta s_k < 0 \end{cases}, \quad \text{where} \quad \underline{\mathscr{L}}_{ij}^{(k)} = \begin{cases} 1 & ,\text{if} \quad j = i + (\bar{q}^{(k-1)}+1)|\Delta s_k| \\ 0 & ,\text{otherwise} \end{cases}$$

and

$$\overline{\mathscr{L}}_{ij}^{(k)} = \begin{cases} 1, & \text{if} \quad i = (\bar{q}^{(k-1)}+1)r + q,\ j = (\bar{q}^{(k-1)}+1)(r-\Delta s_k)^+ + q - (\Delta s_k - r)^+, \\ & \quad q = 0,1,\ldots,\bar{q}^{(k-1)},\ r = 0,1,\ldots,\bar{r}^{(k-1)} \\ 0, & \text{otherwise} \end{cases}.$$

Due to the two-dimensional state space, $\mathscr{L}^{(k)}$ now has size $(\bar{r}^{(k-1)}+1)\cdot(\bar{q}^{(k-1)}+1)$ by $(\bar{r}^{(k)}+1)\cdot(\bar{q}^{(k)}+1)$. Finally, we must account for the changes in state space size. When the staffing level decreases,

the number of levels increases from $\bar{r}^{(k)}$ to $\bar{r}^{(k)} + |\Delta s_k|$, while the maximum queue length $\bar{q}^{(k)}$ remains unchanged. When the staffing level increases, the number of levels decreases from $\bar{r}^{(k)}$ to $(\bar{r}^{(k)} - \Delta s_k)^+$, and the maximum queue length decreases from $\bar{q}^{(k)}$ to $\bar{q}^{(k)} - (\Delta s_k - \bar{r}^{(k)})^+$. Specifically, we have

$$\bar{q}^{(k)} = \bar{q}^{(k-1)} - (\Delta s_k - \bar{r}^{(k-1)})^+, \qquad \bar{r}^{(k)} = (\bar{r}^{(k-1)} - \Delta s_k)^+.$$

Remark 2 (Simultaneous removal and addition) So far, at t_k, we assume that either $|\Delta s_k|$ working servers are to be removed from the staffing level when $\Delta s_k < 0$, or Δs_k new servers are to be added to the staffing level when $\Delta s_k > 0$. In practice, it is often possible that the removal and addition may occur at the same time. Let $\Delta s_k^- \geq 0$ and $\Delta s_k^+ \geq 0$ be the numbers of servers to be removed and added, respectfully. So the net staffing change is $\Delta s_k = \Delta s_k^+ - \Delta s_k^-$ (which can be either positive or negative). Under PE and EH, the arriving servers will instantaneously replace the departing ones by taking over their customers (if any). Hence, under both PE and EH, it suffices to focus on the net staffing change by simply assuming that $|\Delta s_k|$ (Δs_k) servers are to be removed (added) when $\Delta s_k < 0$ ($\Delta s_k > 0$). However, some adjustment is needed for the treatment of the EC case. First, the new effective service rate after time t_k (i.e., the death rate of the updated Markov chain) is again determined only by the net change Δs_k, because the Δs_k^- departing servers will no longer contribute to the advancing the waiting queue and the Δs_k^+ arriving servers immediately will begin serving new customers from the queue. Hence, the updating rule for the transition rate as defined in (3) remains unchanged. On the other hand, the adjustment of the queueing position q should be based only on the number of the incoming servers Δs_k^+ instead of Δs_k. Thus, it suffices to replace the Δs_k by Δs_k^+ in (5) and (6).

4 SIMULATION EXPERIMENTS

In this section, we conduction computer simulation experiments to confirm the effectiveness of our prediction methods under all three policies. We compare results of DETS-PE, DETS-EC, and DETS-EH to their corresponding ground truth values estimated by Monte-Carlo simulations.

4.1 Decreasing Staffing Function

As noted earlier, the three policies are indistinguishable when the staffing level increases with additional servers added to the server pool, but complications arise when busy servers are to be set free. Therefore, in our first example we consider the case of a decreasing staffing level: one server is to be removed every 0.1 time units (top panel in Figure 5). We run our DETS algorithms to generate the predicted waiting time distribution functions with different initial queue size q_0. In Figure 5, we see that all DETS results agree with their ground truth values: they all fall into the simulated 95% confidence interval envelops (shaded bands). Every ground truth result is estimated using 1000 independent simulation runs.

The waiting time CDFs are smooth except when the staffing level decreases. The non-smoothness is more pronounced in the cases of PE and EH, as shown in the second and last panels of Figures 5. Here, we see that when the staffing function decreases, the CDFs experience a short period of flattening out before they continue to increase. This is essentially a result of the "pause" in productivity when a server departs. Though these two policies appear to have similar CDFs, we note that their CDFs will only be identical when $\mu = \theta$. This is essentially because productive service resumes at the same rate only in this case.

4.2 Other Patterns of Time-Varying Staffing

We next consider other profiles of the time-varying staffing level. We consider three cases: (a) an increasing staffing function (the opposite case to the previous example), (b) a sinusoidal staffing function where the number of servers first decreases and then increases, and (c) an alternating staffing which periodically adds and removes one server (see the top panels in Figure 6). We then use DETS to compute the predicted waiting time CDFs under the three policies (see bottom panels in Figure 6).

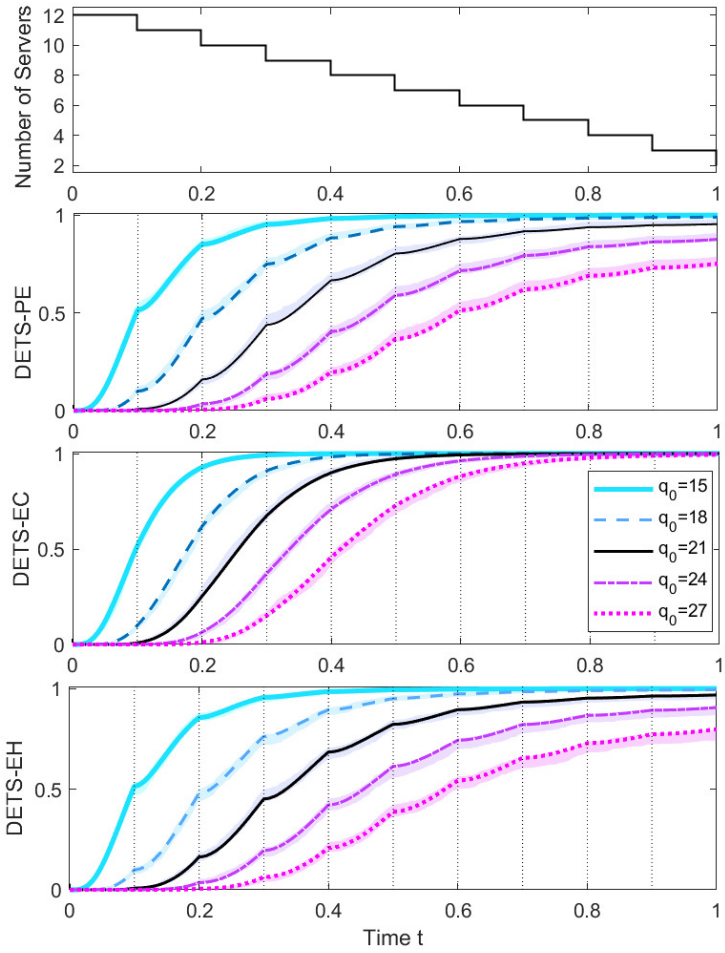

Figure 5: Comparing DETS to simulations: Waiting time distributions under decreasing staffing and three work-releasing policies (i) PE (top), (ii) EC (middle) and (iii) EH (bottom), with $\mu = 3$ and $\theta = 1$.

All DETS results agree with their simulated ground truth values, with all DETS curves being covered by the 95% confidence intervals (shaded bands). When the staffing is increasing (case (a)), the three policies have identical results because the nuances in these policies arise only when busy servers attempt to depart. In addition, the waiting time CDF exhibits positive probability masses whenever the staffing level increases (see the instantaneous jumps in the left panel), because it is possible for a customer to be received by the newly added server.As long as there are periods of time in which the staffing level decreases (as in Cases (b) and (c)), we can clearly distinguish the waiting time CDFs under the three policies. Under EC, the departing servers finish their customer's service prior to departure; this does not affect the waiting time for the remaining customers in queue. On the other hand, under both PE and EH, the process of advancing the waiting queue is temporarily 'paused' until a relevant new event occurs (service completion or abandonment). This explains why EC induces a stochastically shorter waiting time with its CDF significantly higher than those under these other two policies (as exhibited in Figure 6) when the staffing function experiences decreases.

5 CONCLUDING REMARKS

In this paper, we develop a new framework to compute the waiting time distribution in a Markovian queueing system having time-varying staffing levels. We also study the impact of three work-releasing policies on the

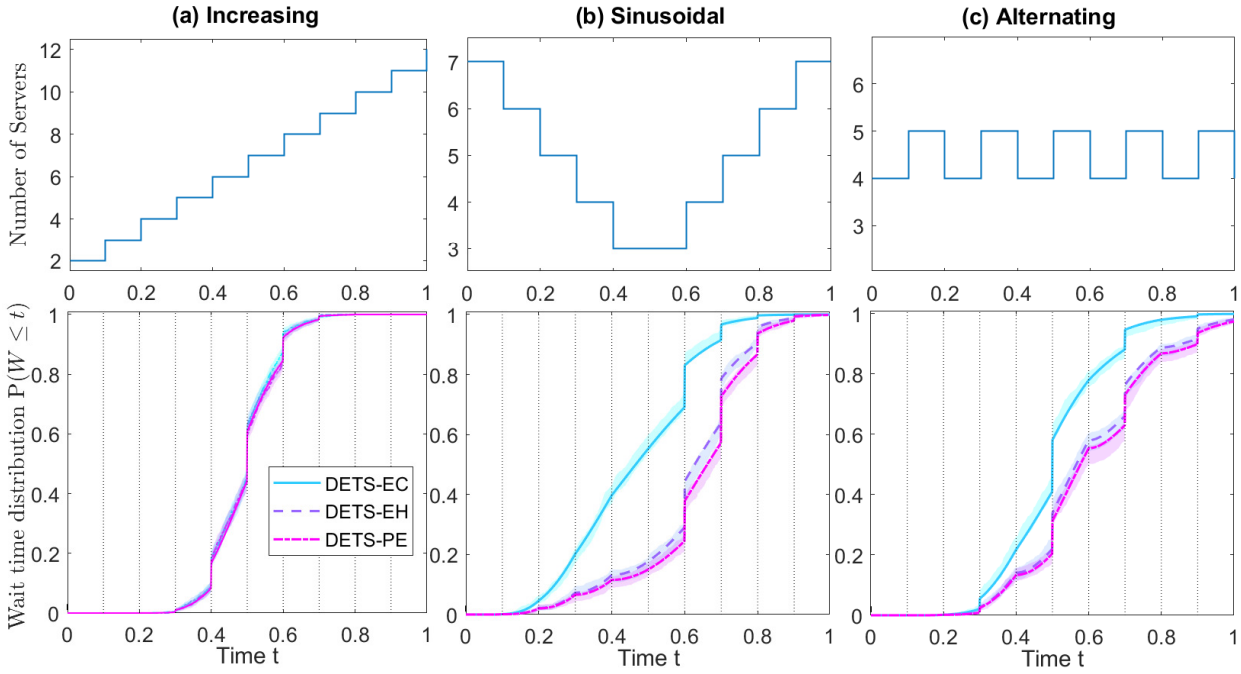

Figure 6: Comparing DETS to simulations: Waiting time distributions under (a) increasing, (b) sinusoidal, and (c) alternating staffing function and three work-releasing policies, with $q_0 = 15$, $\mu = 3$ and $\theta = 1$.

predicted waiting times. We argue that all future queueing simulations with time-varying staffing should explicitly identify the work-releasing policy they use, which can be influential on their results.

There are several avenues for future research in this area. First, we intend to extend the waiting time prediction methodology for the Markovian queue to the non-Markovian queue having nonexponential service and abandonment times, drawing from the method in Whitt (2005) that approximates the instantaneous abandonment rate by the aggregated hazard rates from all customers under different ages. Second, we plan to develop a more comprehensive version of DETS for other important metrics, such as the probability of abandonment and mean waiting time. These extensions will follow similar ideas of the base DETS except for a reconstruction of the transition rate matrices. Finally, an important future direction is to use this prediction methodology to develop an offline reinforcement learning algorithm that prescribes appropriate time-varying staffing levels in order to meet desired SL targets, such as the TPoD (Liu 2018). Drawing inspirations from Feldman et al. (2008), we will develop a recursive method that alternates between (a) the **exploitation** of the queue length distribution Q under the present staffing rule to provide DETS predictions for the Q-conditional wait-time distribution at future times, and (b) the **exploration** of improved staffing rules by keeping the DETs waiting time in check (satisfying the desired SL target).

REFERENCES

Borst, S., A. Mandelbaum, and M. I. Reiman. 2004. "Dimensioning Large Call Centers". *Operations Research* 52(1):17–34.

Feldman, Z., A. Mandelbaum, W. A. Massey, and W. Whitt. 2008. "Staffing of Time-Varying Queues to Achieve Time-Stable Performance". *Management Science* 54(2):324–338.

Garnett, O., A. Mandelbaum, and M. Reiman. 2002. "Designing a Call Center with Impatient Customers". *Manufacturing & Service Operations Management* 4(3):208–227.

Green, L., and P. Kolesar. 1991. "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals". *Management Science* 37(1):84–97.

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.

He, B., Y. Liu, and W. Whitt. 2016. "Staffing a Service System with Non-Poisson Nonstationary Arrivals". *Probability in the Engineering and Informational Sciences* 30(2):593–621.

Ibrahim, R. 2018. "Sharing Delay Information in Service Systems: a Literature Survey". *Queueing Systems* 89(1):49–79.

Ibrahim, R., and W. Whitt. 2011. "Wait-Time Predictors for Customer Service Systems with Time-Varying Demand and Capacity". *Operations Research* 59(5):1106–1118.

Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li, and X. Wu. 2007. "A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline". *INFORMS Journal on Computing* 19(2):201–214.

Jennings, O. B., A. Mandelbaum, W. A. Massey, and W. Whitt. 1996. "Server Staffing to Meet Time-Varying Demand". *Management Science* 42(10):1383–1394.

Latouche, G., and V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, Pa: Society for Industrial and Applied Mathematics.

Li, A., W. Whitt, and J. Zhao. 2016. "Staffing To Stabilize Blocking In Loss Models With Time-Varying Arrival Rates". *Probability in the Engineering and Informational Sciences* 30(2):185–211.

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. 2021. "Scheduling to Differentiate Service in a Multiclass Service System". *Operations Research* 70(1):527–544.

Liu, Y., and W. Whitt. 2011. "A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment". *Operations Research* 59(4):835–846.

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

Liu, Y., and W. Whitt. 2012b. "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(1):59–73.

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

Liu, Y., and W. Whitt. 2017. "Stabilizing Performance in a Service System with Time-Varying Arrivals and Customer Feedback". *European Journal of Operational Research* 256(2):473–486.

Mandelbaum, A., W. A. Massey, and M. I. Reiman. 1998. "Strong Approximations for Markovian Service Networks". *Queueing Systems* 30(1):149–201.

Mandelbaum, A., and G. Pats. 1998. "State-Dependent Stochastic Networks. Part I: Approximations and Applications with Continuous Diffusion Limits". *The Annals of Applied Probability* 8(2):569–646.

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

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 17^{th}.

Sun, X., and Y. Liu. 2021. "Staffing Many-Server Queues with Autoregressive Inputs". *Naval Research Logistics* 68(3):312–326.

Whitt, W. 1992. "Understanding the Efficiency of Multi-Server Service Systems". *Management Science* 38(5):708–723.

Whitt, W. 2005. "Engineering Solution of a Basic Call-Center Model". *Management Science* 51(2):221–235.

Whitt, W. 2018. "Time-Varying Queues". *Queueing Models and Service Management* 1(2):79–164.

Whitt, W., and J. Zhao. 2017. "Many-Server Loss Models with Non-Poisson Time-Varying Arrivals". *Naval Research Logistics (NRL)* 64(3):177–202.

Yom-Tov, G. B., and A. Mandelbaum. 2014. "Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing". *Manufacturing & Service Operations Management* 16(2):283–299.

AUTHOR BIOGRAPHIES

KURTIS KONRAD is a PhD candidate in the Edward P. Fitts Department of Industrial Engineering at North Carolina State University. He holds a masters degree in Industrial and Systems Engineering from North Carolina State University. His research interest is in stochastic modeling, simulation, and online learning, with interests in applying these methodologies to health care, transportation networks, and other service systems. His email address is kekonrad@ncsu.edu.

YUNAN LIU is an associate professor in the Department of Industrial and Systems Engineering at North Carolina State University. He earned his Ph.D. in Operations Research from Columbia University. His research interests include queueing theory, stochastic modeling, simulation, applied probability, online learning, and optimal control, with applications to call centers, healthcare, and transportation. His work was awarded first place in the INFORMS Junior Faculty Interest Group Paper Competition in 2016. His email address is yliu48@ncsu.edu. His website is https://yunanliu.wordpress.ncsu.edu/.