Computing time-dependent waiting time probabilities in $M(t)/M/s(t)$ queueing systems

Linda V. Green • João Soares

Graduate School of Business, Columbia University, New York, New York 10027, USA
Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal
lvg1@columbia.edu • jsoares@mat.uc.pt

In this note we present algorithms that compute, exactly or approximately, time dependent waiting time tail probabilities and the time dependent expected waiting time in $M(t)/M/s(t)$ queueing systems.

1. Introduction

In many service systems, the performance measure of interest is a function of the tail probability of the waiting time. For example, in many telephone call centers, the service target is a maximum fraction of customers delayed for more than a given number of seconds, e.g. the probability that a customer waits more than twenty seconds is less than fifteen percent. Another example is a hospital emergency department (ED), in which the goal is to limit the fraction of patients who experience a delay of more than, e.g., an hour in receiving care from a physician. In both of these examples, as well as in many other systems, the customer arrival rate varies over the day, and managers vary the staffing over the day in order to meet the desired performance standard.

In this note, we consider an $M(t)/M/s(t)$ queueing system with arrival rate $\{\lambda(t), t > 0\}$, service rate μ, and the number of servers given by a piecewise constant function $\{s(t), t \geq 0\}$ (of nonnegative integers). Let $W_q(t)$ denote the waiting time in queue until service commences of a customer that arrives to the system at time t. We are interested in computing the tail probability $P(W_q(t) > x)$, where x is a given time parameter and the expected waiting time in queue to begin service is $E(W_q(t))$. This latter measure is important in many applications such as call centers, where it is commonly referred to as average speed of answer (ASA).

When $x = 0$, $P(W_q(t) > x)$ reduces to the probability of delay, which is dependent only on the state probabilities at time t and the number of servers at time t and not on the number subsequent to time t. But when $x > 0$, the derivation is complicated by the fact that the event '$W_q(t) > x$' depends not only on $s(t)$ but also on the number of servers available after t, i.e., $s(u), u \in (t, t + x]$. Similarly, the derivation of the expected waiting time in queue, denoted $E(W_q(t))$, is problematic since it depends upon the tail probabilities.

In our derivations we assume that the infinite dimensional vector $\vec{p}(t) = [p_n(t)]$, where $p_n(t)$ denotes the probability of n customers in the system at time t, is known. For example, this vector $\vec{p}(t)$ may have been obtained numerically as the solution of the Chapman-Kolmogorov differential equations that describe the queueing system at hand, see e.g. Green at al. (2001). Let $W_q^n(t)$ denote the waiting time until service commences for a customer that arrives at time t and sees n people in the system. Then,

$$P(W_q(t) > x) = \sum_{n=s(t)}^{+\infty} P(W_q^n(t) > x) p_n(t), \tag{1}$$

and

$$E(W_q(t)) = \sum_{n=s(t)}^{+\infty} E(W_q^n(t)) p_n(t). \tag{2}$$

In this note, we present exact expressions for $P(W_q^n(t) > x)$ in the important special case where the number of servers changes at most once in the interval $[t, t + x]$, and we present an algorithm for the general case. Easy-to-compute lower and upper bounds are also derived for the general case. We do a similar analysis for $E(W_q^n(t))$, for any n, so that the desired quantities follow from (1) and (2).

Since the departure process behaves as a non-homogeneous Poisson process with rate $\mu s(u)$, for $u \geq t$ (assuming all servers are busy throughout the interval), the number of departures over the time period $[t, t + x]$ is Poisson distributed with mean

$$a = \mu \int_t^{t+x} s(u)\, du. \tag{3}$$

Thus, when $n \geq s(t)$, we may be tempted to say that the event '$W_q^n(t) > x$' is equivalent to the event '$n - s(t)$ or fewer departures over $[t, t + x]$' so that $P(W_q^n(t) > x)$ would be given by

$$P(\text{`}n - s(t) \text{ or fewer departures over } [t, t + x]\text{'}) = \sum_{j=0}^{n-s(t)} \frac{a^j e^{-a}}{j!}. \tag{4}$$

This is not true in general. For example, suppose the number of servers changes exactly once over the time period $[t, t + x]$ at the epoch $t + \Delta t$, where $\Delta t < x$, and the resulting number

of servers is reduced to a level that is less than the number of customers in service. To maintain consistency with the Chapman-Kolmogorov equations, we must assume that the "excess" customers being served at the epoch $t + \Delta t$ will rejoin the queue. (This standard assumption is reasonable in situations where service times are long relative to shift lengths such as hospital EDs. It is likely to be an approximation in some contexts such as call centers. See Ingolfsson et al. 2005 for the case where servers finish serving any customers in progress at shift changes.) Therefore, the $(n+1)$st customer at time t may have to see more than $n - s(t)$ departures over $[t, t + x]$ prior to starting service, namely, $n - s(t + \Delta t)$ if there are not enough departures in $[t, t + \Delta t]$. Thus, $P(W_q^n(t) > x)$ is not always given by (4), contrary to the exposition in the appendix of Ingolfsson et al. (2002). Similarly, if the number of servers increases during $[t, t + x]$, fewer than $n - s(t)$ departures may result in $W_q^n(t) < x$.

In Section 2, we derive precise and simple formulae for $P(W_q^n(t) > x)$ when the number of servers changes at most once in the interval $[t, t + x]$. In many actual settings this is a valid assumption. For example, in a call center, x is likely to be measured in seconds, while staffing levels are typically changed in intervals ranging from 15 minutes to 2 hours. In Section 3, we study the general case, i.e., when the number of servers changes more than once. The general case is more likely to occur in a system like a hospital ED where x is likely to range from a quarter-hour to more than 2 hours and overlapping shifts may cause staffing changes from one hour to the next, see Green et al. (2005). In Section 4, we develop results for $E(W_q^n(t))$.

We note that our results could readily be extended to the case when the service rate also varies with time. In addition, since our approach is based on the assumption that the vector $\vec{p}(t) = [p_n(t)]$ is known, our results can be used for some other Markovian time-varying queueing systems such as the finite capacity $M(t)/M/s(t)/K$ system.

2. The simplest cases for $P(W_q^n(t) > x)$

First, consider the case where the number of servers does not change during the time period $[t, t + x]$, i.e., $s(u) = s_0, u \in [t, t + x]$. Then, $a = \mu s_0 x$ and using standard queueing theory results such as Gross and Harris (1998), page 67, $W_q^n(t)$ is either zero, when $n < s_0$, or the sum of $n - s_0 + 1$ independent and identically distributed (i.i.d.) exponential random

variables with mean $1/(\mu s_0)$, when $n \geq s_0$. Mathematically,

$$P(W_q^n(t) > x) = \begin{cases} \sum_{i=0}^{n-s_0} \frac{a^i e^{-a}}{i!} & \text{if } n \geq s_0, \\ 0 & \text{if } n < s_0. \end{cases} \tag{5}$$

Now, consider the case where the number of servers changes exactly once in $[t, t+x]$, i.e., there exists some $\Delta t \leq x$ such that

$$s(u) = \begin{cases} s_0 & \text{if } u \in [t, t+\Delta t), \\ s_1 & \text{if } u \in [t+\Delta t, t+x]. \end{cases} \tag{6}$$

In this case, $a = y_0 + y_1$ where $y_0 = \mu s_0 \Delta t$ and $y_1 = \mu s_1 (x - \Delta t)$. Notice that $P(W_q^n(t) > x) = 0$ when $n < \max(s_0, s_1)$ because the $(n+1)$st customer will begin service either immediately upon arrival or no later than time $t + \Delta t$. When $n \geq \max(s_0, s_1)$ then $P(W_q^n(t) > x)$ will have a positive value that depends upon whether the number of servers increases or decreases at time $t + \Delta t$.

Suppose $s_0 < s_1$, i.e., the number of servers increases at time $t + \Delta t$. Then, for $n \geq s_1$, the events '$W_q^n(t) > x$' and '$n - s_1$ or fewer departures over $[t, t+x]$' are equivalent. Thus,

$$P(W_q^n(t) > x) = \begin{cases} \sum_{j=0}^{n-s_1} \frac{a^j e^{-a}}{j!} & \text{if } n \geq s_1, \\ 0 & \text{if } n < s_1. \end{cases} \tag{7}$$

Now, suppose that $s_0 > s_1$, i.e., the number of servers decreases at time $t + \Delta t$. For any $n \geq s_0$, let $K_{n,t}(u)$ denote the number of departures over $[t, t+u]$. The event '$W_q^n(t) > x$' can be expressed as the union of two disjoint events

$$A \equiv \{K_{n,t}(x) \leq n - s_0\}, \qquad B \equiv \{K_{n,t}(\Delta t) \leq n - s_0, n - s_0 < K_{n,t}(x) \leq n - s_1\}.$$

A is the event that there were not enough departures for the customer to have entered service even if the number of servers had not been reduced. B is the event that not enough departures occurred before the shift change and the number of departures after the shift change left more than s_1 people in the system at time $t + x$. These two sub-events in B are not independent and so $P(B)$ is computed by conditioning on the number of departures that occurred before the shift change epoch. Mathematically,

$$\begin{aligned} &P(W_q^n(t) > x) = \\ &= \begin{cases} \sum_{j=0}^{n-s_0} \frac{a^j e^{-a}}{j!} + \sum_{j=0}^{n-s_0} \left(\left(\frac{y_0^j e^{-y_0}}{j!} \right) \left(\sum_{i=(n-j-s_0)+1}^{n-j-s_1} \frac{y_1^i e^{-y_1}}{i!} \right) \right) & \text{if } n \geq s_0, \\ 0 & \text{if } n < s_0. \end{cases} \end{aligned} \tag{8}$$

Note that, for any pair of real numbers y_0, y_1,

$$\sum_{j=0}^{n} \frac{(y_0+y_1)^j}{j!} = \sum_{j=0}^{n} \left(\frac{y_0^j}{j!} \sum_{i=0}^{n-j} \frac{y_1^i}{i!} \right), \tag{9}$$

since after multiplying each side by $e^{-(y_0+y_1)}$ (9) becomes equivalent to stating that the distribution function of the sum of two independent Poisson random variables is Poisson with parameter equal to the sum of the individual parameters. Using this identity, we get the following equivalent expression for (8)

$$\begin{aligned} &P(W_q^n(t) > x) = \\ &= \begin{cases} \displaystyle\sum_{j=0}^{n-s_1} \frac{a^j e^{-a}}{j!} - \sum_{j=(n-s_0)+1}^{n-s_1} \left(\left(\frac{y_0^j e^{-y_0}}{j!} \right) \left(\sum_{i=0}^{n-j-s_1} \frac{y_1^i e^{-y_1}}{i!} \right) \right) & \text{if } n \geq s_0, \\ 0 & \text{if } n < s_0, \end{cases} \end{aligned} \tag{10}$$

that allows for the computation of $P(W_q^{n+1}(t) > x)$ from $P(W_q^n(t) > x)$ in $\mathcal{O}(s_0 - s_1)$ operations. Formula (10) also has an intuitive explanation. Due to the fact that the $(n+1)$st customer arriving at time t will see no more than $n - s_1$ departures before starting service (but may see fewer), the event '$W_q^n(t) > x$' is a subset of the event $C \equiv \{K_{n,t}(x) \leq n - s_1\}$, which is the event 'not enough departures in $[t, t+x]$ if the number of servers was kept at its lower level throughout'. But, to compute $P(W_q^n(t) > x)$ from $P(C)$ we have to exclude the probability of the event

$$C \cap \{K_{n,t}(\Delta t) > n - s_0\}$$

which is the event 'not enough departures in $[t, t+x]$ (if the number of servers was kept at its lower level throughout) but enough departures in $(t, \Delta t]$'.

Formula (8) will be the basis for a lower bound for the general case, while formula (10) will be the basis for an upper bound. This will be explained in the next section.

3. The general case for $P(W_q^n(t) > x)$

In general, $s(u), u \in [t, t+x]$ is piecewise constant, i.e., for some finite K, there are $K+1$ positive integers $s_0, s_1, \ldots, s_K$ and $K+1$ real numbers satisfying $0 = \Delta t_0 < \Delta t_1 < \Delta t_2 < \ldots < \Delta t_K \leq \Delta t_{K+1} = x$ such that, for every $u \in [t, t+x]$,

$$s(u) = \begin{cases} s_0 & \text{if } u \in [t+\Delta t_0, t+\Delta t_1), \\ s_1 & \text{if } u \in [t+\Delta t_1, t+\Delta t_2), \\ \ldots & \\ s_K & \text{if } u \in [t+\Delta t_K, t+\Delta t_{K+1}]. \end{cases}$$

Define the following quantities that will be used in the results that follow. For each $i = 0, 1, \ldots, K$, let

$$\begin{aligned} S_i &= \max\{s_i, s_{i+1}, \ldots, s_K\}, \\ y_i &= \mu s_i \left(\Delta t_{i+1} - \Delta t_i\right) \\ a_i &= \mu \int_{t+\Delta t_i}^{t+x} s(u)\, du = \sum_{j=i}^{K} \mu s_j \left(\Delta t_{j+1} - \Delta t_j\right) = \sum_{j=i}^{K} y_j. \end{aligned}$$

The quantity a_i is the mean parameter of the number of departures over the time period $[t + \Delta t_i, t + x]$, so that a_0 equals the value of a defined in (3). Theorem 1 below provides a characterization of $P(W_q^n(t) > x)$ from which an easy-to-compute lower bound will be derived. Note that the quantity $\eta_i(n)$, used in the theorem, is the probability of, given n customers in system at epoch $t+\Delta t_i$, not enough departures in the interval $[t+\Delta t_i, t+\Delta t_{i+1}]$ and not enough departures in later intervals to start service, but not so few so as to overlap with event A. Theorem 2 provides a similar characterization for an easy-to-compute upper bound. The proofs of both theorems appear in the appendix.

Theorem 1 *For every $i = 0, 1, \ldots, K$,*

$$P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = \begin{cases} e^{-a_i} \left(\displaystyle\sum_{j=0}^{n-S_i} \frac{a_i^j}{j!} + \sigma_i(n) \right) & \text{if } n \geq S_i, \\ 0 & \text{if } n < S_i, \end{cases} \tag{11}$$

where, $\sigma_K(n) = 0$, for every n, and, for every $i = 0, 1, \ldots, K-1$, and $n \geq S_i$,

$$\begin{aligned} \sigma_i(n) &= \eta_i(n) + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \sigma_{i+1}(n-j) \\ \eta_i(n) &= \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=(n-j-S_i)+1}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} \right). \end{aligned}$$

Theorem 1 implies that $P(W_q^n(t + \Delta t_i) > x - \Delta t_i) \geq l_i(n)$ where, for any $i = 0, 1, \ldots, K$ and $n \geq S_i$,

$$\begin{aligned} l_i(n) &\equiv e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{a_i^j}{j!} + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=(n-j-S_i)+1}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} \right) \right) \\ &= e^{-a_i} \left(\sum_{j=0}^{n-S_{i+1}} \frac{a_i^j}{j!} - \sum_{j=(n-S_i)+1}^{n-S_{i+1}} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} \right) \right), \end{aligned}$$

which is an interesting expression because it allows for the computation of $l_i(n+1)$ from $l_i(n)$ in $\mathcal{O}(\max\{S_i - S_{i+1}, 1\})$ operations. A necessary and sufficient condition for this lower bound to be tight for any n is that $S_{i+1} = S_K$.

Theorem 2 *For every* $i = 0, 1, \ldots, K$,

$$P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = \begin{cases} e^{-a_i} \left(\sum_{j=0}^{n-S_K} \frac{a_i^j}{j!} - \epsilon_i(n) \right) & \text{if } n \geq S_i, \\ 0 & \text{if } n < S_i, \end{cases} \tag{12}$$

where, $\epsilon_K(n) = 0$, *for every* n, *and, for every* $i = 0, 1, \ldots, K-1$, *and* $n \geq S_i$,

$$\begin{aligned} \epsilon_i(n) &= \delta_i(n) + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \epsilon_{i+1}(n-j) \\ \delta_i(n) &= \sum_{j=(n-S_i)+1}^{n-S_K} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_K} \frac{a_{i+1}^k}{k!} \right). \end{aligned}$$

Theorem 2 implies that $P(W_q^n(t + \Delta t_i) > x - \Delta t_i) \leq u_i(n)$ where, for any $i = 0, 1, \ldots, K$ and $n \geq S_i$,

$$u_i(n) \equiv e^{-a_i} \left(\sum_{j=0}^{n-S_K} \frac{a_i^j}{j!} - \sum_{j=(n-S_i)+1}^{n-S_K} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_K} \frac{a_{i+1}^k}{k!} \right) \right).$$

As before, this expression is interesting because it allows for the computation of $u_i(n+1)$ from $u_i(n)$ in $\mathcal{O}(\max\{S_i - S_K, 1\})$ operations. A necessary and sufficient condition for this upper bound to be tight for any n is that $S_{i+1} = S_K$.

It is not clear whether it would be better to use (11) or (12) to compute the exact value of $P(W_q^n(t) > x)$. We saw in the previous section one specific case in which (12), namely (10), was preferable. From Theorems 1 and 2, we also get the following bounds,

$$P(W_q^n(t) > x) \leq \sum_{j=0}^{n-S_K} \frac{a^j e^{-a}}{j!}, \quad \text{for every } n \geq S_K, \tag{13}$$

and

$$P(W_q^n(t) > x) \geq \sum_{j=0}^{n-S_0} \frac{a^j e^{-a}}{j!}, \quad \text{for every } n \geq S_0. \tag{14}$$

In particular, the upper bound (13) will be used in the next section when dealing with an infinite number of breakpoints in the number of servers.

4. The computation of $E(W_q^n(t))$

We start by considering the case of finitely many shift changes over the planning horizon. For the most common situation when the arrival rate is periodic and therefore so are staffing levels, this assumption is reasonable because the planning horizon can be chosen to be

long enough relative to the period so that staffing changes beyond the planning horizon are unlikely to have any impact on delays during the period. For example, if staffing decisions are made on a daily basis, then the planning horizon can be chosen to be two days particularly if the expected delay is likely to be much less than 24 hours. Later in the section we address the issue of infinitely many shift changes.

Suppose that, for some finite K, there are $K+1$ positive integers $s_0, s_1, \ldots, s_K$, and $K+1$ real numbers satisfying $0 = \Delta t_0 < \Delta t_1 < \Delta t_2 < \ldots < \Delta t_K$ such that, for every $u \in [t, \infty)$,

$$s(u) = \begin{cases} s_0 & \text{if } u \in [t + \Delta t_0, t + \Delta t_1), \\ s_1 & \text{if } u \in [t + \Delta t_1, t + \Delta t_2), \\ \ldots & \\ s_K & \text{if } u \in [t + \Delta t_K, \infty). \end{cases} \tag{15}$$

The quantities $E(W_q^n(t))$, for every n, can be computed recursively. We use conditioning on the number of service completions to derive the recursion formula. The formula is initiated with

$$E(W_q^n(t + \Delta t_K)) = \begin{cases} \dfrac{n - s_K + 1}{\mu s_K} & \text{if } n \geq s_K, \\ 0 & \text{if } n < s_K. \end{cases} \tag{16}$$

For a generic $i \in \{0, 1, \ldots, K-1\}$ we have the following derivation. For any $n < s_i$, $E(W_q^n(t + \Delta t_i)) = 0$, while, for any $n \geq s_i$,

$$\begin{aligned} &E(W_q^n(t + \Delta t_i)) = \\ &= A_{in}(t) \left(1 - \sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!}\right) + \sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!} \left(\Delta t_{i+1} - \Delta t_i + E(W_q^{n-j}(t + \Delta t_{i+1}))\right), \end{aligned} \tag{17}$$

where $y_i = \mu s_i(\Delta t_{i+1} - \Delta t_i)$ and $A_{in}(t)$ denotes the expected value of $W_q^n(t + \Delta t_i)$ given that there are more than $n - s_i$ service completions in the interval $[t + \Delta t_i, t + \Delta t_{i+1})$. Thus,

$$A_{in}(t) = E(X | X \leq \Delta t_{i+1} - \Delta t_i),$$

where X is the sum of $n - s_i + 1$ i.i.d. exponential random variables of parameter μs_i. From Lemma 2 in the appendix,

$$A_{in}(t) = \frac{n - s_i + 1}{\mu s_i} \left(\frac{1 - \sum_{j=0}^{(n-s_i)+1} \frac{y_i^j e^{-y_i}}{j!}}{1 - \sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!}} \right). \tag{18}$$

Now, we may apply the recursion formula (17) in the order $i = K-1, K-2, \ldots, 0$. In the end, we obtain the exact values for $E(W_q^n(t)) \equiv E(W_q^n(t + \Delta t_0))$, for any n.

If the number of shift changes is not finite and, as an approximation, we assume that $\{s(u), u \geq t\}$ has the form (15) for some finite K then

$$E(W_q^n(t + \Delta t_i)) \approx C^n(t + \Delta t_i), \quad \text{for every } i, n,$$

where $C^n(t + \Delta t_i) = 0$, for every $n < s_i$, and

$$\begin{aligned} C^n(t + \Delta t_i) &= \frac{n - s_i + 1}{\mu s_i} \left(1 - \sum_{j=0}^{(n-s_i)+1} \frac{y_i^j e^{-y_i}}{j!}\right) + (\Delta t_{i+1} - \Delta t_i) \left(\sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!}\right) + \\ &\quad + \sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!} C^{n-j}(t + \Delta t_{i+1}), \end{aligned} \tag{19}$$

for every $n \geq s_i$. Applying this recursive formula starting from

$$C^n(t + \Delta t_K) = \begin{cases} \dfrac{n - s_K + 1}{\mu s_K} & \text{if } n \geq s_K, \\ 0 & \text{if } n < s_K. \end{cases} \tag{20}$$

and in the order $i = K - 1, K - 2, \ldots, 0$ leads us to $C^n(t) \equiv C^n(t + \Delta t_0)$, an approximation of $E(W_q^n(t))$, for any n. Our next result, which is proved in the appendix, presents a bound on the approximation error.

Theorem 3 *Assume that $\{s(u), u \geq t\}$ is always positive. For every $T \geq 0$, let*

$$C(t) = \sum_{n=s(t)}^{\infty} C^n(t) p_n(t)$$

where $\{C^n(t), n = 0, 1, \ldots\}$ is the final output of the recursive formula (19) assuming that $\{s(u) \equiv 1, u \geq t + T\}$. Then

$$0 \leq C(t) - E(W_q(t)) \leq h(T)$$

where

$$h(T) \equiv \sum_{n=s(t)}^{\infty} \left(\frac{n}{\mu}\left(e^{-\mu T} \sum_{j=0}^{n-1} \frac{(\mu T)^j}{j!}\right) - T\left(e^{-\mu T} \sum_{j=0}^{n-2} \frac{(\mu T)^j}{j!}\right)\right) p_n(t). \tag{21}$$

Moreover, if $m(t) = \sum_{n=0}^{\infty} n p_n(t) < \infty$ then $\lim_{T \to \infty} h(T) = 0$.

We are using, by convention, that a sum is zero when the upper limit of a sum is smaller than its lower limit.

It follows from Theorem 3 that, under a mild assumption, the bound can be arbitrarily close to zero. Thus, for every $\epsilon > 0$, there exists $T = T_\epsilon \geq 0$ such that, if we compute $C(t)$ as explained in Theorem 3 then $0 \leq C(t) - E(W_q(t)) \leq \epsilon$. Hence, the approximation can be made arbitrarily accurate by assuming a large enough number of shift changes.

Appendix

Lemma 1 *For any real $y_0 \geq 0$ and integer $n \geq 0$,*

$$\int_0^{y_0} \left(\sum_{i=0}^{n} \frac{y^i e^{-y}}{i!} \right) \, dy = (n+1) \left(1 - \sum_{j=0}^{n} \frac{y_0^j e^{-y_0}}{j!} \right) + y_0 \left(\sum_{j=0}^{n-1} \frac{y_0^j e^{-y_0}}{j!} \right) \tag{22}$$

$$\int_{y_0}^{\infty} \left(\sum_{i=0}^{n} \frac{y^i e^{-y}}{i!} \right) \, dy = (n+1) \left(\sum_{j=0}^{n} \frac{y_0^j e^{-y_0}}{j!} \right) - y_0 \left(\sum_{j=0}^{n-1} \frac{y_0^j e^{-y_0}}{j!} \right) \tag{23}$$

and, for any pair of integers $m, n \geq 0$ such that $m \geq n$,

$$\int_0^{\infty} \left(\sum_{i=n}^{m} \frac{y^i e^{-y}}{i!} \right) \, dy = m - n + 1 \tag{24}$$

Proof: As it may be checked

$$\frac{d}{dy} \left(-(n+1) \left(\sum_{j=0}^{n} \frac{y^j e^{-y}}{j!} \right) + y \left(\sum_{j=0}^{n-1} \frac{y^j e^{-y}}{j!} \right) \right) = \sum_{i=0}^{n} \frac{y^i e^{-y}}{i!}.$$

Thus, (22) and (23) follow from the Fundamental Theorem of Calculus. The sum of (22) and (23) equals $n+1$. Thus,

$$\int_0^{\infty} \left(\sum_{i=n}^{m} \frac{y^i e^{-y}}{i!} \right) \, dy = \int_0^{\infty} \left(\sum_{i=0}^{m} \frac{y^i e^{-y}}{i!} \right) \, dy - \int_0^{\infty} \left(\sum_{i=0}^{n-1} \frac{y^i e^{-y}}{i!} \right) \, dy = m - n + 1.$$

$\diamondsuit$

Lemma 2 *If $\{E_i, i = 1, 2, \ldots, n\}$ are n i.i.d. exponential random variables with mean b and $X \equiv \sum_{i=1}^{n} E_i$ then, for every $t > 0$,*

$$E(X \mid X \leq t) = nb \left(\frac{1 - \sum_{i=0}^{n} \frac{(t/b)^i e^{-t/b}}{i!}}{1 - \sum_{i=0}^{n-1} \frac{(t/b)^i e^{-t/b}}{i!}} \right) \tag{25}$$

Proof: Let N be the number of arrivals during $(0, t]$ in a Poisson process with rate $1/b$ and with interarrival times $\{E_i, i = 1, 2, \ldots\}$. Denote the probability mass function and cumulative distribution function of N by $p_N(j)$ and $F_N(j)$, respectively, and let $G_N(j) =$

$1 - F_N(j) = P\{N \geq j+1\}$. Then, noting that the event $X \equiv \sum_{i=1}^{n} \leq t$ is the same as $N \geq n$, we get

$$\begin{aligned}
E(X|X \leq t) &= \sum_{j=0}^{\infty} E(X|X \leq t, N=j) P\{N=j|X \leq t\} \\
&= \sum_{j=0}^{\infty} E(X|N \geq n, N=j) P\{N=j|N \geq n\} \\
&= \sum_{j=n}^{\infty} E(X|N=j) \frac{p_N(j)}{G_N(n-1)} \\
&= \sum_{j=n}^{\infty} \frac{nt}{(j+1)} \frac{p_N(j)}{G_N(n-1)} \\
&= \frac{nt}{G_N(n-1)} \sum_{j=n}^{\infty} \frac{(t/b)^j e^{-t/b}}{(j+1)!} \\
&= \frac{nb}{G_N(n-1)} \sum_{j=n+1}^{\infty} \frac{(t/b)^j e^{-t/b}}{j!} \\
&= nb \frac{G_N(n)}{G_N(n-1)}
\end{aligned}$$

which one can verify to be equal to the formula in (25). ♢

Note that, when $n = 1$ (25) becomes

$$E(X \mid X \leq t) = b - t\left(\frac{e^{-t/b}}{1 - e^{-t/b}}\right).$$

Proof: (of Theorem 1) We prove this statement by mathematical induction. When $i = K$, (5) applies, so that

$$P(W_q^n(t + \Delta t_K) > x - \Delta t_K) = \begin{cases} \sum_{i=0}^{n-s_K} \frac{y_K^i e^{-y_K}}{i!} & \text{if } n \geq s_K, \\ 0 & \text{if } n < s_K. \end{cases}$$

Since $a_K = y_K$ and $S_K = s_K$, we conclude that the statement is true when $i = K$. Now, suppose that (11) holds for some $i+1$. We will prove that it also holds for i. Conditioning on the system state at time $t + \Delta t_{i+1}$,

$$\begin{aligned}
&P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = \\
&= \begin{cases} \sum_{j=0}^{n-s_i} \frac{y_i^j e^{-y_i}}{j!} P(W_q^{n-j}(t + \Delta t_{i+1}) > x - \Delta t_{i+1}) & \text{if } n \geq s_i, \\ 0 & \text{if } n < s_i, \end{cases}
\end{aligned} \tag{26}$$

because the number of service completions over $[t+\Delta t_i, t+\Delta t_{i+1})$ is Poisson distributed with parameter y_i.

First, consider the case where $s_i > S_{i+1}$ which, in particular, implies $s_i = S_i > S_{i+1}$. For every $n < S_i = s_i$, $P(W_q^n(t+\Delta t_i) > x - \Delta t_i) = 0$ because there are idle servers or servers enough to serve the $(n+1)$-customer on time (see (26)). For every $n \geq S_i$, we have that $n - j \geq S_i > S_{i+1}$, for every $j = 0, 1, \ldots, n - S_i$, so that, from (26), from the induction hypothesis and from the identity (9),

$$
\begin{aligned}
&P(W_q^n(t+\Delta t_i) > x - \Delta t_i) = \\
&= \sum_{j=0}^{n-S_i} \left(\frac{y_i^j e^{-y_i}}{j!} \left(e^{-a_{i+1}} \left(\sum_{k=0}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} + \sigma_{i+1}(n-j) \right) \right) \right) && (27) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} \right) + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \sigma_{i+1}(n-j) \right) && (28) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{a_i^j}{j!} + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=(n-j-S_i)+1}^{n-j-S_{i+1}} \frac{a_{i+1}^k}{k!} \right) + \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \sigma_{i+1}(n-j) \right) && (29) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{a_i^j}{j!} + \sigma_i(n) \right). && (30)
\end{aligned}
$$

Thus, (11) follows when $s_i > S_{i+1}$. Finally, consider the case where $s_i \leq S_{i+1}$, which implies $s_i \leq S_i = S_{i+1}$. Then, from the induction hypothesis,

$$
\begin{aligned}
n < S_i \;&\Rightarrow\; n - j \leq n < S_{i+1} && (j = 0, 1, \ldots, n - s_i) \\
&\Rightarrow\; P(W_q^{n-j}(t+\Delta t_{i+1}) > x - \Delta t_{i+1}) = 0 && (j = 0, 1, \ldots, n - s_i),
\end{aligned}
$$

so that $P(W_q^n(t+\Delta t_i) > x - \Delta t_i) = 0$, follows from (26). Moreover, $n \geq S_i$ implies

$$
n - j \leq S_i - 1 < S_{i+1} \quad (j = n - S_i + 1, n - S_i + 2, \ldots, n - s_i)
$$

which, in turn, also implies $P(W_q^{n-j}(t+\Delta t_{i+1}) > x - \Delta t_{i+1}) = 0$, for every $j = n - S_i + 1, n - S_i + 2, \ldots, n - s_i$. Furthermore, $n \geq S_i$ implies $n - j \geq S_i = S_{i+1}$, for every $j = 0, 1, \ldots, n - S_i$. Thus, (27) and the chain of equalities (28)-(30) is again valid so that the desired result follows also when when $s_i \leq S_{i+1}$. $\diamondsuit$

Proof: (of Theorem 2) As before, we prove this statement through mathematical induction. The statement is true when $i = K$, as shown in the proof of Theorem 1. Now, suppose that the statement is true for some $i+1$. We prove that it is also true for i. As in (26), the

exact value of $P(W_q^n(t+\Delta t_i) > x - \Delta t_i)$ can be derived through conditioning on the system state at time $t + \Delta t_{i+1}$.

First, consider the case where $s_i > S_{i+1}$ which, in particular, implies $s_i = S_i > S_{i+1} \geq S_K$. For every $n < S_i = s_i$, $P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = 0$ because there are idle servers or servers enough to serve the $(n+1)$-customer on time. For every $n \geq S_i$, we have that $n - j \geq S_i > S_{i+1}$, for every $j = 0, 1, \ldots, n - S_i$ so that, from (26), from the induction hypothesis and from (9),

$$
\begin{aligned}
&P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_K} \frac{a_{i+1}^k}{k!} - \epsilon_{i+1}(n-j) \right) \right) && (31) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_K} \frac{a_{i+1}^k}{k!} \right) - \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \epsilon_{i+1}(n-j) \right) && (32) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_K} \frac{a_i^j}{j!} - \sum_{j=(n-S_i)+1}^{n-S_K} \frac{y_i^j}{j!} \left(\sum_{k=0}^{n-j-S_K} \frac{a_{i+1}^k}{k!} \right) - \sum_{j=0}^{n-S_i} \frac{y_i^j}{j!} \epsilon_{i+1}(n-j) \right) && (33) \\
&= e^{-a_i} \left(\sum_{j=0}^{n-S_K} \frac{a_i^j}{j!} - \epsilon_i(n) \right). && (34)
\end{aligned}
$$

Thus, (12) follows when $s_i > S_{i+1}$. Finally, consider the case where $s_i \leq S_{i+1}$, which implies $s_i \leq S_i = S_{i+1} \geq S_K$. Then, from the induction hypothesis,

$$
\begin{aligned}
n < S_i \;&\Rightarrow\; n - j \leq n < S_{i+1} && (j = 0, 1, \ldots, n - s_i) \\
&\Rightarrow\; P(W_q^{n-j}(t + \Delta t_{i+1}) > x - \Delta t_{i+1}) = 0 && (j = 0, 1, \ldots, n - s_i),
\end{aligned}
$$

so that $P(W_q^n(t + \Delta t_i) > x - \Delta t_i) = 0$, follows from (26). Moreover, $n \geq S_i$ implies $n - j \leq S_i - 1 < S_{i+1}$, for every $j = n - S_i + 1, n - S_i + 2, \ldots, n - s_i)$, which, in turn, implies $P(W_q^{n-j}(t + \Delta t_{i+1}) > x - \Delta t_{i+1}) = 0$, for every $j = n - S_i + 1, n - S_i + 2, \ldots, n - s_i)$. Thus, the chain of equalities (31)-(33) is again valid so that the desired result follows also when $s_i \leq S_{i+1}$. $\diamondsuit$

Lemma 3 *For every $n \geq 1$, the function $f(y) \equiv e^{-y} \sum_{j=0}^{n-1} y^j/j!, y \geq 0$, is nonincreasing.*

Proof: For every $n \geq 1$ and for every $y \geq 0$,

$$
f'(y) = -e^{-y} \sum_{j=0}^{n-1} \frac{y^j}{j!} + e^{-y} \sum_{j=1}^{n-1} \frac{y^{j-1}}{(j-1)!} = -e^{-y} \sum_{j=0}^{n-1} \frac{y^j}{j!} + e^{-y} \sum_{j=0}^{n-2} \frac{y^j}{j!} = -e^{-y} \frac{y^{n-1}}{(n-1)!} \leq 0.
$$

$\diamondsuit$

Actually, as pointed out by one of the referees, Lemma 3 could be rephrased to say that a Poisson random variable is stocastically increasing in its mean, and this is proven in Ross (1983), page 256.

Lemma 4 *For any nonnegative sequence* $\{p_n\}$ *such that* $\sum_{n=0}^{\infty} p_n = 1$ *and positive integer* s,

$$\int_0^{\infty} \sum_{n=s}^{\infty} \left(\sum_{j=0}^{n-1} \frac{e^{-x} x^j}{j!} \right) p_n \; dx = \sum_{n=s}^{\infty} \left(\int_0^{\infty} \sum_{j=0}^{n-1} \frac{e^{-x} x^j}{j!} \; dx \right) p_n = \sum_{n=s}^{\infty} n p_n. \tag{35}$$

Proof: The sequence $\{f_n(x)\}$, defined by $f_n(x) \equiv \sum_{j=0}^{n-1} (e^{-x} x^j / j!) p_n$, for $n = 0, 1, \ldots$, is such that $f_n(x) \leq p_n$, for every $x > 0$ and every n. Since $\sum_{n=s}^{\infty} p_n < +\infty$ then $\sum_{j=s}^{n} f_j$ converges uniformly in $(0, +\infty)$, see Rudin (1976), page 148, Theorem 7.10. The first equality in (35) follows from Theorem 7.16 on page 151 of the same book. The second equality follows from (24). $\diamondsuit$

Proof: (of Theorem 3) For any fixed $T \geq 0$, $E(W_q(t))$ equals

$$\int_0^T \sum_{n=s(t)}^{\infty} P(W_q^n(t) > x) p_n(t) \; dx + \int_T^{\infty} \sum_{n=s(t)}^{\infty} P(W_q^n(t) > x) p_n(t) \; dx. \tag{36}$$

Now, consider a new $M(t)/M/s(t)$ model which only differs from the original $M(t)/M/s(t)$ model in the values of staffing function $s(\cdot)$ over the interval $[t+T, \infty)$, in which it is reset to one. For this new model, the expected waiting time until service commences of a customer that arrives to the system at time t, denoted $C(t)$, can be computed exactly through the recursive formula (19). Since the new staffing $s(\cdot)$ function kept the original values over $[t, t+T)$ then $C(t)$ equals

$$\int_0^T \sum_{n=s(t)}^{\infty} P(W_q^n(t) > x) p_n(t) \; dx + g(T). \tag{37}$$

Note that $C(t) - E(W_q(t)) \geq 0$ due to the way the staffing function differs in both models.

Assume that the new staffing function has K breakpoints. From (13),

$$g(T) \leq \int_T^{\infty} \sum_{n=s(t)}^{\infty} \left(e^{-a_K(x)} \sum_{j=0}^{n-S_K} \frac{a_K(x)^j}{j!} \right) p_n(t) \; dx,$$

where

$$\begin{aligned} a_K(x) &\equiv \mu s_0(\Delta t_1 - \Delta t_0) + \ldots + \mu s_{K-1}(\Delta t_K - \Delta t_{K-1}) + \mu(x - \Delta t_K) \\ &\geq \mu(\Delta t_1 - \Delta t_0) + \ldots + \mu(\Delta t_K - \Delta t_{K-1}) + \mu(x - \Delta t_K) \\ &= \mu x, \end{aligned}$$

Thus, from Lemma 3 and using the fact that $S_K = 1$,

$$g(T) \leq \int_T^\infty \sum_{n=s(t)}^{\infty} \left(e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!} \right) p_n(t)\ dx,$$

a bound that we denote by $h(T)$. From Lemma 4,

$$h(T) = \sum_{n=s(t)}^{\infty} \left(\int_T^\infty e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ dx \right) p_n(t)$$

and from Lemma 1, $h(T)$ equals the expression in (21). Moreover, since the second integral in (36) is nonnegative ,

$$C(t) - E(W_q(t)) \leq g(T) \leq h(T).$$

This completes the first part of the proof. It remains to be shown that $\lim_{T\to\infty} h(T) = 0$. From (24),

$$h(0) = \sum_{n=s(t)}^{\infty} \left(\int_0^\infty e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ dx \right) p_n(t) = \sum_{n=s(t)}^{\infty} n p_n(t) < +\infty.$$

On the other hand, for any nonnegative T,

$$\begin{aligned} h(0) &= \int_0^\infty \left(\sum_{n=s(t)}^{\infty} e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ p_n(t) \right) dx \\ &= \int_0^T \left(\sum_{n=s(t)}^{\infty} e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ p_n(t) \right) dx + \int_T^\infty \left(\sum_{n=s(t)}^{\infty} e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ p_n(t) \right) dx \\ &= h(T) + \int_0^T \left(\sum_{n=s(t)}^{\infty} e^{-\mu x} \sum_{j=0}^{n-1} \frac{(\mu x)^j}{j!}\ p_n(t) \right) dx. \end{aligned}$$

As T goes to infinity, the integral in the right-hand-side goes to $h(0) < +\infty$ and, therefore, $h(T)$ must go to zero. Thus, the second part of the proof is complete. $\diamondsuit$

5. Acknowledgements

The authors are very grateful for the helpful comments of the senior editor and the referees. We'd particularly like to note the contribution of one of the referees in suggesting the proof of Lemma 2 that appears in this note.

References

L. Green, J. Giulio, R. Green and J. Soares. 2005. Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine* , to appear.

L. Green, P. Kolesar and J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49**, 549–564.

Donald Gross and Carl M. Harris. 1998. Fundamentals of queueing theory. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York, third edition.

A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li and X. Wu. 2005. A Survey and Experimental Comparison of Service Level Approximation Methods for Non-Stationary $M(t)/M/s(t)$ Queueing Systems with Exhaustive Discipline. *INFORMS Journal on Computing*, to appear.

A. Ingolfsson, M. Haque and A. Umnikov. 2002. Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research* **139**, 585–597.

S. Ross. 1983. Stochastic Processes. Wiley, New York.

W. Rudin. 1976. Principles of mathematical analysis. McGraw-Hill Book Co., New York, third edition. International Series in Pure and Applied Mathematics.